Libros de Cátedra

Introducción al análisis estadístico de datos geológicos

Marta Alperin

FACULTAD DE
CIENCIAS NATURALES Y MUSEO

naturales

Edulp
Editorial de la Universidad de La Plata

UNIVERSIDAD NACIONAL DE LA PLATA

INTRODUCCIÓN AL ANÁLISIS ESTADÍSTICO DE DATOS GEOLÓGICOS

Marta Alperin

2013

Alperin, Marta

Introducción al análisis estadístico de datos geológicos / Marta Alperin ; con prólogo de Luis Spalletti. - 1a ed. - La Plata : Universidad Nacional de La Plata, 2013.

E-Book: ISBN 978-950-34-1029-5

1. Geología. 2. Estadísticas. I. Luis Spalletti, prolog. II. Título

CDD 551

Fecha de catalogación: 28/10/2013

Diseño de tapa: Dirección de Comunicación Visual de la UNLP

Universidad Nacional de La Plata – Editorial de la Universidad de La Plata

47 N.º 380 / La Plata B1900AJP / Buenos Aires, Argentina
+54 221 427 3992 / 427 4898
editorial@editorial.unlp.edu.ar
www.editorial.unlp.edu.ar

Edulp integra la Red de Editoriales Universitarias Nacionales (REUN)

Primera edición, 2013
ISBN: 978-950-34-1029-5

a Belén y Maite,
mis hijas

AGRADECIMIENTOS

En primer lugar quiero agradecer al Ms. Carlos Skorupka y al Dr. Luis Castro de la Facultad de Ciencias Naturales y Museo de la UNLP, a la Lic. Mónica Berisso del Ministerio de Desarrollo Social de la provincia de Buenos Aires y a la Ms. Norma Chhab de Statistics Canada por la lectura crítica y sugerencias que permitieron mejoraron la redacción de algunos capítulos y a la Lic. Carolina Cabrera y el Dr. Horacio Echeveste quienes realizaron todo el material gráfico de esta obra.

Especiales agradecimientos para a los estudiantes quienes despertaron mí la vocación docente y el deseo de mostrarles toda la potencialidad que tienen los análisis estadísticos en la resolución de los problemas geológicos y paleontológicos.

Mi gratitud al Dr. Luis Spalletti y al Dr. Mario Hernández por las elogiosas palabras vertidas en el Prologo y la Presentación del libro.

A mis amigos, Gabriela, Raquel, Clarita y Mario por su constante apoyo.

Por último agradezco a mi esposo Horacio por su prolongado y permanente apoyo durante los últimos veinticinco años y especialmente durante la redacción de este libro.

Marta I. Alperin

La Plata, Buenos Aires

Mayo de 2013

ÍNDICE

INFERENCIA

ANÁLISIS DE LA VARIANZA

RELACIONES ENTRE DOS VARIABLES

PRESENTACIÓN

La utilización de la estadística en el medio científico ha tomado sin duda un ritmo aceleradamente creciente, señalando una tendencia que, de forma más atenuada, es acompañada cotidianamente por el resto de las actividades del hombre.

A tal extremo que tanto en economía, como en política, arte, periodismo, deporte, espectáculos, oficios, religiones y en todas las demás, se utilizan estimadores, rankings, porcentuales, frecuencias, tendencias y otras expresiones estadísticas del más diverso carácter y profundidad, generalmente acompañadas de los códigos que corresponden a cada actividad.

Las geociencias no escapan a esta realidad, ya que se ha convertido en una herramienta de primer orden, facilitada en su empleo por el advenimiento de la informática. Es por lo tanto habitual que los geólogos la utilicen en sus tareas profesionales, de investigación o de gestión.

Pero es también habitual que lo hagan para cubrir una necesidad y porque no, una "moda", sin el pleno conocimiento del contexto de esta verdadera ciencia formal y de la mayor parte de su oferta. Es como adquirir un instrumental o un software y aplicarlo sin haber leído acabadamente su correspondiente manual.

Es aquí donde la obra de la Dra. Alperin adquiere su mayor dimensión y apunta al enriquecimiento de la labor del geólogo, con el mayor conocimiento posible de la dimensión de las bases teóricas para llegar a aprovechar su utilidad, ya se trate de la componente descriptiva o la referencial. O también el caso de poblaciones relativamente estáticas o dinámicas, como el agua subterránea, donde por ejemplo, el análisis de series resulta básico para reconstruir los sucesos hidrológicos y poder formular pronósticos.

Luego de una muy valiosa introducción analítica general a la aplicación de la ciencia en Geología, Marta recorre los caminos de la Estadística descriptiva con la teoría de las probabilidades y sus modelos, resaltando específicamente el valor de los muestreos y la distribución de los consecuentes estadísticos.

Se adentra seguidamente en el dominio inferencial, con la estimación de los parámetros de una distribución y las pruebas de hipótesis, para abordar a continuación el análisis de las medias muestrales, las correlaciones y las comparaciones múltiples, para alcanzar el abordaje de las pruebas no paramétricas.

Como se anticipara párrafos más arriba, el estudio de las series de tiempo, cronológicas y observaciones secuenciales fue de especial interés para quien presenta el libro, avanzando luego la autora con soltura en una introducción al análisis de datos composicionales, direccionales y espaciales con el cual concluye la obra.

El potencial lector podría creer que los contenidos hasta aquí comentados guardan un formato y/o un estilo abrumadoramente matemático, o que corre el riesgo de sucumbir a manos de algún complicado y perverso algoritmo. Pues no es así.

Le sorprenderá que en cambio, el marco esencial es familiarmente geológico, incluyendo todas las componentes imaginables de las geociencias y abundando en ejemplos, que hacen sumamente interesante no sólo ya la lectura, sino la comprensión de conceptos que no resultan difíciles cuando las explicaciones se ofrecen en el marco de la propia actividad de quién lee. La acertada y profusa selección de ilustraciones, muy oportunas en todos los casos, ayuda sin duda a lograrlo plenamente.

Podría decirse que no se trata de una Geoestadística, sino en realidad de una ordenada estadística para geólogos no desprevenidos, que sólo tendrán que usarla a sabiendas.

Es esa al menos la sensación que experimenta quien reconoce, además de la excelencia del producto, el enorme esfuerzo de poder escribirlo de tal manera y con éxito.

Los esfuerzos cuando provienen de la convicción nunca son vanos y seguramente llenan espacios que lo cotidiano no logra, porque carece de fortaleza.

Mario A. Hernández
Profesor Titular Cátedra de Hidrogeología
Doctor en Ciencias Naturales

La Plata, Mayo de 2013

PRÓLOGO

Durante la segunda mitad del siglo XX las Ciencias Naturales en general y la Geología en particular experimentaron importantes cambios metodológicos. Entre ellos, se pasó de descripciones cualitativas a la colección sistemática de información que permitiera contar con datos valiosos, esenciales para definir más acabadamente los atributos de los objetos de estudio y explicar con mayor consistencia los fenómenos naturales. Esta variación en los enfoques metodológicos constituyó un importante desafío para los profesionales e investigadores en las Ciencias de la Tierra, quienes se vieron obligados a la incorporación de nuevas herramientas para el desenvolvimiento de sus tareas.

El proceso de obtención de datos geológicos que puedan ser expresados numéricamente es por una parte costoso y arduo, y por otra bastante limitado. Por tal motivo, resulta crítica la evaluación de la calidad y representatividad de la información, así como la ponderación de su significado. De ello se ocupa la Estadística, y a la Estadística está dedicado este texto de la Dra. Marta Alperin.

Aun cuando la ciencia moderna no puede prescindir de los estudios probabilísticos y estadísticos, es muy frecuente que los estudiantes de las licenciaturas en Geología y Geoquímica los consideren materias complejas cuando no de difícil comprensión, quizás debido a sus fundamentos matemáticos. Por ello, disponer de un texto adaptado a los niveles de instrucción de nuestros estudiantes de grado resulta esencial para la mejor comprensión de los fundamentos de la Estadística y para lograr que los futuros profesionales estén capacitados para la obtención, organización e interpretación de datos numéricos.

El libro de la Dra. Alperin hace hincapié en los muy diversos métodos estadísticos que son aplicables a las Ciencias de la Tierra y comprende tanto a sus fundamentos como a las principales técnicas para el procesamiento y evaluación de la información. Con una visión ágil, moderna y muy actualizada la autora nos brinda un aporte muy comprensivo sobre cómo las distintas herramientas de la estadística pueden auxiliar en la resolución de problemas geológicos. El contenido de la obra es muy amplio e incluye desde conceptos de estadística descriptiva, probabilidades, muestreo y análisis de varianza y correlación hasta tratamientos más complejos, tales como métodos no paramétricos, series de tiempo y cadenas de Markov, análisis de datos composicionales, direccionales y geoestadísticos. Digno de destacar es el equilibrio entre el soporte teórico de los conceptos, los procedimientos estadísticos y sus aplicaciones mediante el empleo de muy adecuados ejemplos.

Marta Alperin es licenciada en Geología (1982) y doctora en Ciencias Naturales (1988) de la Facultad de Ciencias Naturales y Museo (UNLP). En los primeros años de su carrera científica se dedicó a los estudios micropaleontológicos, y en estas investigaciones ya mostró su particular interés por los métodos estadísticos, tal como lo ilustra su trabajo sobre interpretaciones paleoambientales de radiolarios publicado por la Asociación Geológica Argentina en 1993. Fue en ese mismo año que Marta se incorporó a la cátedra de Estadística de la Facultad de Ciencias Naturales y Museo, y es en esa área del conocimiento donde ha desarrollado su ininterrumpida tarea docente y en la que en la actualidad se desempeña como profesora adjunta. En el campo profesional ha volcado su experiencia en el manejo de diversas técnicas estadísticas particularmente en investigaciones sobre medio ambiente y gestión territorial.

En este libro, Marta ha sabido amalgamar su vocación para la educación universitaria con sus amplios conocimientos sobre Estadística. Como resultado disponemos ahora de una obra que tiende a derribar las habituales barreras entre la Geología y las Matemáticas mostrándonos a nivel comprensible los principales conceptos y técnicas estadísticas.

De esta forma, *Introducción al Análisis Estadístico de Datos Geológicos* constituye un excelente soporte para los cursos destinados a la utilización de métodos cuantitativos en Geología y disciplinas relacionadas. Aún cuando está dirigido a los estudiantes de grado, el texto es también una obra de actualización y consulta para los profesionales e investigadores que quieran adentrarse en los empleos de la Estadística como una herramienta esencial en la obtención, procesamiento y análisis de datos geológicos.

Es entonces para mi un motivo de orgullo presentar este trabajo para el que auguro una cálida recepción. Por ello agradezco profundamente a Marta por su cálida invitación y felicito a los responsables de la Editorial de la Universidad Nacional de La Plata por haberlo seleccionado para su publicación.

Luis Spalletti
Profesor Titular Emérito de Sedimentología
Facultad de Ciencias Naturales y Museo (UNLP)
Doctor en Ciencias Naturales

La Plata, mayo de 2013

APLICACIONES DE LA ESTADÍSTICA EN LA GEOLOGÍA

DATOS GEOLÓGICOS

Introducción

La estadística aplicada en la actualidad es parte del lenguaje científico cotidiano. Particularmente en las Ciencias de la Tierra los métodos estadísticos se aplican en trabajos científicos y profesionales tan diversos como los de hidrogeología, hidrología, petrología ígnea y sedimentaria, pedología, evaluación de recursos naturales (gas, petróleo y minerales), evaluación de impacto ambiental, gestión ambiental, teledetección, entre otras. Cualquiera sea el objeto de estudio los geólogos, geoquímicos y paleontólogos se enfrentan a los desafíos que surgen cuando se estudian formaciones de rocas, yacimientos minerales, especies fósiles, y cualquier otro objeto o fenómeno que implique grandes volúmenes de material espacialmente disperso y/o de difícil acceso.

Si bien la geología es una ciencia principalmente observacional, algunas veces existe la necesidad de planear un diseño experimental para obtener la información del fenómeno que se quiere estudiar. Los investigadores se ven forzados a tomar muestras (subconjunto del total) y a partir de la información que de ellas se obtiene caracterizar todo el fenómeno que se estudia (la población). Es necesario conocer como tomar la muestra para que represente a la población, y de este modo permita deducciones válidas sobre el fenómeno que se estudia. Los experimentos deben estar diseñados de forma tal que se minimice la variabilidad introducida en el muestreo y solo se obtenga las variaciones naturales. Además se requiere que los datos sean tomados con exactitud y las medidas realizadas con precisión para eliminar las fuentes de error.

Cuando se aborda el estudio de un fenómeno geológico alguno de los objetivos e interrogantes más frecuentes son estimar el valor medio y la variabilidad de alguna propiedad en el cuerpo de roca, detectar si existen diferencias geológicas importantes entre las propiedades de dos o más afloramiento o si los afloramientos pertenecen a la misma unidad litológica, evaluar el grado de asociación (correlación estadística) y descubrir patrones de variación espacial (tendencias) de todo tipo de propiedades mapeables de la población (i.e. sustancias contaminantes, estructuras, metales). Es difícil saber si los valores medios y la variabilidad de una muestra son los mismos que los de la población,

así como decidir si las diferencias entre muestras de distintos afloramientos o de diferentes experimentos son producto del azar o realmente se trata de muestras de la misma población, incluso si las propiedades que en una muestra se presentan asociadas lo están en la población.
La estadística, con el cálculo de probabilidades permite conocer esas diferencias así como saber cuál es la magnitud del error en esa estimación. Provee metodologías para tomar muestras representativas que permitan extraer el máximo de información de un conjunto limitado de datos que sean una base para la interpretación de los fenómenos geológicos. Permite testear las hipótesis geológicas con objetividad y, a partir de los resultados, retroalimentar la generación de nuevas hipótesis que den cuenta del fenómeno estudiado.

Breve historia

La cuantificación de la Geología ha sido siempre un tópico fascinante que se remonta a su inicio como ciencia y los métodos estadísticos se utilizan desde hace más de ciento cincuenta años. Uno de los trabajos pioneros fue el de Lyell quien, en 1830, subdividió la edad de rocas del Terciario utilizando la proporción de especies de bivalvos actuales y fósiles. Antes de terminar el siglo XIX inician los estudios petrográficos en rocas sedimentarias e ígneas, se mide el tamaño y forma de clastos, se describe la composición de las rocas, se cuentan cristales o granos bajo el microscopio, etcétera. Hacia 1930 la Estadística se expande rápidamente en algunas ramas de la Geología y es desde 1940 que los análisis estadísticos se incluyen en todas sus ramas.
A mediados del siglo pasado se conjugaron una serie de acontecimientos que impulsaron el desarrollo y utilización de modelos matemáticos y estadísticos que facilitan la comprensión de los complejos sistemas naturales geológicos y resolución de problemas relacionados con su manejo. Durante la década de 1950 surge un importante avance en la utilización de la estadística en las Ciencias de la Tierra que coincidió con el desarrollo de técnicas estadísticas multivariadas y la aparición de las computadoras que facilitó los cálculos.
Fue en la década de 1960 que las aplicaciones se hicieron cada vez más frecuentes en las industrias del Petróleo y Minera. Para resolver problemas propios de estas ramas se diseñaron y adaptaron especialmente nuevas metodologías que permiten la interpretación de secuencias de perforaciones, la descripción de reservorios, analizar la distribución de propiedades de roca madre y roca portadoras de petróleo, efectuar simulaciones y cubicar reservorios, entre otros. Entre estos métodos se encuentran los que abordan el análisis de datos espacializados y/o georeferenciados (obtenidos en tareas de campo o de información satelital) y por otro los datos composicionales (obtenidos de análisis modal o análisis químicos de roca). En los últimos años se han desarrollado metodologías específicas para el análisis de estos datos: la geoestadística y el análisis estadístico composicional.

Población y Muestra Geológica. Población y Muestra Estadística

La Estadística es una ciencia metodológica cuyos principios y técnicas se aplican a datos de algún tipo. Las aplicaciones de la estadística en la Geología abarcan dos de sus tres ramas, la estadística **descriptiva** o deductiva y la **estadística inferencial** o inductiva. La estadística descriptiva se ocupa del ordenamiento, la presentación, el cálculo de promedios y porcentajes y la síntesis de los datos. La estadística inferencial utiliza la información de una muestra estadística para formular conclusiones o realizar predicciones acerca de la población relacionando los modelos matemáticos y la práctica. A estas dos ramas se agrega una tercera, en la cual los geólogos prácticamente no se involucran que es la **teoría de probabilidades**. La teoría de probabilidades se ocupa del estudio de los fenómenos que ocurren al azar, formula teoremas, proposiciones y modelos que permiten describirlos y son los que utiliza la estadística inferencial.

Cualquiera sea la aplicación de la estadística en los trabajos geológicos, un número limitado de observaciones o datos, llamados muestra estadística, permiten hacer inferencias a la totalidad de las posibles medidas, valores o cualidades estudiadas, la población estadística.

Antes de proseguir, es necesario aclarar algunos conceptos:

Población estadística es todo el grupo posible de medidas, valores o cualidades que son motivo del estudio.

Unidades de muestreo o **ejemplares** son los miembros individuales de la población.

Datos son los valores medidas o cualidades que se obtienen de la observación y/o medición de las unidades de muestreo.

Muestra estadística la forman un número limitado de datos. La muestra debe ser representativa de la población y ser obtenida por un procedimiento que permita que cualquier individuo de la población tenga la misma probabilidad de ser elegido.

Población geológica, dada la naturaleza peculiar y la variedad de los fenómenos geológicos, la población geológica comprende diferente clase de objetos (ej. cristales, pozos de agua y petróleo, unidades litológicas, emanaciones de gases), eventos (ej. erupciones volcánicas, inundaciones, precipitaciones, terremotos) o simplemente números (ej. producción de barriles de petróleo, número de manifestaciones minerales en un distrito minero, medidas de rumbo de estructuras, longitudes de onda de olas de diferente tipo, profundidades), que son de interés para el estudio que se va a desarrollar. Es necesario tener el control conceptual de la población lo que implica distinguir entre la población hipotética, la existente, la disponible y la objetivo o blanco.

La población **hipotética**, representa el volumen total de material formado por algún proceso que existe en algún punto inicial del tiempo geológico. La población **existente** es la parte remanente de la población hipotética. La mayoría de los individuos no están disponibles para el muestreo porque están enterrados o porque fueron devastados por procesos geológicos posteriores. La población **disponible** representa el volumen de material de la población existente que es accesible a ser muestreada. Por

último, la población **objetivo** o **blanco** es aquella sobre la que se hacen las inferencias con base a los datos que se obtienen en el muestreo. La población a ser muestreada debe coincidir con la población objetivo. Cuando la población muestreada es más limitada y difiere mucho, se debe tener en cuenta que cualquier conclusión que se alcance sólo podrá aplicarse a la población muestreada. Para finalizar es importante recalcar que una única población geológica puede dar origen a varias poblaciones estadísticas (diferentes conjuntos de números o cualidades).

Datos geológicos, son diferentes a los que estudian otras disciplinas. La mayoría se obtienen a partir de manifestaciones o afloramientos dispersos en el espacio y producidos por procesos que el geólogo (investigador) no puede controlar. Generalmente se observa un producto fijo, a lo que se suman otros procesos naturales superficiales que pueden haber removido o enmascarado parte de los productos originales, a esto se agrega que puede existir evidencia no accesible por estar oculta en el subsuelo.

Muestra geológica, informalmente se puede definir como una cantidad finita de roca o sedimentos consolidados o inconsolidados, muestreados al azar de la parte del cuerpo de roca que está disponible. La muestra geológica se puede obtener de un cuerpo de roca que puede ser una Formación, un aflramiento, un cilindro de roca (testigo) obtenido durante una perforación, o *cutting*. La posible confusión entre muestra geológica y muestra estadística se puede minimizar si, para referirse a una muestra geológica se utiliza la palabra **espécimen** (Rock, 1988) y dejar el término muestra exclusivamente en sentido estadístico (Cuadro 1).

Población geológica: conjunto de todos los elementos, objetos, números o eventos acotados en un tiempo y en un espacio determinado, con alguna característica común observable o medible objeto de estudio.
Población estadística: todos los valores o datos derivados de las características de los elementos de la población geológica. La población estadística puede ser infinita o finita.
Unidad de observación, espécimen ó unidad muestral: entidad física sujeta a medición o caracterización (roca, sedimentos, agua, hidrocarburo, etc.).
Datos: conjunto de valores o de atributos que se obtienen midiendo, contando o clasificando cada espécimen.
Muestra: subconjunto de datos extraídos por algún método aleatorio representativos de la población estadística.

Cuadro 1. Definiciones de conceptos básicos.

En suma, para iniciar un estudio geológico se debe tener claro no solo el objetivo de la investigación tan específicamente como sea posible, sino también definir la población objetivo. Se recomienda **delimitar la población en tiempo y espacio**, definir las unidades de muestreo, seleccionar cuidadosamente la característica o características a ser tomadas y determinar el volumen de material a recolectar, todo esto permite planificar el diseño de muestreo evitando omitir datos cuando se muestrea. Se requiere del **juicio geológico** no sólo para elegir la característica a relevar, observación o

medida a realizar, sino también para evaluar las fuentes de variabilidad que presentan los datos y la precisión y exactitud requeridas en el estudio. En el quinto capítulo se profundizan estos temas.

Un ejemplo concreto de las implicancias que tiene definir claramente a la población geológica acotándola en tempo y espacio se plantea por ejemplo ante el objetivo estudiar la mineralógica de un cuerpo granítico específico, o de todos los cuerpos graníticos producidos durante la Orogenia Andina, o incluso de todos los granitos de la Orogenia Andina que se hayan formado durante el Oligoceno superior en la Argentina. Esto, condicionará el diseño de muestreo y la elección de los sitios para tomar las muestras. En el primer caso es suficiente con muestrear solamente el cuerpo granítico específico, en el segundo se deberán muestrear granitos de toda América producidas durante la Orogenia Andina, y en el tercer caso se deberá limitar el muestreo en tiempo y espacio a los granitos del Oligoceno superior de la Orogenia Andina argentina. Cualquier sea el caso y generalizando, los programas de muestreo se adecuan a la población disponible.

Una vez que el objetivo está claro y la población delimitada, está claro que resulta imposible examinar todos los cristales de los minerales que lo componen ya que nunca se puede acceder a la totalidad del cuerpo aunque estuviera totalmente expuesto y disponible para tal observación. Esto obliga a que el estudio se realice a partir de los datos de un número limitado de especímenes de granito (las muestras de roca) de forma que, con la información obtenida de estos especímenes, sea posible caracterizar la mineralogía del cuerpo entero.

Suponga ahora que uno de los objetivos específicos es estudiar los cristales de feldespato potásico del stock. El estudio podrá realizarse estudiando en el laboratorio algunos cristales de feldespato potásico de los especímenes de granito muestreados. Para este objetivo específico los cristales de feldespato son los especímenes o unidades observacionales. A ojo desnudo o con el microscopio se podrá medir el largo, ancho, forma, ángulo de clivaje, presencia o ausencia de alteraciones, ancho de las alteraciones, tipo de alteración, color, orientación, etcétera. Las mediciones y observaciones se hacen de un número *n* limitado de cristales y la muestra estadística tendrá *n* datos lo que permitirán inferir las características de los cristales de todo el cuerpo granítico. Se tiene así una única población geológica, todos los cristales de feldespato del cuerpo granítico, pero varias poblaciones estadísticas, una para cada propiedad del cristal de feldespato medida o descripta. Se señala que una única población de objetos puede dar origen a varias poblaciones de números que representan cada una de las características medidas o descriptas: el largo, el ancho, la forma, la orientación, etcétera.

Muestreo

La naturaleza de los datos geológicos, como se mencionó, permite sólo en algunos casos puntuales examinar alguna característica en la población entera (tener un censo), por ello en la gran mayoría de los casos se analiza parte de la población, la población disponible, y sobre la base de la información

relevada en esa porción se hacen inferencias sobre toda la población. Desafortunadamente es difícil para los geólogos tomar una muestra representativa porque solo se puede acceder a los afloramientos de rocas. Se debe tener presente que las características del aforamiento no siempre son las mismas que las del cuerpo entero pues solo tienen rocas que son resistentes a la meteorización. Ahora bien, para que las inferencias acerca de la población sean válidas es necesario que en el procedimiento de la recolección de datos, llamado muestreo, los especímenes que forman la muestra hayan tenido igual probabilidad de ser elegidos y sean representativos de la población. Aún si el afloramiento se muestrea con ambos criterios se está asumiendo que la muestra es representativa de toda la población y no solo de la población disponible. Se profundizará este tema en el capítulo 5.

Si se excluyen o incluyen especímenes con ciertas características de la muestra sistemáticamente, deliberadamente o inadvertidamente, se dice que la muestra es **sesgada**. Por ejemplo suponga que interesa caracterizar la porosidad de una unidad arenosa particular. Si se excluyen las muestras friables o fisuradas porque es difícil medir su porosidad, los resultados se alteran, el rango de porosidad está truncado en las muestras de elevada porosidad, los valores están sesgados hacia el extremo de menor porosidad y se incorporan errores en la estimación de la media y de la variabilidad de porosidad en la unidad arenosa. Si en un estudio granulométrico de sedimentos sólo se incluyen datos de una fracción también se obtiene una muestra sesgada.

Además, durante la recolección de la muestra hay que evitar que no se pierda nada del material recolectado y que desde el proceso de muestreo hasta el lugar de estudio, en el laboratorio, la muestra se disturbe lo menos posible.

Datos geológicos

Naturaleza de los datos

Las observaciones o mediciones sobre los elementos de una población constituyen la materia prima con la cual trabaja la estadística. Interesa estudiar las características que van variando de individuo en individuo. Cada medición o asignación de valores de una característica efectuada en un espécimen origina un **dato**. El conjunto de *n* datos constituye una **muestra estadística**.

Los datos geológicos se pueden clasificar de acuerdo con la forma de obtención. Así se identifican dos grupos: los **datos observacionales**, incluyen observaciones y mediciones de objetos naturales y eventos en el campo o en el laboratorio, y los **datos experimentales**, comprende las medidas que surgen de experiencias de laboratorio.

Otro criterio de clasificación de datos considera el método de recolección. Se distinguen cuatro tipos: mediciones, recuentos, identificaciones y ordenamientos o ranqueos.

Los datos derivados de **mediciones**, involucran operaciones de medidas tales como deflección de una aguja sobre un dial, la amplitud de una línea de rayos-X, o el espesor de una capa sedimentaria. Medidas de rumbo e inclinación de un plano o de rumbo y buzamiento de una línea, distancias medidas a escala o por alidadas, medidas microscópicas, incluyendo ángulos ópticos, ángulos de extinción y de clivaje, e índices de refracción.

Una segunda clase de datos deriva de los **recuentos**. Ejemplos son el número de circones reconocidos en un campo del microscopio, el número de foraminíferos levógiros o dextrógiros, el número de pozos de petróleo en un distrito.

Tanto las mediciones como los recuentos son medidas de **datos métricos** y están constituidas de tal forma que es posible que un objeto pueda identificarse por cantidades relativas entre grado o cantidad.

El tercer tipo de datos geológicos de interés proviene de la **identificación**. Estos datos **no métricos**, **cualitativos** son atributos, características o propiedades categóricas que identifican o describen al objeto de estudio. Describen diferencias en tipo o clase, indican la presencia o ausencia de una característica o propiedad. Son ejemplos el reconocimiento de un determinado tipo de resto fósil en un sedimento, o de una especie mineral en una roca, o la existencia o ausencia de cierta estructura en una capa sedimentaria.

A estas tres clases de datos podemos agregar una cuarta, los datos que surgen del **ordenamiento** (establecer un ranquin: mayor a menor, menor a mayor, más claro a más oscuro, más blando a más duro, etc.) donde es difícil, si no imposible, asignar una escala de medidas. Ejemplos de datos de esta categoría son descripciones de color, grado de aptitud de rocas portadoras de minerales o petróleo, etcétera.

Los estudios geológicos en los cuales se utilizan métodos estadísticos se centran en las características que cambian su estado o expresión entre los diferentes elementos de la población. Se define formalmente el término **variable** como aquella característica, propiedad o atributo, con respecto a la cual los elementos de una población difieren de alguna forma. Se utilizan letras mayúsculas para referirse a la variable (ej. X), y la misma letra en minúscula se emplea para señalar el valor de un elemento de la población, al dato (x). En el caso particular en que la característica no cambie en los elementos de la población recibe el nombre de **constante**. Por ejemplo el periodo de semidesintegración de un isótopo radiactivo es constante.

Las variables pueden ser cuantitativas (**métricas**) y **cualitativas** (**no métricas**). Las variables cuantitativas permiten diferenciar los individuos por diferencias de grado o cantidad relativas. Se distinguen dos tipos de variables cuantitativas, las continuas y las discretas y tres tipos de variables cualitativas, las dicotómicas, las nominales y las ordinales (Cuadro 2).

Se llama **variable continua** a aquella característica cuyas observaciones pueden asumir cualquier valor entre dos valores posibles, o lo que es lo mismo, cualquier valor dentro de los infinitos valores

de un intervalo. Formalmente una variable es continua si toma sus valores en un conjunto continuo, es decir, un intervalo del eje de los números reales.

Se llama **variable discreta** a aquella característica que asumen un número finito o infinito numerable de posibles resultados. Así las variables discretas surgen de recuentos. Una variable estadística es discreta, si el conjunto de valores posibles se puede poner en la forma $X = \{ a_1, a_2, \ldots , a_k)$, en que k es el número de valores diferentes que pueden tomar los elementos de la muestra.

Se llama **variable categórica**, en contraposición con las variables cuantitativas no se expresan con números, a las que permiten identificar a los individuos a partir de una cualidad, atributo, características o propiedades. Describen diferencias en tipo o clase indicando la presencia o ausencia de la propiedad. Entre ellas distinguir al menos tres subtipos: a) las **dicotómica**s o **binarias** son las más simples, diferencian sólo dos estados como si/no, presente/ausente, contaminado/no contaminado, alta ley/baja ley; b) las categóricas **nominales** como la orientación de los vientos, Norte, Sur, Este y Oeste, listas de especies de fósiles presentes en una muestra; c) las categóricas **ordinales**, se enuncian siguiendo un orden ascendente o descendente como el grado de alteración de una roca que podrá ser severo, moderado o leve, o el color cuando se describe como oscuro, mediano y claro.

Variables cualitativas Los individuos se diferencian por los atributos que poseen. No permiten realizar operaciones algebraicas.	**Binarias:** describe sólo dos estados (si/no, contaminado/sin contaminar)
	Nominales: describen los estados del atributo (estratificación entrecruzada, plana, cruzada, etc.).
	Ordinales: los estados tienen orden ascendente o descendente (muy seleccionado, bien seleccionado, pobremente seleccionado, mal seleccionado).
Variables cuantitativas Los individuos se diferencian por grado o cantidad relativa expresada en un valor numérico. Permiten realizar operaciones algebraicas.	**Discretas**: sólo pueden tomar valores enteros (recuentos: 1, 2,..., 25).
	Continuas: pueden tomar cualquier valor real dentro de un intervalo (mediciones: 1,25; 1,32;.., 5,2845).

Cuadro2. Síntesis y ejemplos de tipos de variables.

Procesos de medición

Asociado estrechamente con la toma de datos se encuentran los procesos de medición. La medición es un proceso que consiste en asignar valores numéricos a cantidades, grados, extensión, magnitudes, etc., o alguna cualidad de un elemento de la población. La asignación se puede hacer utilizando escalas de medidas no métricas y escalas de medidas métricas.

Escalas de medidas no métricas*:* a) **Escalas nominales** los elementos se clasifican en términos de igualdad de sus atributos usando categorías o valores arbitrarios que no mantienen una relación de

orden entre sí. Por ejemplo al color de una roca pelitica se le asigna el valor 1 si es gris, 2 si es roja, 3 si es negra. b) **Escalas ordinales** se utilizan para elementos que puedan ser dispuestos o clasificados en algún orden, jerarquía o ranking entre las categorías, por ejemplo ordenados en forma ascendente o descendente. Cada clase se compara con otra en términos de **mayor que** o **menor que**. Los números sucesivos utilizados en estas escalas no necesariamente deben estar igualmente espaciados pues indican sólo posiciones relativas en la serie ordenada. Algunos ejemplos son la asignación de la dureza mineral utilizando la escala de dureza de Mohs, la asignación a la presencia de fósiles en un estrato como 0: ausente; 1: raro; 2, común, 3: abundante, la escala de selección de sedimentos de Mcmanus.

Escalas de medidas métricas*:* a) **Escalas de intervalos** se utilizan cuando es posible establecer igualdad de intervalos (unidades constantes de medida) de forma que la diferencia entre puntos adyacentes de cualquier parte de la escala es la misma. Estas escalas tienen un cero arbitrario, por ejemplo las escalas de temperatura como la Celsius y Fahrenheit. Además el valor cero no indica una cantidad cero o ausencia de temperatura dado que existen temperaturas bajo cero. b) **Escalas de razón** se utiliza cuando es posible demostrar igualdad de proporciones con respecto a la cualidad que se analiza. Esto implica explícitamente la identificación del punto cero. Longitudes y masas son ejemplos de números de escalas de razón. Los escala ordinal de números (1, 2, 3, ... , *N*) para el recuento de objetos y el cálculo de porcentajes a partir de recuentos son ejemplos de este tipo de escala. Muchos números que se usan en geología están en una escala de razón: medidas de campo como espesores de capas, ángulos de inclinación y buzamiento, número de capas en una unidad estratigráfica, elevación con respecto al nivel del mar. De forma semejante mediadas del tamaño de partículas, forma y orientación y porosidad o permeabilidad de las rocas; velocidad del agua, altura y periodo de las olas, profundidad del agua. Este tipo de escala es la más importante por sus implicancias en la cuantificación de los procesos geológicos.

Propiedades de los datos

Además de las diferencias relacionadas con las escalas de medida para diferenciar las observaciones se pueden utilizar otras propiedades. Algunos datos son dimensionales otros adimensionales; algunos son escalares otros vectoriales; algunas se expresan en sistemas cerrados, otros en sistemas abiertos.

Al realizar observaciones suele resultar que la característica estudiada presenta **dimensiones** tales como longitud, tiempo o masa o alguna combinación de estas. Por ejemplo el espesor de un estrato, el tamaño de una partícula tienen la dimensión longitud (L), la superficie de una cuenca L^2, el periodo de una ola oceánica se mide en tiempo (T), la velocidad de una partícula L/T, la densidad se expresada como masa (M) por centímetro cúbico M/L^3, entre otras. Sin embargo otro grupo de datos son **adimensionales**. Entre ellos se encuentran las mediciones de la gravedad específica que es la relación

entre la densidad de un objeto y la densidad del agua $(M/L^3)/(M/L^3)$ y las funciones angulares comos seno y coseno.

Otro criterio se refiere a considerar si el dato es de tipo vectorial o escalar. Recuerde que un vector se define con un segmento orientado que tiene dirección y longitud. Datos **vectoriales** como el azimut (a dirección es un ángulo que se forma respecto al Norte tomado arbitrariamente como referencia) son muy comunes en geología podemos mencionar entre muchas la orientación de clastos en una capa sedimentaria, trenes de ondulas, rumbos de fallas y diaclasas, etcétera. En contraste con los datos vectoriales que requieren de dos cantidades para poder ser definidos, los datos **escalares** se definen solamente con un número.

Otro tipo de dato muy frecuente en los trabajos geológicos son los porcentajes o proporciones que se obtienen en los estudios mineralógicos o de geoquímica de roca, composición química de rocas y aguas, composición granulométrica de sedimentos, etcétera. En porcentajes y proporciones la suma de las partes es constante y acotada a 100% o a 1, se definen como datos **cerrados** y se suelen denominar datos composicionales (ver Capítulo 12). Por otra parte, otro gran número de observaciones, en las cuales la relación anterior no se cumple se denominan **abiertos**.

Medición del error

Cuando se recolectan datos a través de algún proceso de medición suelen aparecer valores inconsistentes. Para explicar estas inconsistencias se utiliza el concepto de error. Los errores se introducen por el operador u observador, a causa de errores del instrumento de medición, por falta de precisión en la definición operacional o en el proceso de medición.

Los errores **determinados**, llamados también **groseros**, se atribuyen principalmente al instrumental o a reactivos en el caso de análisis químicos, también pueden ser operativos, debidos a distracciones por parte del observador o personales, o de método. Generalmente son grandes en magnitud e irregulares en ocurrencia.

Los **errores sistemáticos** se producen cuando las medidas tienden a ser siempre más grandes o más pequeñas. Suelen originarse por errores en la calibración de los aparatos, aunque también pueden deberse a condiciones externas como por ejemplo cambios de humedad.

Los **errores de método** se introducen si existen discrepancias entre la definición conceptual de la cualidad a ser medida y la definición operacional utilizada para efectuar esa medida.

Aún si los procesos de medidas están libres de errores determinados, sistemáticos y de método pueden existir fluctuaciones en los valores numéricos que se obtienen al repetir la medida. Estos errores impredecibles son los **errores aleatorios**, que si se producen en un gran número de observaciones tienden a anularse, es decir las desviaciones positivas o negativas del valor verdadero, en promedio, tienden a compensarse aproximándose al valor verdadero.

Los procesos de medición y las mediciones del error encuentran su correspondencia en los conceptos de precisión y exactitud. Así, la **precisión** (es lo cerca que los valores medidos están uno de otros), se relaciona con el proceso de medida. Se logra una alta precisión en la medida en que los errores aleatorios sean lo más pequeños posible. La **exactitud** (es la proximidad de un valor medido o calculado al valor verdadero), es externa al proceso de medición. La exactitud implica la ausencia de errores sistemáticos. Por otra parte, como se verá más adelante, precisión y exactitud se vinculan con dos conceptos estadísticos clave, la exactitud está relacionada al valor medio y la precisión está relacionada a la varianza.

ESTADÍSTICA DESCRIPTIVA

Introducción

Cualquiera sea el trabajo geológico que se aborda y el objetivo que se persiga se requiere observar el fenómeno natural y tomar datos midiendo, contando o registrando la presencia de algún carácter. En los últimos años la recolección de información se ha visto favorecida por los avances tecnológicos que en algunos casos va acompañada de una disminución de los costos de obtención de datos. Es así que la cantidad de datos que están disponibles para analizar suelen aumentar a ritmo acelerado. Si los geólogos quieren sacar provecho de esta información necesitan organizar y sintetizar los datos. La **estadística descriptiva** ayuda en este aspecto pues ofrece métodos que permiten resumir la información contenida en un conjunto de datos de la manera más concisa y completa posible. Esto se logra con tablas y gráficos y con unas medidas resumen llamadas **estadísticos** ó **parámetros** según se trate de medias de la muestra ó de la población respectivamente. Además, estadísticos y parámetros permiten no solo tener una apreciación del conjunto total de los datos, sino que posibilita resolver problemas prácticos como se verá más adelante.

Ordenamiento de datos. Descripción mediante gráficos

Existen diferentes gráficos que son útiles para describir la muestra, algunos son apropiados para representar datos de variables cualitativas, como color, forma o alguna otra cualidad, y otros son útiles cuando se trata de variables cuantitativas, que provienen de recuentos o mediciones (Cuadro 1).

Variable		Gráfico
Cualtativas	Nominales	Gráficos circulares Gráficos de barras
	Ordinales	Gráficos de barras
Cuantitativas	Discretas	Gráfico de barras
	Continuas	Histograma Polígono de frecuencias Ojiva Diagramas bivariados (2 variables Box plot Diagramas ternarios (3 variables)

Cuadro 1. Tipos de variables y sus gráficos.

Variables categóricas

Las variables categóricas, como se mencionó en el capítulo anterior, son aquellas cuya escala de medidas es un conjunto de categorías ordinales o nominales. Para graficar datos categóricos se usan diagramas de tortas, también llamados gráficos por sectores, y gráficos de barra. Para realizar los gráficos se requieren conocer el número de datos de cada categoría.

Diagrama de tortas (pie de tortas) o **diagramas de sectores**. Se utiliza un círculo que representa el 100% de las unidades. El círculo se divide en tantos sectores como categorías posean los datos. El tamaño de cada sector se dibuja proporcional al porcentaje de unidades que pertenecen a cada categoría (Fig. 1a).

Gráfico de barras. Cada categoría se representa con una barra, generalmente vertical aunque puede también ser horizontal. La longitud de cada barra es proporcional al porcentaje de unidades que pertenecen a cada categoría. El ancho de la barra es el mismo para todas (Fig. 1b). Si se necesita mostrar en el mismo gráfico más de dos variables categóricas simultáneamente son útiles los gráficos de **barras adyacentes** o **agrupadas** (Fig. 1c) y los de **barra segmentadas** o **apiladas** (Fig. 1d).

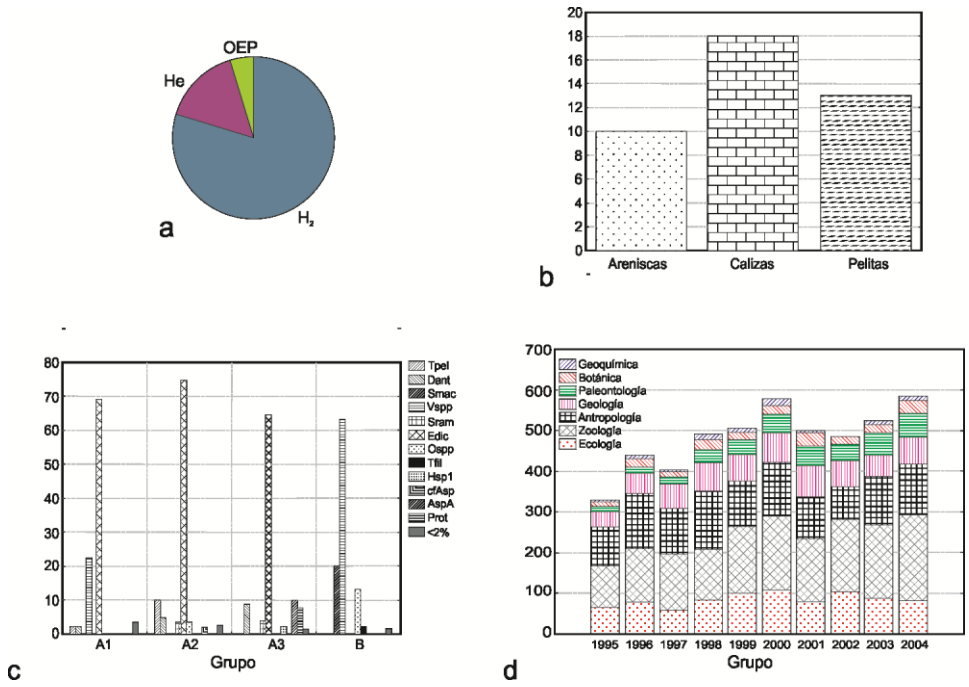


Figura 1.Gráficos de datos cualitativos. a) Gráfico de tortas. b) Gráfico de barras. c) Gráfico de barras adyacentes. d) Gráfico de barras apiladas.

Variables cuantitativas

Analizar una muestra con pocos datos es una tarea relativamente fácil, sobre todo si están ordenados, pero si la muestra grande las cosas se complican. Cuando las muestras son grandes, con el número de

datos *n* mayor que 30, conviene construir una **distribución de frecuencias**. Una distribución de frecuencias, se puede representar con una tabla y/o mediante un gráfico. En ella los datos se acomodan en clases. La **clase** refiere a los número de valores diferentes a_i que toman los datos de una variable discreta *A*, o a los intervalos disjuntos c_i, que cubren el dominio de definición de la variable continua *X*. A cada clase le corresponde una **frecuencia absoluta**, *fi*, que es el número de veces que se repite un dato. La suma de las frecuencias de las *i*-ésimas clases de una muestra es igual al tamaño de la muestra $\left(\sum_{i=1}^{n} f_i\right) = n$, en donde *fi* es la frecuencia absoluta del valor a_i o del intervalo c_i.

Los intervalos disjuntos de la variable continua *X* son llamados **intervalos de clase**. Estos intervalos, se eligen de igual amplitud (*C*). Tomar pocos intervalos en el dominio de la variable implica pérdida de información en el sentido que se disipa la variabilidad y tomar muchos intervalos tampoco es conveniente pues se tergiversa la idea de síntesis. Para calcular el número adecuado de intervalos de clase, *m*, de una distribución se puede utilizar una aproximación empírica ($m=\sqrt{n}$) o la fórmula de Sturgers (*m = parte entera de* [*1+log n/log 2*)]). Conocido el número de intervalos se calcula la amplitud de los mismos (*(valor más alto – valor más bajo)/m*). Los límites de cada intervalo se definen de manera tal que no existan dudas de donde ubicar a los datos que caen exactamente sobre el borde. El criterio más utilizado consiste en incluir en un intervalo los datos iguales o mayores que el límite inferior y menores que el límite superior. Por ejemplo si un fósil mide 5mm y los límites de dos intervalos adyacentes son 4 - 5mm y 5 - 6mm se incluye en el intervalo 5 - 6 milímetros.

La tabla de frecuencias tiene al menos dos columnas, en la primera columna se ubican los límites inferior y superior de cada intervalo de clase de la variable continua *X* o las valores diferentes que toma la variable discreta *A* en estudio. En la segunda columna se ponen las frecuencias absolutas *fi* correspondiente a cada clase. Se pueden poner dos columnas opcionales mas, una con las frecuencias relativas y otra las frecuencias acumuladas.

La **frecuencia relativa**, fr_i, resulta del cociente entre la frecuencia absoluta y el número *n* de datos de la muestra ($fr_i = fi/n$). Se cumple que la suma de todas las frecuencias relativas de los datos de una muestra es igual a uno o a 100%, según se expresen como proporciones o como porcentaje $\left(\sum_{i=1}^{n} fr_i = 1\right)$.

La **frecuencia acumulada**, fa_i, es el valor que surge de la suma o acumulación de todas las frecuencias absolutas (o relativas) precedentes a la frecuencia de la clase en cuestión. La frecuencia acumulada en la última clase es igual al tamaño de la muestra *n* o a 100 si se trata de frecuencias relativas acumuladas.

Gráficos de barras e histogramas

El diagrama de barras, también llamados de bastones, es el gráfico que muestra la distribución de frecuencias de la variable discreta. Sobre el eje horizontal se representan los valores discretos que

toman los datos y sobre cada uno de ellos se coloca una barra vertical de longitud (altura) proporcional a la frecuencia (Fig. 2a). Si se levantan rectángulos en lugar de barras estos deben estar separados unos de otros.

El histograma es el gráfico que representa la tabla de frecuencias de variables continuas agrupadas en intervalos, en abscisas se representan los intervalos de clase y en ordenadas la frecuencia absoluta. Sobre cada intervalo se yergue un rectángulo de altura igual a la frecuencia. Es importante remarcar que el área de cada rectángulo es proporcional a la clase que representa y no su altura. Se recomienda que en ordenadas se represente la frecuencia absoluta y no porcentajes o proporciones porque se enmascara el tamaño de la muestra (Fig. 2b).

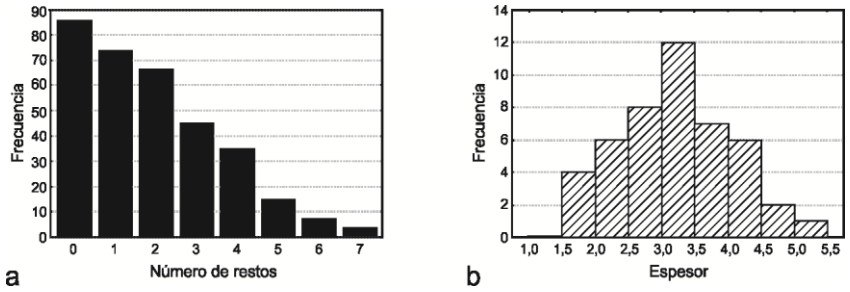


Figura 2. a) Gráfico de barras. b) Histograma.

Los histogramas y gráficos de barra proporcionan mucha información sobre la distribución de los datos, son sumamente útiles en las etapas iniciales de un trabajo pues permiten familiarizarse rápidamente con los datos. En las etapas más avanzadas del trabajo, aunque los histogramas hayan revelado patrones más o menos claros y sugestivos, se deben realizar pruebas estadísticas para cotejarlos.

Polígono de frecuencias

Otro gráfico que se suele emplear es el polígono de frecuencias que se construye uniendo **el centro** de de la parte superior de cada rectángulo del histograma con una curva (Fig. 3). El primer segmento se extiende desde la marca de clase de un intervalo menor al de los datos con frecuencia 0 y el ultimo intervalo se une con la marca de clase de un intervalo mayor al último que también tiene frecuencia 0, de modo que el polígono quede cerrado. El área debajo del polígono de frecuencias es igual a la suma del área de cada rectángulo del histograma.

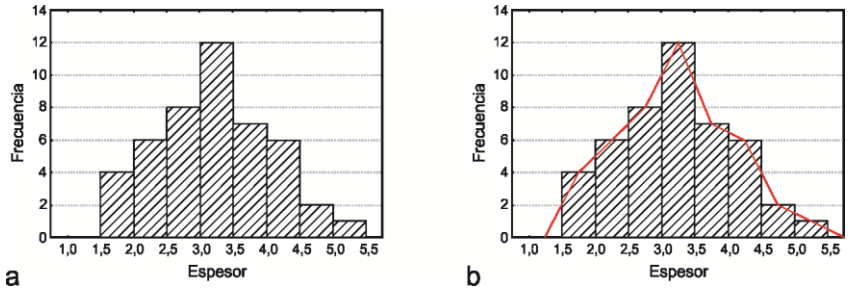


Figura 3. Relación entre a) Histograma y el b) Polígono de frecuencias.

Histograma de frecuencias acumuladas y Polígonos de frecuencias acumuladas (ojivas)

El tercer gráfico muy útil que se deriva de la tabla de frecuencias es el histograma de frecuencias acumuladas. Esté gráfico tiene en abscisas los intervalos de clase y en ordenadas las frecuencias relativas acumuladas (Fig. 4a). El polígono de frecuencias acumuladas, llamado ojiva, se dibuja uniendo los puntos del **extremo superior** de cada intervalo con segmentos de recta. El primer intervalo tiene 0 frecuencia acumulada y el último intervalo termina en 100% (Fig. 4b).

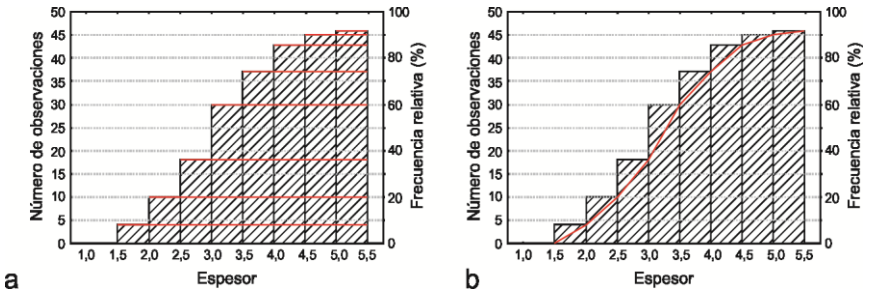


Figura 4. Relación entre a) Histograma de frecuencias acumuladas y el b) Polígono de frecuencias acumuladas.

Gráficos para representar simultáneamente dos o más variables

En el caso que se midan dos variables en el mismo espécimen y se necesite representar como es la variación de ambas simultáneamente se puede construir un **grafico de dispersión**, también llamado **diagrama bivariado** o ***scaterplot***. En este gráfico cada observación está representada por un par ordenado x_i,y_i cuyas coordenadas están dadas por los valores registrados en ambas variables (Fig. 5a).
El **diagramas de líneas** es una modificación del diagrama de dispersión al que se agregan segmentos de rectas uniendo los puntos según un orden dado, por ejemplo abscisas (Fig. 5b).

Los **gráficos cuantil - cuantil**, conocidos a veces como **Q-Q Plot** son gráficos bivariados especiales que permiten compara la distribución de frecuencias de una variable observada con una las de un modelo o distribución teórica. El Q-Q Plot normal es muy utilizado, en este caso el par x_i,y_i corresponden a los valores observados (ordenados de menor a mayor) y a los valores esperados según el modelo normal para una distribución con los mismos parámetros poblacionales (se verá con más profundidad que es un modelo Normal en el Capítulo 3). Cuando la distribución de la variable coincide con el modelo propuesto (distribución teórica) los puntos se alinean en una recta a 45º (Fig. 5c) y cuando presentan cambios de pendiente indican la presencia de más de una población. Cabe aclarar que se grafican los valores correspondientes a los cuantiles (este concepto se explicara más adelante).

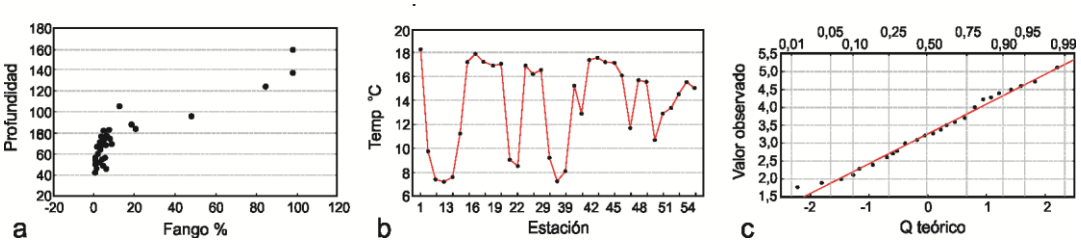


Figura 5. a) Diagrama de dispersión. b) Gráfico de líneas. c) Q-Q plot.

Los **diagramas de caja**, también llamados de **caja y bigote** ó ***box plot*** en inglés, se utilizan para presentar sintéticamente los aspectos más importantes de una distribución de frecuencias: posición, dispersión, asimetría, longitud de las colas, puntos anómalos. Son especialmente útiles para comparar varios conjuntos de datos. En los diagramas de caja se presenta en abscisas las categorías y en ordenadas los valores que toma la variable. Se construye una caja para cada categoría. En algunos *box plot* los extremos de la caja son los cuartiles (corresponden al 25% y 75% de los datos) y dentro de ella un punto o una línea horizontal correspondiente a la mediana (50% de los datos). A partir de los bordes superior e inferior de la caja se trazan líneas hasta los datos más alejados estos segmentos se llaman bigotes (Fig. 6). En otros diagramas de caja el punto medio es el promedio, los extremos de la caja el promedio mas/menos el desvío estándar y los bigotes son los datos anómalos.

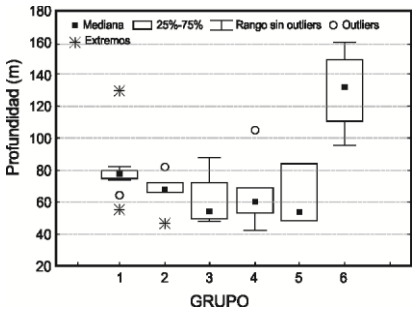


Figura 6. Diagramas de caja.

Los **diagramas ternarios** o **triangulares** indican simultáneamente la distribución de tres variables diferentes. Son muy útiles dado que permiten trabajar condiciones de proporcionalidad y de relaciones mutuas. La figura 7 muestra un diagrama ternario, cada vértice representa 100% de la variable en él indicada y la base opuesta el 0% de la misma. Los puntos interiores del triangulo muestran la mezcla de los tres componentes: A, B y C. Las líneas de proporcionalidad son paralelas a uno de los lados y mantienen fija en todos sus puntos la proporción del elemento que ocupa el vértice opuesto, independientemente de la relación entre las proporciones de los otros dos elementos.

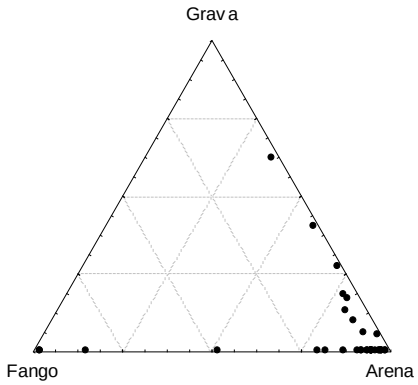


Figura 7. Diagrama ternario.

Entre los diagramas ternarios más difundidos se encuentran las clasificaciones de Streckeisen (1965) asumida por la IUGS para la clasificación de las rocas ígneas, el de clasificación de psamitas de Folk *et al*. (1970), los de Dott (1964) modificado por Pettijohn *et al*. (1972), los diagramas QFL y QmFLt de discriminación tectónica de áreas de aporte (Dickinson *et al*., 1983) y los diagramas Piper de composición química de aguas.

Descripción de conjuntos de datos: métodos numéricos

Los métodos gráficos son muy útiles para obtener una descripción general rápida de los datos recolectados y para su presentación, sin embargo no sirven para hacer inferencias porque no están bien definidos en el sentido que se pueden realizar, por ejemplo, diferentes histogramas a partir del mismo conjunto de datos. Para hacer inferencias se necesitan medidas rigurosamente definidas. Se buscan números transmitan una imagen de la distribución de frecuencias, medidas que informen sobre el valor central de la distribución, otras que describan la dispersión o variabilidad de los datos, medidas que indiquen la posición relativa de los datos y otras que den cuenta de la forma de la distribución.

Para describir la muestra y la población se utilizan las mismas funciones. Se habla de **estadística** o **estadístico** para referirse a cualquier función muestral, **parámetro** es el término que se usa para la

población. El tamaño de muestra se representa con *n*, los de la población con *N*. Los estadísticos se representan con mayúsculas con adornos ($\bar{X}$, $\tilde{Y}$, *S*, etc.), los parámetros con letras griegas (σ, ρ, μ, Φ, γ, etc.).

Medidas de tendencia central: Moda, Mediana, Media, Media Geométrica

Moda

La moda ($\hat{X}$) de una serie de datos es el valor que aparece con más frecuencia que cualquier otro. Una serie de datos puede no tener moda o tener más de una, si tiene dos modas se dice bimodal y polimodal si tiene más de dos. En los **datos sin agrupar** la moda se observa claramente cuando se ordenan los datos de menor a mayor. Cuando los **datos están agrupados** la moda se encuentra en la clase de mayor frecuencia, llamada clase modal (Fig. 8). Su valor se halla a partir de la siguiente expresión:

$$\hat{X} = L_{imo} + \left(\frac{\Delta 1}{\Delta 1 + \Delta 2} \right) C \qquad (2.1)$$

dónde L_{imo} es el límite inferior de la clase modal, $\Delta 1$ el valor absoluto de la diferencia entre la frecuencia de la clase **premodal** (inmediatamente anterior a la clase modal) y modal $|f_{mod} - f_{premod}|$, $\Delta 2$ el valor absoluto de la diferencia entre la frecuencia de la clase **posmodal** (inmediatamente posterior a la clase modal) y modal $|f_{posmod} - f_{mod}|$ y *C* la amplitud del intervalo de clase.

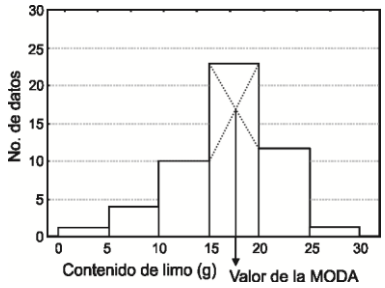


Figura 8. El valor de la moda es el x *correspondiente a la intersección de la línea que une el límite superior del intervalo premodal con el límite superior del intervalo modal con otra que une el límite inferior del intervalo modal con el límite inferior del intervalo posmodal.*

La moda es inestable ya que puede cambiar con el método de redondeo de los datos. En distribuciones que aumentan o disminuyen continuamente y a ritmo constante, la moda podrá ser un valor extremo más que un valor de tendencia central.

Mediana

La mediana ($\tilde{X}$) es el valor medio de una serie cuando los valores se ordenan de menor a mayor. Divide la serie de tal forma que el 50% de los valores son menores a él y el otro 50% de los valores son mayores a él. Una característica importante de la mediana es que no está influenciada con la magnitud de los valores de las colas de la distribución. Si el número de datos de la serie es impar el valor coincide con el valor central y cuando es par, la mediana se encuentra entre los dos valores centrales. Si los datos están agrupados la mediana se ubica en la clase mediana (la clase cuya frecuencia acumulada supera primero el valor [(n + 1) / 2]) (Fig. 9). Su valor se halla a partir de la siguiente expresión:

$$\tilde{X} = Lme + \left(\frac{[(n+1)/2] - fap}{fme} \right) C \qquad (2.2)$$

dónde *Lme* es el límite inferior de la clase mediana, *fap* la frecuencia acumulada en la clase que precede inmediatamente a la clase que tiene a la mediana, *fme* la frecuencia de la clase que tiene a la mediana y *C* la amplitud del intervalo.

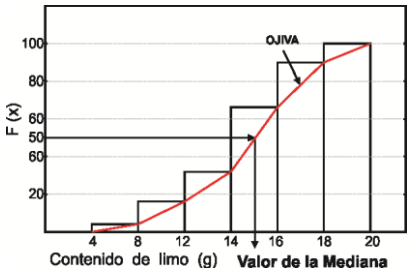


Figura 9. La mediana gráficamente se encuentra utilizando el polígono de frecuencias acumuladas. Desde el valor del 50% (eje y*) se traza una línea paralela a las* x *hasta cortar la ojiva y desde este punto se traza una línea que corte en forma perpendicular al eje de las* x.

Media aritmética

La media aritmética, promedio o simplemente media, es la medida de tendencia central más común y útil. El símbolo $\overline{X}$ se usa para la media de la muestra y la media de la población se representa con la

letra μ. Para un conjunto de *n* observaciones {x_1, x_2 ..., x_n}, es igual a las suma de las *n* observaciones dividido el número total de datos *n*,

$$\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \; . \tag{2.3}$$

Si los datos estén agrupados la media se calcula como:

$$\overline{X} = \sum_{i=1}^{n} \frac{x_i f_i}{n} \tag{2.4}$$

donde x_i es el valor de la variable discreta o el punto medio de intervalo de clase (marca de clase) si se trata de una variable continua y f_i es la frecuencia de la variable o intervalo de clase según corresponda.

La media tiene algunas propiedades importantes:

1° Es un valor típico. Esto significa que es el centro de gravedad, un punto de equilibrio. Su valor puede sustituir al valor de cada dato de la serie sin cambiar el total dado que

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \rightarrow \qquad \overline{X} \cdot n = \sum_{i=1}^{n} x_i \; .$$

2° La suma algebraica de las desviaciones con relación a la media es 0.

$$\sum_{i=1}^{n} \left(x_i - \overline{X}\right) = 0$$

3° La suma del cuadrado de las desviaciones de los datos respecto a la media es menor que las desviaciones al cuadrado de cualquier otro punto.

$$\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 = \min \; .$$

Una característica de la media es su inestabilidad pues con el agregado de datos extremos su valor cambia sustancialmente.

Uno de los de los objetivos de los estudios geológicos más frecuentes es estimar la media poblacional μ desconocida utilizando la información de la muestra. La precisión en la estimación de μ se basa en la independencia de las observaciones, como se verá más adelante, de esto depende la ley fundamental del error estándar de la media ($\sqrt{\sigma/n}$, dónde la varianza, σ^2 es la varianza de la población). Para lograr un error estándar pequeño solo son posibles dos caminos, reducir la varianza o incrementar el número de observaciones.

EJEMPLO 1

Cálculo de la moda, mediana y media

Se tienen datos de permeabilidad de una arena obtenidos de registros de pozo expresados en 10^5 miliDarcys.

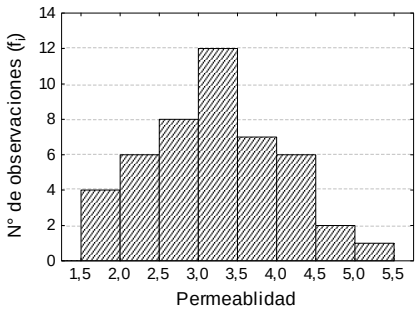
Límites de clase	*Marca de clase c_i*	f_i	fa_i	
1,5 - 2,0	1,75	4	4	7,00
2,0 - 2,5	2,25	6	10	13,50
2,5 - 3,0	2,75	8	18	22,00
3,0 - 3,5	3,25	12	30	39,00
3,5 - 4,0	3,75	7	37	26,25
4,0 - 4,5	4,25	6	43	25,50
4,5 -5,0	4,75	2	45	9,50
5,0 - 5,5	5,25	1	46	5,25

n= 46

Promedio

$$\sum_{i=1}^{n} c_i f_i = (1{,}75 \cdot 4) + \cdots + (5.25 \cdot 1) = 148$$

$$\overline{X} = \sum_{i=1}^{n} \frac{c_i f_i}{n} = \frac{148}{46} = 3{,}2 \; miliDarcys$$



Moda

Clase que contiene la moda: 3,0 – 3,5

Li_{mo}: 3,0 $\quad \Delta 1 = |8 - 12| = 4$

C = 2,5 $\quad \Delta 2 = |7 - 12| = 5$

$$\hat{X} = Li_{mo} + \left(\frac{\Delta 1}{\Delta 1 + \Delta 2}\right) C = 3{,}0 + \left(\frac{4}{4+5}\right) 2{,}5 = 3{,}6 \; miliDarcys$$

Mediana

n+1/2=47/2=23,5

Clase que contiene a la median: 3,0 – 3,5

Li_{me} = 3,0 $\quad$ *fme = 12*

fap = 18 $\quad$ *C = 2,5*

$$\tilde{X} = Li_{me} + \left(\frac{(n+1)/2 - fap}{fme}\right) C = 3{,}0 + \left(\frac{23{,}5 - 18}{12}\right) 2{,}5 = 4{,}1 \; miliDarcys$$

Media y moda se encuentran en el mismo intervalo de clase. La mediana es mayor que la moda y la moda es mayor que la media.

Media Geométrica

La media geométrica es otra medida de tendencia central cuyo uso más frecuente es el de promediar variables tales como porcentajes, tasas, números índices etcétera. Es la única medida de tendencia central que describe bien datos cerrados (Capítulo 12). Además tiene la virtud de ser mejor estimador de tendencia central que la media aritmética cuando la distribución de frecuencias es de asimetría a la derecha (ej. distribución log-normal).

La media geométrica, *G*, para un conjunto de *n* observaciones $\{x_1, x_2, ..., x_n\}$, es igual a la raíz *n*-ésima del producto de esos datos.

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}, \qquad G = \sqrt[n]{x_1^{f1} x_2^{f2} \ldots x_n^{fn}}. \qquad (2.5 \text{ y } 2.6)$$

Solamente se puede calcular la media geométrica si todos los datos son positivos, si uno de ellos es 0, entonces el resultado es 0. Aunque su cálculo es más complicado que el de la media aritmética, tiene la ventaja que se ve menos influenciada por los valores extremos. La media geométrica de un conjunto de números positivos es siempre menor a la media aritmética.

EJEMPLO 2

Cálculo de la media geométrica

La Tonalita la Ovejería aflora en la ladera oriental de la Sierra del Aconquija en la provincia de Tucumán. Los datos corresponden al contenido Na_2O (%) de las tonalitas.

{6,2; 9,3; 4,8; 7,2; 5,5}

$$G = \sqrt[5]{6{,}2 \cdot 9{,}3 \cdot 4{,}8 \cdot 7{,}2 \cdot 5{,}5} = \sqrt[5]{10960{,}0} = 6{,}4$$

Relaciones entre la media, la mediana y la moda

La mediana se ve afectada sólo ligeramente por datos atípicos que, como se verá más adelante, son valores alejados del resto de los datos, mientras que la media muestral se ve muy afectada por este tipo de valores. La media tiene la ventaja de tener en cuenta todas las observaciones, mientras que la mediana sólo tiene en cuenta el orden de las mismas y no su magnitud. La moda es una medida estable que no cambia con el agregado o perdida de un dato. La media y la mediana son muy distintas cuando la distribución es asimétrica, lo que implica la heterogeneidad de los datos.

En distribuciones simétricas, media, mediana y moda coinciden (moda = mediana = media) (Fig. 10a). En distribuciones con gran número de datos pequeños y pocos datos grandes, la asimetría es positiva o de cola derecha. La moda se encuentra a la izquierda, seguida por la mediana y la media (moda > mediana > media) (Fig. 10b). En distribuciones con muchos datos grandes y pocos datos pequeños, la asimetría es negativa o de cola izquierda. La media se encuentra a la izquierda, seguida por la mediana y la moda (media > mediana > moda) (Fig. 10c).

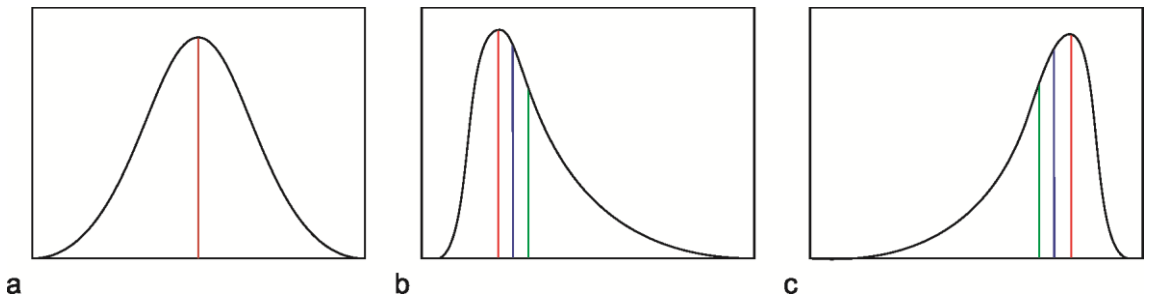


Figura 10. a. Distribución simétrica (moda = mediana = media). b. Distribución de cola derecha (moda > mediana > media). c. Distribución de cola izquierda (media > mediana > moda). En verde la media, en azul la mediana y en rojo la moda.

Medidas de dispersión: Rango, Varianza, Desvío estándar y Coeficiente de Variación

Rango

El rango, también llamado amplitud ó recorrido, para un conjunto de *n* observaciones $\{x_1, x_2, \ldots, x_n\}$, es la diferencia entre el valor máximo y el mínimo. El rango se aprecia tanto en los histogramas como en los diagramas de frecuencia acumulada (Fig. 11).

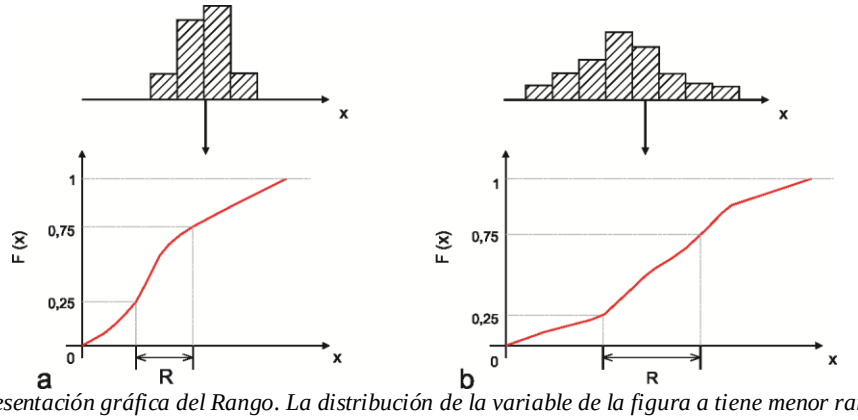


Figura 11. Representación gráfica del Rango. La distribución de la variable de la figura a tiene menor rango que la de la figura b.

Varianza y Desvío estándar

Una forma de medir la variación de los datos con respecto a la media es calculando la diferencia entre cada dato y la media, y luego promediar las diferencias:

$$(x_1 - \bar{X}), (x_2 - \bar{X}), \ldots, (x_n - \bar{X}) \rightarrow \frac{x_1 - \overline{X} + x_2 - \overline{X} + \ldots + x_n - \overline{X}}{n}, \text{ pero } \frac{x_1 + x_2 + \ldots + x_n}{n} - \overline{X} = 0$$

luego, la desviación promedio siempre es nula (ver 2º propiedad de la media).

Para salvar el problema se toma el promedio del cuadrado de las diferencias que es la **varianza**

$$\frac{(x_1 - \bar{X})^2, (x_2 - \bar{X})^2, \ldots, (x_n - \bar{X})^2}{n},$$

a veces llamada **variancia**. La varianza de todas las mediciones de la población, el parámetro, se representa con σ^2 y la varianza de la muestra, el estadístico, con S^2.

La varianza de *n* observaciones (x_1, x_2,…, x_n) se define como el promedio del cuadrado de las desviaciones con respecto a la media.

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{N} \text{ , o bien } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 . \qquad (2.7)$$

La varianza muestral no es un buen estimador de la varianza poblacional, la subestima. Esa desviación se puede corregir disminuyendo el denominador del cociente, por esto, en el cálculo del promedio de las desviaciones se resta una unidad el tamaño de la muestra (n - 1)[1].

$$S^2 = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + ... + (x_n - \overline{X})^2}{n-1} \quad . \qquad (2.8)$$

Debido a que la varianza es una suma de cuadrados la unidad de S^2 y σ^2 es igual a la unidad de la variable elevada al cuadrado, por ejemplo si la variable se mide en milímetros, las unidades de S^2 son mm^2. Para expresar la dispersión en la misma unidades que la variable y simplificar la interpretación se define la **desviación** ó **desvío estándar**. El desvío estándar de n observaciones (x_1, x_2,..., x_n), es la raíz cuadrada positiva de la varianza, $\sigma = \sqrt{\sigma^2}$ y $S = \sqrt{S^2}$.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad , \qquad S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad (2.9 \text{ y } 2.10)$$

En el caso en que los datos estén agrupados se consideran las frecuencias de modo que

$$S = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{X})^2 f_i}{n-1}} \quad , \qquad (2.11)$$

donde x_i es el valor de la variable discreta o el punto medio del intervalo de clase en las variables continuas y fi es la frecuencia.

El desvío estándar se registra con un decimal más que los datos, sin antes redondear la varianza. La varianza se expresa con dos decimales más que los datos.

Algunas propiedades interesantes del desvío estándar son:

1° El desvío estándar no se modifica cuando se suma una constante a todos los valores de la variable.

2° El desvío estándar se incrementa el valor de la constante cuando todos los valores de la variable se multiplican por una constante.

Otro objetivo importante en los trabajos geológicos, además del cálculo del promedio, es la medición de la variabilidad. La variabilidad puede ser causada, como mencionamos en el capítulo anterior, durante la formación del cuerpo de roca.

Coeficiente de variación

El coeficiente de variación para una muestra de valores x_1, x_2, ... , x_n es la razón entre su desvío estándar y la media de esos datos.

$$CV = \frac{S}{\overline{X}} \qquad (2.12)$$

El Coeficiente de Variación no posee unidades aunque suele expresarse en forma porcentual.

Este cociente da cuenta de la desviación estándar como una proporción de la media, y es a veces un indicador bastante útil. Por ejemplo, un valor S = 10 no es significativo a menos que se lo compare con algo diferente. Si S = 10 y $\overline{X}$ = 1000, entonces la variación es muy pequeña con respecto a la media. Sin embargo, si S = 10 y $\overline{X}$ = 5, la variación con respecto a la media es grande. Por ejemplo, si se estudia la precisión (variación en mediciones repetidas) de un instrumento de medición, por ejemplo una Estación Total, el primer caso CV = 10/1000 = 0,01 presentaría una precisión aceptable, pero en el segundo caso CV = 10/5 = 2 sería totalmente inaceptable.

En el caso de datos de campo o de laboratorio, el coeficiente de variación refleja una mezcla desconocida de la variabilidad natural, la variabilidad introducida durante el proceso de muestreo y de causas aleatorias. El coeficiente de variación de una población homogénea es típicamente menor que la unidad. Si es mayor que 1,5 conviene investigar posibles fuentes de heterogeneidad en los datos, también puede indicar la existencia de valores extremos. A pesar de esto, en numerosas variables geológicas el coeficiente de variación toma valores entre 2,5 y 0,2.

El coeficiente de variación también resulta útil para comparar la variabilidad entre varias muestras, incluso la variabilidad entre mediciones realizadas en diferentes unidades. Pero cuando la media es cercana a cero no es útil calcular el coeficiente de variación.

EJEMPLO 3

Cálculo de la varianza desvío estándar y coeficiente de variación

Se utilizan los datos del ejemplo 1, los datos de permeabilidad de una arena obtenidos de registros de pozo expresados en 10^5 miliDarcys.

Límites de clase	*Marca de clase* c_i	f_i	$(c_i - \overline{X})^2 f_i$
1,5 - 2,0	1,75	4	8,6129
2,0 - 2,5	2,25	6	5,6151
2,5 - 3,0	2,75	8	1,7476
3,0 - 3,5	3,25	12	0,0128
3,5 - 4,0	3,75	7	1,9857
4,0 - 4,5	4,25	6	6,3977
4,5 -5,0	4,75	2	4,6978
5,0 - 5,5	5,25	1	4,1315

n = 46

$\overline{X} = 3{,}2\ miliDarcys$

Varianza

$$S^2 = \frac{(c_1 - \overline{X})^2 f_1 + (c_2 - \overline{X})^2 f_2 + \ldots + (c_n - \overline{X})^2 f_n}{n-1} = \frac{33{,}2011}{45}$$

$S^2 = 0{,}7378\ miliDarcys^2$

S

Desvío estándar

$S = \sqrt{S^2} = \sqrt{0{,}7378}$

$S = 0{,}86\ miliDarcys$

Coeficiente de variación

$$CV = \frac{S}{\overline{X}} = \frac{0{,}86}{3{,}2}$$

$$CV = 0{,}27$$

La distribución tiene baja variabilidad.

Medidas de localización

Los cuantiles son valores de la distribución que la dividen en partes iguales. Los más usados son los **cuartiles** (*Q*), los **deciles** (*D*) y los **percentiles** (*P*), que dividen a los datos en 4, 10 y 100 partes iguales respectivamente. El segundo cuartil (Q_2 ó $Q_{50\%}$), quinto decil (D_5) y percentil 50 (P_{50}) son equivalentes a la mediana.

De igual manera que la mediana, un cuantil x_δ divide a la muestra de datos en dos partes, el δ% de los valores es menor que δ y el (1 – δ) de los valores es mayor que x_δ.

El cálculo de estos cuantiles es similar al de la mediana. Para datos no agrupados la serie se ordena de menor a mayor y se buscan los valores que satisfacen la condición buscada. Si los datos están agrupados, se determina el intervalo en el cual se encuentra la medida buscada, luego se recorre la columna de frecuencias acumuladas y el primer valor que sobrepasa el % de *n* a la izquierda de la medida indica el intervalo mencionado. Después se efectúa la interpolación aplicando

$$x_{\delta\%} = L\delta\% + \left(\frac{\delta\% - fap}{fm\delta}\right)C \qquad (2.13)$$

dónde *Lδ%* es el límite inferior de la clase *δ%*, δ% es el total de observaciones que quedan a la izquierda de *δ%*, *Fap* la frecuencia acumulada en la clase que precede inmediatamente a la clase que tiene al *δ%*, *fmδ* la frecuencia de la clase que tiene al *δ%* y *C* la amplitud del intervalo (Fig. 12).

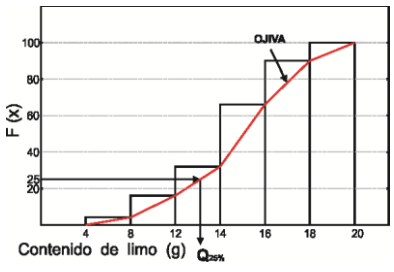


Figura 12: Se indica cómo encontrar el valor del primer cuartil $Q_{25\%}$.

Rango intercuartilico. Algunas medidas de dispersión toman como referencia ciertos cuartiles o deciles de orden δ. El rango intercuartilico es la diferencia entre el tercer cuartil ($Q_{75\%}$) y el segundo cuartil ($Q_{25\%}$) ($RI = Q_{75\%} - Q_{25\%}$). Otros proponen utilizar la diferencia entre el percentil noventa ($P_{90\%}$) y el percentil 10 ($P_{10\%}$) ($RI = X_{90\%} - X_{10\%}$).

Medidas de forma

Coeficiente de Simetría

El coeficiente de simetría (*CS*) se utiliza para caracterizar el comportamiento de la distribución respecto a la media (Fig. 13). La simetría es importante para saber si los valores de la variable se encuentran en una determinada zona del recorrido de la variable.

Existen varias formas de medir la simetría. La primera es comparando la media con la moda. Si la diferencia media menos moda $(\overline{X} - \hat{X})$ es positiva, la asimetría se dice positiva o derecha, en caso que la diferencia sea negativa la asimetría es negativa o a izquierda. Aunque simple esta manera de medir la asimetría es poco práctica pues depende de las unidades de la variable. El segundo modo de medir la asimetría es calculando el Coeficiente de simetría de Pearson que considera el cubo de las desviaciones a la media. Se define como:

$$CS = \frac{\sum_{i=1}^{n} \frac{(x_i - \overline{X})^3 f_i}{n}}{S^3}. \qquad (2.14)$$

Se ha demostrado que cuando el coeficiente es negativo ($CS < 0$) la asimetría es negativa (Fig. 13a), cuando es positivo ($CS > 0$) la asimetría es positiva (Fig. 13b) y cuando es igual a cero ($CS = 0$) la distribución es simétrica (Fig. 13c).

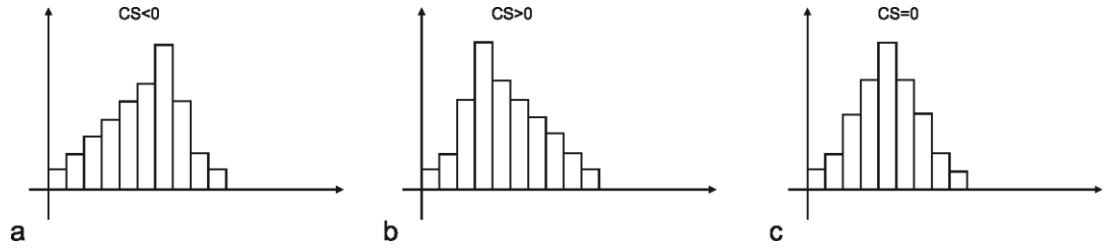


Figura 13. Ejemplos de distribuciones de frecuencias con diferentes asimetrías. a) Histograma con asimetría negativa. b) Histograma con asimetría positiva. c) Histograma simétrico.

Coeficiente de Kurtosis

El coeficiente de Kurtosis, *K*, también llamado de Exceso o de Curtosis, mide el grado de achatamiento de la distribución con respecto al modelo teórico Normal (o de forma de campana al que nos referiremos más adelante) (Fig. 14). Se define como:

$$K = \frac{\sum_{i=1}^{n} \frac{(x_i - \bar{X})^4 f_i}{n}}{S^4} - 3. \qquad (2.15)$$

Se ha demostrado que cuando en las distribuciones normales la kurtosis es cero ($K = 0$) (Fig. 14a). La kurtosis es positiva ($K > 0$) en las distribuciones más puntiagudas que la del modelo normal, se dice que son leptocurticas (Fig. 14b). Si la distribución es más achatada que la del modelo normal, la kurtosis es negativa ($K < 0$) y se las llama platicurticas (Fig. 14c).

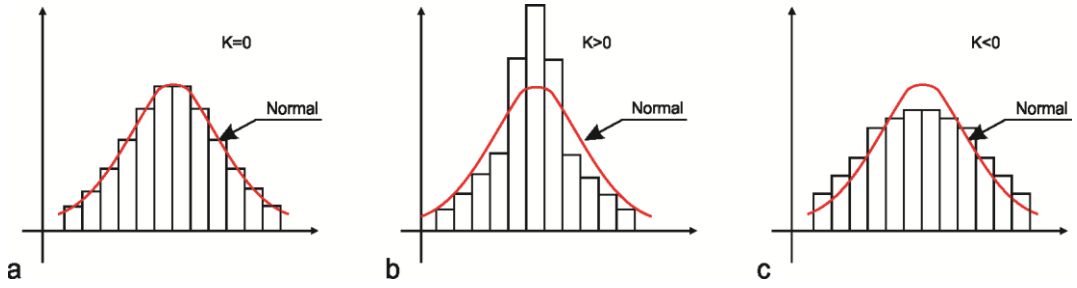


Figura 14. Ejemplo de distribuciones de frecuencias con diferentes kurtosis. a) Histograma normal o mesocurtico. b) Histograma puntiagudo o leptocurtico. c) Histograma achatado o platicurtico.

EJEMPLO 4

Cálculo del coeficiente de simetría y la kurtosis

Se utilizan los datos del ejemplo 1, los datos de permeabilidad de una arena obtenidos de registros de pozo expresados en 10^5 miliDarcys.

Límites de clase	*Marca de clase* c_i	f_i	$(c_i - \bar{X})^3 f_i$	$(c_i - \bar{X})^4 f_i$
1,5 - 2,0	1,75	4	-12,6386	18,5457
2,0 - 2,5	2,25	6	-5,4320	5,2548
2,5 - 3,0	2,75	8	-0,8168	0,3818
3,0 - 3,5	3,25	12	0,0004	0,0001
3,5 - 4,0	3,75	7	1,0576	0,5633
4,0 - 4,5	4,25	6	6,6063	6,8217
4,5 -5,0	4,75	2	7,1999	11,0346
5,0 - 5,5	5,25	1	8,3977	17,0693

$n = 46$

$S = 0{,}87$ *miliDarcys*

Coeficiente de simetría

$$CS = \frac{\sum_{i=1}^{n} \frac{(c_i - \bar{X})^3 f_i}{n}}{S^3} = \frac{4{,}3745}{0{,}6337} = 0{,}15$$

$CS = 0{,}15$

Kurtosis

$$K = \frac{\sum_{i=1}^{n} \frac{(c_i - \bar{X})^4 f_i}{n}}{S^4} - 3 = \frac{59{,}6712}{0{,}5444} - 3$$

$K = -0{,}62$

La distribución tiene ligero sesgo positivo y es ligeramente platicurtica.

Datos Anómalos

Los datos anómalos, datos atípicos u *outliers* se pueden producir por un error de medición o de recuento, un error de transcripción al momento de volcar los datos al papel, o bien pueden ser causados por algún suceso sumamente extraño.

Un criterio que se utiliza para detectar los datos anómalos de un conjunto determinado de datos tiene en cuenta los cuartiles (*Q*). Así se define a los datos anómalos moderados como aquellos que son menores al valor de $Q_{25\%} - 1{,}5\ (Q_{75\%} - Q_{25\%})$ o mayores que $Q_{75\%} + 1{,}5\ (Q_{75\%} - Q_{25\%})$. En tanto los anómalos extremos son los menores a $Q_{25\%} - 3\ (Q_{75\%} - Q_{25\%})$ o los mayores a $Q_{75\%} + 3\ (Q_{75\%} - Q_{25\%})$, donde $Q_{25\%}$ es el primer cuartil, $Q_{75\%}$ el tercer cuartil y $(Q_{75\%} - Q_{25\%})$ el rango intercuartil.

Otros criterios para identificar datos extremos son gráficos, usando box-plot, Q-Q plots e histogramas.

Es muy importante identificar los datos atípicos que influencian fuertemente los resultados de un análisis estadístico clásico, sin ir muy lejos, se mencionó el efecto que tienen sobre la media. Se debe inspeccionar cuidadosamente los valores anómalos para ver si son producto de un error y deben ser eliminados o si, por el contrario, el dato corresponde a un individuo que tiene algo particular y debe permanecer en el análisis. Por otra parte, si se encuentran muchos datos anómalos pueden estar indicando que la escala elegida no es la más adecuada.

Tratamiento de datos cero (0)

Existen tres tipos de datos cero. En el primero la variable toma el **valor 0**, por ejemplo no se encuentran prospectos mineros en un distrito o no existen pozos contaminados con mercurio en una región rural. En estos casos el cero se incluyen para los cálculos de los estadísticos con la jerarquía de cualquier otro valor. El segundo tipo se conoce como **0 por redondeo**. Estos ceros son muy comunes en geología, suelen estar relacionados con el límite de detección del aparato o metodología utilizada para cuantificar los valores de la variable. Su aparición es frecuente en datos de geoquímica de roca, tanto de elementos mayoritario como traza. En las tablas de datos se indican como "< valor" (< 0,05) o "-valor" (por ejemplo -0,05). Una estrategia habitual para el cálculo de los estadísticos, es reemplazar los ceros por redondeo utilizando la mitad del valor del límite de detección, por ejemplo si el límite de detección es 0,01 se reemplazan por 0,005. En el último tipo, los valores 0 indican **dato ausente**, no se trata de ceros verdaderos, sólo es una forma de indicar que no se midió la variable en el espécimen o que se perdió la información. Los estadísticos se calculan omitiendo los ceros, sólo con los valores disponibles, es decir se disminuye el tamaño de la muestra.

Aplicaciones especiales

Estadísticos y gráficos como el polígono de frecuencias acumulado y gráficos bivariados son utilizados en todas las ramas de la Geología. Una aplicación muy difundida para estudios granulométricos de psamitas fue propuesta por Folk y Ward (1957). Estos autores utilizan los gráficos de frecuencia acumulada que surgen de los análisis granulométricos de sedimentos actuales (en abscisas se indica el tamaño de grano en escala ϕ (-$\log_2$ diámetro mm) y en ordenadas el peso retenido en el tamiz), y calculan a partir de ellos, los estadísticos de la siguiente manera: $\hat{X} = \phi 50$,

$\overline{X} = \frac{\phi 16 + \phi 84 - \phi 50}{3}$, $S = \frac{\phi 84 - \phi 16}{4} + \frac{\phi 95 - \phi 5}{6{,}6}$, $CS = \frac{\phi 16 + \phi 84 - 2\phi 50}{2(\phi 84 - \phi 16)} + \frac{(\phi 5 + \phi 95 - 2\phi 50)}{2(\phi 95 - \phi 5)}$, $K = \frac{\phi 95 - \phi 5}{2{,}44(\phi 75 - \phi 25)}$.

Estos estadísticos se utilizan en el análisis textural de sedimentos. Con ellos también se construyen gráficos bivariados, por ejemplo a base de graficar kurtosis vs. asimetría Folk y Ward (1957) interpretaron detalles ambientales y Mason y Folk (1958) diferenciaron arenas de playa, duna y planos eólicos en una barrera litoral. Passega (1957, 1964) graficó *C* (percentil 1, aproximadamente el valor del máximo tamaño de grano) en función de *M* (media) para tratar de determinar los agentes o procesos depositacionales. Visher (1969) utilizó los gráficos de frecuencias acumuladas probabilísticos para distinguir subpoblaciones de sedimentos a partir de los cambios de pendiente y las vinculó con mecanismos de transporte (diferenció tres segmentos, que de fino a grueso corresponden a suspensión, saltación y tracción).

Por otra parte, los histogramas de altura de una cuenca y los polígonos de frecuencia acumulada, llamadas curvas hipsométricas en los estudios de hidrogeología y geomorfología cuantitativa, permiten caracterizar la fisiografía de la cuenca (Fig. 15). Por ejemplo una curva hipsométrica con concavidad hacia arriba indica una cuenca con valles extensos y cumbres escarpadas y una con concavidad hacia abajo revela valles profundos y sabanas planas.

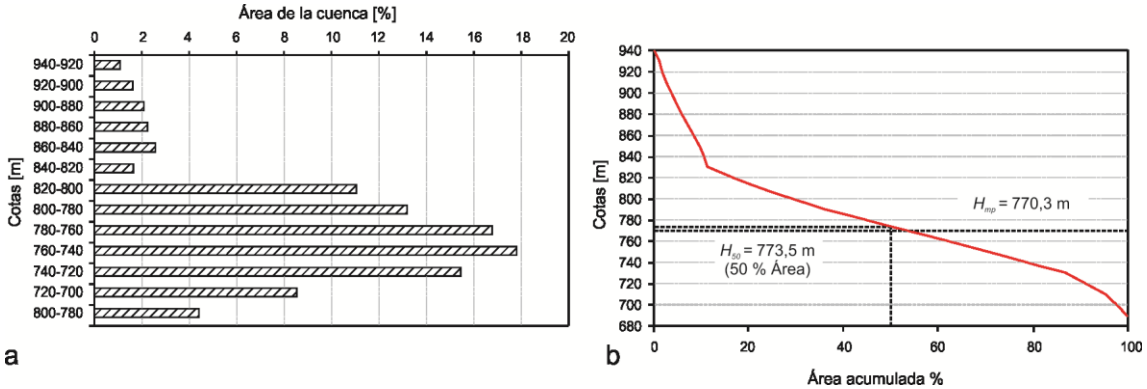


Figura 15. a. Polígonos de frecuencia de cotas en una cuenca. b) Curva hipsométrica.

PROBABILIDES

FENÓMENOS GEOLÓGICOS EN EL CONTEXTO DE LA TEORÍA DE PROBABILIDADES

Introducción

Gran parte de los trabajos geológicos, profesionales o de investigación, se basan en la observación y análisis de procesos geológicos en los que intervienen variables físicas, químicas y ambientales para originar todo el vasto conjunto de productos geológicos (cuerpos de rocas, estratos, depósitos minerales, erupciones volcánicas, procesos de remoción en masa, inundaciones, entre otras). Desafortunadamente el conocimiento de la génesis de los productos y procesos es limitado, se desconoce gran parte de las complejas interacciones que ocurren entre ellos, además se suma el desconocimiento causado por la inaccesibilidad o lo errático de los afloramientos. Si bien se pueden clasificar contar y medir los productos originados durante estos procesos no es posible predecir con exactitud el valor, tamaño, cantidad, tipo, etc. que tendrá una observación lo que impide la utilización de **modelos determinístcos**[2].

Sin embargo aunque los procesos no siempre son exactamente iguales, en términos generales, es posible encontrar en ellos algunas regularidades que permitan usar conceptos estadísticos y modelos probabilísticos. La utilización de modelos probabilísticos en geología permite analizar la información como el resultado de un proceso aleatorio y entonces estimar los valores desconocidos y predecir, con un margen de error conocido, el comportamiento que tendrán procesos y/o productos en sitios no muestreados. Es importante señalar que aleatorio no significa impredecible, tampoco significa que al considerar que un problema geológico como un fenómeno aleatorio se admitan estimaciones carentes de sentido.

Por otra parte, tanto la estimación de los parámetros como la toma de decisiones se sustentan en la teoría de probabilidades debido a que la probabilidad es simultáneamente el lenguaje y la medida de la incertidumbre y los riesgos asociada a ella. Antes de pasar a los procedimientos estadísticos en la toma de decisiones es entonces necesario describir la teoría de probabilidades.

Probabilidades

Es posible arribar al concepto de probabilidades desde diferentes aproximaciones, se verán tres: basada en la experiencia, clásica y axiomática.

Definición empírica de probabilidad

Es claro que los estudios geológicos se basan fundamentalmente en las observaciones. En las observaciones hay una gran parte de incertidumbre que impide que sean predecibles. Por ejemplo al realizar observaciones o mediciones de un cristal de cuarzo de un granito no se puede predecir con certeza su tamaño, ni es posible saber cual será el contenido metálico en cobre de un depósito, ni la precipitación caída en una cuenca en un lapso dado, o el número de bloques explotables por depósito, ni el tonelaje por sobre una ley de corte en una mina. Sin embargo, la experiencia muestra que en algunas situaciones prácticas, las frecuencias de ocurrencia repetidas de un determinado fenómeno son, en términos generales parecidas. Por ejemplo el tamaño de cristales de cuarzo cuerpos graníticos similares se desarrollaran aproximadamente lo mismo, la precipitación caída en una cuenca anualmente será semejante año a año, la ley de corte de los bloques de un yacimiento en explotación será similar, etcétera. Esta regularidad es mayor si se consideran series que comparten una gran cantidad de observaciones. Cuando esto sucede, la frecuencia de ocurrencia, expresada como frecuencia relativa, se transforma en la **probabilidad de ocurrencia**. Es esta verificación experimental que conduce a interpretar **la probabilidad como el límite de la frecuencia relativa que resultaría de considerar una serie infinita de experiencias**.

Dado que la probabilidad de un suceso cuaquiera, *A*, tiende a coincidir con la frecuencia experimental cuando el número de repeticiones del experimento es lo suficientemente grande, es que la probabilidad de *A* se define como el límite de la frecuencia relativa que resultaría de considerar una serie infinita de experiencias.

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}, \qquad (3.1)$$

donde n_A es el número de veces que ocurre *A* en una serie de *n* repeticiones del experimento.

Esta definición frecuentista de la probabilidad se llama también probabilidad *a posteriori* ya que sólo se puede dar la probabilidad de un suceso después de repetir y observar un gran número de veces el experimento. Algunos autores las llaman probabilidades teóricas.

Definir las probabilidades de esta manera permiten vincularlo intuitivamente con lo visto en el capítulo anterior. Imagine una distribución de frecuencias empíricas de un número muy grande de observaciones con intervalos de clase infinitamente pequeños. Para esa distribución el polígono de frecuencia sería sumamente suave y representaría todas las potenciales observaciones, por lo tanto, se podría decir que se trata de la distribución de la Población. Esta curva, entonces puede considerarse

una representación teórica que se define a partir de un modelo matemático[3] de las potenciales observaciones. Contar con modelos permite desarrollar métodos estadísticos y, de este modo, a partir del análisis de las probabilidades asociadas con los eventos poder estimar el comportamiento pasado o futuro de los objetos o de los fenómenos bajo estudio.

Definición clásica de probabilidad de La Place[4] (1812)

El concepto de probabilidades empíricas es necesario en circunstancias geológicas como las descriptas más arriba así como en muchas otras circunstancias en donde existe una variedad (a veces infinitas) de posibles resultados, que impiden que no se pueda predecir con exactitud qué sucederá. Sin embargo, en otras ocasiones esta definición no es aplicable, por ejemplo cuando se cuenta la presencia de ejemplares de una cierta especie de bivalvo en un estrato que poseen ornamentación respecto al número total de bivalvos de esa especie. En circunstancias parecidas a esas se puede recurrir a la aproximación de probabilidades desde la **frecuencia relativa** (definición de La Place, 1812). Esta aproximación es intuitivamente utilizada por los geólogos ya que se trata de un concepto cercano al del Uniformitarismo (Charles Lyell[5], 1797-1875).

Según La Place la probabilidad de *A* es igual al número de casos favorables a *A* sobre número de casos totales.

$$P(A) = \frac{k}{n}, \tag{3.2}$$

dónde *k* es el número de casos posibles a *A* en una serie de *n* casos totales siempre que todos los casos sean **equiprobables**.

La definición de La Place también se conoce con el nombre de probabilidad *a priori* pues, para calcularla, es necesario conocer antes de realizar el experimento aleatorio, el número de resultados posibles.

Este concepto de probabilidades de frecuencias relativas se relaciona con lo que en geología se conoce como probabilidades geométricas. Las **probabilidades geométricas** *P*(*A*) se calculan como una razón de longitudes, áreas o volúmenes. Suponga que la figura 1 representa un mapa geológico donde el área blanca corresponde a una litología y la gris a otra diferente, si se elige al azar (sin apuntar) un punto, la probabilidad que corresponda a la litología gris es el cociente entre el área ocupada por la litología gris *s* y el ocupado por la litología blanca *S*, $P(A) = s/S$ siempre que los casos sean equiprobables (Fig. 1).

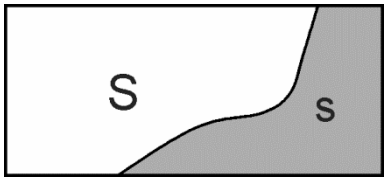


Figura 1.Probabilides geométricas.

Definición axiomática de la probabilidad

El cálculo de probabilidad que proviene de axiomas[6] fue desarrollado por Kolmogorov[7] en 1933. Según este matemático, se llama probabilidad de un suceso *A* a un número real, P(*A*), que verifica los siguientes axiomas:

AXIOMA 1: La probabilidad de *A* es mayor o igual a cero, $P(A) \geq 0$. Es decir la probabilidad es un valor que se encuentra entre cero y uno, $0 \leq P \leq 1$. Cuando la probabilidad de *A* es cero, el suceso es imposible. Cuando la probabilidad de *A* es uno, el suceso es seguro.

AXIOMA 2: La probabilidad de todos los puntos que forman el espacio muestral es igual a uno, P(S) = 1.

AXIOMA 3: Si *A* y *B* pueden ocurrir simultáneamente, la probabilidad que ocurra *A* o *B* está dada por $P(A \cup B) = P(A) + P(B) - P(A \cap B)$; $(A \cap B) \neq \emptyset$. Si A y B son sucesos incompatibles, es decir no puede ocurrir simultáneamente, $A \cap B = \emptyset$, entonces $P(A \cup B) = P(A) + P(B)$.

Incertidumbre, proceso aleatorio y conceptos relacionados

Para introducir en la teoría de probabilidades es necesario presentar algunos conceptos.

Se entiende por **Fenómenos aleatorios**[8] aquellos en los que las mismas causas dan lugar a resultados diferentes. Si bien los datos geológicos no son ciertamente el resultado de procesos aleatorios, pensarlos como fenómenos aleatorios resulta una herramienta útil al momento de resolver problemas de estimación. Por ejemplo si se considera los procesos que dan lugar a un depósito mineral o a un reservorio de petróleo, es innegable que son extremadamente complicados, y que el conocimiento de ellos es tan pobre que esta complejidad se presenta como una conducta aleatoria, aunque está claro que no significa que el fenómeno sea regido por el azar, es más bien una muestra del desconocimiento que se tiene del fenómeno.

Los **Experimentos aleatorios** son experiencias cuyo resultado depende del azar, es decir, pueden variar cuando el experimento se repite en condiciones supuestamente idénticas. Se considera a los datos obtenidos de una unidad experimental como el resultado de un proceso o fenómeno aleatorio. Por ejemplo tirar un dado y mirar el número que aparece en su cara superior es un experimento

aleatorio y lo es también medir el tamaño de cristales de feldespato que se presentan en un corte delgado en una riolita, pues no todos los cristales medirán exactamente lo mismo, así como inspeccionar los resultados de los análisis químicos por cobre de minerales de un yacimiento ya que los resultados de los análisis de cada muestra mineral no serán exactamente los mismos, sino que variaran de muestra en muestra.

El **Resultado** es la información aportada por la realización de una experiencia, el conjunto de todos los resultados posibles de un experimento se llama **Espacio Muestral**. El espacio muestral también se define como el conjunto de todos los sucesos simples o puntos de muestra que resultan de un experimento. Se suele designar con *S* y representar con un conjunto, diagrama de Venn, sistema de ejes cartesianos, o con un árbol.

$S = \{1, 2, 3, 4, 5, 6\}$ para el lanzamiento de un dado.

$S = \{x: x > 0\}$ si se mide el tamaño de los feldespatos o el contenido metálico de un mineral.

Estrechamente relacionado al experimento se encuentra la noción de variable aleatoria. **Variable aleatoria** es una función que asigna a cada punto del espacio muestral un número real $\Re$ (se aborda este concepto detalladamente más adelante en este capítulo).

Una **muestra** de *n* resultados es el conjunto de valores tomados por la variable aleatoria durante *n* experimentos, ejemplos:

$M = \{3, 5, 6, 6, 5, 2, 3, 2\}$ si se tira un dado 8 veces,

$M = \{2{,}0; 3{,}3; 2{,}1; 2{,}5; 2{,}9\}$ si se miden 4 cristales de feldespato de un granito,

$M = \{0{,}39; 1{,}02; 0{,}50; 0{,}30; 0{,}89\}$ si se analizan 5 muestras por cobre.

Suceso o **Hecho** es cualquier subconjunto del espacio muestral. Un suceso es un subconjunto específico de puntos de muestra. Para un dado que en una tirada salga 6 (E_6), que salga par (A_{par}) o que salga menor a 4 ($B_{>4}$) en su cara superior:

$E_6 = \{6\}$ $A_{par} = \{2, 4, 6\}$ $B_{>4} = \{1, 2, 3\}$

Entre los sucesos se distinguen varios tipos:

a) **Suceso simple** es el resultado de un experimento que no puede ser descompuesto, es decir posee un único elemento del espacio muestral, para el dado el suceso, por ejemplo, E_6.

b) **Suceso compuesto** es el resultado de un experimento que puede ser descompuesto en sucesos simples. Contiene más de un elemento del espacio muestral como A_{par} y $B_{>4}$.

c) **Suceso seguro** es aquel que siempre ocurre; se ve claramente que el espacio muestral y suceso seguro son lo mismo $P(S) = 1$.

d) **Suceso imposible** es aquel que nunca ocurre, es el conjunto vacio Ø.

$I = \emptyset \rightarrow P(I) = 0$

E_7= *{lanzar un dado y sacar 7}*

e) **Suceso complementario** o **contrario** son todos los puntos de muestra que están en *S* y no están en el suceso *A*. Es el suceso contrario de *A*, ocurre cuando *A* no ocurre, se simboliza como $\bar{A}$.

A_{par}= {2, 4, 6} $\rightarrow \bar{A}_{no\ par}$ = {1, 3, 6}

Como $\sum$P = 1 $\rightarrow$ P(A) + P(Ā) = 1

Cuando $A = \bar{A} \rightarrow S = \emptyset$

f) **Sucesos equiprobables** son aquellos que poseen igual probabilidades asociadas a cada uno de los puntos del espacio muestral. Por ejemplo en un dado todas las caras tienen la misma probabilidad de ocurrencia. Esta probabilidad se puede calcular apelando a la definición de probabilidades clásica. Ya que el dado tiene 6 cara, son 6 los casos posibles y el espacio muestral S = {1, 2, 3, 4, 5, 6}, la probabilidad de una cara cualquiera es 1/6.

g) **Sucesos no equiprobables** son aquellos en los que los puntos del espacio muestral tienen diferentes probabilidades. Por ejemplo tirar dos dados y registrar el número mayor. El espacio muestral y las probabilidades asociadas son:

S		**Dado 1**					
		1	**2**	**3**	**4**	**5**	**6**
Dado 2	**1**	1	2	3	4	5	6
	2	2	2	3	4	5	6
	3	3	3	3	4	5	6
	4	4	4	4	4	5	6
	5	5	5	5	5	5	6
	6	6	6	6	6	6	6

X	**P(x)**
1	1/36
2	3/36
3	5/36
4	7/36
5	9/36
6	11/36

h) **Sucesos mutuamente excluyentes** son dos sucesos que no pueden ocurrir simultáneamente. Por ejemplo si al lanzar un dado se definen los sucesos A_{par} (2, 4 , 6) y B_{impar} (1, 3, 5) ambos no tienen elementos en común (A $\cap$ B = $\varnothing$).

Sumar probabilidades

Cuando se quiere conocer la probabilidad que ocurra un suceso *A* u ocurra un suceso *B* y si *A* y *B* pueden ocurrir simultáneamente, entonces la probabilidad de que ocurra ***A ó B*** es la suma de las probabilidades individuales de esos eventos menos la probabilidad que ambos ocurran simultáneamente.

$$P(A \text{ ó } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B), \text{ para } A \cap B \neq \varnothing. \quad (3.3)$$

Por ejemplo si el suceso A_{par} = {2, 4 , 6} y el $B_{>3}$ = {4, 5, 6} ambos tienen dos elementos en común, A $\cap$ B = {4, 6}. La probabilidad de que ocurra un número par o un número mayor a 3 es

$P(A \cup B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3}$.

Si A y B son **sucesos incompatibles**, es decir no puede ocurrir simultáneamente, A $\cap$ B = Ø, la probabilidad de que ocurra **A ó B** es la suma de las probabilidades individuales de esos eventos es

$$P(A \cup B) = P(A) + P(B). \quad (3.4)$$

Por ejemplo si el suceso $A_{<3}$ = {1, 2} y el $B_{>3}$ = {4, 5, 6} no existen elementos en común (A ∩ B = ∅). La probabilidad de que ocurra un número menor a 3 o un número mayor a 3 es $P(A \cup B) = \frac{2}{6} + \frac{3}{6} = \frac{5}{6}$.

Estas situaciones pueden extenderse a más de dos eventos.

Multiplicar probabilidades, Probabilidad condicional y Sucesos independientes

Si se quiere conocer la probabilidad de dos sucesos que ocurren simultáneamente se hace foco en los eventos que ambos sucesos tienen en común, esto es la intersección. La fórmula que se emplea para el cálculo lleva a la definición de **probabilidad condicional** y a la de **sucesos dependientes**. Dos sucesos son dependientes cuando la probabilidad de ocurrencia de uno cambia, afecta o depende de la ocurrencia de otro. Sea *B* un suceso que sabemos que ha ocurrido, la probabilidad condicional de un suceso *A* dado que ha ocurrido *B*, se escribe (A|B), se define como

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{que puede reescribirse como } P(A \cap B) = P(A|B) \cdot P(B). \qquad (3.5)$$

Y se llama probabilidad condicional de un suceso *B* dado que *A* ha ocurrido

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad \text{que puede reescribirse como } P(A \cap B) = P(B|A) \cdot P(A). \qquad (3.6)$$

De las expresiones 3.5 y 3.6 se deduce la **regla de la multiplicación** de **sucesos dependientes**.

Como ejemplo si se extrae una carta de un mazo de 40 cartas la probabilidad de extraer un As sabiendo que las cartas son espadas $P(As|Espadas) = \frac{P(As\ de\ espadas)}{P(Espadas)} = \frac{1/40}{10/40} = \frac{1}{10}$.

Observe que en cambio, la probabilidad de que la carta sea una espada sabiendo que se trata de un As es $P(Espadas|As) = \frac{P(As\ de\ espadas)}{P(As)} = \frac{1/40}{4/40} = \frac{1}{4}$.

Dos sucesos *A* y *B* son **independientes** cuando la ocurrencia o no ocurrencia de uno de ellos no cambia la probabilidad de ocurrencia del otro. Es decir $P(B|A) = P(A)$ entonces al reemplazar esto en la ecuación 3.5 se tiene

$$P(A \cap B) = P(A) \cdot P(B) \qquad (3.7)$$

Se llega a la expresión 3.7 cuando se considera $P(A|B) = P(B)$.

Evidentemente, si *A* y *B* son mutuamente excluyentes $P(A \cap B) = 0$, entonces $P(A \cap B) = P(A) \cdot P(B) = 0$ de ahí que si la probabilidad conjunta de dos eventos es cero, los eventos son independientes.

Por ejemplo la probabilidad que una moneda lanzada al aire caiga con la cara hacia arriba P(*C*) y la probabilidad que al lanzarla nuevamente caiga con la cara hacia es arriba P(*C*) es $P(C\ y\ C) = P(C \cap C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Es bueno aclara en este punto la **diferencia entre sucesos mutuamente excluyentes y sucesos independientes**. Si interesa conocer si dos sucesos son mutuamente excluyentes se analiza la probabilidad de que por lo menos ocurra uno o varios sucesos, es decir se analiza lo que ocurre en la unión. Pero cuando interesa conocer si dos sucesos son independientes se considera la intersección, o la probabilidad de que ocurran todos los sucesos. Además, si los dos sucesos son mutuamente excluyentes y si ningún suceso tiene cero probabilidades de ocurrencia, los dos sucesos son estadísticamente dependientes. Sin embargo que dos sucesos sean dependientes no necesariamente indica que sean mutuamente exclusivos. Finalmente, si dos sucesos son independientes no pueden ser mutuamente excluyentes.

EJEMPLO 1
Cálculo de probabilidades
Se ponen al azar tres personas A, B y C, en una fila. Sean los sucesos:
R = A está a la izquierda de B
T = C está a la izquierda de B
Encontrar P(R), P(T), P(R ∩ T), P(R|T), ¿Son independientes R y T?

Respuesta:
S = {ABC, ACB, CAB, BAC, BCA, CBA}
R = {ABC, ACB, CAB}
T = {CAB, ACB, CBA}
R ∩ T = {ACB, CAB}

Luego: P(R)= 3/6 = 1/2
P(T)= 3/6 = 1/2
P(R ∩ T) = 2/6 = 1/3
P (R|T) = P(R ∩ T) / P(T) = (1/3) / (1/2) = 2/3

Entonces R y T no son independientes porque P(R|T) ≠ P(R).

Teorema de Bayes[9]

El teorema de Bayes, formulado en 1763, vincula la probabilidad de un suceso condicionado por la ocurrencia de otro suceso. Empleando un vocabulario estadísticos expresa la probabilidad condicional de un suceso aleatorio *A* dado *B* en términos de la distribución de probabilidad condicional del evento *B* dado *A* y la distribución de probabilidad marginal de sólo *A*.
El teorema expresa que si A_1, A_2, ..., A_n son un conjunto de sucesos mutuamente excluyentes y exhaustivos ($\sum P(A_i) = 1$) con probabilidad de cada uno de ellos distinta de cero (0), y B es un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$, entonces la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)\cdot P(A_i)}{\sum P(B|A_i)\cdot P(A_i)}, \qquad (3.8)$$

donde $P(A_i)$ es la probabilidad de A_i, $P(B|A_i)$ es la probabilidad de B cuando se conoce que A_i ha sucedido y $P(A_i|B)$ son las probabilidades de A_i cuando se conoce que sucedió B.

EJEMPLO 2

Teorema de Bayes

Un geólogo recolecta 60 fósiles de las especies B y C de tres afloramientos A_1, A_2 y A_3:

Afloramiento	*N° de observaciones*	*N° de fósiles*	
		Especie B	*Especie C*
A_1	10	7	3
A_2	20	10	10
A_3	30	10	20

a) ¿Cuál es la probabilidad de elegir al azar un fósil del afloramiento A_1? ¿uno del afloramiento A_2? ¿uno del afloramiento A_3?

$$P(A_1) = \frac{10}{60} = \frac{1}{6}$$

$$P(A_2) = \frac{20}{60} = \frac{1}{3}$$

$$P(A_3) = \frac{30}{60} = \frac{1}{2}$$

b) ¿Dado que el fósil es del afloramiento A_1, cuál es la probabilidad que pertenezca a la especia B? ¿Cuál es la probabilidad que pertenezca a la especie B si se conoce que el fósil es del afloramiento A_2? ¿y sabiendo que es del afloramiento A_3?

$$P(B|A_1) = \frac{P(B \cap A_1)}{P(A_1)} = \frac{7/60}{10/60} = \frac{7}{10}$$

$$P(B|A_2) = \frac{P(B \cap A_2)}{P(A_2)} = \frac{10/60}{20/60} = \frac{1}{2}$$

$$P(B|A_3) = \frac{P(B \cap A_3)}{P(A_3)} = \frac{10/60}{30/60} = \frac{1}{3}$$

c) ¿Dado que el fósil es de la especie B, cuál es la probabilidad que provenga del afloramiento A_1?

$$P(A_1|B) = \frac{P(B|A_1) \cdot P(A_1)}{P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + P(B|A_3) \cdot P(A_3)}$$

$$= \frac{7/10 \cdot 1/6}{7/10 \cdot 1/6 + 1/2 \cdot 1/3 + 1/3 \cdot 1/2} = \frac{7}{27} = 0{,}259$$

Variable aleatoria

Como se ha visto, los hechos elementales resultado de los experimentos aleatorios se expresan comúnmente en forma cualitativa: pozo contaminado o no contaminado, bloque por encima de la ley de corte o por debajo de la ley de corte, series finitas de símbolos, entre otros. Ahora bien, para aplicar la teoría de probabilidades a situaciones prácticas es más conveniente trabajar con números que con resultados cualitativos porque los números reales permiten análisis matemáticos. A esto se suma que en ciertos experimentos aleatorios no es factible identificar todos los puntos de muestra posibles,

aunque se puede determinar el hecho elemental a cierto conjunto de hechos del espacio de muestra y asociarlos con un valor numérico. Estas observaciones inducen a considerar lo que se conoce como variable aleatoria y función de probabilidades.

Se llama **Variable Aleatoria** a una función que asigna un numero real $\Re$ a cada punto del espacio muestral. Es decir es una función que permite expresar los resultados de un experimento aleatorio, como el de lanzar una moneda, en un número. Formalmente la variable aleatoria es la transposición teórica de una variable estadística. En símbolos $fn(x)$: $S \Rightarrow \Re$.

Se utilizan letras mayúsculas para representar variables estadísticas:

X = el resultado obtenido al lanzar un dado,

Y = el tamaño de cristales de feldespato,

W = el contenido metálico de un mineral.

Un ejemplo geológico sencillo, suponga el experimento aleatorio que consiste en registrar la presencia de *Escherischia coli* en dos pozos del acuífero Pampeno. Cuando la enterobacteria está presente se registra C (con *E. coli*) y cuando no está presente S (sin *E. coli*). El espacio muestral de este experimento es S = {CC, SS, CS, SC}.

Dado que los resultados del experimento no son números este no es una variable aleatoria. Pero se puede definir una función que transforme los resultados de este experimento en números. Por ejemplo número de pozos con *E. coli*. La variable aleatoria es entonces *X = Nº de pozos contaminados (C)* (Fig. 2).

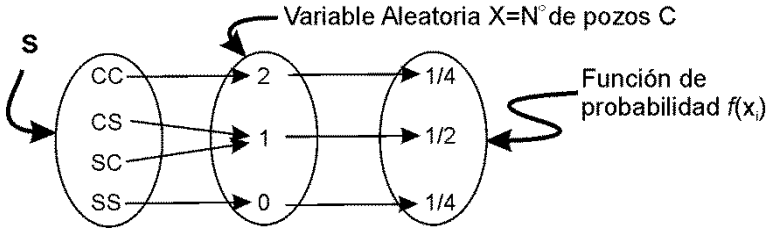


Figura 2. Espacio muestral, Variable aleatoria y Función de probabilidad asociadas del experimento registrar la presencia de E. coli en dos pozos.

Como los números 0, 1 y 2 provienen de un experimento aleatorio correspondiente a observar el número de pozos con *E. coli*, C, en los dos pozos del Pampeano, los números {0, 1, 2} corresponden al espacio de muestra que está formado por tres subconjuntos mutuamente excluyentes y colectivamente exhaustivos.

X = Nº de pozos con C *R = {0, 1, 2}*

Se llama **Rango** *R* de una variable aleatoria *X* al conjunto de todos los valores que puede tomar *X*. Por ejemplo

X = el número que aparece en la cara superior al tirar un dado *R = {1, 2, 3, 4, 5 ,6}*

X = longitud de un cristal de feldespato potásico *R = {long.: ≥0}*

Dado que una variable aleatoria puede tomar cualquier valor dentro del campo de los $\Re$ es posible distinguir variable aleatoria discretas y continuas.

Una **variable aleatoria discreta** es la que tiene el espacio muestral asociado con un número finito de elementos o con una cantidad infinita numerable. El rango en los $\Re$ es de la forma:

$R = \{x_1, x_2, x_3, ..., x_n\}$, (Fig. 3).

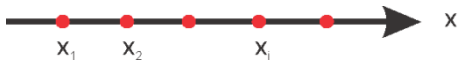


Figura 3. Rango de una variable aleatoria discreta.

En las **variables aleatorias continuas** el espacio muestral está asociado con un número infinito de puntos. El rango en $\Re$ es de la forma:

$R = \{x: a \leq x \leq b\}$, a y b puede ser eventualmente $-\infty$ y $+\infty$ (Fig. 4).

Figura 4. Rango de una variable aleatoria continua.

Descripción probabilística de una variable aleatoria

Como las variables aleatorias son hechos que provienen de un experimento cada hecho tiene una probabilidad asociada. Para el ejemplo de los pozos hay una probabilidad asociada a 0, a 1 y a 2 que serán P(0), P(1) y P(2) respectivamente. Las variables aleatorias se describen con su función de probabilidades y con la función de probabilidades acumuladas.

Variables aleatorias discretas. Función de Probabilidad y-Función Acumulada de Probabilidades

Si X una **Variable Aleatoria Discreta**, cada valor posible de x_i tiene asociado un valor de probabilidad $P(x_i)$. Dicha probabilidad es la misma que origina el suceso y el valor de la variable. Son variables aleatorias discretas el número de terremotos al mes que se producen en una región, el número de bloques encima de la ley de corte en un yacimiento por ejemplo. La función que permite asociar a cada valor de la variable aleatoria discreta su probabilidad se llama **función de probabilidad** (Fig. 2), en símbolos

$$f(x_i) = P(X=x_i), \qquad (3.9)$$

donde X es la variable aleatoria y x_i es un valor particular observado. La expresión $(X = x)$ se lee el conjunto de todos los puntos del espacio muestral a los que la variable X les asigno el valor x.

El conjunto de pares ordenados *x, f(x)* es una función de probabilidad (también llamada función de masa de probabilidad o distribución de probabilidad) si se cumple que:

1. La función $f(x_i)$ asume el valor numérico para todo x_i que se encuentran entre$1 \leq i \leq N$.

   Esto significa que $f(x_i) = \begin{cases} P(X = x_i), & si\ x \in S \\ 0 & en\ cualquier\ otro\ caso \end{cases}$.

Además $f(x_i) \geq 0$ para cualquier valor posible de x.

En palabras, siempre existe un valor de probabilidad para cada x_i que pertenezca al rango de la variable.

2. La suma de todas las probabilidades que conforman el rango de la variable aleatoria es igual a uno.

   $\sum f(x_i) = 1$,

esto es, $P(x_1) + P(x_2) + ... + P(x_n) = 1 \Leftrightarrow \Sigma P(x_i) = 1$

La función de probabilidad se representar con una tabla, una ecuación o una gráfica (Fig. 5).

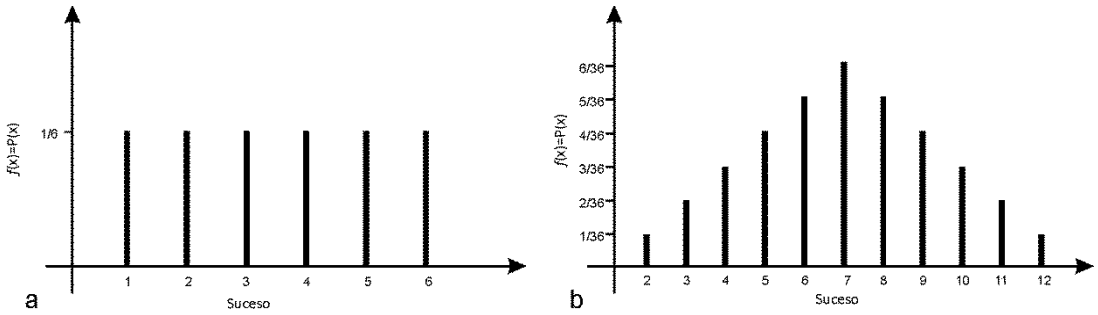


Figura 5. a) Función de probabilidades del experimento aleatorio lanzar un dado y observar es el número que aparece, si el dado no está cargado, los seis números tienen una oportunidad igual de aparecer en cualquier jugada. b) Función de probabilidades del experimento lanzar dos dados simultáneamente con el mismo número y la función de probabilidad de la variable aleatoria suma de los números de ambos dados, en este caso los sucesos no son equiprobables.

Otra función útil para caracterizar probabilísticamente a una variable aleatoria *X* es la **Función Acumulada de Probabilidades** también llamada Distribución de Acumulada (CDF). Esta función permite calcular la probabilidad de que la variable aleatoria asuma un valor menor o igual que un número particular.

Para una **Variable Aleatoria Discreta** la distribución de probabilidad acumulada es una función escalonada, puesto que se incrementa por saltos o escalones en cada uno de los posibles valores de *X* (en un conjunto numerable de puntos) (Fig. 6).

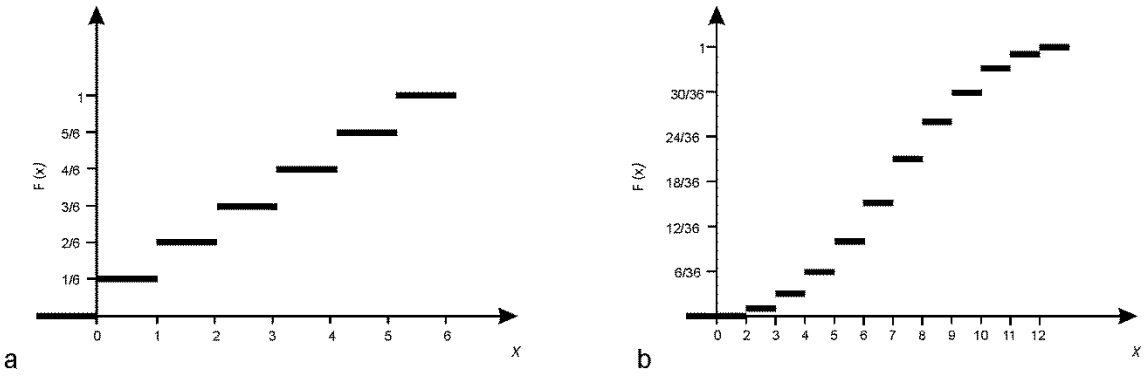


Figura 6. a) Función acumulada de probabilidades de la variable número que aparece al lanzar un dado. b) Función acumulada de probabilidades de la variable aleatoria suma de los números de dos dados lanzados simultáneamente.

EJEMPLO 3

Función de probabilidad y función acumulada de una variable aleatoria discreta

Se detecta la presencia de *Escherischia coli* en dos pozos de agua del acuífero Pampeano. Se conoce que la probabilidad de encontrar *E. coli* en un pozo es de 15% y que su presencia es independiente en cada pozo. Entonces se puede calcular la probabilidad de hallar ninguno, uno o ambos pozos contaminados.

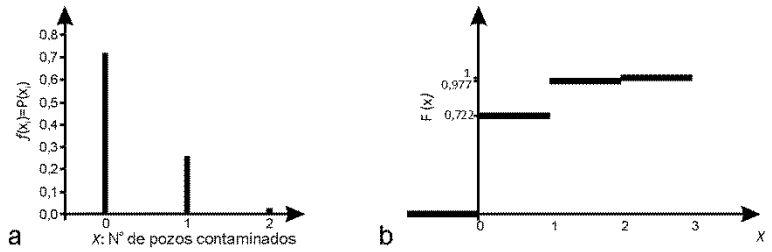
Ya que la presencia de *E. coli* en un pozo no está relacionada a la presencia en el otro, esta situación puede ser analizada como si se lanzaran dos monedas al aire y registrar el número de caras. Siguiendo esta analogía, los pozos se pueden representar como dos monedas, cada uno tiene sólo dos valores posibles: 1 (el pozo está contaminado) y 0 (el pozo no está contaminado). En este caso las monedas son defectuosas. El valor 1 sale sólo 15% de las veces mientras que el valor 0 sucede 85% de las veces. Las probabilidades de los tres posibles resultados (0, 1, o 2 pozos contaminados) se pueden determinar multiplicando las probabilidades asociadas con los posibles resultados para cada pozo, contaminado o no contaminado.

Espacio muestral	*X*: N° de pozos contaminados	Función de Probabilidad $f(x_i)$	Método de Calcular la Probabilidad
CC	2	0,023	Igual a la probabilidad de que el pozo 1 esté contaminado (0,15) y el pozo 2 esté contaminado (0,15) = (0,15)·(0,15)
CS SC	1	0,255	Igual a la probabilidad de que pozo 1 esté contaminado y el 2 no lo esté (0,15)*(0,85) además de la probabilidad de que el pozo 1 no esté contaminado y el pozo 2 sí lo esté (0,85)*(0,15) = (0,15) · (0,85)+ (0,85) · (0,15)
SS	0	0,722	Igual a la probabilidad de que el pozo 1 (0,85) y 2 (0,85) no estén contaminados =(0,85) · (0,85)

Se observa que: $p(x_0) + p(x_1) + p(x_2) = 1$

La función de probabilidad acumulada para esta variable aleatoria discreta *X* = Nº de pozos con *E. coli*, C y su gráfica son:

X: N° de pozos contaminados	Función de Probabilidad $f(x_i)$	Función de probabilidad acumulada $F(x_i)$
0	0,722	0,722
1	0,255	0,977
2	0,023	1,000



Variable aleatoria continua, Función de Densidad de probabilidades y Función acumulada de probabilidades

Si *X* es una **Variable Aleatoria Continua** como la ley de un yacimiento, el valor de concentración tóxica de un elemento en el agua, la longitud de un cristal o cualquier medida de tiempo, longitud, peso y volumen, tiene un número incontable de valores posibles. La función de probabilidades es entonces continua en el sentido que su gráfico es suave, no presenta saltos. Debido a la continuidad, la probabilidad de que *X* asuma cualquiera de sus valores posibles es cero, así que la función de probabilidades no es práctica. En su lugar se utiliza la función *f*(x) **Densidad de Probabilidad** (PDF). La función de densidad de probabilidad, al igual que una función como la velocidad, acumula la probabilidad al pasar rápidamente de derecha a izquierda o de izquierda a derecha sobre el eje de los valores de la variable aleatoria continua *X*. Por lo tanto, la función de densidad es una medida de la concentración de probabilidad dentro de un intervalo. Esta probabilidad se interpreta como un área (una integral) bajo la curva de *f*(*x*) llamada densidad de probabilidad. La función de densidad permite calcular la probabilidad que tomará la variable entre dos valores de interés comprendido en el rango de la variable (Fig. 7 a y b).

La Función de Densidad de Probabilidad de una variable continua definida en el conjunto de números reales $\Re$ satisface:

1. La función $f(x_i)$ asume el valor numérico para todo x_i mayor o igual a cero.

   $f(x) \geq 0 \quad \forall x \in \Re$

2. El área bajo la curva *f*(*x*) es igual a uno (Fig. 7 a y b).

   $\int_{-\infty}^{\infty} f(x)dx = 1$

3. La probabilidad que un valor se encuentre comprendido en el intervalo [*a*,*b*], entonces

   $P(a \leq X \leq b) = \int_a^b f(x)dx.$

   (Fig. 7c).

En particular en el caso $X = a$ la probabilidad de *a* es nula ya que no hay área bajo la curva, lo que no significa que el suceso *a* sea imposible.

$P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)dx = 0.$

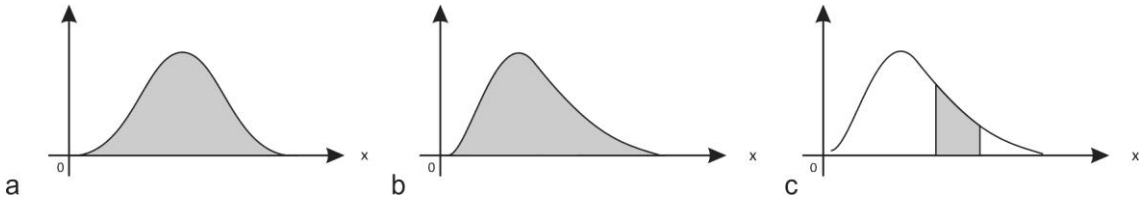


Figura 7. Función densidad de variables aleatorias continuas.

Como en el caso discreto, la **función de de distribución acumulada** es

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt. \tag{3.10}$$

Observe que $f(x)$ se puede interpretar como la velocidad de cambio de $F(x)$. Esta función acumulada tiene las siguientes propiedades:

1. $\forall x \in \Re, F_x(x) \in [0,1]$
2. Dado que $F(x)$ no es decreciente, se deduce que $f(x) \geq 0$.
3. Por otra parte, si $a < b$, se tiene

   $P(a < X < b) = F(b) - F(a)$.
4. Si $-\infty < X < \infty$

   $\lim_{x \to -\infty} F_X(x) = P(-\infty) = 0$,

   $\lim_{x \to \infty} F_X(x) = P(X < \infty) = 1$, es decir el área bajo la curva de densidad es 1, la probabilidad total de la distribución de X.

EJEMPLO 4

Función de densidad y función acumulada de una variable aleatoria continua

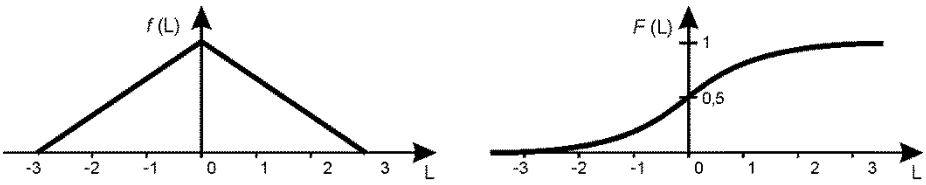
Se adaptan los valores numéricos presentados por Chou (1977) a un problema geológico.

Uno de los pozos del acuífero Pampeano tiene una bomba que permite bombear entre 9.700 hasta 10.300 litros al día. La variable continua es L = litros bombeados por día, en este caso particular el rango se ubica entre los 9.700 y los 10.300 litros. La frecuencia relativa de muchos días sugiere la función de densidad de probabilidad de la variable L, litros, se aproxima muy bien a un triángulo isósceles. Para facilitar los cálculos suponga que la amplitud de L se encuentra entre -3 y 3 (-3 equivale al mínimo, 9.700 l; 3 equivale al máximo, 10.300 l y 0 equivale a 10.000 l). Para satisfacer la condición que el área bajo la curva es 1, el vértice arriba de $L = 0$ debe tener una altura $h = 1/3$ $(1/2(h)\ 6 = 1)$.

Para derivar la función de distribución acumulada F(L) se puede recurrir a las propiedades de triángulos similares se logra la siguiente descripción:

$$F(l) = P(L \leq l) = \begin{cases} 0, & si \quad l \leq -3 \\ \dfrac{(l+3)^2}{18}, & si \quad -3 < l \leq 0 \\ 1 - \dfrac{(3-l)^2}{18}, & si \quad 0 \leq l < 3 \\ 1, & si \quad l \geq 3 \end{cases}$$

La curva de densidad de probabilidad y la curva de la función acumulada de L son:



Suponga que se elige al azar un día cualquiera, la **probabilidad** de que los litros bombeados sean:

a) menor o igual a 9.900 litros

$$F(-1) = \frac{(-1+3)^2}{18} = \mathbf{0,222}$$

b) menor o igual a 10.150 litros

$$F(1{,}5) = \frac{(3-1{,}5)^2}{18} = \mathbf{0,875}$$

c) entre 9.000 y 10.150 litros

P(-1 ≤ l ≤1,5) = F(1,5) – F(-1)
= 0,875 – 0,222
= **0,653**

El Valor Esperado y la Varianza de una variable aleatoria

La distribución de probabilidades para una variable aleatoria es un modelo teórico para la distribución experimental de datos asociados a una población real. Si el modelo coincide con la realidad, las distribuciones teórica y empírica son equivalentes. Para describir la distribución teórica (distribución de probabilidad y distribución de densidad) se utiliza la media y la varianza.

El valor medio de una variable aleatoria se conoce como la **Esperanza Matemática** o **Expectativa**. Se utiliza la notación *E(X)* para representar la esperanza matemática.

La esperanza para una **variable discreta** es un promedio ponderado de los valores que puede asumir la variable *X* con probabilidades para los valores de *X* como pesos,

$$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_N p_N$$
$$E(X) = \sum_{i=1}^{N} x\, p(x) \qquad (3.11)$$

donde p(*x*) es a función de probabilidad de la variable aleatoria *X*.

Si la función de probabilidades describe exactamente la distribución de frecuencias entonces el valor esperado es igual a la media poblacional, $E(X) = \mu$.

La esperanza, como una medida de tendencia central, indica donde se ubica el centro de gravedad de la distribución de probabilidad de la variable aleatoria y es el valor medio si el experimento se repite una y otra vez. Cabe aclarar que no es necesariamente un valor posible de la variable aleatoria como se observa en el ejemplo 5 donde la variable aleatoria es *X* = N° de pozos con *E. coli*.

EJEMPLO 5

Esperanza y varianza de una variable aleatoria discreta

La esperanza y la varianza de la variable *aleatoria discreta X = Nº de pozos con E. coli, C* del ejemplo de dos pozos de agua del acuífero Pampeano son:

$E(X) = 0 \cdot 0{,}722 + 1 \cdot 0{,}255 + 2 \cdot 0{,}023 = 0{,}301\ pozos\ contaminados$
$V(X) = (0^2 \cdot 0{,}722 + 1^2 \cdot 0{,}255 + 2^2 \cdot 0{,}023) - (0{,}301)^2 = 0{,}256$

La esperanza para una **variable continua** es

$$E(X) = \int_{-\infty}^{\infty} f(x)dx. \tag{3.12}$$

E(X) representa la abscisa del centro de gravedad de la masa de la curva bajo la curva *f*(x). Para obtener el valor precisos de *E(X)* se requiere conocimientos de cálculo fuera del alcance de este libro, sólo se menciona que se puede obtener una aproximación subdividiendo el rango en *n* partes iguales y calculando las probabilidades de que la variable se encuentre en el intervalo comprendido por cada subintervalo. Se muestran los cálculos de la aproximación a la esperanza con el ejemplo de los litros bombeados para un pozo del acuífero Pampeano (Ejemplo 4).

La **varianza** es de una variable aleatoria es una medida de la dispersión de los valores que toma con respecto a la esperanza matemática.

La varianza de una **variable aleatoria discreta** es

$$V(X) = \sigma_X^2 = \sum_{i=1}^{n} [(x_i - \mu)^2 f(x_i)]$$

$$V(X) = E(X^2) - \mu^2 \tag{3.13}$$

La varianza de una **variable aleatoria continua** es

$$V(X) = \sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx$$

Por el mismo razonamiento que para el cálculo de la esperanza, se puede obtener una aproximación a la varianza con la siguiente expresión

$$V(X) = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i)\ c \tag{3.14}$$

donde *c* es la amplitud del intervalo de clase.

En tanto el desvío estándar de para ambos casos, es $\sigma_X = \sqrt{V(X)}$.

EJEMPLO 6

Esperanza y varianza de una variable aleatoria continua

La esperanza de la variable continua *L* = litros bombeados por día del pozo del acuífero Pampeano se calculó subdividiendo el rango de la variable en 6 intervalos. Para simplificar los cálculos se considera que la amplitud de *L* se encuentra entre -3 y 3 (-3 equivale al mínimo, 9.700 l; 3 equivale al máximo, 10.300 l y 0 equivale a 10.000 l), la amplitud de cada intervalo es 1.

Las probabilidades de F(l) = P(L ≤ l) fueron derivadas en el ejemplo anterior.

Para calcular la probabilidad de cada intervalo se recurre a la propiedad *P(a < X < b) = F(b) – F(a)*, donde *a* es el límite inferior y *b* el límite superior del intervalo.

Luego se calcula la media siguiendo para datos agrupados con las probabilidades de cada intervalo como peso. La tabla que se presenta a continuación muestra los cálculos realzados.

Intervalo	*1* *P(I ≤ b)*	*2* *P(b) – P(a))*	*3* *l_i= marca de clase*	*4* *2 x 3*	*5* *$(l_i - \mu)^2$*	*6* *5 x 3*
-3; -2	0,056	0,056	-2,5	-0,139	6,250	0,347
-2; -1	0,222	0,167	-1,5	-0,250	2,250	0,375
-1; 0	0,500	0,278	-0,5	-0,139	0,250	0,069
0; 1	0,778	0,278	0,5	0,139	0,250	0,069
1; 2	0,944	0,167	1,5	0,250	2,250	0,375
2; 3	1,000	0,056	2,5	0,139	6,250	0,347
Total		1,000		0,000		1,583

La Esperanza es 10.000 litros (valor que equivale a 0).
La Varianza es 10.158,3 litros^2.

MODELOS PROBABILÍSTICOS ÚTILES EN GEOLOGÍA

Introducción

Las distribuciones de frecuencias de las variables geológicas han despertado interés desde dos puntos de vista. El primero se centra en identificar si la distribución de los datos observados se asimila con algún modelo probabilístico. Lo más corriente es confrontarla la variable geológica con el modelo de distribución Normal pues, si esta condición se cumple, entonces es posible realizar predicciones e inferencias robustas utilizando métodos estándar. La otra perspectiva analiza los modelos de distribuciones empíricas con el objeto de indagar cuales son los factores geológicos que intervienen, sin embargo, aconsejan cautela al realizar este tipo de análisis, pues se ha probado que en algunos casos, el tipo de distribución está controlada por causas ajenas a los factores geológicos. Por ejemplo las definiciones operacionales usadas para generar los valores numéricos asociados a la medida del atributo pueden controlar la forma de la distribución de nuestras variables. Esto sucede cuando expresamos el "contenido de granos de minerales pesados en un subconjunto de *n* granos", los minerales se distribuyen según un modelo Poisson, mientras que si se trata del contenido de minerales abundantes estos siguen un modelo binomial. Si se trata de la forma, por ejemplo esfericidad de gravas sin especificar que tipo de gravas, se abarca un rango granulométrico muy grande y la distribución tiende a ser un modelo asimétrico, en tanto si se restringe el análisis a un rango granulométrico dentro del tamaño gravas, la distribución tiende a ser simétrica. De forma similar cuando se miden ángulos sobre un rango relativamente limitado, la distribución puede acercarse a un modelo normal.

Por otra parte, contar con un modelo probabilístico que describa la variable de estudio permite decir que, dadas ciertas condiciones, sucederán tales o cuales hechos conjuntamente con las probabilidades asociadas a cada uno. En este capítulo se describen los principales modelos probabilísticos con los que se pueden cotejar las variables geológicas.

Modelos de variables aleatorias discretas

Modelo Bernoulli

El modelo Bernoulli es el más sencillo de todos los modelos. Es apropiado para describir experimentos que sólo resultan en un hecho o en su opuesto tales como hallar petróleo (éxito) o o no hallarlo (fracaso) al perforar un pozo, proposiciones si o no (presencia de fósiles, estructuras), entre otros.

Entonces una variable aleatoria Bernoulli *X* que toma solamente los valores 0 y 1, con probabilidades respectivamente p y $q=1-p$ ($0 \leq p \leq 1$).

La función de distribución de la variable Bernoulli es $f(x) = (a \cdot p)\ y\ (b \cdot q)$ (Fig. 1).

Se puede probar que la esperanza y la varianza de este modelo son:

$E(X) = \sum_i x_i f(x_i) = p,$

$V(X) = p\ q.$

El modelo Bernoulli tiene sólo un parámetro p, aunque la mayoría de las veces p es conocido, puede suceder que no lo sea, más adelante se verá como estimar p.

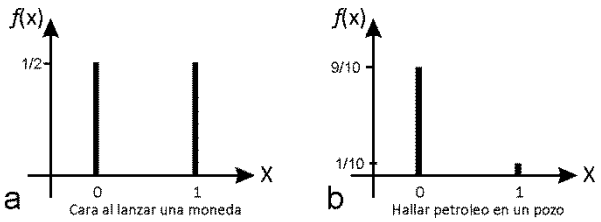


Figura 1. a. Función de probabilidad Bernoulli para la variable X= cara al lanzar una moneda (p=q=1/2)
b. Función de probabilidad para la variable P= hallar petróleo al perforar un pozo (p=0,1; q=0,9).

Variable Aleatoria Geométrica

El modelo geométrico se origina con el mismo experimento repitiéndose un número indefinido de veces. Los experimentos son independientes, en cada repetición caben dos alternativas: que ocurre un suceso *A* con probabilidad *p,* o que ocurre un suceso *B* con probabilidad *1 – p*. El experimento se detiene cuando ocurre por primera vez el suceso *A*.

La variable aleatoria geométrica es *X*, el número de repeticiones necesarias para que ocurra *A*.

El conjunto de resultados posibles es entonces infinito, por ejemplo S = {A, BA, BBA, BBBA, …}.

En este caso la variable *X* será 1 si *A* ocurre en el primer experimento, 2 si ocurre en el segundo, 3 si ocurre en el tercero, 4 si ocurre en el cuarto y así siguiendo.

Las probabilidades asociadas con cada valor de la variable son:

P(*x = 1*) = P(*A*) = *p*

P(*x = 2*) = P(*BA*) = P(*B*)·P(*A*) = (*1 - p*)·*p* (por la independencia)

P(*x=3*) = P(*BBA*) = P(*B*)·P(*B*)·P(*A*) = $(1 - p)^2 \cdot p$

La función de distribución de la variable geométrica es $f(x) = \mathrm{P}(x) = 1 - p^{x-1} \cdot p$ (Fig. 2).

Se puede probar que la esperanza y la varianza de este modelo son:

$E(X) = \frac{1}{p}$,

$V(X) = \frac{1-p}{p^2}$.

El modelo geométrico, al igual que el Bernoulli, tiene sólo un parámetro, p.

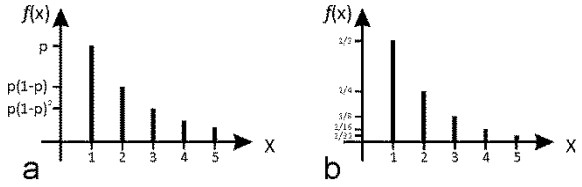


Figura 2. a) Función de probabilidad de la variable geométrica. b) Función de probabilidad de X = número de lanzamientos necesarios de una moneda hasta que aparezca cara, (parámetro p=1/2).

EJEMPLO 1

Modelo geométrico

Un estudiante tiene un examen a las 8.00 de la mañana. Ese día se queda dormido y llega a la parada del colectivo que lo lleva a la facultad a las 7.40 hs. El colectivo tiene una frecuencia de 10 minutos en horas pico (desde las 7.00 hasta las 9.00 hs.) y el recorrido demanda 10 minutos. En este horario la probabilidad que el colectivo pase lleno y no pare es 75%.

¿Cuál es la probabilidad de que logre subir en el tercer colectivo que pase y llegue a horario al examen?

La probabilidad que el colectivo pase vacio es la que interesa pues si pasa vacio el estudiante logra subir y se acaba el experimento. La probabilidad que pase lleno es dato, es 0,75 = 1 – *p*, la probabilidad que pase vacio es $p = 0{,}25$,

x = 3

$P(x) = (1 - p)^{x-1} \cdot p$

$P(x=3) = (1 - 0{,}25)^{3-1} (0{,}25) = 0{,}14$

La probabilidad que llegue a horario a rendir el examen es de 0,14.

Modelo Binomial

Algunos datos geológicos provienen de poblaciones de datos nominales que tienen solamente dos categorías. Cada pozo perforado puede tener petróleo o no tenerlo, cada alumno puede aprobar o desaprobar una prueba, etc., es decir son variables Bernoulli. Considere los experimentos que consisten en la observación de una serie de *n* variables Bernoulli, o sea pruebas idénticas e independientes que generar solamente dos resultados. Experimentos de este tipo originan variables que se adecuan al modelo **binomial** y reúnen las siguientes características:

a) El experimento consta de ***n*** pruebas idénticas.

b) Cada prueba tiene solamente dos resultados posibles. Se llama a uno éxito *E* y al otro fracaso *F*.

c) La probabilidad de tener **éxito** en una sola prueba es igual a *p*, y permanece constante de prueba en prueba. La probabilidad de **fracaso** es igual a *q = (1 - p)*.

d) Las pruebas son independientes, es decir el resultado de una prueba no influye sobre el de las otras.

e) La variable aleatoria bajo estudio es *X*, **el número de éxitos observados en las *n* pruebas**.

f) La variable aleatoria **binomial** es **discreta** y tiene **n + 1** valores posibles.

g) La función de probabilidad que permite calcular la probabilidad de obtener exactamente *x* éxitos en las *n* pruebas independientes de un experimento, con *p* como la probabilidad de éxito es:

$$B(x,n,p) = \binom{n}{x} p^x \, q^{n-x} \qquad (4.1)$$

donde $\binom{n}{x}$ son las combinaciones posibles de *n* elementos tomados en grupos de *x* elementos, $\binom{n}{x} = \left(\frac{n!}{x!(n-x)!}\right)$.

La distribución binomial es simétrica cuando $p = 0{,}5$ y asimétrica cuando $p \neq 0{,}5$. La asimetría es derecha cuando $p < 0{,}5$ y es a la izquierda para $p > 0{,}5$ (Fig. 3 a, b y c). Además, la asimetría se reduce al aumentar *n*.

h) La función se probabilidad binomial se define con dos parámetros *n* y *p*. Además, el modelo binomial deriva su nombre del hecho que es un término de la expansión del binomio $(q + p)^n$.

i) La Esperanza de una variable binomial es $E(X) = n \cdot p$.

j) La varianza de una variable binomial es $V(X) = n \cdot p \cdot q$.

k) La función de distribución acumulada de una variable binomial, como la de cualquier variable aleatoria discreta, da la probabilidad de obtener *r* éxitos ó menos en *n* pruebas, con $r \leq n$, y se obtiene sumando las probabilidades individuales para todos los valores binomiales iguales o menores a *r*, es decir:

$$\begin{aligned} b(r,n,p) &= P(x \leq r) \\ &= b(0,n,p) + b(1,n,p) + \cdots + b(r,n,p) \\ &= \sum_{x=0}^{r} b(x;n;p)\,. \end{aligned}$$

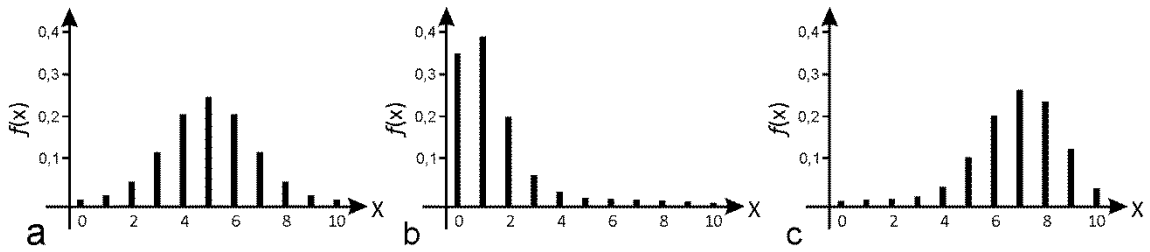


Figura 3. Función de probabilidades binomial para n*=10. a) p=0,5. b) p=0,1. c) p=0,7.*

EJEMPLO 2

Modelo binomial

Una empresa petrolera decide financiar 4 pozos exploratorios en un bloque. La probabilidad de hallar petróleo en un pozo es 0,1. Suponga que se trata de sucesos independientes.

Se conoce que al perforar un pozo pueden ocurrir sólo dos resultados posibles que se representan como *F* si en el pozo no se encuentra petróleo y *E* cuando se encuentra petróleo. La probabilidad de hallar petróleo es $p = 0{,}1$ y la de no hallarlo es $q = 1 - p = 0{,}9$. Las probabilidades de todos los posibles resultados (0, 1, 2, 3, 4, pozos con petróleo) se pueden calcular usando la función de probabilidades binomial. La variable es "número de pozos donde se halla petróleo".

X = número de pozos exploratorios con petróleo
$n = 4$
$p = 0{,}1$ (probabilidad de hallar petróleo, estado 0)
$q = 0{,}9$ (probabilidad de no hallar petróleo, estado 1)

a) ¿Cuál es la probabilidad de que se encuentre petróleo en los 4 pozos?

$x = 4$

$$b(4;4;0{,}1) = \frac{4!}{4!(4-4)!}\ 0{,}1^4\ 0{,}9^{(4-4)}$$

$$= 1 \cdot 0{,}0001 \cdot 1 = 0{,}0001$$

La probabilidad de encontrar petróleo en los 4 pozos es 0,0001.

b) ¿Cuál es la probabilidad de encontrar petróleo en 1 pozo?

$x = 1$

$$b(4;1;0{,}1) = \frac{4!}{1!(4-3)!}\ 0{,}1^1\ 0{,}9^{(4-1)}$$

$$= 4 \cdot 0{,}1 \cdot 0{,}729 = 0{,}2916$$

La probabilidad de encontrar petróleo en un pozo es 0,2916.

c) Suponga que la empresa tiene un costo fijo de U$ 20.000 para preparar el equipo de perforación, que hallar un pozo con petróleo cuesta U$ 2.10^6. ¿Cuánto dinero está dispuesta a invertir la empresa?

$E(X) = n.p$
$E(X) = 4\ (0{,}1) = 0{,}4$
El costo será de 0,4 x 2.10^6 + 20.000 = U$ 42.000

Algunas variables geológicas que siguen el modelo binomial son las proporciones granos de un mineral de arena, de fósiles en submuestras de un tamaño dado, la ocurrencia de estructuras sedimentarias (0= ausente, 1= presente) y todos aquellos estudios que se puedan describir en forma dicotómica (si/no). A esto se suman las variables continuas que, por el objetivo del trabajo, se transforman en dicotómicas como el porcentaje de carbonato que define si una roca es o no una caliza o la concentración de un metal pesado límite para precisar si los sedimentos de un río están o no contaminados.

Modelo uniforme discreto

El modelo uniforme discreto se origina a partir de experimentos cuyos *N* resultados posibles tienen todos las mismas probabilidades de ocurrencia. Por ejemplo la distribución de probabilidades de los puntos de la cara superior de un dado perfecto tienen igual probabilidad de ocurrencia, p = 1/6.
La función de probabilidad de este modelo es

$$u(x;N) = \frac{1}{N}, \qquad x = 1, 2, \cdots, N. \tag{4. 2}$$

Una de las variables geológicas que se coteja con el modelo uniforme es en el contexto de los datos direccionales describe la situación donde la probabilidad de ocurrencia de todos los puntos es la misma en todas direcciones como se verá en el capítulo 13.

Modelo Poisson

Muchos hechos no ocurren como resultado de un número definido de pruebas de un experimento, sino en puntos de tiempo, espacio o volumen al azar. El hecho puede ser el número de ocurrencias de fósiles con ciertas características en *x* cm^3 de sedimentos, el número de partículas radiactivas emitidas por *x* cantidad de sustancia, el número de terremotos en *x* años, etcétera. Experimentos de este tipo se conocen como experimentos Poisson (atribuido al matemático francés S.D. Poisson, 1781-1840) y reúnen las siguientes características:

a) El número de ocurrencias del hecho es independiente de una unidad especificada a otra. La unidad especificada puede ser un intervalo de tiempo, de espacio o un volumen.
b) El valor esperado de la variable es **proporcional** al tamaño de la unidad especificada.
c) La probabilidad de más de una ocurrencia del hecho en una unidad especificada muy pequeña es despreciable en comparación con la probabilidad de una sola ocurrencia; por lo tanto puede considerarse nula.
d) Las pruebas son independientes.
e) La variable aleatoria bajo estudio es *X*, el **número de ocurrencias por unidad especificada**.
f) La variable aleatoria **Poisson** es **discreta** y tiene **infinitos** valores posibles.
g) La función de probabilidad Poisson, del número de ocurrencias por unidad especificada queda completamente definida por su **promedio de ocurrencia** en esa unidad especificada, el parámetro llamado **Lambda (λ)**.
h) La función de probabilidad que permite calcular la probabilidad de obtener exactamente ***x*** **ocurrencias** en la unidad especificada es

$$P_t(x, \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}. \tag{4.3}$$

La distribución es asimétrica positiva pero la asimetría disminuye al aumentar λ (Fig. 4).

i) La Esperanza de una variable Poisson es $E(X) = \lambda$.

j) La varianza de una variable Poisson es $V(X) = \lambda$.

k) La función de distribución acumulada de una variable Poisson, como la de cualquier variable aleatoria discreta, da la probabilidad de obtener ***r* ó menos ocurrencias**, se obtiene sumando las probabilidades de obtener las probabilidades individuales para todos los valores Poisson iguales o menores a *r*, es decir

$$P_t(r;\lambda) = P(x \le r)$$

$$P_t = \sum_{x=0}^{r} P_t(x;\lambda) \cdot$$

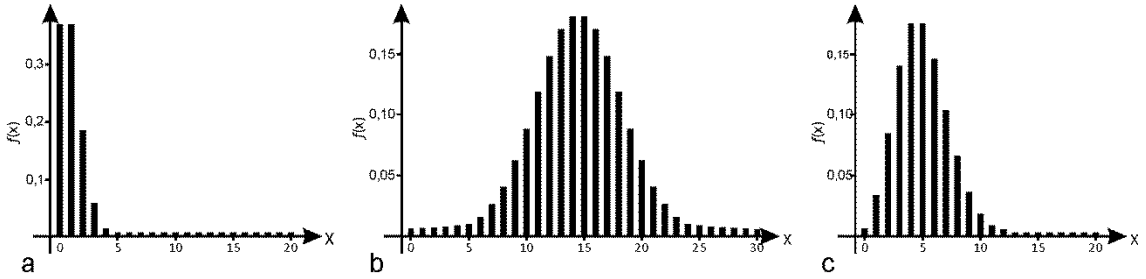


Figura 4. Distribuciones de Poisson para: a) λ = 1, b) λ =15 y c) λ = 5.

EJEMPLO 3

Modelo Poisson

Un evento que ocurre más de una vez, o se espera que ocurra más de una vez, se dice que es recurrente. El periodo de retorno es el intervalo de tiempo esperado entre ocurrencias. El Popoctepetl es el volcán activo más alto de México. A partir del estudio del historial de las erupciones se estimó el periodo de retorno en 0,0202 años^{-1}.

a) Calcular las probabilidades de tener 1, 2 y ninguna erupción en 20 años.

El dato inicial λ = 0,0202 erupciones en un año. Como λ es proporcional a la unidad entonces para 20 años λ = 0,404

Ninguna erupción x = 0, una erupción x=1 y dos erupciones x=2.

$$P_t(0;0{,}404) = \frac{e^{-0{,}404} \cdot 0{,}404^0}{0!} = 0{,}668 \qquad \text{e} = 2{,}71828$$

$$P_t(1;0{,}404) = \frac{e^{-0{,}404} \cdot 0{,}404^1}{1!} = 0{,}270$$

$$P_t(2;0{,}404) = \frac{e^{-0{,}404} \cdot 0{,}404^2}{2!} = 0{,}054$$

a) ¿Cuántas erupciones se espera que ocurran en 50 años?

Para este caso λ=1,01. Dado que la esperanza de la distribución de probabilidades Poisson es E(X) = λ, se es pera que ocurra una erupción cada 50 años.

Algunas variables geológicas que siguen el modelo Poisson son: minerales raros en rocas expresados como número de granos en muestras de un tamaño dado, número de pozos de petróleo por área en un yacimiento, el número de partículas α emitidas por unidad de tiempo de sedimentos

radioactivos y todo evento u objeto que se pueda contar en unidades iguales de área, volumen o intervalos de tiempo iguales. También suele describir adecuadamente a los eventos raros como inundaciones, tsunamis, actividad volcánica, etcétera.

Relaciones entre los modelos discretos Binomial y Poisson

La función de probabilidades Poisson sirve para aproximar las probabilidades de la función Binomial para valores pequeños de *p*. Algunos autores consideran que prácticamente la aproximación es aceptables si $p < 0{,}1$ y $n \cdot p < 5$. Otros sugieren considerar $n > 100$ y $p < 0{,}01$. Cualquiera sea el criterio adoptado para aproximar una función binomial con una Poisson, el parámetro Lambda se calcula como $\lambda = n \cdot p$.

Modelos de variables aleatorias continuas

Modelo Uniforme continuo

Una variable aleatoria cuyo valor solo se encuentra dentro de cierto intervalo delimitado por los números *a* y *b* sigue una distribución uniforme o rectangular si su función de densidad de probabilidad es constante en el intervalo de *a* a *b* (Fig. 5). Entonces, la probabilidad de cualquier subintervalo de [a,b] es el cociente entre su longitud y la del intervalo [a,b].
La función de densidad *f*(*x*) está dada por:

$$f(x) = \begin{cases} \frac{1}{b-a} & si\ a \leq x \leq b \\ 0 & de\ otro\ modo \end{cases} \qquad (4.4)$$

La función de distribución acumulada es

$$\begin{cases} 0 & si\ x \leq a \\ \frac{x-a}{b-a} & si\ a < x < b \\ 1 & si\ x \geq b \end{cases}$$

Los parámetros del modelo son *a* y *b*.
La esperanza y la varianza son $E(X) = \frac{a+b}{2}$ y $V(X) = \frac{(b-a)^2}{12}$ respectivamente.

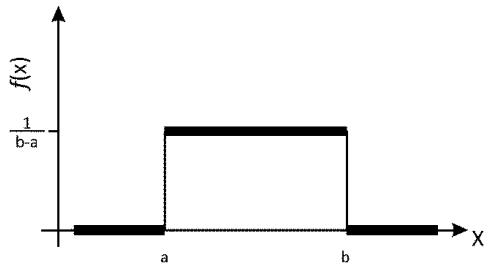


Figura 5. Función de densidad de una variable aleatoria uniforme.

EJEMPLO 4

Modelo uniforme continuo

La línea de colectivo que un estudiante toma para ir a la Facultad tiene una frecuencia de 30 minutos. ¿Cuál es la probabilidad de que si llega a la parada a una hora al azar espere menos de 5minutos?

La variable aleatoria T = tiempo hasta el siguiente colectivo, está uniformemente distribuida $0 \leq T \leq 30$.

La probabilidad que tenga que esperar menos de 5 minutos es

$$P(T < 5) = \frac{5-0}{30-0} = \frac{1}{6} = 0{,}17$$

Modelo Normal o Gaussiano

Muchas variables geológicas siguen una distribución de tipo normal o gaussiano[6]. Son algunos ejemplos el relieve topográfico, la esfericidad y redondez para un tamaño de grano determinado, el nivel de agua en acuíferos a través del tiempo, la densidad de drenaje (km/km^2), la densidad de empaquetamiento de granos de arena, ciertas dimensiones de especímenes de invertebrados fósiles, la humedad en sedimentos y los errores de medición casuales.

Por otra parte, el modelo normal es el más importante en el análisis estadístico pues, como se verá en los capítulos siguientes, las distribuciones de muchas estadísticas de muestra se aproximan a la distribución normal como un límite cuando el tamaño de la muestra es grande.

Una variable aleatoria *X*, tiene una distribución normal general, también se dice está **normalmente distribuida**, si:

a) Es continua.

b) Existen las constantes μ (promedio poblacional) y σ (desvío estándar poblacional), con μ entre $-\infty$ y $+\infty$, ($-\infty < \mu < +\infty$) y el desvío estándar mayor que cero ($\sigma > 0$).

c) La función de densidad se puede calcular de la siguiente manera:

$$n(x;\mu;\sigma) = n(\mu;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty < x < +\infty) \tag{4.5}$$

donde *e* = 2,718 y π = 3,142. Los dos parámetros que definen la distribución normal son μ y σ.

El exponente $-\frac{(x-\mu)^2}{2\sigma^2}$ tiene *x*, un valor particular de la variable, y a los parámetros μ y σ. Cuanto mayor es la desviación de un valor *x* con relación a la media μ, tanto menor (más negativo) es el numerador de este exponente. La desviación, σ, está elevada al cuadrado, por lo que dos valores de *x* que muestren la misma desviación respecto a μ tienen la misma probabilidad lo que implica que se una distribución **simétrica** (Fig. 6).

El signo negativo del exponente indica que cuanto más alejado se encuentra *x* de μ, menor es el área bajo la curva de densidad de probabilidad. Por otra parte, cuando *x* = μ, el exponente es cero y la densidad es $1/\sigma\sqrt{\pi}$, el valor más grande de la densidad normal. Además la distribución tiene una sola moda con valor *x* = μ.

La curva tiene una amplitud infinita, nunca toca el eje *X*, no obstante si *x* está muy alejado de μ, su probabilidad es despreciable. El 99% del área bajo la curva queda comprendido por valores de *x* que se alejen 3 ± σ de μ. Conviene recordar también que el 95% del área bajo la curva se encuentra limitado por valores de $x \leq \mu \pm 2\sigma$ y el 68% por valores de $x \leq \mu \pm \sigma$ (Fig. 6).

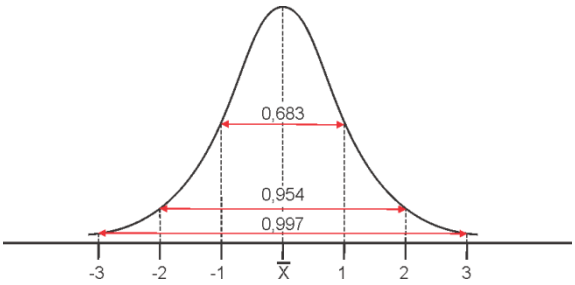


Figura 6. Porcentajes de probabilidad bajo la curva normal. 68% del área bajo la curva por valores queda comprendido por valores de $x \leq \mu \pm \sigma$, 95% del área bajo la curva se encuentra limitado por valores de $x \leq \mu \pm 2\sigma$ y 99% x que se alejen $3 \pm \sigma$ de μ.

Un cambio en el valor de μ, desplaza la distribución hacia la derecha o la izquierda (Fig. 7a). Un cambio en el valor de σ cambia la forma, cuanto más pequeño es aumenta al máximo el valor de *f*(*x*) (Fig. 7b).

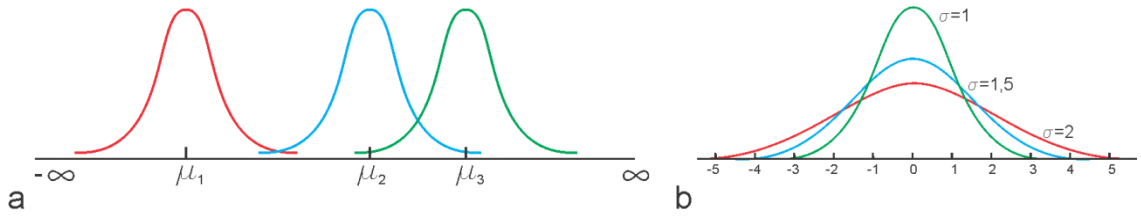


Figura 7. Distribuciones normales: a) con la misma desviación estándar y diferentes medias; b) con la misma media y diferente desviaciones estándares.

Por último se ha probado que la esperanza de cualquier variable normalmente distribuida es $E(X) = \mu$ y la varianza $V(X) = \sigma^2$.

Modelo Normal Estándar

Dado que existe un número infinito de posibles distribuciones normales una para cada par posible de valores de μ, σ y que el cálculo de la función densidad comprende integrales que son difíciles de resolver, es más eficiente y rápido trabajar con la **Variable Normal Estándar** (*Z*).
La distribución normal estándar tiene media cero, $\boldsymbol{\mu = 0}$, y varianza desvío estándar igual a uno, $\boldsymbol{\sigma = 1}$.
La función de densidad de la distribución normal general (Fig. 8a) entonces se reduce para la variable normal estándar, *Z*

$$N(z;0;1) = N(0;1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad (-\infty < x < +\infty)\,. \tag{4.6}$$

La función de distribución acumulada correspondiente a la densidad estándar *N*(0,1), como en otros casos, da la probabilidad de que la variable normal estándar asuma un valor igual o menor que *z* es $F(x) = P(Z \leq z)$ (Fig. 8b).

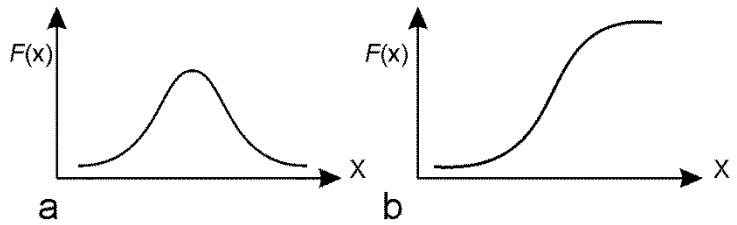


Figura 8. a) Densidad de probabilidad normal estándar. b) Función de distribución acumulada de la normal estándar.

Es importante señalar que toda variable *X* con distribución normal, $N(\mu, \sigma)$, se puede transformar en la variable normal estándar *Z* calculando el cociente de la diferencia entre cualquier valor y la media y el desvío estándar

$$Z = \frac{x - \mu}{\sigma} \qquad \text{es } N(0, 1). \tag{4.7}$$

Esta transformación de *X* en *Z* reduce *X* a unidades en términos de desviaciones estándares alejadas de la media. En otras palabras, dado un valor *x*, el correspondiente valor de *Z* indica cuan alejada está *x* de su media μ, y en que dirección, en términos de desviación estándar, σ. De esta manera, una sola tabla de probabilidades *N*(0,1) (Tabla 1 del Anexo) permite evaluar probabilidades normales para cualquier $N(\mu, \sigma)$ ya que

$$n(x;\ \mu;\sigma) = P(X \leq x)$$
$$= P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$
$$= N\left(\frac{x-\mu}{\sigma}; 0; 1\right) \qquad (4.8)$$

Así, para cualquier $N(\mu, \sigma)$ y dos números reales a y b, con $a < b$, se tiene

$$P(a \leq X \leq b) = N(b) - N(a)$$
$$= N\left(\frac{b-\mu}{\sigma}\right) - N\left(\frac{a-\mu}{\sigma}\right) \qquad (4.9)$$

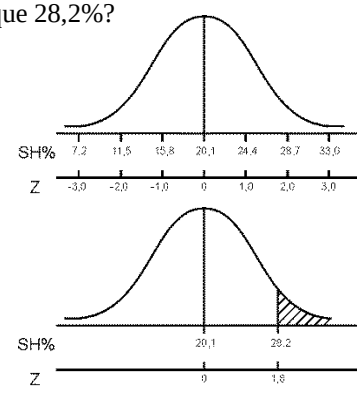
EJEMPLO 5

Cálculo de probabilidades de una variable normalmente distribuida

Se ha medido la saturación de hidrocarburos (SH) en 1145 muestras de testigos extraídos de la Cuenca Austral. Dado el gran número de datos los estadísticos pueden considerarse parámetros. La saturación media hallada es de 20,1% y su desviación estándar de 4,3%. Si se asume que la variable está normalmente distribuida:

a) ¿Cuál es la probabilidad de que una muestra posea una SH mayor que 28,2%?

P($x \geq 28,2$) = 1 – P($x \leq 28,2$) = ?



Dado que P($x \leq 28,2$) = P($z \leq Z$)

Se realiza un cambio de variable de X a Z utilizando la expresión 4.7

$$Z = \frac{28,2 - 20,1}{4,3} = 1,88$$

De la Tabla 1 del Anexo se tiene

P($z \leq 1,88$) = 0,9699

Entonces P($x \geq 28,2$) = 1 – P($z \leq 1,88$)

= 1 – 0,9699

= 0,0301

La probabilidad de que una muestra posea una SH mayor a 28,2% es 0,0301

b) ¿Cuántas muestras de esa población poseen una SH mayor que 28,2%?

El número de muestras se obtiene multiplicando P($x \geq 28,2$) por el tamaño de la muestra (N = 1145).

(0,0301) (1145) = 34,4645

Treinta y cuatro muestras de las 1145 analizadas tendrán una SH mayor o igual a 28,2%.

c) Calcular, aproximadamente, la proporción de muestras que tienen SH entre 18 y 24%

P($18 \leq x \leq 24$) = P($Z_{(x=18)} \leq z \leq Z_{(x=24)}$)

P($x \leq 24$) - P($x \leq 18$) = P($z \leq Z_{(x=24)}$) - P($z \leq Z_{(x=18)}$)

Se debe hallar el área bajo la curva N(0,1) equivalente al área de N(20,1;4,3) para ello se realiza un cambio de variable de X a Z para ambos valores de x: 18 y 24. Luego se obtienen las la probabilidad asociada a cada valor de Z utilizando Tabla 1 del Anexo.

P($18 \leq X$)= P($Z_{(X=18)}$)

$Z_{(X=18)}$= (18-20,1)/4,3 = -0,55

P ($Z \leq -0,55$) = 0,2912

P($X \leq 24$) = P($z \leq Z_{(X=24)}$)

$Z_{(X=24)}$= (24-20,1)/4,3 = 0,90

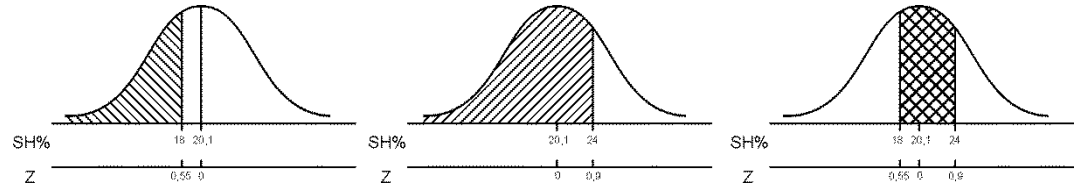
P ($Z \leq 0,90$)= 0,8159

Utilizando la expresión 4.9

P($x \leq 24$)- P($x \leq 18$) = P($z \leq Z_{(X=24)}$) - P($z \leq Z_{(X=18)}$)

= 0,8159 – 0,2912 = 0,5247

Aproximadamente el 52% de las muestras tendrán valores de SH entre 18 y 24%.



d) ¿Cuál es el valor de la variable tal que la probabilidad de observar un valor de SH mayor que él valga 0,10?

$P(X \geq x?) = 0,10 = 1 - 0,90$

$P(Z \geq z?) = 0,10 = P(Z \leq z?) = 0,90$

En esta ocasión se conoce es el valor de P(0,90). Se debe encontrar el valor de Z que hace que se cumpla que $P(Z \leq z?)=0,90$. Ese valor de Z se busca en la Tabla 1, entrando por el cuerpo de la tabla.

$P(Z \leq 1,285) = 0,90$

Luego, dado que el área bajo la curva N(0,1) equivalente al área de n(20,1; 4,3), utilizando la expresión 4.7

$$1,285 = \frac{x-20,1}{4,3}$$

$$x = (1,285 \cdot 4,3) + 20,1$$

$$= 25,625$$

La SH tal que la probabilidad de observar un valor mayor que él es 0,1 es 25,6%.

Relaciones entre el modelo Normal y los modelos discretos Binomial y Poisson

El modelo normal da aproximaciones buenas a modelos discretos como el binomial y el Poisson. La distribución binomial se aproxima a la normal cuando *np* y *n*(*1-p*) son suficientemente grandes, mayor o igual a cinco, pues el estadístico que se usa para realizar inferencias sobre proporciones $\frac{\bar{X}-np}{\sqrt{n\,p\,(1-p)}}$ sigue una distribución normal. De igual forma cuando el parámetro λ de la distribución Poisson es grande ($\lambda > 5$) la distribución de probabilidades aproxima a la distribución normal con 90% de confianza.

MUESTREO Y DISTRUBUCIONES DE ESTADISTICOS MUESTRALES

Introducción

Uno de los propósitos esenciales tanto en trabajos de investigación como profesionales es conocer algún aspecto de una población específica. Las características de los fenómenos geológicos (eventos y rocas), como se mencionó en el primer capítulo, impiden contar con la información de todos los individuos de la población por esa razón se toma una muestra, para que con esfuerzos y costos razonables, se obtengan conclusiones tan válidas como las que se habrían tenido con un censo.
La información conseguida en la muestra se utiliza para estimar los parámetros poblacionales pues se conocen las relaciones entre ambos, μ - $\bar{X}$, $\sigma^2 - S^2$, π - p, etcétera (Fig. 1).

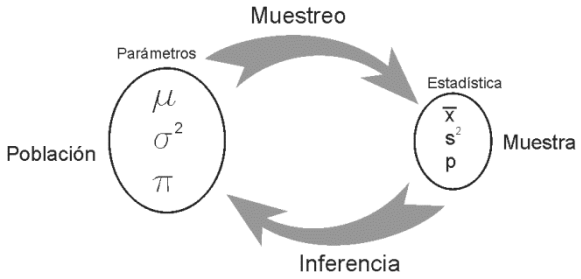


Figura 1. Los estadísticos obtenidos con los datos del muestreo permiten inferir los parámetros poblacionales.

Muestreo

Existen muchas razones para tomar una muestra, en general son exploratorios (estimar el grado de contaminación de un acuífero, la porosidad de una arena o la selección de un sedimento, entre otros), pero otras veces apunta a la estimación de un recurso (ley media, variabilidad, cubicación). El objetivo más usual es estimar el valor medio de uno o más constituyentes de una roca. También suele ser estimar la variabilidad que posee una sustancia ya sea para compararla con la que presentan las

rocas que se han formado en ambientes geológicos semejantes o bien para estudiar las relaciones entre los constituyentes o cuáles son las causas y los procesos que producen esas variaciones.

Independientemente de cuál sea el objetivo, los vincula el hecho que no alcanza sólo caracterizar la muestra con algún estadístico como la media y la varianza, sino que es necesario utilizar la información que provee la muestra para extender o realizar inferencias sobre la población. Sintéticamente se trata de estimar un parámetro de la población de *N* elementos con la información de una muestra de tamaño menor *n*.

Antes de avanzar conviene recordar y aclarar que se entiende por población, muestra y por muestreo. La **población estadística** es la totalidad de medidas, valores o cualidades que son motivo del estudio. Se trata de la población de la cual se harán las inferencias, llamada población objetivo. Una **muestra** es una parte representativa del todo, o lo que es lo mismo, es una parte representativa de la población. Recuerde que se habla de muestra en sentido estadístico, es decir la muestra está representada por un número *n* de atributos (medidas, valores o cualidades) observados en cada unidad muestral y no se trata de una muestra de roca o un pequeño volumen de agua.

En esta definición**, representativo** hace referencia a dos cualidades. Por un lado sus características deben coincidir con los de la población esto es, la proporción y distribución de las características que se investigan deben ser iguales en la población y en la muestra (Fig. 2). Esto, de alguna manera, remite al tamaño de la muestra, que debe ser lo suficientemente grande como para que todas las características de la población estén representadas en ella. Por otro lado, alude a la **equiprobabilidad**, lo que significa que todos los elementos de la población deben tener igual probabilidad de ser elegidos. Naturalmente ambos aspectos están estrechamente vinculados con el conocimiento que se tenga de la población muestreada, especialmente el con la homogeneidad o heterogeneidad de la población. La muestra debe ser representativa si se va a usar para estimar las características de la población.

El **muestreo**, según el diccionario de la real academia española, tiene dos acepciones. En la primera se refiere a todas las operaciones que conducen a establecer los parámetros principales de una población. La segunda dice textualmente que "muestreo es una operación estadística mediante la cual se eligen *n* individuos con objeto de representar a una población *N* mucho mayor".

Premisas para un buen muestreo

El plan de muestreo debe formularse en las etapas iniciales del trabajo debido al rol fundamental que ocupa. Cualquier investigación fracasa si no se tiene un plan de muestreo, un muestreo inadecuado no se reemplaza con ningún procedimiento experimental o estadístico. El cuidado que se tenga en la planificación del muestreo trae aparejadas el ahorro de dos insumos básicos para los trabajos de

investigación y profesionales, tiempo y dinero. Un muestreo bien planeado evita que se tome información redundante y se olvide relevar información vital cuando se muestrean las unidades.

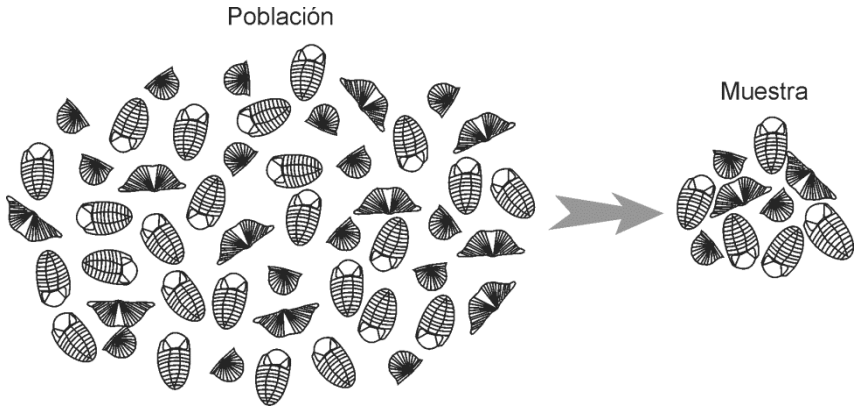


Figura. 2: Población y Muestra. Suponga que la población está formada por 50% de trilobites, 30% de bivalvos y 20 % de braquiópodos; es de esperar que si la muestra es representativa se guarden estas proporciones de los individuos de cada tipo en la muestra.

Un muestreo planificado correctamente contempla:

1°. Establecer explícitamente el objetivo del estudio, incluyendo la formulación de las hipótesis del trabajo (lo que se espera que indiquen los datos), permite la elección del método de muestreo.

2°. Delimitar perfectamente en espacio y tiempo la población objetivo que se va a muestrear.

3°. Definir, describir y listar los elementos de la población, esto se conoce como marco muestral.

4°. Especificar los datos a recolectar y que observaciones y/o medidas que se van a realizar. Por ejemplo si se muestrea la roca entera para determinar la composición química o propiedades físicas. La composición, por su parte se podrá expresar en óxidos de elementos químicos o minerales. Las propiedades físicas a relevar pueden ser dureza, velocidad de transmisión de las ondas sísmicas y otras ondas, resistividad, paleomagnetismo, entre otras. También se encuentran contenido fosilífero, presencia de estructuras, propiedades estructurales como rumbo y buzamiento y dirección de corriente.

5°. Evaluar las fuentes de variabilidad que presentan los datos ya sea a través de experiencias anteriores o por algún marco teórico.

6°. Convenir la cantidad o peso de material que se va a recolectar en cada punto de muestreo. A esto se suma la necesidad de definir, en forma precisa, la operación de recolección de muestra geológica, es decir fijar el método a seguir en la toma de la muestra. Por ejemplo si se toman medidas de rodados *in situ* de donde a donde se toman ancho, largo y alto, si se trata de rumbos y buzamientos si se va a emplear la regla de la mano derecha. Acordar también el instrumento de medida: cinta métrica, calibre, carta de colores, etcétera. Por último pactar lugar y frecuencia con que deben tomarse las muestras.

7°. Estipular la precisión y exactitud requeridas en el estudio.

Población objetivo

Para iniciar un estudio geológico se deben tener claro no solo el objetivo de la investigación tan específicamente como sea posible, sino también definir cuidadosamente la población objetivo. La **población objetivo** o blanco es aquella sobre la que se realizan las inferencias con base a los datos que se obtienen en el muestreo. La población a ser muestreada debe coincidir con la población objetivo. Sin embargo, como sucede cuando se mapea en el campo el contacto de dos unidades geológicas, o se datan fósiles de una asociación, o se clasifican rocas, las relaciones entre población objetivo y la población que se va a muestrear no son simples. En los dos primeros casos, al igual que en los ensayos biomédicos, la población objetivo no se puede identificar y se debe establecer claramente la relación entre población objetivo y población muestreada. En el tercero, como sucede con la causa de cáncer, la relación entre la población objetivo y su relación con la muestreada es poco clara. Cuando la población muestreada es más limitada y difiere mucho, se debe tener en cuenta que cualquier conclusión que se alcance sólo podrá aplicarse a la población muestreada.

La población objetivo debe estar delimitada en tiempo y espacio. Además, la definición debe tener la descripción de los elementos de la población que serán incluidos. También es necesario definir las unidades de muestreo y seleccionar cuidadosamente la característica o características para ser tomadas, esto permite planificar el diseño de muestreo para evitar omitir datos cuando se muestrea.

Factores geológicos que afectan el muestreo

La variabilidad tiene un rol importante en la planificación y en la elección del método de muestreo. Se describirá someramente la variabilidad natural y la originada durante la recolección de la muestra geológica. La variabilidad natural de una sustancia en la roca se relaciona con aspectos genéticos, con cuestiones de escala de trabajo y con los caracteres geométricos del cuerpo de roca. Su conocimiento puede provenir de muestreos preliminares o bien se puede estimar a partir de conocimientos geológico obtenido de experiencias en rocas similares o, en algunos casos, proceder de la interpretación de una teoría geológica.

En sentido muy general se ha descripto que las rocas formadas a altas temperaturas son menos variables que aquellas formadas a bajas temperaturas, debido a que las primeras se encontrarían en equilibrio con el medio. Los basaltos, diabasas, gabros, anortositas, mármoles, cuarcitas y eclogitas son relativamente homogéneas, en tanto las andesitas, dioritas, riolitas, granodioritas, granitos, hornfels, milonitas esquistos, gneises y rocas piroclásticas en general tienden a ser menos uniformes. Las rocas sedimentarias, como grupo, son más variables que las rocas ígneas y metamórficas, sin

embargo algunas, como las areniscas muy maduras, tienen una composición extremadamente uniforme. Las calizas, cherts, diatomitas son más homogéneos que las gravas, waques, dolomitas, till y rocas de composición intermedia como las margas.

La variabilidad/homogeneidad de las rocas también depende de la variabilidad de la composición mineralógica, por ejemplo la composición química de los granates es altamente variable, pero en un único cuerpo de roca la composición química suele ser más o menos uniforme o variar gradualmente de un lugar a otro.

Además de la variabilidad natural, inherente del cuerpo de roca o de otro objeto geológico, existen otras fuentes de variabilidad relacionadas con el muestreo. Entre ellas se encuentra las introducidas durante la **recolección de la muestra** geológica, en la **preparación de la muestra** (i.e. triturado, submuestreo, etc.) y la variabilidad **analítica** producidas en los laboratorios donde se determina la geoquímica de la roca. La variabilidad introducida durante la preparación de la muestra y la originada en los ensayos químicos son tratados extensamente por Koch y Link (1980).

Las cuatro fuentes de variabilidad se expresan en un modelo estadístico en el caso de una única variable es $X = \mu + \varepsilon_n + \varepsilon_s + \varepsilon_p + \varepsilon_a + e$, donde ε_n es la variabilidad natural, ε_s es la variabilidad de muestreo, ε_p es la variabilidad de la preparación y ε_a de la muestra, es la variabilidad analítica y *e* es la variabilidad del azar que no está contemplada por las otras fuentes de variabilidad.

Métodos de muestreo

Una vez se han especificado la población objetivo y las características, el punto siguiente es elegir un método para obtener una muestra representativa de toda la población. Elegir el método de muestro apropiado es fundamental para emplear correctamente los métodos estadísticos e inferir los parámetros poblacionales a partir de una muestra así como para estimar valores medios y la variabilidad.

Los métodos para seleccionar una muestra representativa son numerosos. Establecer el método adecuado se basa en conocer la naturaleza de la población (origen, estructuras presentes, forma y posición del cuerpo de roca, etc.), la accesibilidad al lugar de muestreo, el tipo de información buscada, la dificultad para tomar una muestra, el costo del equipamiento y el tiempo y dinero disponibles.

Los métodos de muestreo se clasifican según el número de muestras y de acuerdo a la manera usada para seleccionar los elementos de la población que integren la muestra.

De acuerdo con el **número de muestras** tomadas de una población se distinguen el muestreo simple, el doble y el múltiple. En el **muestreo simple** se toma únicamente una muestra de la población, cómo

es solo una, el tamaño debe ser lo suficientemente grande para extraer una conclusión. El problema es que tomar una muestra grande muchas veces cuesta demasiado dinero y demanda mucho tiempo. El **muestreo doble** se realiza cuando el resultado del estudio de la primera muestra no es concluyente. Entonces se saca una segunda muestra de la misma población. Las dos muestras se combinan para analizar los resultados. La ventaja del muestreo doble es que permite comenzar con una muestra relativamente pequeña para ahorrar costos y tiempo. Si la primera muestra arroja una resultado definitivo, no hace falta tomar la segunda muestra. Por último, el **muestreo múltiple** es semejante al procedimiento del muestreo doble, excepto que el número de muestras sucesivas requerido para llegar a una decisión es más de dos muestras.

Por otra parte se encuentran los procedimientos para seleccionar cuales serán los elementos de una muestra. Se distinguen entre ellos dos grandes categorías: los muestreos no probabilísticos y los muestreos probabilísticos.

Muestreos no probabilísticos

En los **muestreos no probabilísticos**, es el que se usa en forma empírica, intervienen opiniones y criterios personales o simplemente razones de comodidad al momento de elegir los elementos que conforman la muestra. La información obtenida **no sirve para hacer inferencias** sobre la población porque no se puede utilizar la teoría de probabilidad para medir el error de muestreo, sólo sirven para describir la muestra. Entre los muestreos no probabilísticos se encuentran los muestreos de juicio, los muestreos por cuotas, los accidentales, los incidentales y los muestreos bola de nieve.

a) **Muestreos de juicio** u opinión, en ellos los elementos de la muestra son seleccionados mediante juicio personal. La persona que selecciona los elementos de la muestra, usualmente es un experto, sin embargo como usa su criterio para elegir los elementos de la muestra y no un método probabilístico, la información no se puede usar para hacer inferencias de la población. Las ventajas de este muestreo son comodidad y bajo costo. El muestreo de juicio, lamentablemente, suele ser muy popular entre los científicos pero se recomienda enfáticamente evitarlo.
b) **Muestreos por cuotas** requieren conocer la población y/o los individuos más representativos o adecuados para la investigación. En este muestreo se fijan cuotas que consisten en número de individuos con determinadas condiciones. Por ejemplo para una encuesta sobre el lugar conveniente para ubicar desechos urbanos se realiza a veinte mujeres entre 25 y 40 años residentes en La Plata que no sean nacidas en La Plata y luego se administra la encuesta a los primeros elementos que reúnan estas características para integrar la muestra.
c) **Muestreo accidental**, los individuos de la muestra se obtienen sin ningún plan, son elegidos producto de circunstancias casuales. Por ejemplo para un sondeo de opinión se entrevistan los 50 primeros transeúntes que pasan por una esquina a las 12 del mediodía.

d) **Muestreo incidental** o de conveniencia se seleccionan directa e intencionalmente a los individuos de la población que formaran la muestra. Se usa en estudios exploratorios y en pruebas piloto.
e) **Muestreo bola de nieve** la premisa es que los elementos se relacionen entre sí. Se localizan algunos individuos de la población y estos conducen a otros que llevan a otros y así hasta tener una muestra de tamaño suficiente. Se usa en estudio de poblaciones de sectas, enfermos, oficios, etcétera.

Muestreos probabilísticos

En los muestreos probabilísticos, también llamados aleatorios o estadísticos, la selección de los elementos que conforman la muestra debe efectuarse de manera tal que cada elemento de la población tenga igual oportunidad de ser elegido. La muestra se dice que es aleatoria o probabilística pues la obtención de los elementos de la población es objetiva. Los muestreos aleatorios **permiten realizar inferencias** dado que se puede medir el error muestral como probabilidad bajo la curva normal.
Los tipos comunes de muestreo aleatorio son el muestreo aleatorio simple, muestreo sistemático, muestreo estratificado y muestreo de conglomerados.
a) **Muestreo aleatorio simple**. En este muestreo las muestras están distribuidas aleatoriamente en el espacio o en el tiempo. La selección se realiza de modo que todas las muestras posibles de tamaño *n* tengan la misma probabilidad de ser elegidas y que cada elemento de la población tenga una oportunidad igual de ser incluido en la muestra. Es condición para aplicarlo que la **población sea homogénea** respecto a la variable de interés. Los muestreos aleatorios simples se pueden realizar en superficies o en estructuras lineales como vetas, transectas y ríos. En los muestreos de superficie se suele superponer una cuadrícula o grilla cuadrada, pero puede ser rectangular si se sospecha que los datos tienen estructura lineal, o con coordenadas polares para patrones radiales como las rocas piroclásticas generadas por los volcanes. Cada unidad de muestreo o cada nodo de la cuadrícula, según corresponda, se identifica con un número y utilizando una tabla de números aleatorios[11] (o la función *random* de las maquinas de calcular, o la función *aleatorio* de Excel) se eligen las unidades de muestreo de donde se tomar los datos.
El muestreo aleatorio se puede utilizar siempre y cuando no se requiera cobertura areal y cuando los datos no estén estratificados.
b) **Muestreo sistemático**. Este muestreo se utiliza cuando el objetivo del trabajo requiere cobertura areal. La distancia entre los puntos de muestreo es uniforme en el espacio y/o tiempo (Fig. 4). Si bien es fácil de llevar a cabo y está menos expuesto a los errores de selección de los puntos de muestreo, cuando la población tiene un comportamiento cíclico como el producido por plegamientos la muestra puede ser poco representativa o sesgada.

Muestreos sistemáticos se usan corrientemente en perforaciones y en muestreos de suelos. También se utilizan en estructuras y cuerpos lineales como las arenas de playa, en estos casos las transectas se

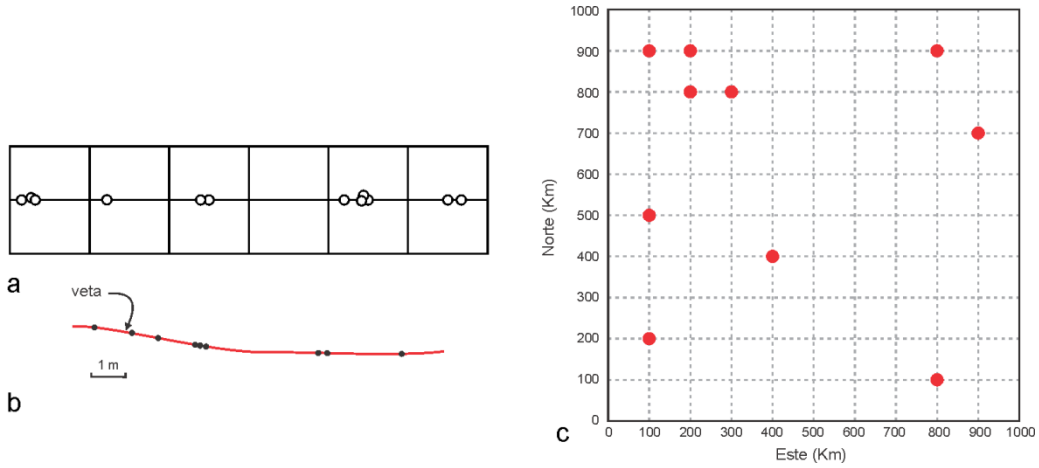


Figura 3. Muestro Aleatorio Simple. a. Esquema de muestreo en una línea. b. Muestreo de una veta. c. Esquema de muestreo de superficie con una malla cuadrada.

ubican perpendiculares a la estructura, regularmente distanciadas y se toman muestras en cada transecta a distancia también regulares.

A veces en los muestreos geológicos se combinan métodos aleatorios y sistemáticos.

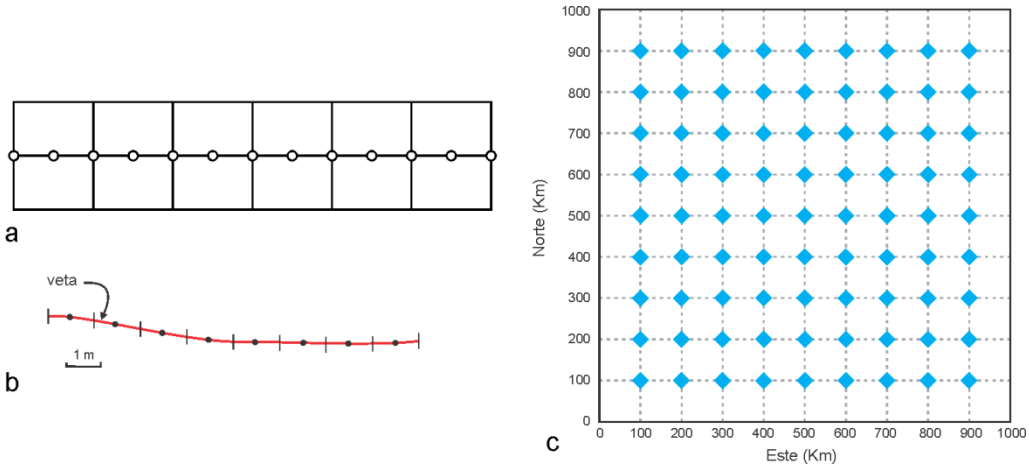


Figura 4. Muestro sistemático. a. Esquema de muestreo en una línea. b. Muestreo de una veta. c. Esquema de muestreo de superficie con una malla cuadrada.

c) **Muestreo estratificado simple**. Se usa cuando se realiza una investigación en una población que naturalmente está dividida en estratos o capas de diferentes tamaños (subpoblaciones). Los estratos son **homogéneos al interior** de cada uno y **diferentes entre sí**, además no se superponen o traslapan. La división en estratos puede ser litológica (unidades de rocas estratificadas), pero también se puede efectuar con base en la topografía, los horizontes de suelo, cambios de color de suelo, en rocas ígneas

y metamórficas los estratos pueden ser diferentes rocas que definen facies o zonas metamórficas, litológicas aflorantes en una región, etcétera (Fig. 5). Como los estratos son más homogéneos que la población como un todo y resulta importante que en la muestra haya representación de todos y cada uno de los estratos, se debe utilizar una estrategia de muestreo que garantice la inclusión de elementos de cada estrato. Una muestra aleatoria estratificada se obtiene seleccionando una muestra aleatoria simple en cada estrato. El tamaño de la muestra de cada estrato puede ser proporcional o no proporcional al tamaño del estrato.

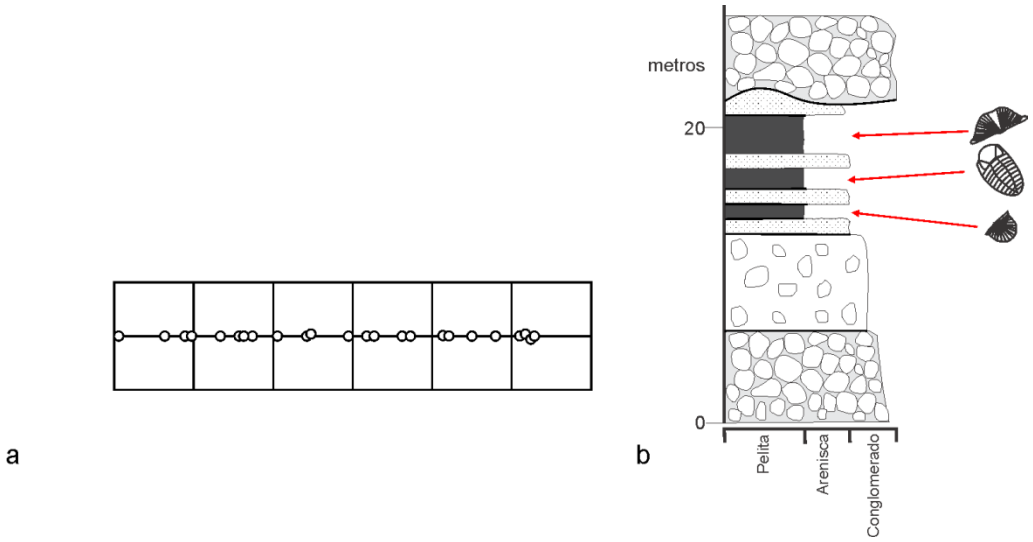


Figura 5. Muestreo estratificado simple. a. Esquema de muestreo en una estructura lineal (las líneas verticales representan estratos). b. Esquema de muestro en un perfil, cada estrato fosilífero tiene solo un tipo de fósil.

d) **Muestreo de conglomerados**. Se utiliza en poblaciones que naturalmente están divididas en grupos o subpoblaciones llamadas conglomerados. Cada conglomerado presenta la misma variabilidad de la población, por eso **los conglomerados son muy parecidos entre sí**. El concepto de conglomerado es opuesto al de estrato dado que los estratos son homogéneos al interior y diferentes entre sí (Fig. 6). Debido a la heterogeneidad interna, la totalidad de los elementos del conglomerado representan fielmente a la población, de modo que conviene incluirlos a todos en la muestra. Como todos los conglomerados son equivalentes, la selección de los conglomerados que integran la muestra se realiza al azar o con un método sistemático.

Sin embargo, los elementos individuales dentro de cada conglomerado tienden a ser iguales y, si bien no se muestrean todos los conglomerados, la variación entre los elementos obtenidos de los seleccionados es, por lo tanto, frecuentemente mayor que la que se obtiene muestreando la población entera mediante muestreo aleatorio simple.

Es recomendable un muestreo de conglomerados, por ejemplo, si la población objetivo tiene gran extensión areal con muchos afloramientos de la misma unidad y la variabilidad local y regional son aproximadamente iguales. En un caso extremo se podría incluso obtener la muestra de un solo

afloramiento. De forma semejante, se podría muestrear verticalmente hacia abajo en un único punto cuando la estratificación es horizontal, como en un yacimiento de petróleo, en un manto de carbón o en el agua de mar.

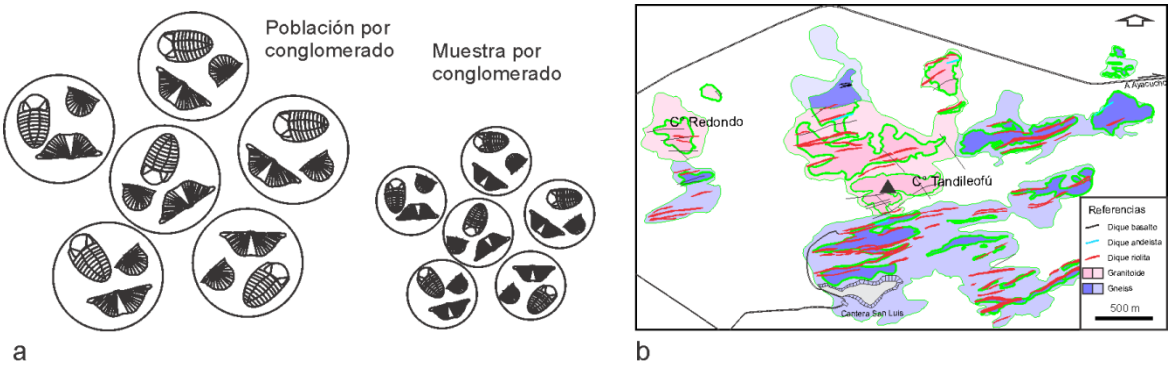


Figura 6. Muestreo por conglomerados. a. Muestreo esquemático. b. Mapa geológico.

Tamaño de la muestra y precisión en la estimación

Simultáneamente con la elección del método adecuado de muestreo se debe establecer el tamaño de la muestra y la precisión en la estimación de los parámetros que demanda el trabajo.

Tamaño y precisión están íntimamente ligados. El nivel de precisión en la estimación generalmente es 10% a 15% en investigación y 20% a 25% en toma de decisiones. El número de unidades de muestreo que formaran la muestra debe estar determinado por la variación existente entre las unidades de muestreo. Cuanto mayor es la variabilidad o heterogeneidad de la población, mayor debe ser el tamaño de la muestra.

En este sentido, contar con información de una muestra exploratoria permite estimar el tamaño de la muestra utilizando la ecuación

$$n = \left(\frac{S}{Pr\,\bar{X}}\right)^2, \qquad (5.1)$$

donde n es el tamaño de la muestra, S el desvío estándar y $\bar{X}$ la media de una muestra preliminar de la población y Pr el nivel de precisión requerido. Sin embargo se debe confirmar si el tamaño de la muestra es compatible con el presupuesto con el que se cuenta para el estudio.

Por otra parte, es importante destacar la relación que existe entre el volumen de la muestra y el volumen de la población, conocida geológicamente como **Razón de Muestreo**. Para poner en evidencia este punto suponga un yacimiento de cobre cubicado en 160 millones de toneladas fue reconocido a través 40.000 metros de sondajes con un peso total de 103 toneladas, la Razón de Muestreo es $103/(160 \cdot 10^6) = 0{,}000064\%$. Esto quiere decir que se estimó la ley del yacimiento de ciento sesenta millones de toneladas conociendo sólo sesenta y cuatro millonésimas del yacimiento. Razonamientos del mismo tipo pueden hacerse al pensar cualquier tipo de estudio y usarlos como argumento para obtener fondos destinados al muestreo.

Volumen de la muestra y distancia entre muestras

Volumen de material y distancia entre muestras también se estipulan cuando se elige el método de muestreo. Conviene aclarar en este punto cual es la representatividad del material que se va a recolectar, es decir la muestra geológica que, como se mencionó previamente, está vinculado con la variabilidad que presente la variable (cualidad o cantidad) que se investiga. Cuando los elementos de la población son homogéneos un cierto número de ellos la representa y si la variación es regular en el espacio, el problema es fácil. Pero la toma de muestra de una población heterogénea se dificulta a medida que aumenta la variabilidad y el problema se agudiza cuando se suman variaciones espaciales. En casos extremos, si la variación es aleatoria, la muestra representa sólo su propio volumen y el valor de la variable es independiente de la ubicación de la muestra y en consecuencia, la distancia entre muestras sucesivas no es importante.

Toma de la muestra

Indudablemente la mayoría de las muestras geológicas son **muestras de mano** de afloramientos, son excepciones las muestras tomadas para minería, petróleo o agua. Si bien muchas teorías geológicas y conclusiones se apoyan en la información obtenida de las muestras de mano, estas suelen ser muy sesgadas. El sesgo se produce por factores relacionados con la naturaleza del ser humano y por causas geológicas. Es innegable que los coleccionistas se interesan por objetos raros, los comunes se descarta, la recolección de rocas no escapa de esto. La tendencia de los investigadores es a recolectar muestras de mano que contengan especímenes que sean conspicuos por color, tamaño o por alguna otra razón. Esto tiene implicancias importantes en los trabajos geológicos y paleontológicos pues entonces la muestra de mano es tomada de forma subjetiva lo que impide utilizar los datos de la muestra para realizar inferencias sobre la población. Por otra parte las muestras de mano son tomadas en afloramientos y las propiedades de las rocas en los afloramientos son diferentes a los de las partes no expuestas del cuerpo de roca. Además un afloramiento puede tener unos pocos lugares donde sea factible tomar muestras de mano. Sin embargo, a pesar de estas limitaciones, la mayor parte del conocimiento geológico se sustenta en datos obtenidos de afloramientos y si las muestras de manos se obtienen con algún muestreo probabilístico pueden ser usadas para obtener datos válidos para realizar inferencias.

Las muestras de **canaleta** se obtienen cortando una canaleta en la superficie de la roca. Los fragmentos que se producen durante el corte de la canaleta constituyen la muestra geológica. Para no introducir sesgo se deben mantener las dimensiones de la canaleta (usualmente entre 5 y 10 cm de ancho y 2 cm de profundidad) y la distancia entre canaletas sucesivas (muestreos sistemáticos). Las muestras de **esquirlas** (chips) se sacan de forma semejante al de canaleta con la diferencia que la

muestra geológica la forman las pequeñas esquirlas de roca y no se intenta cavar una canaleta. La separación entre canaletas suele ser de 2 metros pero es menor cuando la mineralogía presenta cambios y particularmente si los minerales poseen diferente dureza o friabilidad. Ambos tipo de muestras, canaletas y esquirlas, son comunes en muestreos de vetas ya sea en minas subterráneas o en vetas superficiales, no obstante se pueden utilizar en afloramientos y en perfiles. Los datos que se obtienen en los muestreos de canaletas son mejores dado que el volumen de roca recogido es mayor que el recuperado en las muestras de esquirlas.

El **análisis modal** es un método que se utiliza para estimar la composición mineralógica de una roca. Los datos se obtienen superponiendo una cuadricula (grilla) sobre un corte delgado y contando o eventualmente midiendo los minerales que se encuentran en los nodos de la cuadricula. Se trata entonces de un muestreo sistemático donde existe la componente aleatoria, el análisis estadístico de los datos está asociado con la distribución binomial. Los problemas del análisis modal se encuentran en la identificación de los minerales y en la variabilidad introducida por el operador ya sea porque diferentes personas estudien diferentes cortes delgados de la misma roca o por cansancio en el mismo operador.

Distribuciones en el muestreo

Con los métodos de muestreo se ha resuelto el problema de la representatividad y de la aleatoriedad, sin embargo resta el problema de la inferencia a partir de los estadísticos de muestra de los parámetros poblacionales. Es decir averiguar, si es que existe, cuál es la relación entre los estadísticos y los parámetros. Suponga que se tiene una población finita de tamaño N y que interesa estudiar el parámetro cualquiera que se simboliza con θ (por ejemplo μ, σ^2, π). Para ello se extrae, de manera aleatoria, k muestras de las M posibles de tamaño n tal como se ilustra en la figura 7. Se puede observar que cada muestra aporta un valor estimado del parámetro θ, simbolizado con $\hat{\theta}$. Este valor varía de muestra en muestra. La variabilidad en los estadísticos de muestra provienen de lo que se llama el **error de muestreo** debido al azar, esto es las diferencias entre cada muestra y la población y entre las diferentes muestras se debe únicamente a las características de los individuos que fueron seleccionados por azar para integrar la muestra.

La distribución de todos los valores que puede tomar el estadístico calculado a partir de muestras del mismo tamaño seleccionadas de forma aleatoria de la misma población se llama **distribución muestral** de esa estadística. Se describirán detalladamente las distribuciones de la media muestral y de la varianza muestral y sintéticamente la de diferencia de medias.

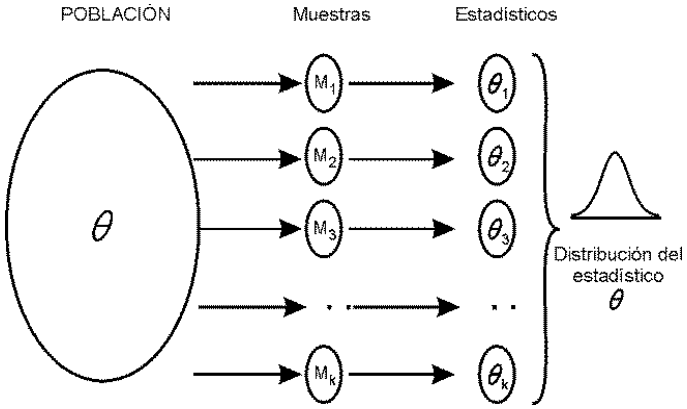


Figura 7. La distribución del estadístico θ.

Distribución de las Medias Muestrales

Con el objeto de ver la relación entre la media muestral y la media poblacional se propone analizar una población hipotética formada sólo por tres individuos, el número 2, el 4 y el 6. Esta población puede ser descripta a través de una gráfica como la de la figura 8, con su parámetro poblacional μ= 4 y con su varianza poblacional $\sigma^2 = 2{,}66$.

$X = \{2, 4, 6\}$

$$\mu = \frac{2+4+6}{3} = \mathbf{4}$$

$$\sigma^2 = \frac{(2-4)^2 + (4-4)^2 + (6-4)^2)}{3} = \frac{8}{3} = \mathbf{2}$$

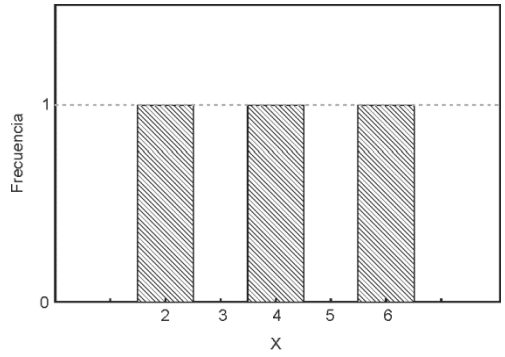
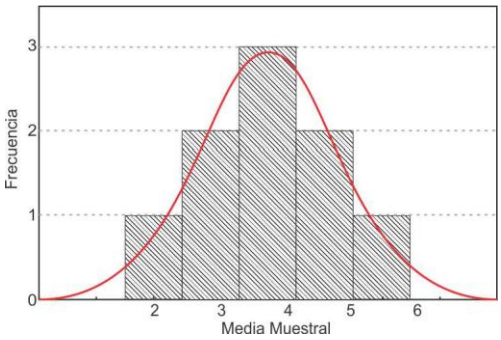


Figura 8: Distribución de frecuencias y parámetros de la población de la variable X.

Suponga que se realiza un muestreo exhaustivo de esta población de todas las muestras de tamaño dos, en un muestreo con reposición, que es análogo a lo que ocurre en una población de tamaño infinito, como se considera son la gran mayoría de las poblaciones geológicas. Se obtienen las siguientes muestras: (2 – 2), (2 – 4), (2 – 6), (4 - 2), (4 – 4), (4 – 6), (6 – 2), (6 -4) y (6 -6). A continuación se calcula la media muestral ($\overline{X}_i$) de cada una de esas muestras (Fig. 9). Dado que las medias muestrales calculadas de esta población no son todas iguales, la media muestral $\overline{X}_i$ se ha convertido en una nueva variable. Es conveniente describir la distribución de frecuencias y calcular los parámetros poblacionales ($\mu_{\overline{X}}$ y $\sigma^2_{\overline{X}}$, note el subíndice de ambos) de esta nueva variable media muestral $\overline{X}$ (Fig. 9).

Así, la media de la distribución de $\bar{X}$ es $\mu_{\bar{X}} = 4$ y la varianza es $\sigma^2_{\bar{X}} = 1{,}33$.

Muestra	$\bar{X}$	$(\bar{X} - \mu)^2$	$(\bar{X} - \mu)^2$
2 ; 2	2	$(2 - 4)^2$	4
2 ; 4	3	$(3 - 4)^2$	1
2 ; 6	4	$(4 - 4)^2$	0
4 ; 2	3	$(3 - 4)^2$	1
4 ; 4	4	$(4 - 4)^2$	0
4 ; 6	5	$(5 - 4)^2$	1
6 ; 2	4	$(4 - 4)^2$	0
6 ; 4	5	$(5 - 4)^2$	1
6 ; 6	6	$(6 - 4)^2$	4
Total	36		12



$\mu_{\bar{X}} = 36/9 = 4$

$\sigma^2_{\bar{X}} = 12/9 = 1{,}33$

Error típico: $\sqrt{\sigma^2_{\bar{X}}} = \sqrt{1{,}33} = 1{,}15$

Figura 9. Distribución de frecuencias y parámetros de la población de la variable $\bar{X}$de un muestro con reposición de muestras tamaño 2 de la variable X.

De todo lo anterior se puede concluir que:

1º. La media de las medias muestrales coincide con la media de la población,

$$\mu_{\bar{X}} = \mu. \tag{5.2}$$

Para el ejemplo el valor de ambas es 4.

2º. La varianza de las medias muestrales es igual a la varianza de la población dividido el tamaño de la muestra *n*,

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}. \tag{5.3}$$

Para el ejemplo $\sigma^2_{\bar{X}} = 1{,}33$ y $\sigma^2 = 2{,}66$.

3º. El desvío estándar de la distribución de medias muestrales, llamado **Error Estándar**, es

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}. \tag{5.4}$$

4º. La distribución de las medias muestrales es simétrica, aunque la distribución de la variable no lo sea.

Esta distribución simétrica induce a pensar en el modelo de distribución de una variable normal, en este caso una normal cuyos parámetros son μ y $\sigma^2_{\bar{X}}$. Por lo que se puede definir en forma informal y luego formal del Teorema central del Límite, también conocido como Teorema del límite Central.

Teorema Central Del Límite (Definición Informal)

Cuando se efectúa un muestreo aleatorio de tamaño fijo *n* de una población tiene forma arbitraria, pero media y varianza finita, la distribución de las medias muestrales tiende aproximadamente hacia una distribución de frecuencias normal a medida que el tamaño de la muestra aumenta.

Teorema Central Del Límite (Definición Formal)

Si $X_1, X_2, \ldots X_n$ son variables aleatorias independientes y tienen todas la misma distribución, con esperanza matemática, $E(X) = \mu$ y varianza $V(X) = \sigma^2$ finitas y ≠0. Se define una nueva variable aleatoria

$$Z_n = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}; \qquad \text{donde } \overline{X} = \frac{1}{n}\sum_{1}^{n} x_i \tag{5.5}$$

Entonces la función de distribución Z_n converge a una función de distribución normal estándar cuando $n \to \infty$.

Recapitulando, la diferencia entre **desvío estándar** y **error estándar** es que el primero se refiere a la distribución de la variable original, mientras que el último está relacionada con la distribución del estadístico. Si se analiza la distribución de las medias muestrales resulta claro que, a medida que el tamaño de la muestra aumenta, aumenta la probabilidad de incluir en la muestra la variabilidad inherente a los individuos que forman la población, por lo que esa media muestral, será mucho más parecida a la media poblacional que si su tamaño fuera pequeño. Esta reflexión apunta al denominador del error estándar de la distribución de media muestrales, que tiene como divisor del cociente a *n*, el tamaño de la muestra. A medida que *n* disminuye el error estándar aumenta.

Suponga que la curva A de la figura 10 es la distribución de la población de la variable *X* = largo de valva de *Ostrea maxima*, cuya media poblacional es 100 mm, μ=100 y desvío estándar poblacional igual a 15 mm, σ = 15. Si se realiza un muestreo con muestras de tamaño 10 (*n = 10*), la media de la distribución de las medias muestrales es $\mu_{\overline{X}(10)} = 100$ y el error estándar es $\sigma_{\overline{X(10)}} = 4{,}7$ (Fig. 10, curva B). Pero si el muestreo se realza con muestras de tamaño 100 (*n = 100*), obviamente la distribución de las medias muestrales tiene la media poblacional $\mu_{\overline{X}(100)} = 100$, pero el error estándar se reduce, $\sigma_{\overline{X(100)}} = 1{,}5$ (Fig. 10, curva C).

Es claro entonces que el tamaño de la muestra establece el parecido de las medias muestrales a la media poblacional debido a que existen más probabilidades de incluir individuos de la población y con ellos la variabilidad. Esto determina que la distribución del estadístico media muestral se encuentre apretada en torno al parámetro poblacional.

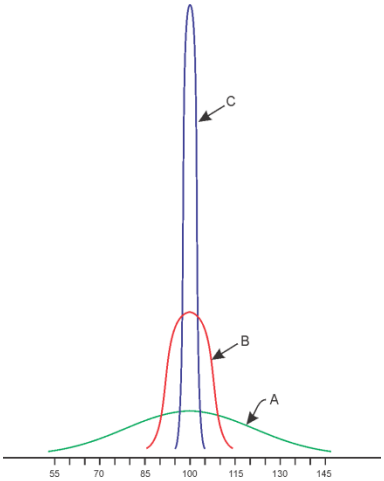


Figura 10. A: distribución de la población de la variable X=largo de valva de Ostrea maxima, μ=100 y σ=15.

B: distribución de medias muestrales de n=10, μ=100 y $\sigma_{\bar{X}(10)} = 4{,}7$

C: distribución de medias muestrales de n=100. μ=100 y $\sigma_{\bar{X}(100)} = 1{,}5$.

Por otra parte, cuando se selecciona una sola muestra de todas las posibles de una población, existe una probabilidad de error en la inferencia del parámetro dado que, aunque con probabilidad baja, una muestra puede tener valores medios alejados de la media muestral ya sea por muy pequeña o por muy grande. Recuerde que en el caso de la población de tres individuos muestreados con muestras de tamaño 2, sólo dos muestras tenían promedios bajos o altos, 2 y 6 con frecuencia igual a uno ($\mu_{\bar{X}} = \mu = 4$), en casos como este pueda ser que induzca a pensar que esa muestra, por tener esa media no pertenece a la población con media 4. La probabilidad de error se puede calcular utilizando la distribución Normal estándar recordando el teorema central de límite.

Suponga que se ha muestreado esa población y la media obtenida fue 6, entonces $Z = \frac{6-4}{\sqrt{\frac{2{,}66}{2}}} = 1{,}73$,

de la Tabla 1 del anexo, la P($z \geq 1{,}73$) = 0,0418. Existe 4,18% de probabilidad de encontrar en muestras de tamaño 2 medias mayores o iguales a 6, más adelante se verá si esta probabilidad es alta o baja en términos estadísticos.

Se señala en este punto que la diferencia entre el resultado obtenido de una muestra (un estadístico) y el resultado que se debería haber obtenido de al analizar la población (el parámetro correspondiente) se llama el error muestral o **error de muestreo**. El error de muestreo es medido por el error estadístico, en términos de probabilidad, bajo la curva normal. El valor del error muestral indica la precisión de la estimación de la población basada en el estudio de la muestra. Precisión es el alejamiento máximo que el investigador está dispuesto a permitir entre el estadístico y el parámetro correspondiente. Mientras más pequeño el error muestral, mayor es la precisión de la estimación.

Para explicar la relación entre la varianza poblacional y la muestral se retoma la población hipotética de los tres números {2, 4, 6} en la que se realiza un muestreo con reposición de muestras de tamaño dos (*n = 2*). Dado que el interés está puesto en la varianza muestral, $S^2 = \frac{\sum(x-\overline{X})^2}{n-1}$, se calculan las varianzas muestrales de todas las muestras. La figura 11 muestra la distribución de las varianzas muestrales. Se trata de una distribución asimétrica cuya media poblacional es $\mu_{S^2} = 24/9 = 2{,}66$.

Muestra	$\overline{X}$	$\sum_1^n (x-\overline{X})^2$	S^2
2 ; 2	2	$(2–2)^2+(2–2)^2$	0
2 ; 4	3	$(2-3)^2+(4–3)^2$	2
2 ; 6	4	$(2-4)^2+(6–4)^2$	8
4 ; 2	3	$(4–3)^2+(2–3)^2$	2
4 ; 4	4	$(4-4)^2+(4–4)^2$	0
4 ; 6	5	$(4-5)^2+(6-5)^2$	2
6 ; 2	4	$(6–4)^2+(2–4)^2$	8
6 ; 4	5	$(6-5)^2+(4–5)^2$	2
6 ; 6	6	$(6-6)^2+(6–6)^2$	0
Total	36		24

$\mu_{S^2} = 24/9 = 2{,}66$

$\sigma^2 = 2{,}66$

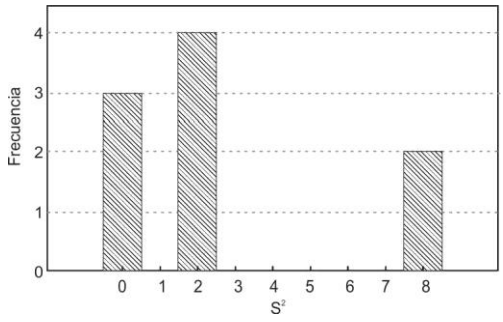


Figura 11. Distribución de frecuencias y parámetros de la población de la variable S^2 de un muestro con reposición de muestras tamaño 2 de la variable X.

El ejemplo demuestra que en muestras de tamaño dos (*n = 2*) la varianza muestral, S^2, coincide con la varianza poblacional, σ^2. Pero a medida que el tamaño de la muestra aumenta el cociente entre la varianza muestral S^2 y la varianza poblacional σ^2 se aleja de 1. Se ha encontrado que la relación entre la varianza muestral y la varianza poblacional está dada por el estadístico χ^2 (**chi cuadrado**)

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} . \qquad (5.6)$$

La cantidad (*n - 1*) se llama **grados de libertad** (ν). Entonces, la media de la distribución de la varianza muestral depende de la cantidad (*n – 1*). El único parámetro de la **distribución** χ^2, la media, es igual a los grados de libertad. Hay una curva diferente para diferentes grados de libertad. A medida que cambian los grados de libertad la forma de la distribución cambia, las curvas tiene asimetría a la derecha, la asimetría disminuye a medida que *n* aumenta (Fig. 12). Como la varianza nunca puede ser menor que cero, la curva es siempre positiva. En la Tabla 2 del anexo se encuentra la función de densidad de χ^2 para varios grados de libertad ν.

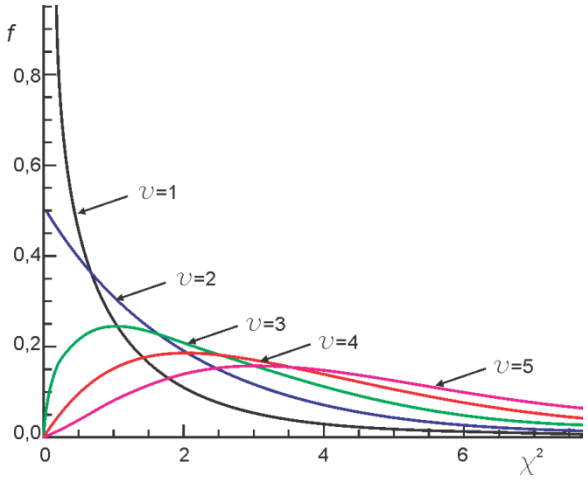


Figura 12. Distribución χ^2 para diferentes grados de libertad.

Distribución Muestral de Diferencia de Medias muestrales con varianzas conocida

Suponga ahora que se tienen dos poblaciones distintas y que sus medias y varianzas poblacionales son μ_1 y σ_1^2 y μ_2 y σ_2^2 respectivamente. Se elige una muestra aleatoria de tamaño n_1 de la primera población y una muestra independiente aleatoria de tamaño n_2 de la segunda población; se calcula la media muestral para cada muestra ($\overline{X_1}$ y $\overline{X_2}$) y la diferencia entre dichas medias ($\overline{X_1} - \overline{X_2}$). La colección de todas esas diferencias se llama **distribución muestral de las diferencias entre medias** o la **distribución muestral del estadístico** $\overline{X_1} - \overline{X_2} = \Delta_{\bar{X}}$. Como se vio, debido al proceso de muestreo, las medias muestrales son distintas y por lo tanto cada diferencia de medias variará, surge así la variable diferencia de medias, $\Delta_{\bar{X}}$. Se ha probado que la distribución de $\Delta_{\bar{X}}$ es aproximadamente normal para tamaños de muestra mayor o igual a 30 ($n_1 \geq 30$ y $n_2 \geq 30$). Si las poblaciones son normales, entonces la distribución muestral de medias es normal sin importar los tamaños de las muestras.

Por otra parte, se ha demostrado que $\mu_{\bar{X}} = \mu$ y que $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, en forma análoga se puede deducir que:

1°. La media de las diferencias de medias muestrales coincide con la media de la diferencia de la población $\mu_{\bar{X}1} - \mu_{\bar{X}2} = \mu_1 - \mu_2$.

2°. El error estándar de la distribución de diferencia de medias es $\sigma_{\bar{X}1} - \sigma_{\bar{X}2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

3°. La distribución de la variable diferencia de medias sigue un modelo Normal, que en su forma estandarizada es

$$z_{\Delta_{\bar{X}}} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \qquad (5.7)$$

INFERENCIA

Introducción

La inferencia estadística es la rama de la estadística que comprende los métodos y procedimientos para deducir propiedades de una población estadística (de algún parámetro o de la forma de la distribución) a partir de datos de una muestra a través de un razonamiento inductivo. Se puede definir la inferencia como una evaluación, juicio o generalización, derivada de observaciones empíricas que permite arribar a una implicancia lógica.

Por otra parte, al igual que en otras disciplinas, el conocimiento geológico avanza a base de formular hipótesis que se contrastan con la información que aportan los datos, luego, es necesario decidir si los datos apoyan o refutan la hipótesis. Es en este punto en el que la estadística inferencial ofrece su aporte dado que permite tomar decisiones objetivas cuando se abordan tanto problemas geológicos teóricos como prácticos. Sin embargo, como se mencionó en el capítulo precedente, para arribar a conclusiones válidas acerca de la población, la muestra debe ser representativa.

En este capítulo se abordan dos metodologías de la inferencia estadística: la estimación de los parámetros de una distribución y las pruebas de hipótesis.

Estimación de parámetros

La estimación consiste en elegir un valor que represente el parámetro poblacional. Existen dos alternativas para estimar los parámetros que son conocidas con el nombre de estimación puntual y estimación por intervalos.

Estimación puntual

En la estimación puntual se utiliza el valor de un estadístico de muestra para inferir el parámetro poblacional, por ejemplo $\bar{X}$ para estimar μ. El estadístico que se utiliza para estimar el parámetro

poblacional se llama **estimador**, en forma genérica se simboliza $\hat{\theta}$ en tanto el parámetro a estimar con θ. No hay ningún estimador perfecto pues se trata de variables aleatorias que tienen una función de densidad correspondiente a las distribuciones muestrales como se vio en el capítulo precedente. Un buen estimador debe ser insesgado, con varianza mínima, consistencia y suficiencia.

Son estimadores **insesgados** los que poseen la esperanza matemática igual al parámetro que tratan de estimar ($E(\hat{\theta})=\theta$). Por ejemplo la media muestral $\overline{X}$ es un estimador insesgado de la media poblacional μ pues E($\overline{X}$) = μ (Capítulo 5). Si un estimador es sesgado, el sesgo es la diferencia entre la esperanza matemática del estimador y el parámetro poblacional que se trata de estimar (Fig.1).

Un estimador insesgado $\widehat{\theta}$" es **eficiente** si comparado con otro estimador insesgado, $\hat{\theta}$, del mismo parámetro θ tiene menor varianza ($V(\widehat{\theta}") \leq V(\hat{\theta})$) (Fig.1).

Un estimador insesgado $\hat{\theta}$ del parámetro θ es **consistente** si cuando el tamaño de la muestra tiende a infinito el valor del estimador es igual al del parámetro ($n \to \infty: p(\hat{\theta} = \theta) \to 1$).

Por último $\hat{\theta}$ es un estimador **suficiente** de θ, si toda la información acerca del parámetro se obtiene de todos los datos de la muestra.

Aunque los estimadores reúnan todas las propiedades descriptas, cuando se realiza una estima por puntos casi siempre se cometen errores, por ejemplo cuando se utiliza solo una media muestral $\overline{X}$ para estimar la media poblacional μ. Recuerde el ejemplo desarrollado en el capítulo anterior para el muestro con reposición de una población de sólo tres individuos, en solo tres ocasiones de nueve, la media muestral $\overline{X}$ es 4. Es simple ver que en las estimas por puntos los resultados pueden ser cuestionados fácilmente pues basta tomar otra muestra y calcular sus estadísticos para objetar la estima, a esto se suma la imposibilidad de medir el error que se comete al realizarla.

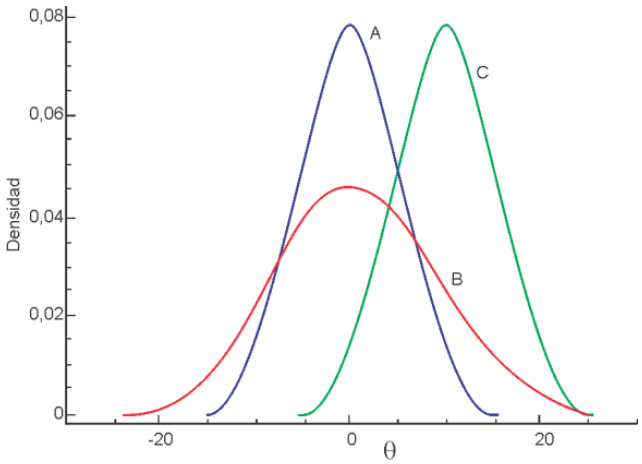


Figura 1. Función de densidad de tres estimadores del mismo parámetro. A y B son insesgados, C tiene sesgo, sobreestima el valor del parámetro. A es más eficiente que B (V(A) < V(B)).

Estimación por intervalos

Dado que las estimaciones puntuales pocas veces coinciden con los parámetros poblacionales es preferible determinar un rango dentro del cual se encuentre el valor parámetro que se va a estimar. Suponga que se preguntan el nivel de concentración de arsénico de un acuífero [As]. Primero se podrá hacer una estimación puntual, esto es calcular la el promedio de una muestra de pequeño tamaño (n = 6) y obtener $\bar{X} = 4$ mg/l. Luego hacer afirmaciones del tipo:

a) puede ser que la concentración se encuentra $(3{,}5 \leq [As] \leq 4{,}5)$ error 0,5;

b) casi seguro que la concentración se encuentre $(3 \leq [As] \leq 5)$ error 1;

c) seguro la concentración está entre $(2 \leq [As] \leq 6)$ error 2.

Cada afirmación tiene una medida de la seguridad de que la concentración esté comprendida en el intervalo y un error asociado.

Se define un intervalo de confianza

$$P[Li \leq \theta \leq Ls] = 1 - \alpha. \qquad (6.1)$$

El intervalo [*Li*, *Ls*] recibe el nombre de **Intervalo de confianza** del 100(1 - α) % para el parámetro desconocido θ. *Li* y *Ls* son los **Límites de confianza inferior y superior** respectivamente. ($1-\alpha$) es el **Nivel de Confianza o Coeficiente de confianza** (*CC*) asociado a este intervalo.

Está claro que se pueden calcular intervalos de confianza para cualquier parámetro, sin embargo en este capítulo se describe el procedimiento para determinar intervalos de confianza de algunos parámetros cuya distribución poblacional se conoce y que son los más utilizados en las geociencias.

En general, para cualquier parámetro θ y su correspondiente estimador $\hat{\theta}$, el intervalo de confianza tiene la forma:

$$P\left(\hat{\theta} - k \cdot \sigma_{\hat{\theta}} < \theta < \hat{\theta} + k \cdot \sigma_{\hat{\theta}}\right) = 1 - \alpha, \qquad (6.2)$$

donde: $\hat{\theta} - k \cdot \sigma_{\hat{\theta}}$ y $\hat{\theta} + k \cdot \sigma_{\hat{\theta}}$ son los límites inferior y superior del intervalo de confianza respectivamente, *k* es una constante relacionada a la distribución del estimador puntual y a 1 - α, α es la probabilidad de que el intervalo no incluya al verdadero valor del parámetro.

La expresión $P\left(\hat{\theta} - k \cdot \sigma_{\hat{\theta}} < \theta < \hat{\theta} + k \cdot \sigma_{\hat{\theta}}\right) = 1 - \alpha$ se lee: "el intervalo de límites inferior y superior tiene la probabilidad 1 – α de contener al parámetro θ".

Por tratarse de una probabilidad el **Coeficiente de Confianza** ($1 - \alpha$) es un valor entre 0 y 1, aunque corrientemente se expresa en forma porcentual, $(1 - \alpha) \cdot 100$. Para que la estimación sea buena el coeficiente de confianza debe ser grande, es decir 1 - α debe ser próximo a 1. Para el caso en que α = 10%, entonces 1 - α = 90%, se tiene un intervalo de confianza del 90% esto significa que la probabilidad de que el intervalo contenga al verdadero valor del parámetro es del 90%. Por ejemplo se toman diez muestras del mismo tamaño (*n*) de una población y se construyen los diez intervalos de

confianza para el parámetro con la información de cada una de esas muestras, nueve intervalos contendrán al parámetro y uno no lo harán (Fig. 2).

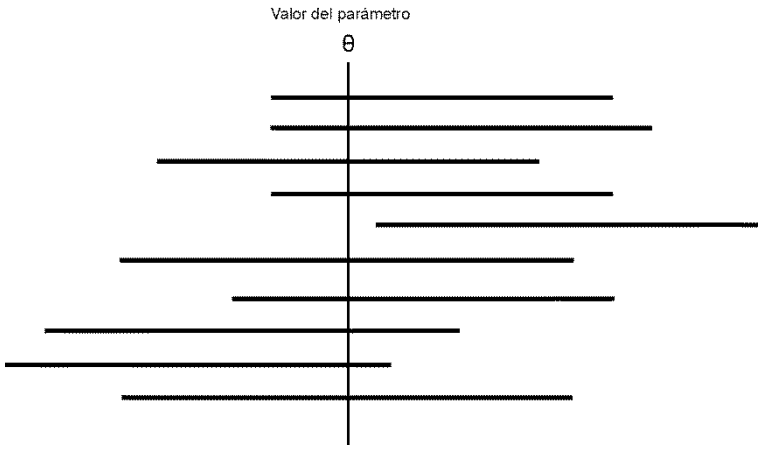


Figura 2. Intervalos de confianza para el parámetro θ (90%). Se utiliza el mismo procedimiento de construcción del intervalo para 10 muestras aleatorias independientes de tamaño n, entonces 9 (m(1 – α)) intervalos contienen el valor del parámetro y una no lo contiene.

El Coeficiente de Confianza (CC) lo elige el investigador. En los trabajos geológicos se utilizan CC de 90%, 95% y 99%. Algunos autores (Koch y Link 1980) recomiendan utilizar CC 90% argumentando que la variabilidad de los datos geológicos es mayor que a aquella controlada por los experimentos de laboratorio o manufacturas pues muchos cuerpos geológicos presentan discontinuidades estructurales.

La estimación de cualquier parámetro poblacional por el método de los intervalos de confianza requiere: 1°) fijar el coeficiente de confianza, 2°) extraer la muestra y calcular el estadístico y 3°) conocer la distribución que tiene el estimador del parámetro.

Intervalos de confianza para la media poblacional (μ) conocida la varianza poblacional (σ^2)

Para estimar la media poblacional μ el mejor estimador es la media muestral $\bar{X}$. Se conoce que la media muestral se distribuye normalmente con valor esperado $E = \mu$ y varianza σ^2 y la distribución asociada es $z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$ (Capítulo 5). Si se fija el CC como 1 - α, entonces la probabilidad α se divide en dos partes, una parte se asocia con el límite inferior, α/2, y la otra con el superior, α/2. Si $-z_{\alpha/2}$ y $z_{\alpha/2}$ son los valores de la distribución normal estándar que tienen probabilidades acumuladas α/2 y 1-α/2 respectivamente (Fig. 3). Entonces cuando se calcula el intervalo de confianza de μ conocida σ^2 la expresión 6.2 toma la forma

$$P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \qquad (6.3)$$

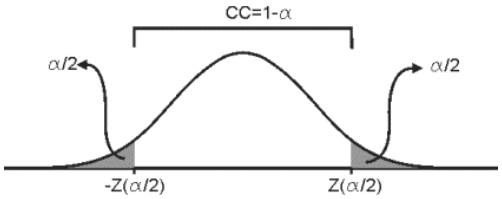


Figura 3. Probabilidades asociadas al Coeficiente de Confianza en la distribución normal estándar.

EJEMPLO 1

Límites de confianza para la media poblacional con varianza poblacional conocida
Los siguientes datos corresponden a concentración de plomo en efluentes de una industria.

{8415 – 8100 - 8820 – 9215 – 7875 – 9800 – 8235 – 10305 – 8325 – 9430}
Se conoce $\sigma = 900$
$n = 10$
$\bar{X} = 8852$, $S = 765$
CC = 1 - 0,05 = 0,95
$\alpha = 0,05$; $\alpha/2 = 0,025$
De la Tabla 1 del Anexo, $Z_{0,025} = -1,96$; $Z_{0,925} = 1,96$

$$P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(8852 - 1{,}96\frac{900}{\sqrt{10}} < \mu < 8852 - 1{,}96\frac{900}{\sqrt{10}}\right) = 1-0{,}05$$

$P(8294 < \mu < 9410) = 0,95$

El intervalo de (8294, 9410) tiene 95% de probabilidad de contener a la media poblacional.

El intervalo de confianza para la media poblacional μ se calcula sobre la base de la distribución normal de la media muestral $\bar{X}$. Luego, siempre que se calcula el intervalo de confianza para otro parámetro cuyo estimador (estadístico) esté normalmente distribuido, se puede utilizar la expresión 6.3 reemplazando $\bar{X}$ por el estadístico adecuado y $\sigma/\sqrt{n}$ por el error estándar del correspondiente estadístico (Capítulo 5). Por ejemplo un **intervalo de confianza para la diferencia de medias** $\Delta\mu = \mu_1 - \mu_2$ es

$$P\left(\bar{X}_1 - \bar{X}_2 - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \Delta\mu < \bar{X}_1 - \bar{X}_2 + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha \cdot \qquad (6.4)$$

Intervalos de confianza para la media poblacional (μ) con varianza poblacional estimada con la varianza muestral (S^2) *y/o para tamaño de muestra chico*

En la práctica cuando se quiere estimar la media poblacional tampoco se conoce la varianza poblacional σ^2 y no se tienen los datos para calcular el error estándar muestral ($\sigma_{\bar{X}} = \sigma/\sqrt{n}$) entonces no se puede usar *Z* en el cálculo del intervalo de confianza. Una alternativa para es estimar $\sigma_{\bar{X}}$ a partir de $S_{\bar{X}}$. Cuando el tamaño de la muestra es grande entonces $S_{\bar{X}}$ es buen estimador de $\sigma_{\bar{X}}$ y es posible utilizar *Z*. Sin embargo la mayoría de las veces el tamaño de la muestra es insuficiente y se requiere utilizar el estadístico ***t* de Student** para calcular el intervalo de confianza.

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \qquad (6.5)$$

La **distribución *t* de Student**, al igual que la distribución χ^2 (Capítulo 5), tiene diferentes formas para los diferentes grados de libertad ν (Fig. 4). La distribución es leptocurtica, los valores se concentran alrededor de la media, pero a medida que aumentan los grados de libertad la forma de la distribución *t* tiende a ser similar a la de la distribución normal y para ν = ∞, las dos distribuciones son idénticas. En la Tabla 3 del Anexo se encuentran probabilidades asociadas a algunos puntos para diferentes grados de libertad. Si no se requiere mucha precisión, cuando los valores de *t* no están tabulados, se puede usar el menor valor tabulado más próximo, pero si se necesita precisión se debe efectuar una interpolación lineal.

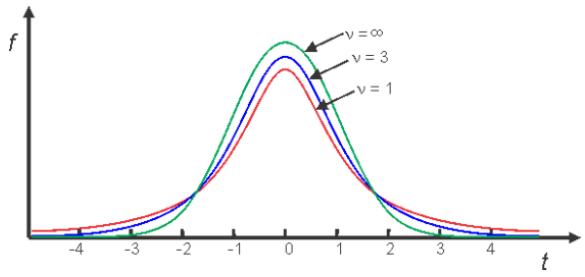


Figura 4. La distribución t *para varios grados de libertad ν. Para ν = ∞, la distribución* t *es idéntica a la distribución normal.*

La expresión para calcular el intervalo de confianza de μ cuando se desconoce σ es análoga a la expresión 6.3, sólo se reemplaza σ por *S* y *Z* por *t*.

$$P\left(\bar{X} - t_{\nu;\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\nu;\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha\,, \qquad (6.6)$$

t tiene $\nu = n - 1$ grados de libertad. Es necesario resaltar en este punto que para usar la expresión 6.6 se requiere que **los datos se distribuyan normalmente**.

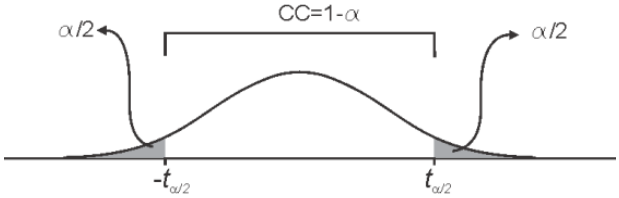


Figura 5. Probabilidades asociadas al Coeficiente de Confianza en la distribución t..

EJEMPLO 2

Límites de confianza para la media poblacional con varianza poblacional desconocida

Los siguientes datos de concentración de arsénico (mg l^{-1}) en el agua fueron obtenidos de un acuífero del sector central de la provincia del Chaco que se suponen normalmente distribuidos.

{6,9 – 3,8 – 4,9 – 6,5 – 3,5 – 3,7}

$n = 6$

$\overline{X} = 4{,}88$

$S = 1{,}49$

CC = *1* - α = 1 - 0,90 = 0,10

α/2 = 0,05

$\sqrt{6} = 2{,}45$

$\frac{S}{\sqrt{6}} = 0{,}61$

$\nu = n - 1 = 5$

De la Tabla 3 del Anexo $t_{5;\alpha/2} = 2{,}05$

$$P(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$P(4{,}88 - 2{,}05 \cdot 0{,}60 < \mu < 4{,}88 + 2{,}05 \cdot 0{,}60) = 0{,}90$$

$$P(3{,}65 < \mu < 6{,}11) = 0{,}90$$

El contenido medio de arsénico del acuífero pertenece al intervalo (3,65;6,11) mg l^{-1}, con un nivel de confianza del 90%.

Para el caso en que se deba calcular un **intervalo de confianza para la diferencia de dos medias poblacionales**, $\Delta\mu = \mu_1 - \mu_2$, y si también se desconocen los desvío estándar poblacionales σ_1 y σ_2 pero se estiman con sus respectivos desvíos estándar muestrales S_1 y S_2 se utiliza una expresión análoga a la 6.4.

$$P(\Delta\overline{X} - t_{\alpha/2;\nu}S_{\Delta\mu} < \Delta\mu < \Delta\overline{X} + t_{\alpha/2;\nu}S_{\Delta\mu}) = 1-\alpha \qquad (6.7)$$

Cuando las **varianzas muestrales son iguales** se calcula una varianza ponderada en la que intervienen los tamaños de muestra y las varianza muestrales de la siguiente manera:

$$Sp^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \qquad (6.8)$$

y el error estándar de la distribución de la diferencia de medias $S_{\Delta\mu}$ es

$$S_{\Delta\mu} = \sqrt{Sp^2(1/n_1 + 1/n_2)}\,. \qquad (6.9)$$

Pero cuando las **varianzas muestrales son diferentes** Sp^2 es

$$Sp^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}, \tag{6.10}$$

y el error estándar de la diferencia de medias

$$S_{\Delta\mu} = \sqrt{(S_1^2/n_1 + S_2^2/n_2}\,. \tag{6.11}$$

En ambos casos para tamaños de muestra diferentes *t* tiene $\nu = n_1 + n_2 - 2$ grados de libertad y si ambas muestras tienen el mismo tamaño ($n_1 = n_2 = n$) *t* tiene $\nu = n - 1$ grados de libertad.

EJEMPLO 3

Intervalo de confianza para la diferencia de medias poblacionales con varianzas desconocidas

Se presentan datos de concentración de arsénico (mg l^{-1}) en el agua subterránea de dos localidades del Chaco central.

	Localidad 1	Localidad 2
	6,9	4,8
	3,8	3,9
	4,9	5,9
	6,5	5,0
	3,5	4,7
	3,7	6,0
		3,7
n	*6*	*7*
$\overline{X}$	*4,883*	*4,857*
S^2	*2,234*	*0,783*

$$\Delta\overline{X} = \overline{X}_1 - \overline{X}_2 = 4{,}883 - 4{,}857 = 0{,}026$$

Como se muestreo el mismo acuífero se supone que las varianzas de ambas poblaciones son iguales de modo que se calcula $Sp^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$

$$Sp^2 = \frac{(6-1)2{,}234 + (7-1)0{,}783}{6+7-2} = 1{,}4427$$

$$S_{\Delta\mu} = \sqrt{Sp^2(1/n_1 + 1/n_2)} \qquad S_{\Delta\mu} = \sqrt{1{,}4427(1/6 + 1/7)} = 0{,}6683$$

$CC = 1-\alpha = 1-0{,}90 = 0{,}10$
$\alpha/2 = 0{,}05$

$\nu = n_1 + n_2 - 2 = 6 + 7 - 2 = 11$
De la Tabla 3 del Anexo $t_{\alpha/n;\nu} = t_{0.05;11} = 1{,}796$

$$P(\Delta\overline{X} - t_{\alpha/2;\nu}S_{\Delta\mu} < \Delta\mu < \Delta\overline{X} + t_{\alpha/2;\nu}S_{\Delta\mu}) = 1-\alpha$$

P(0,026 - 1,796 · 0,6683 < $\Delta\mu$ < 0,026 + 1,796 · 0,6683) = 0,90

P(-1,174 < $\Delta\mu$ < 1,226) = 0,90

Con un nivel de confianza del 90% la diferencia entre contenido medio de arsénico de ambos sitios se encuentra en el intervalo (-1,174; 1,226) mg/l.

Además como el intervalo contiene el valor cero, se puede afirmar con una confianza del 90% que no existen diferencias estadísticas entre el contenido medio de arsénico de ambos sitios.

Cálculo de tamaño muestral para obtener un intervalo de confianza para μ de amplitud definida

El cálculo del tamaño de muestra que se debe tomar para obtener una estimación por intervalos de confianza de la media poblacional μ, de una amplitud específica, requiere datos de un muestreo preliminar. La expresión 6.6 (cálculo de límites de confianza de la media poblacional) permite definir la **Amplitud** del Intervalo de confianza como la diferencia entre el Límite superior y el Límite inferior ($A = Ls - Li$). La Amplitud se expresa entonces como:

$$A = Ls - Li = \bar{X} + t_{\upsilon;\,\alpha/2}\frac{S}{\sqrt{n}} - \bar{X} + t_{\upsilon;\,\alpha/2}\frac{S}{\sqrt{n}}$$

$$A = 2 \cdot t_{\upsilon;\,\alpha/2}\frac{S}{\sqrt{n}}\,. \qquad (6.12)$$

Si se llama δ a la amplitud específica del intervalo de confianza que se requiere en la estimación de la media poblacional μ, el tamaño de la muestra *n* se despeja de la expresión 6.12,

$$n = \left(\frac{2\,t_{\alpha/2,\nu}}{\delta}\right)^2. \qquad (6.13)$$

El procedimiento para calcular el tamaño de muestra requiere definir el coeficiente de confianza de la estimación y proponer un tamaño de muestra inicial n_0, con ambos se establece el valor de $t_{\alpha/2,\,\nu 0}$ ($\nu_0 = n_0 - 1$) utilizado en la expresión 6.13. Como el tamaño de la muestra se calcula partiendo de un muestreo preliminar de tamaño n_0, el coeficiente $t_{\nu 0;\,\alpha/2}$ depende de n_0, se recomienda rehacer el cálculo con el $t_{n-1;\,\alpha/2}$ correspondiente al tamaño de muestra *n* obtenido.

EJEMPLO 4

Tamaño de la muestra

Cálculo del tamaño de la muestra para disminuir de 2,46 a δ = 0,5 el intervalo de confianza de μ de los datos de concentración de arsénico (mg l^{-1}) del ejemplo 1 (acuífero del sector central de la provincia del Chaco).

Se necesita disminuir el error a δ = 0,5
$n_0 = 6$
CC = $1 - \alpha = 1 - 0{,}90 = 0{,}10$
$\alpha/2 = 0{,}05$
$\nu = n - 1 = 5$

De la Tabla 3 del Anexo, $t_{5;\alpha/2} = 2{,}05$

$$n = \left(\frac{2\,t_{\alpha/2,\nu}}{\delta}\right)^2$$

$$n = \left(\frac{2 \cdot 2{,}05}{0{,}5}\right)^2 = 67{,}24$$

De la Tabla 3 del Anexo, $t_{68;\alpha/2}$= 1,67

$$n = \left(\frac{2 \cdot 1{,}67}{0{,}5}\right)^2 = 44{,}62$$

Para disminuir el error de la estima a 0,5 manteniendo el coeficiente de confianza en 90% se debe tomar una muestra de tamaño 45.

Intervalo de confianza para la varianza poblacional (σ^2)

En los trabajos geológicos muchas veces es necesario estimar la variabilidad de la población, por ejemplo es importante conocer la variabilidad del contenido de fósforo en los yacimientos de hierro porque éste está ligado a la siderurgia. En el Capítulo 5 se vio que S^2 es un buen estimador de σ^2 y que ambas se tienen una relación que se describe con el estadístico χ^2 con (n - *1*) grados de libertad del modo

$$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}. \qquad (6.14)$$

Por lo tanto para obtener un intervalo de confianza para la varianza poblacional σ^2 conociendo la varianza muestral S^2 se aplica la expresión 6.14 (Fig. 6).

$$P\left(\frac{(n-1)S^2}{\chi^2_{(n-1),\,(1-\alpha/2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1),\,(\alpha/2)}}\right) = 1-\alpha \qquad (6.15)$$

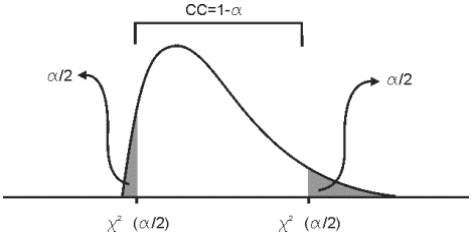


Figura 6. Probabilidades asociadas al Coeficiente de Confianza en la distribución χ^2.

EJEMPLO 5

Intervalos de confianza para la varianza poblacional

Se utilizan los datos de concentración de arsénico del acuífero del Chaco en el Sitio 1 del ejemplo 2.

{6,9 – 3,8 – 4,9 – 6,5 – 3,5 – 3,7}

n = 6

$\overline{X}$ = 4,88

S = 1,49

S^2 = 2,234

CC = *1* – α = 1 - 0,90 = 0,10

α/2 = 0,05

(n - 1) S^2=11,17

De la Tabla 2 del Anexo, $\chi^2_{5;0.95} = 1{,}145$ y $\chi^2_{5;0.05} = 11{,}070$

$$P\left(\frac{(n-1)S^2}{\chi^2_{(n-1),(1-\alpha/2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1),(\alpha/2)}}\right) = 1-\alpha$$

$$P\left(\frac{11{,}17}{11{,}070} < \sigma^2 < \frac{11{,}17}{1{,}145}\right) = 1-0{,}1$$

$$P(1{,}00 < \sigma^2 < 9{,}71) = 0{,}90$$

La varianza poblacional de la concentración de arsénico del acuífero en la localidad del Chaco se encuentra en el intervalo (1,00; 9,71) $(mg/l)^2$, con un nivel de confianza del 90%.

Intervalos de confianza de una sola cola

Si bien en el cálculo de todos los intervalos de confianza explicados son limitados por dos lados (**intervalos de confianza de dos colas**), se pueden calcular intervalos de confianza que tengan un sólo límite, el inferior o el superior (**intervalos de confianza de una cola**). En los intervalos a dos colas el error α se reparte en ambos límites, en cambio en los intervalos a una cola el límite no calculado no tiene riesgo de error y α se encuentra sólo en el límite que se calcula. En los intervalos de **cola superior** el error está a la derecha $\mathbf{P(-\infty \leq \theta \leq Ls) = 1 - \alpha}$, en los intervalos de **cola inferior** el error se acumula a la izquierda $\mathbf{P(Li \leq \theta \leq \infty) = 1 - \alpha}$ (Fig.7).

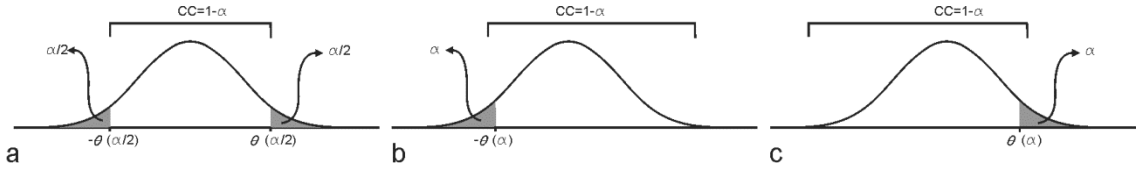


Figura 7. a. Intervalo de confianza de dos colas. b. Intervalo de confianza de cola inferior. c. Intervalo de confianza de cola superior.

Prueba de hipótesis

Las pruebas de hipótesis estadísticas se enmarcan perfectamente en lo que Popper (1968) describió como el método científico. Según este filósofo austro-británico (1902-1994), el conocimiento científico tiene características que lo diferencian de otro tipo de conocimientos: tiene origen empírico, se somete permanentemente a revisiones, las teorías e hipótesis se contrastan con la realidad para descubrir falsedad ya que la verdad es circunstancial o sustentada en probabilidades. Suele decirse que el método científico es un camino que recorre gradualmente para dilucidar un problema específico. Se trata de un método hipotético-deductivo donde el investigador observa o muestrea un hecho natural y usa toda la información disponible para formular una suposición sobre el mismo o sobre su funcionamiento, esto es la hipótesis. La suposición se realiza sólo sobre la base de la intuición. El

investigador no tiene manera de conocer si su hipótesis es correcta. Realiza predicciones. Para testear la hipótesis efectúa otros muestreos o un experimento. Si los resultados son consistentes con las predicciones entonces se retiene la hipótesis. Si los resultados no son consistentes, se rechaza y se formula una nueva hipótesis.

He aquí un ejemplo. Al noroeste de la provincia de Buenos Aires el agua subterránea destinada a consumo humano tiene elevadas concentraciones de Arsénico. Un grupo de geólogos piensa que si el agua se mezcla con arcillas, estas pueden retener el arsénico debido a su peculiar estructura. El tratamiento es un proceso sencillo y de muy bajo costo. La hipótesis que se necesita testear es simple: las arcillas retienen el arsénico del agua.

Para está hipótesis la predicción es lineal, se espera menor concentración de arsénico después que el agua circule y se mezcle con la arcilla. La predicción se puede testear realizando un experimento. Entonces se mezclan 10 gr de arcillas en 5 litros de agua subterránea. Debido a que la concentración cambia según la localidad, se replica el tratamiento varias veces. Se elijen tres localidades al azar de todas las que presentan el problema y se recolectan 10 l de agua. Cinco litros se someten al tratamiento y 5 l se dejan para control. Cada 24 horas durante 10 días se toma una alícuota del agua que se analiza químicamente para obtener los datos para testear la hipótesis.

Este experimento tiene al menos cuatro resultados posibles:

1° La concentración de arsénico del agua después del tratamiento es mucho menor que la del agua sin el tratamiento. El resultado es consistente con la hipótesis. Como la hipótesis supera la prueba sobrevive provisoriamente.

2° La concentración de arsénico del agua después del tratamiento no cambia, es la misma que la del agua sin tratar. El resultado contradice la hipótesis. Como la hipótesis no supera la prueba se descarta.

3° Hay una pequeña disminución de la concentración de arsénico después del tratamiento. Es difícil saber si el resultado se relaciona con la hipótesis, por ejemplo 10 gr de arcilla no son suficientes para remover el arsénico o es poco el tiempo que dura el experimento y el arsénico no alcanza a quedar retenido. La hipótesis no se puede rechazar ni aceptar.

4° La concentración de arsénico después del tratamiento es mayor que la del agua sin tratar. Esto es un resultado inesperado y no consistente con la hipótesis, y debe ser rechazado.

Este tipo de resultados experimentales, donde se debe decidir si la arcilla reduce o elimina el arsénico del agua o si las diferencias entre el agua sin tratamiento o tratada sucede por azar, se puede resolver utilizando una prueba de hipótesis estadística.

Cuando se somete a prueba una hipótesis pueden suceder dos cosas: los resultados son consistentes con la hipótesis entonces esta se retiene, o los resultados no son consistente con la hipótesis y se rechaza. Si la hipótesis se rechaza se debe proponer otra. Si la hipótesis no se rechaza hay que someterla repetidamente a otras pruebas pues siempre hay posibilidades de desecharla en el futuro.

Es una convención que en un test de hipótesis se formulen dos hipótesis complementarias. Por ejemplo: *La arcilla retiene el Arsénico del agua* y *La arcilla no retiene el Arsénico del agua*. Estas hipótesis son llamadas **alternativa** y **nula** respectivamente.

La **hipótesis nula** enuncia lo contrario de lo que considera como verdadero, siempre es la hipótesis de **no diferencias** o no efectos. La hipótesis nula se expresa normalmente en símbolos por H_0. El cero en referencia al efecto nulo o que los datos que se comparan no presentan diferencias entre sí, la diferencia es cero.

La **hipótesis alternativa** es la hipótesis contra la cual se contrasta la hipótesis nula, es la hipótesis que el investigador quiere probar, se denota H_A o H_1. Generalmente es mejor argumento encontrar evidencias contrarias a la hipótesis nula y tener que rechazarla, que aceptar la hipótesis nula pues se puede sospechar que se buscan evidencias para probarla.

Definiciones

El objetivo de una prueba de hipótesis permite distinguir al menos dos categorías: las Pruebas de hipótesis sobre parámetros que se usan para determinar si un parámetro poblacional toma o no un valor fijo, y las Pruebas de Bondad de Ajuste que se emplean para definir si un conjunto de datos se puede modelar con una distribución teórica. A continuación se presentan conceptos y definiciones que se utilizaran en adelante:

Hipótesis estadística es una proposición con respecto a una o más variables poblaciones, ya sea uno o varios de sus parámetros desconocidos (simbolizados con θ), o a la forma de su distribución de probabilidad ($f(x, \theta)$).

Hipótesis nula (H_0): generalmente se plantea que no existen diferencias entre los valores a comparar y de modo tal que especifique el valor exacto del parámetro.

Hipótesis alternativa (H_A o H_1): recuerde que es la hipótesis de investigación, que se contrasta contra la hipótesis nula.

Hipótesis nula y alternativa pueden incluir el 0, como en este caso H_0: $\theta_A - \theta_B = 0$ vs. H_A: $\theta_A - \theta_B < 0$ o referirse a un valor particular como en este otro caso H_0: $\theta_A = \theta_0$ vs. H_A: $\theta_A \neq \theta_0$, donde θ_0 es un valor específico.

Por ejemplo, siguiendo con el problema del arsénico en el agua subterránea, la concentración tolerable según las leyes de la provincia de Buenos Aires es 0,01 mg/l. Si se quiere demostrar que la concentración de arsénico en el agua de una localidad es mayor que la tolerable, la hipótesis nula será: La concentración media de arsénico del acuífero es menor o igual a 0,01 mg/l (H_0: $\mu \leq 0{,}01$) y la hipótesis alternativa: La concentración media de arsénico del acuífero es mayor a 0,01 mg/l (H_A: $\mu > 0{,}01$).

Prueba de Hipótesis: también llamada contraste de hipótesis estadística, es un procedimiento que establece un criterio que permite decidir si se acepta o se rechaza una hipótesis con base en los resultados de una muestra aleatoria de la población de interés. El procedimiento tiene los siguientes pasos:

1° Plantear la hipótesis nula y alternativa.

2° Tomar una muestra aleatoria de tamaño *n* de la población $\{x_1, x_2,..., x_n\}$.

3° Definir el estadístico de prueba adecuado y fijar el nivel de significación de la prueba, llamado α. α es la probabilidad del error que se está dispuesto a tolerar al aceptar la hipótesis nula cuando es falsa.

4° Calcular el Estadístico de prueba a partir de los datos de la muestra.

5° Definir el criterio de aceptación o de rechazo de la hipótesis nula. Es decir, el o los valores del estadístico de prueba (**valor crítico**) que permitan delimitar una **región de aceptación de H_0** y otra **región de rechazo de H_0**.

6° Tomar la decisión de no aceptar o aceptar H_0. Existen varios procedimientos para tomar la decisión estadística. Uno es ver si el estadístico de prueba queda en la región de aceptación o en la región de rechazo de la hipótesis nula. Otro, el que ofrecen los software estadísticos, consiste en calcular la probabilidad *p* asociada con el estadístico de prueba. Este valor *p* corresponde a la hipótesis nula y se interpreta como la probabilidad de error si se rechaza H_0 cuando ésta es cierta. Se rechaza H_0 si $p < \alpha$. Hay que recalcar que rechazar la hipótesis nula sólo indica que los datos no dan evidencia suficiente para concluir que es falsa.

Errores al tomar una decisión: cuando se toma una decisión estadística se pueden cometer dos tipos de errores: rechazar la hipótesis nula cuando es verdadera, es el llamado **Error de tipo I** o α y aceptar la hipótesis alternativa cuando es falsa, denominado **Error de tipo II** o β (Cuadro 1). Se conoce que para un tamaño de muestra dado, el tamaño de ambos errores está inversamente relacionado, al aumentar α disminuye β y viceversa. Esto es, bajas probabilidades de cometer error de tipo I están asociadas a grandes probabilidades de error de tipo II. La única manera de reducir ambos errores es aumentar el tamaño de la muestra. Es importante saber que combinación es aceptable. Generalmente se propone $\alpha = 0,05$ pero eso depende del problema a investigar, cuando el nivel de significación no es 5% se debe justificar y aclarar antes de realizar el muestreo o el experimento.

Decisión	**H_0 verdadera (H_1 falsa)**	**H_0 falsa; (H_1 verdadera)**
Se rechaza H_0	*Decisión incorrecta* *Error tipo I* *Porcentaje de error: α*	***Decisión correcta*** ***No hay error*** ***P(1 - β)***
No se rechaza H_0	***Decisión correcta*** ***No hay error*** ***P(1 - α)***	*Decisión incorrecta* *Error tipo II* *Porcentaje de error: β*

Cuadro 1. Tipos de error de una prueba estadística.

Nivel de significación de la prueba: es la máxima probabilidad de rechazar la hipótesis nula cuando es verdadera. Se utiliza α para indicar el nivel de significación por ejemplo $\alpha = 0,05$ pero a veces se utiliza como probabilidad $P < 0,05$ (se lee la probabilidad es menor que 0,05); NS significa no significativo que es cuando la probabilidad es mayor a 0,05 ($P \geq 0,05$).

Región crítica: es la región de rechazo de la hipótesis nula (1 - α).

Poder o potencia de la prueba estadística: es $1 - \beta$, se interpreta como la probabilidad de rechazar de manera correcta una hipótesis nula falsa. Describe la capacidad de una prueba para detectar diferencias.

Tipos de pruebas de hipótesis: la hipótesis alternativa que se debe formular depende del problema que se analiza y los resultados esperados. Las **pruebas bilaterales** o de **dos colas** se enuncian cuando el problema requiere comprobar si existen diferencias con respecto a un valor poblacional conocido, pero sin especifica el sentido, se trata sólo de decidir si es igual o es diferente, son del tipo H_0: $\theta = \theta_0$ vs. H_A: $\theta \neq \theta_0$. Por otro lado, las pruebas unilaterales o de una cola se formulan cuando se investiga si existen diferencias y además se especifica el sentido, si es mayor o menor a un valor conocido. Las **pruebas unilaterales derecha** o de **cola superior** se enuncian si el interés es encontrar diferencias mayores, en este caso las hipótesis son H_0: $\theta \leq \theta_0$ vs. H_A: $\theta > \theta_0$. Cuando se investiga si existen diferencias y si estas son menores, la **prueba es unilateral izquierda** o de **cola inferior** y las hipótesis son del tipo H_0: $\theta \geq \theta_0$ vs. H_A: $\theta < \theta_0$ (Fig. 8). La idea que subyace es hacer el intervalo de Rechazo de la Hipótesis Nula lo más grande posible.

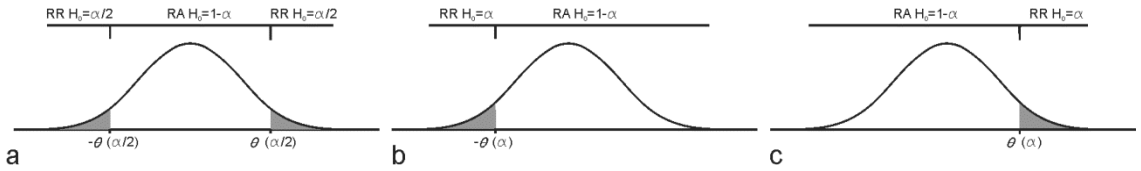


Figura 8. Región de aceptación (RA) y región de rechazo (RR) de la Hipótesis Nula.. a. Prueba de Hipótesis bilateral (H_0: $\theta = \theta_0$ vs. H_A: $\theta \neq \theta_0$). b. Prueba de Hipótesis de cola inferior (H_0: $\theta \geq \theta_0$ vs. H_A: $\theta < \theta_0$). c. Prueba de Hipótesis de cola superior (H_0: $\theta \leq \theta_0$ vs. H_A: $\theta > \theta_0$).

Pruebas de Hipótesis para una muestra

Prueba de Hipótesis para la media muestral

El estadístico de prueba de hipótesis relativas a la media poblacional μ cuando la **varianza muestral es conocida** es el normal estándar Z (expresión 6.16), pero si la **varianza muestral es desconocida** el estadístico es t (expresión 6.17). .Ambos usan la información de la muestra: la media $\bar{X}$, el tamaño de la muestra n y si fuera necesario, la varianza muestral S^2.

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad (6.16) \qquad t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \quad (6.17)$$

Dependiendo del problema que se investiga se pueden plantear las siguientes hipótesis:

H_0: $\mu = \mu_0$ vs. H_A: $\mu \neq \mu_0$, H_0 se rechaza si $|z| > Z_{\alpha/2}$ ó $|t| > t_{\alpha/2,\ n-1}$.

H_0: $\mu \leq \mu_0$ vs. H_A: $\mu > \mu_0$, H_0 se rechaza si $z > Z_\alpha$ ó $t > t_{\alpha,\ n-1.}$

H_0: $\mu \geq \mu_0$ vs. H_A: $\mu < \mu_0$, H_0 se rechaza si $z < -Z_\alpha$ ó $t < -t_{\alpha,\ n-1.}$

Los estadísticos de prueba *Z* y *t* se utilizan también para contrastar hipótesis sobre otros parámetros sustituyendo el error estándar de la distribución de la media muestral ($\sigma/\sqrt{n}$ o $S/\sqrt{n}$, según corresponda) por el correspondiente error estándar del parámetro, como se verá más adelante. El único requisito que debe cumplir el estimador (estadístico) del parámetro es que esté normalmente distribuido.

EJEMPLO 6

Prueba de hipótesis para una media muestral

En el Parque Provincial Copahue el mayor atractivo turístico son el volcán Copahue y las manifestaciones termales. El sitio sufre el impacto del tránsito de pobladores y visitantes. El inadecuado trazado de caminos y el intenso pisoteo contribuyen a la perdida de la cobertura vegetal herbácea y arbustiva junto con la fusión nívea han generando condiciones para que se produzca un intenso carcavamiento que origina procesos de erosión y pérdida de suelo.

Para controlar y revertir los procesos erosivos se construyeron azudes en algunas cárcavas para reducir la energía hídrica. Interesa saber si las tasas de acumulación media de sedimentos producida en los mismos supera el valor de acumulación esperada para el año hídrico estimada en 0,54 m^3/año.

Datos: 0,4 0,3 0,65 0,55 0,84 0,73 0,34 0,83 0,65 0,54 0,71 0,6 0,38 0,52 0,78 0,4

$\bar{X} = 0{,}58$; S = 0,18; n = 16

H_0: $\mu \leq 0{,}54$
H_A: $\mu > 0{,}54$
$\alpha = 0{,}05$

De la Tabla 3 del Anexo, $t_{0,05;\ 16-1} = 1{,}753$

Criterio de rechazo de H_0: $t < -t_{\alpha,\ n-1}$

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{0{,}58 - 0{,}54}{\frac{0{,}18}{\sqrt{16}}} = \frac{0{,}04}{0{,}044} = 0{,}82$$

Dado que 0,82 < 1,753, no existen evidencias para rechazar la hipótesis nula.

Se puede concluir que la acumulación de sedimentos en los azudes es menor que la estimada para el año hídrico y consecuentemente no evitarían el proceso de erosión.

Prueba de Hipótesis para la varianza muestral

Si el problema que se investiga se refiere a la variabilidad que presenta la población de una variable normalmente distribuida se puede realizar una prueba sobre la varianza poblacional σ^2. Se toma una muestra aleatoria de tamaño *n* de una población normal, se calcula la varianza muestral es S^2, el estadístico de prueba se calcula como

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad (6.18)$$

Como se mencionó, el problema que se investiga determina las hipótesis de interés, estas pueden ser:

H_0: $\sigma^2 = \sigma^2_0$ vs. H_A: $\sigma^2 \neq \sigma^2_0$, en este caso la H_0 se rechaza cuando $\chi^2 \leq \chi^2_{1-\alpha/2,n-1}$ ó $\chi^2 \geq \chi^2_{\alpha/2,n-1}$.

H_0: $\sigma^2 \leq \sigma^2_0$ vs. H_A: $\sigma^2 > \sigma^2_0$, se rechaza H_0 si $\chi^2 \geq \chi^2_{\alpha,n-1}$.

H_0: $\sigma^2 \geq \sigma^2_0$ vs. H_A: $\sigma^2 < \sigma^2_0$, se rechaza H_0 si $\chi^2 \leq \chi^2_{1-\alpha,n-1}$.

EJEMPLO 7

Prueba de hipótesis sobre la varianza poblacional

Las arenas que se usan en pozos de *sheil gas* y *sheil oil* deben ser muy bien seleccionada, tener sus granos esféricos, con gran redondez y alta resistencia a la compresión. Uno de los ensayos que se realizan para ver si son aptas consiste en someterlas a presiones de 10.000 psi y analizar la cantidad de material perdido. Las arenas de buena calidad tienen en las pruebas de resistencia a la compresión con varianzas menores a 2.

Se realizan pruebas a 10 muestras de arena con varianza de 1,6. ¿Hay evidencias para afirmar que la arena muestreada podría ser utilizada en pozos de *sheil gas* y *sheil oil*?

H_0: $\sigma^2 \geq 2$

H_A: $\sigma^2 < 2$

$\alpha = 0{,}05$

Criterio de rechazo de H_0 $\chi^2 \leq \chi^2_{1-\alpha/2,n-1}$.

$n = 10$

De la Tabla 2 del Anexo, $\chi^2_{0,95;9} = 3{,}33$

$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{9 \cdot 1{,}6}{2} = 7{,}2$

Dado que 3,33 < 7,2; no existen evidencias para rechazar la hipótesis nula.

Se puede concluir que la resistencia de las arenas a la compresión es mayor a 2 y consecuentemente las arenas no serían aptas para usar en pozos de *sheil gas* y *sheil oil*.

Pruebas de Hipótesis para dos muestras

Prueba de Hipótesis para comparar dos varianzas

En las ocaciones en las que es necesario comparar la variabilidad que presentan dos poblaciones, se puede formular una prueba sobre las varianzas poblacionales σ^2_1 y σ^2_2. Para obtener el estadístico de

prueba se muestrean dos poblaciones normales, población 1 y población 2. Se toman muestras de tamaño n_1 y n_2 y se calculan las varianzas muestrales S_1^2 y S_2^2.

El estadístico de prueba en este caso es

$$F = \frac{S_1^2}{S_2^2}. \tag{6.19}$$

Nuevamente el problema que se investiga determina las hipótesis de interés, que pueden ser:

H_0: $\sigma_1^2 = \sigma_2^2$ vs. H_A: $\sigma_1^2 \neq \sigma_2^2$, la H_0 se rechaza cuando $F \leq F_{1-\alpha,n1-1,n2-1}$.ó $F \geq F_{\alpha,n1-1,n2-1}$.

H_0: $\sigma_1^2 \leq \sigma_2^2$ vs. H_A: $\sigma_1^2 > \sigma_2^2$, se rechaza H_0 si $F \geq F_{\alpha,n1-1,n2-1}$.$H_0$: $\sigma_1^2 \geq \sigma_2^2$ vs. H_A: $\sigma_1^2 < \sigma_2^2$, se rechaza H_0 si $F \leq F_{\alpha,n1-1,n2-1}$.

La **distribución *F*** de Fisher es una familia de distribuciones de frecuencia teóricas para la relación entre dos varianzas obtenidas de un muestreo independiente de dos poblaciones cuyas observaciones están normalmente distribuidas y cuyas varianzas son iguales. Conceptualmente una distribución *F* se obtiene tomando una muestra de tamaño n_1 de una población con varianza σ^2 y calculando la varianza de la muestra S_1^2 con (n_1 – 1) grados de libertad. De la misma manera se obtiene una muestra de tamaño n_2 de una población con varianza σ^2 y se calcula la varianza de la muestra S_2^2 con (n_2 – 1) grados de libertad. El cociente de las dos varianzas muestrales sigue una distribución *F* cuyos dos parámetros son $\nu_1 =$ (n_1 – 1) y $\nu_2 =$ (n_2 – 1) grados de libertad.

Existen diferentes distribuciones *F* para cada combinación de números de grados de libertad en el numerador y en el denominador. *F* es el cociente de dos números positivos, el valor mínimo es 0 y el máximo infinito. A medida que los grados de libertad del numerador y denominador aumentan, la distribución se vuelve más simétrica (Fig. 9).

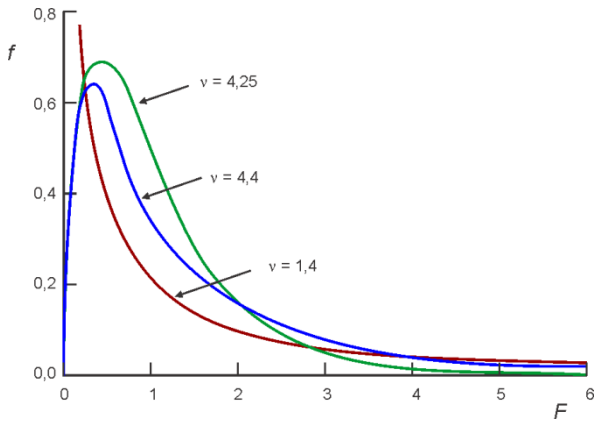


Figura 9. Distribución F *para varios pares de grados de libertad* ν_1 *y* ν_2.

Prueba de Hipótesis para la diferencia de medias muestrales

Muchos problemas geológicos se plantean hipótesis que requieren comparar las medias poblacionales de dos muestras, por ejemplo la precipitación caída en dos localidades, la carga hidráulica de dos

ambientes sedimentarios o los niveles de contaminación de un río donde una industria evacúa sus efluentes, entre otros. En estos casos se toman muestras de tamaño n_1 y n_2 de la población 1 y de la 2 respectivamente, se calculan los promedios, $\overline{X_1}$ y $\overline{X_2}$, las varianzas S_1^2 y S_1^2, y se plantean las hipótesis. Suponga que el interés del trabajo es conocer si ambas poblaciones tienen igual media poblacional, en este caso las hipótesis serán: H_0: $\mu_1 = \mu_2$ y H_A: $\mu_1 \neq \mu_2$. Las hipótesis se pueden escribir también como H_0: $\mu_1 - \mu_2 = 0$ y H_A: $\mu_1 - \mu_2 \neq 0$.

Al igual que en las pruebas de hipótesis relativas a una media poblacional, las pruebas para la diferencia de medias tienen distintos estadísticos de prueba según las varianzas poblacionales son conocidas o desconocidas. Cuando las **varianzas poblacionales son conocidas** el estadístico de prueba es el normal estándar *Z* (expresión 6.20), en tanto si las **varianzas poblacionales son desconocidas** y se estiman con las varianzas muestrales el estadístico de prueba es *t* (expresión 6.21).

Estadístico de prueba para **varianzas poblacionales conocidas**:

$$z = \frac{\overline{X_1} - \overline{X_2} - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{6.20}$$

Estadístico de prueba para **varianzas poblacionales desconocidas**:

$$t = \frac{\overline{X_1} - \overline{X_2} - \delta_0}{S_{\Delta \bar{X}}}, \tag{6.21}$$

para δ_0 cualquier diferencia hipotética de medias poblacionales (incluso 0 como se indicó en el párrafo precedente).

Antes de calcular *t* se debe verificar si las varianzas muestrales son iguales o diferentes con una prueba de Hipótesis *F* para comparar ambas varianzas.

Cuando las **varianzas muestrales son iguales** $S_{\Delta \bar{X}}$ se calcula con

$$S_{\Delta \bar{X}} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{6.22}$$

Pero cuando las **varianzas muestrales son diferentes** $S_{\Delta \bar{X}}$ se calcula con

$$S_{\Delta \bar{X}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}. \tag{6.23}$$

Las diferentes hipótesis que se pueden presentar son las siguientes:

H_0: $\mu_1 - \mu_2 = \delta_0$ vs. H_A: $\mu_1 - \mu_2 \neq \delta_0$, H_0 se rechaza si $|z| > Z_{\alpha/2}$ ó $|t| > t_{\alpha/2, \nu}$.

H_0: $\mu_1 - \mu_2 \leq \delta_0$ vs. H_A: $\mu_1 - \mu_2 > \delta_0$, H_0 se rechaza si $z > Z_\alpha$ ó $t > t_{\alpha, \nu}$.

H_0: $\mu_1 - \mu_2 \geq \delta_0$ vs. H_A: $\mu_1 - \mu_2 < \delta_0$, H_0 se rechaza si $z < -Z_\alpha$ ó $t < -t_{\alpha, \nu}$.

Los grados de libertad de las pruebas de *t* se calculan de diferente forma según el tamaño de las muestras sean iguales o diferentes. Se presentan cuatro casos:

a) Poblaciones con varianzas iguales y tamaños de muestra iguales ($S_1^2 = S_2^2$ y $n_1 = n_2 = n$), ν=2(n-1).

   Cabe aclarar que en este caso $S_{\Delta\bar{X}} = \sqrt{\frac{(n-1)s_1^2 + (n-1)s_2^2}{n-2}}\sqrt{\frac{2}{n}}$.

b) Poblaciones con varianzas iguales y tamaños de muestra diferentes ($S_1^2 = S_2^2$ y $n_1 \neq n_2$), ν=n_1+n_2-2.

c) Poblaciones con varianzas diferentes y tamaños de muestra iguales ($S_1^2 \neq S_2^2$ y $n_1 = n_2$), ν=(n−1).

d) Poblaciones con varianzas diferentes y tamaños de muestra diferentes ($S_1^2 \neq S_2^2$ y $n_1 \neq n_2$),

$$\nu = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1-1)}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2-1)}}$$

Cabe aclarar en este punto que los resultados solamente son válidos si en las dos poblaciones, población 1 y población 2, la variable se distribuye normalmente.

EJEMPLO 8

Prueba de hipótesis para dos muestras independientes

Las emisiones industriales generan y emiten material particulado con capacidad de adsorción hidrocarburos aromáticos polocíclicos, HAP's, que pueden producir cáncer. En la ciudad de La Plata, muy cercana al polo petroquímico, se realizaron muestreos entre el 3/5/2012 y 26/6/2012 con el objeto de saber si existen diferencias entre la cantidad de HAP's (%ng/m^3) adsorbidos por las fracciones granulométrica F2 y F3.

H_0: No existen diferencias en los HAP's adsorbidos por F2 y F3

H_A: Existen diferencias en los HAP's adsorbidos por F2 y F3.

H_0: $\mu_{F1} = \mu_{F2}$ o $\mu_{F1} - \mu_{F2} = 0$

H_A: $\mu_{F1} \neq \mu_{F2}$ o $\mu_{F1} - \mu_{F2} \neq 0$

Nivel de significación α= 0,05

Muestra	*F2*	*F3*
1	0,025	0,008
2	0,030	0,004
3	0,035	0,004
4	0,658	0,357
5	0,200	0,300
6	0,080	0,050
7	0,045	0,156
8	0,223	0,346
9	0,123	0,178
10	0,150	0,200
11	0,067	0,185
12	0,568	1,349
13	0,360	0,128
14	0,200	0,405
15	2,374	2,519
16	0,003	
17	0,201	
18	0,007	
N	**18**	**15**
Promedio	**0,297**	**0,413**
S	**0,535**	**0,646**
S^2	**0,286**	**0,417**

Prueba de Hipótesis para comparar las varianzas

H_0: $\sigma_{F1}^2 = \sigma_{F2}^2$

H_A: $\sigma_{F1}^2 \neq \sigma_{F2}^2$

Criterio de rechazo de H_0 $F \leq F_{1-\alpha,n1-1,n2-1}$.ó $F \geq F_{\alpha,n1-1,n2-1}$

$F = \frac{S_1^2}{S_2^2} = \frac{0,286}{0,417} = 0,6859$

De la Tabla 4 del Anexo, $F_{0,05;\ 17;\ 14} = 2,31$

Dado que 0,6859 ≤ 2,31 existen evidencias para aceptar la hipótesis nula. Las varianzas de ambas poblaciones son iguales.

Prueba de hipótesis para la diferencia de medias

Criterio de rechazo de H_0 $|t| > t_{\alpha/2,\ \nu}$

$\nu = 18+15-2 = 31$

De la Tabla 3 del Anexo, $t_{(0,05;\ 31)} = 1,697$

Dado que las varianzas muestrales son iguales corresponde calcular $S_{\Delta\bar{X}}$

$$t = \frac{(\bar{X}_{F1} - \bar{X}_{F2}) - \delta_0}{S_{\Delta\bar{X}}} = \frac{\bar{X}_{F1} - \bar{X}_{F2} - 0}{S_{\Delta\bar{X}}}$$

$$S_{\Delta\bar{X}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_{\Delta\bar{X}} = \sqrt{\frac{(18-1)0,286 + (15-1)0,417}{18+15-2}} \sqrt{\frac{1}{18} + \frac{1}{15}} = 1,142$$

$$t = \frac{0,297 - 0,413}{1,142} = \frac{-0,116}{1,142} = -0,10$$

Dado que |-0,10 |< 2,73, no existen evidencias para rechazar la hipótesis nula. Se puede concluir que no existen diferencias en la adsorción de HAP's entre F2 y F3.

Prueba de Hipótesis para muestras apareadas

Las muestras apareadas se obtienen con distintas mediciones realizadas sobre los mismos individuos. Uno de los objetivos que se buscan al realizar un muestro apareado es controlar o eliminar la influencia de variables extrañas con el fin de evaluar las diferencias que existen entre las dos observaciones en el mismo individuo. Las diferencias pueden ser producto de diferentes tratamientos (por ejemplo diferentes concentraciones, diferente tiempo de exposición ante un agente, diferentes métodos de análisis químicos). Se toma una muestra aleatoria de una población normal bivariada, se obtienen pares de mediciones para los *i* elementos de la muestra del tipo (x_{Ai}, x_{Bi}). La variable de estudio es la diferencia $d = x_A - x_B$. *d* se distribuye normalmente. La muestra tiene media $\overline{D}$ y varianza S_d^2. La prueba de hipótesis para *d* es una prueba de hipótesis para una sola muestra de la media, entonces el estadístico de prueba es *t* y tiene la siguiente expresión:

$t = \dfrac{\overline{X}_d - \delta_0}{\dfrac{S_d}{\sqrt{n}}}$, para δ_0 cualquier diferencia hipotética de medias poblacionales y n = n° de pares. (6.24)

Dependiendo del problema que se investiga se pueden plantear las siguientes hipótesis:

H_0: $\mu_d = \delta_0$ vs. H_A: $\mu_d \neq \delta_0$, H_0 se rechaza si $|t| > t_{\alpha/2,\ n-1}$.

H_0: $\mu_d \leq \delta_0$ vs. H_A: $\mu_d > \delta_0$, H_0 se rechaza si $t > t_{\alpha,\ n-1}$.

H_0: $\mu_d \geq \delta_0$ vs. H_A: $\mu_d < \delta_0$, H_0 se rechaza si ó $t < -t_{\alpha,\ n-1}$.

EJEMPLO 9

Prueba de hipótesis para datos apareados

Se sospecha que los resultados de los análisis químicos enviados al laboratorio AMM están sesgados por la metodología utilizada. Para corroborarlo se realiza un muestreo testigo de las emisiones industriales del material particulado del aire del área de Ensenada. Cada muestra se divide en dos alícuotas para duplicar los análisis, una se envía al laboratorio AMM y otra al laboratorio BestMar que se sabe no tiene error.

AMM	***BestMar***	***d***
6,5	5,4	***1,1***
5,6	5,8	***-0,2***
6,6	5,4	***1,2***
6,1	5,8	***0,3***
5,8	5,7	***0,1***
6,0	5,4	***0,6***
5,1	5,7	***-0,6***
6,3	6,0	***0,3***
6,1	5,3	***0,8***
6,6	6,0	***0,6***

n=10

$\overline{X}_d = 0{,}52$; $S_d^2 = 0{,}188$; $S_d = 0{,}434$

H_0: No existen diferencias entre los resultados de ambos laboratorios
H_A: Existen diferencias entre los resultados de ambos laboratorios
H_0: $\Delta\mu_d = 0$
H_1: $\Delta\mu_d \neq 0$
Nivel de significación de la prueba $\alpha = 0{,}01$
Criterio de rechazo de H_0 $|t| > t_{\alpha/2,\ n-1}$

De la Tabla 3 del Anexo, $t_{(0{,}05;\ 9)} = 1{,}833$

$$S_{\mu d} = \sqrt{\frac{s_d^2}{n}} = \sqrt{\frac{0{,}188}{10}} = 0{,}137$$

$$t = \frac{\overline{X}_d - \delta_0}{\dfrac{S_d}{\sqrt{n}}} = \frac{0{,}52}{0{,}137} = 3{,}795$$

Dado que 1,833 < 3,795 existen evidencias para rechazar la hipótesis nula.
Se puede concluir que existen diferencias entre los resultados de ambos laboratorios, se deduce que los resultados de *AMM* están sesgados.

Prueba de hipótesis para diferencia de proporciones

La prueba de hipótesis para diferencias de proporciones es análoga a la prueba de diferencia de medias para muestras grandes y/o con varianzas conocidas. Si se toman dos muestras aleatorias e independientes y *p* es la estimación del parámetro poblacional π, el estadístico de prueba es *Z* que se calcula como

$$Z = \frac{(p_1 - p_2) - \delta_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \qquad \text{para } \delta_0 \leq 0. \tag{6.25}$$

El estadístico de prueba que permite contrastar diferentes hipótesis:

H_0: $\pi_1 - \pi_2 = \delta_0$ vs. H_A: $\pi_1 - \pi_2 \neq \delta_0$, H_0 se rechaza si $|z| > Z_{\alpha/2}$.

H_0: $\pi_1 - \pi_2 \leq \delta_0$ vs. H_A: $\pi_1 - \pi_2 > \delta_0$, H_0 se rechaza si $z > Z_\alpha$.

H_0: $\pi_1 - \pi_2 \geq \delta_0$ vs. H_A: $\pi_1 - \pi_2 < \delta_0$, H_0 se rechaza si $z < -Z_\alpha$.

EJEMPLO 10

Prueba de hipótesis para diferencia de proporciones

La proporción de minerales arcillosos presentes en los suelos se vincula directamente a la capacidad de contraerse en seco y de expandirse en húmedo. En términos edafológicos esta propiedad se denomina *Expansión Libre* y se mide en %. Los suelos mejores para la construcción son aquellos que presentan valores menores de Expansión Libre. Se realizan nuestros aleatorios en dos suelos de la provincia de Buenos Aires, la Serie Estancia Chica (EA) y la Serie Esquina Negra (EN). Interesa conocer que suelo es más apto para la construcción.

Serie	*EC*	*EN*
Expansión Libre (%)	69	75
n	5	8

H_0: $\pi_{EC} \leq \pi_{EN}$
H_A: $\pi_{EC} > \pi_{EN}$
$\alpha = 0{,}05$

De la Tabla 1 del Anexo, $z_{0,05} = -1{,}645$
H_0 se rechaza si $z > z_\alpha$

$$Z = \frac{p_{EC} - p_{EN}}{\sqrt{\frac{p_{EC}(1-p_{EC})}{n_{EC}} + \frac{p_{EN}(1-p_{EN})}{n_{EN}}}} = \frac{0{,}69 - 0{,}75}{\sqrt{\frac{0{,}69(1-0{,}69)}{5} + \frac{0{,}75(1-0{,}75)}{8}}} = -0{,}23$$

Dado que -0,23 < -1,645, existen evidencias para aceptar la hipótesis nula. Se puede concluir que los suelos de la Serie Estancia Chica tienen menor expansión libre que los de la Serie Esquina Negra y son mejores para ser utilizados en la construcción.

Relación entre estimación por intervalos de confianza y prueba de hipótesis

Se ha visto que los Intervalos de Confianza se plantean para estimar parámetros, mientras que las Pruebas de Hipótesis para tomar decisiones en relación a los valores postulados para ellos. Sin embargo, los Intervalos de Confianza permiten tomar una decisión estadística analizando la ubicación

del valor parámetro hipotetizado en el intervalo de confianza. La hipótesis nula se rechaza cuando el valor del parámetro está fuera de los límites del intervalo de confianza.

Pruebas de bondad de ajuste

Las Pruebas de Bondad de Ajuste tienen por objetivo determinar si un conjunto de datos se puede modelar con una distribución teórica.

Prueba χ^2 (Chi cuadrado)

La prueba χ^2 es muy versátil pues se puede utilizar para datos nominales tanto como para datos numéricos continuos y discretos. El objetivo de la prueba es inferir si la población, se ajusta a una cierta distribución teórica. Los datos de la muestra deben ser **frecuencias absolutas**. Si los datos son continuos se deben agrupar en intervalos de clase.
El estadístico de prueba es una medida de la desviación entre las frecuencias observadas en la muestra respecto a las frecuencias de una distribución teórica.
El estadístico de prueba es:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}, \tag{6.26}$$

dónde *fo*: frecuencia observada, *fe*: frecuencia teórica o esperada y *k*: número de categorías.
Las hipótesis que se contrastan son:
H_0: La distribución observada se ajusta al modelo teórico.
H_A: La distribución observada no se ajusta al modelo teórico.
H_0: $fo = fe$ (las frecuencias observadas son iguales a las frecuencias esperadas).
H_A: $fo \neq fe$ (las frecuencias observadas son diferentes a las frecuencias esperadas).

Es fácil ver que cuanto mayor es la diferencia entre la frecuencia observada y la esperada, mayor será el valor de χ_c^2. Cuando la frecuencia observada es igual a la frecuencia esperada la diferencia es cero (0) y χ_c^2 tiende a cero (0). La hipótesis nula se rechaza cuando $\chi_c^2 \geq \chi_{\alpha;\nu}^2$. Los grados de libertad ν se calculan como *k* (número de clases) menos el número de parámetros utilizados para estimar las frecuencias esperadas menos 1 *(k – nº de parámetros estimados - 1)*.
Corrientemente se desconocen los parámetros de la población muestreada y se estiman con los estadísticos muestrales. Así cuando los datos se contrastan contra un modelo **Binomial** se estima *p*, en

el modelo **Poisson** se estima λ, en el modelo **Normal** se estiman μ y σ, y finalmente en el modelo **Uniforme** no se estima ningún parámetro.

Se recomienda que las categorías tengan frecuencia esperada de cinco o más ($fe \geq 5$). Cuando las frecuencias esperadas son menores a 5, se deben combinar las categorías contiguas hasta obtener esta cantidad. Por último, para evitar errores hay de calcular las frecuencias esperadas con cuatro decimales y χ^2 con tres decimales.

Se presentan dos problemas aplicados. El ejemplo 11 se prueba la bondad de ajuste a un modelo binomial y en ejemplo 12 a un modelo normal.

EJEMPLO 11

Prueba de bondad de ajuste χ^2 a un modelo binomial

Se postula que una industria papelera evacúa sus efluentes en un río. Durante 109 días se efectúa un muestre aguas debajo de la industria. Se toman 4 muestras de agua por día y se registra si contienen una sustancia contaminante por encima del límite permitido por la OMS. Si la empresa es responsable los niveles de concentración de la sustancia no ocurrirán al azar. El problema se puede resolver con una prueba de bondad de ajuste a un modelo Binomial.

H_0: La población muestreada se adecúa a un modelo binomial.
H_A: La población muestreada no se adecúa a un modelo binomial.
H_0: $fo = fe$
H_A: $fo \neq fe$
Nivel de significación, $\alpha = 0{,}05$

x_i	fo	p_i	$fe=p_i \cdot N$	χ^2
0	20	0,1678	18,2872	0,160
1	41	0,3775	41,1461	0,001
2	33	0,3185	34,7170	0,085
3	11	0,1194	13,0189	0,002
4	4	0,0168	1,8308	

En la distribución binomial $E = \bar{X} = p \cdot n$

$N = 109$
$n = 4$
$\bar{X} = 1{,}43$
$p = 0{,}36$

Las p_i se calculan con $B(x; n; p) = \left(\frac{n!}{x!(n-x)!}\right) p^x q^{n-x}$

Debido a que la $fe_{(x=4)}$ es 1,8308 < 5, se agrupan con las frecuencias de la clase anterior. Se pierde una categoría.

$\nu = k - n°\ parámetros\ estimados - 1 = 4 - 1 - 1 = 2$

De la Tabla 2 del Anexo, $\chi^2_{0{,}05;\,2} = 5{,}99$

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe} = 0{,}247$$

Dado que 0,247 < 5,99 no existen evidencias para rechazar la hipótesis nula.

Se puede concluir que los datos se han muestreado de una población Binomial. No hay evidencias que apoyen que la papelera contamine.

EJEMPLO 12

Prueba de bondad de ajuste de χ^2 a un modelo normal

El caudal del Río Salado de la provincia de Buenos Aires presenta alternancias de mermas y desbordes. Se postula que el caudal se comporta normalmente dentro de un periodo de 300 días. Se realizan mediciones diarios de caudal (m^3/s) en la estación de aforo de General Belgrano para testear la hipótesis.

Intervalo	*MC (x)*	*Fe*	*Intervalo*	Z sup	*P*	*fe=p·n*	
35,5-40,5	38	7	>40,5	-1,8	0,0359	10,7700	1,3197
40,5-45,5	43	54	40,5-45,5	-0,8	0,1760	52,8000	0,0273
45,5-50,5	48	120	45,5-50,5	0,2	0,3674	110,2200	0,8678
50,5-55,5	53	84	50,5-55,5	1,2	0,3056	91,6800	0,6434
55,5-60,5	58	31	55,5-60,5	2,2	0,1012	30,3600	0,0059
60,5-65,5	63	4	<60,5	∞	0,0139	4,1700	

H_0: La población muestreada es normal.
H_A: La población muestreada no es normal.
H_0: *fo = fe*
H_A: *fo* ≠ *fe*
Nivel de significación, $\alpha = 0,05$

Se calcula el punto medio de cada intervalo de clase MC(*x*) para calcular la media y desvío muestral.

$n = 300$, $\bar{X} = 49,5$, $S = 5$

El área bajo la curva normal (*p*) viene dado por ejemplo para el intervalo (40,5 - 45,5)

$$z = \frac{x-\bar{X}}{S}, \quad z_{sup} = \frac{45,5-49,5}{5} = -0,8 \quad , \quad z_{inf} = \frac{40,5-49,5}{5} = -1,8$$

De la Tabla 1 del Anexo
$p(z_{sup} \leq -0,8) - p(z_{inf} \leq -1,8) = 0,4641 - 0,2881 = 0,1760$

El z_{sup} de un intervalo es el z_{inf} del siguiente intervalo. El z_{inf} del primer intervalo es -∞. El z_{sup} del último intervalo es +∞.
Debido a que la *fe* del último intervalo es 4,17 < 5, se agrupan con las frecuencias del intervalo precedente. Se pierde una categoría.

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe} = 2,864$$

Se estiman la media y el desvío estándar poblacional.

$\nu = k$ – n° parámetros estimados – 1 = 5 – 2 – 1 = 2

De la Tabla 2 del Anexo, $\chi^2_{0,05;\,2} = 5,99$

Dado que 2,864 < 5,99 no existen evidencias para rechazar la hipótesis nula.
Se puede concluir que los datos se han muestreado de una población normalmente distribuida.

Método G *de Fisher* (*Log-likelihood ratio*)

La prueba *G* de Fisher tiene las mismas aplicaciones que la prueba de χ^2. En relación a la prueba anterior, no es tan sensible a las frecuencias esperadas bajas por lo que no es necesario agrupar categorías. El estadístico de prueba *G* se calcula con la siguiente expresión:

$$G = 2\sum_{i=1}^{k} fo \cdot \ln\left(\frac{fo}{fe}\right) \quad \text{ó} \quad G = \sum_{i=1}^{k} fo\, ln\, fo - \sum_{i=1}^{k} fe\, lnfe. \tag{6.27}$$

La hipótesis nula se rechaza cuando $\boldsymbol{G \geq \chi^2_{\alpha;\, \nu}}$, los grados de libertad se calculan igual que en la prueba de χ^2, $\nu = k$ – *(número de parámetros estimados) – 1*.

EJEMPLO 13
Prueba de bondad de ajuste *G* de Fisher
Se prueba el supuesto de normalidad de datos diarios de caudal (m^3/s) del Río Salado, provincia de Buenos Aires.

H_0: La población muestreada es normal.
H_A: La población muestreada no es normal.
H_0: $fo = fe$
H_A: $fo \neq fe$
Nivel de significación, $\alpha = 0{,}05$

$$G = 2\sum_{i=1}^{k} fo \cdot \ln\left(\frac{fo}{fe}\right)$$

$$G = 2\left(7\ln\frac{7}{10{,}77} + 54\ln\frac{52}{52{,}8} + ... + 4\ln\frac{4}{4{,}17}\right) = 3{,}06$$

$\nu = k$ – *n° parámetros estimados – 1* = 6 – 2 – 1 = 3
De la Tabla 3 del Anexo, $\chi^2_{0{,}05;\, 3} = 7{,}51$

Dado que 3,06 < 7,51 no existen evidencias para rechazar la hipótesis nula.
Se puede concluir que los datos se han muestreado de una población normalmente distribuida como en la prueba anterior.

Test de Kolmogorov-Smirnov

La prueba de bondad de ajuste de Kolmogorov-Smirnov se puede usar con variables continuas, con variables discretas y con variables que presenten sus datos agrupados. Es una prueba sensible a hallar las diferencias entre cualquier tipo de distribución empírica y teóricas, tiene la virtud de ser útil con tamaño de muestra pequeños y su cálculo es muy simple.

Las hipótesis de esta prueba son las mismas que las de otras pruebas de bondad de ajuste (H_0: la distribución observada se ajusta al modelo teórico, H_A: la distribución observada no se ajusta al modelo teórico).

Para realizar la prueba para **datos continuos** se requiere calcular las cantidades

$$d_i = \left|Fr_i - \widehat{Fr_i}\right| \tag{6.28}$$

$$d_i^{'} = \left|Fr_{i-1} - \widehat{Fr_i}\right| \tag{6.29}$$

donde Fr_i es la frecuencia relativa acumulada observada del *i*-ésimo dato, $\widehat{Fr_i}$ la frecuencia relativa acumulada esperada y Fr_{i-1} la frecuencia relativa del dato anterior al *i*-ésimo.

El estadístico de prueba es

$$d = \max\left[(\max d_i), \left(\max d_i^{'}\right)\right], \tag{6.30}$$

esto significa que el estadístico de prueba *d*, es el máximo valor de cualquiera de las dos valores máximos obtenidos en d_i o $d_i^{'}$. La hipótesis nula se rechaza cuando $d \geq D_{\alpha,\, n}$. Los valores de *D* críticos se encuentran en la Tabla 5 del Anexo.

Para usar la prueba de **Kolmogorov-Smirnov** para el supuesto de **normalidad** se requiere conocer la media μ y la varianza poblacional σ^2, hecho que sucede rara vez. **Lillifords** (1967) propone calcular las frecuencias esperadas utilizando los estimadores muestrales y buscar el valor crítico *D* en una tabla que calculó *ad-hoc* (Tabla 6 del Anexo). En el ejemplo 14 se encuentra una aplicación de ajuste a un modelo normal.

EJEMPLO 14

Prueba de bondad de ajuste de Kolmogorov - Smirnov para datos continuos

Los Rodados Patagónicos, de edad Cenozoico tardío, son gravas cuyo origen se atribuye a zonas pedemontanas o a procesos glacifluvial. En las proximidades del Bajo de San Julián, provincia de Santa Cruz, se mide, en forma expeditiva en el campo, el diámetro máximo de 25 rodados elegidos al azar con una regla graduada cada 5 milímetros. Si bien las mediciones no son exactas se desea saber si la población muestreada el tamaño sigue un modelo normal.

Diam. máx.	fo_i	Fr_i	*z*	$p_i = \widehat{Fr_i}$	|*d*|	|*d'*|
1,5	1	0,04	-3,89	0,0000	0,04	0,04
2	1	0,08	-2,76	0,0029	0,08	0,04
2,5	3	0,20	-1,63	0,0515	0,15	0,03
3	7	0,48	-0,50	0,3092	**0,17**	0,11
3,5	8	0,80	0,63	0,7370	0,06	**0,26**
4	4	0,96	1,77	0,9613	0,00	0,16
4,5	1	1,00	2,90	0,9981	0,00	0,04

H_0: La población muestreada es normal.

H_A: La población muestreada no es normal.

Nivel de significación, $\alpha = 0{,}05$

$n = 25$, $\bar{X} = 3{,}2$, S=0,7

Se estandarizan los datos con: $z = \frac{x - \bar{X}}{S}$

De la Tabla 1 del Anexo, se obtienen $p(z \leq z_0) = \widehat{fr_i}$
Por ejemplo
$d_4 = |0{,}96 - 0{,}9613| = 0{,}00$ y $d_4' = |0{,}8 - 0{,}9613| = 0{,}16$

$$d = \max\left[(\max d_i), (\max d_i')\right]$$
$$d = \max[(0{,}17), (0{,}26)]$$

En este caso se utilizaron los estadísticos $\bar{X}$ y S^2 para estimar μ y σ^2 entonces corresponde buscar el valor crítico en la Tabla de Lillifords.
De la Tabla 6 del Anexo, $D_{0{,}05;\ 25} = 0{,}18$
Dado que 0,26 > 0,18 existen evidencias para rechazar la hipótesis nula.
Se puede concluir que los datos se han muestreado de una población que no es normal.

En la prueba de Kolmogorov – Smirnov para **datos discretos o agrupados** sólo requiere calcular el valor absoluto de las diferencias entre las frecuencias acumuladas observadas y esperadas

$$d_i = \left|fa_i - \widehat{fa_i}\right|, \tag{6.31}$$

donde fa_i es la frecuencia acumulada observada del *i*-ésimo dato y $\widehat{fa_i}$ la frecuencia acumulada esperada según el modelo teórico.
El estadístico de prueba *d* es simplemente el máximo valor d_i,

$$d = max\left|fa_i - \widehat{fa_i}\right|, \tag{6.32}$$

La hipótesis nula se rechaza cuando $d \geq D_{\alpha,;\ n;\ k}$ (*n* =N° de datos, *k* = N° de categorías). Los valores de *D* se encuentran en la Tabla 7 del Anexo. En el ejemplo 15 se encuentra una aplicación de ajuste a un modelo binomial con los datos del ejemplo 11.

EJEMPLO 15
Cálculo una prueba de bondad de ajuste de Kolmogorov-Smirnov para datos discretos
Se realiza la prueba con los datos del ejemplo 11 de la industria papelera que evacúa sus efluentes en un río. Se prueba la bondad de ajuste a un modelo Binomial.

H_0: La población muestreada se adecúa a un modelo binomial.
H_A: La población muestreada no se adecúa a un modelo binomial.
H_0: $fa_i = \widehat{fa_i}$
H_A: $fa_i \neq \widehat{fa_i}$
Nivel de significación, α = 0,05
$d = max|105 - 107{,}17| = 2{,}17$
De la Tabla 7 del Anexo, $D_{0{,}05;\ 100;\ 5} = 9$

x_i	f_i	p_i	$\widehat{f_i} = p_i \cdot N$	fa_i	$\widehat{fa_i}$	d
0	20	0,1678	18,287	20	18,287	1,713
1	41	0,3775	41,146	61	59,433	1,567
2	33	0,3185	34,717	94	94,150	0,150
3	11	0,1194	13,019	105	107,169	2,169
4	4	0,0168	1,831	109	109	0

Dado que 2,17 < 9 no existen evidencias para rechazar la hipótesis nula.
Se puede concluir que los datos se han muestreado de una población Binomial como sucedió con la prueba de χ^2. No hay evidencias que apoyen que la papelera contamine.

Síntesis

PRUEBAS PARA UNA MUESTRA			
	Hipótesis	Estadístico de prueba	Criterio de rechazo
Prueba para una media. Varianza Poblacional conocida y/o n > 30. μ_0= un número fijo conocido	Ho: $\mu = \mu_0$ H1: $\mu \neq \mu_0$.	$z_0 = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$	$\lvert z_0 \rvert > z_{\alpha/2}$ Bilateral
	Ho: $\mu \geq \mu_0$ H1: $\mu < \mu_0$		$z_0 < -z_\alpha$ Unilateral izquierda
	Ho: $\mu \leq \mu_0$ H1: $\mu > \mu_0$		$z_0 > z_\alpha$ Unilateral derecha
Prueba para una media. Varianza Poblacional desconocida y/o n < 30	Ho: $\mu = \mu_0$ H1: $\mu \neq \mu_0$.	$t_0 = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}}$	$\lvert t_0 \rvert > t_{\frac{\alpha}{2};\nu}$ Bilateral
	Ho: $\mu \geq \mu_0$ H1: $\mu < \mu_0$		$t_0 < -t_{\alpha;\nu}$ Unilateral izquierda
	Ho: $\mu \leq \mu_0$ H1: $\mu > \mu_0$		$t_0 > t_{\alpha;\nu}$ Unilateral derecha
Prueba de Varianza	Ho: $\sigma^2 = \sigma_0^2$ H1: $\sigma^2 \neq \sigma_0^2$.	$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$	$\chi^2 \leq \chi^2_{1-\alpha/2,n-1}$ ó $\chi^2 \geq \chi^2_{\alpha/2,n-1}$ Bilateral
	Ho: $\sigma^2 \geq \sigma_0^2$ H1: $\sigma^2 < \sigma_0^2$.		$\chi^2 \geq \chi^2_{\alpha/2,n-1}$ Unilateral izquierda
	Ho: $\sigma^2 \leq \sigma_0^2$ H1: $\sigma^2 > \sigma_0^2$.		$\chi^2 \leq \chi^2_{1-\alpha/2,n-1}$ Unilateral derecha
PRUEBAS PARA DOS MUESTRAS			
Prueba para comparar Varianzas	Ho: $\sigma_1^2 = \sigma_2^2$ H1: $\sigma_1^2 \neq \sigma_2^2$.	$F = \frac{S_1^2}{S_2^2}$	$F \leq F_{1-\alpha,n1-1,n2-2}$.ó $F \geq F_{\alpha,n1-1,n2-2}$ Bilateral
	Ho: $\sigma_1^2 \geq \sigma_2^2$ H1: $\sigma_1^2 < \sigma_2^2$.		$F \leq F_{\alpha,n1-1,n2-2}$ Unilateral izquierda
	Ho: $\sigma_1^2 \leq \sigma_2^2$ H1: $\sigma_1^2 > \sigma_2^2$.		$F \geq F_{\alpha,n1-1,n2-2}$ Unilateral derecha
Prueba de diferencia de medias. Varianzas Poblacionales conocidas	Ho: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$	$z_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\lvert z_0 \rvert > z_{\alpha/2}$ Bilateral
	Ho: $\mu_1 \geq \mu_2$ H1: $\mu_1 < \mu_2$		$z_0 < -z_\alpha$ Unilateral izquierda
	Ho: $\mu_1 \leq \mu_2$ H1: $\mu_1 > \mu_2$		$z_0 > z_\alpha$ Unilateral derecha
Prueba de diferencia de medias. Varianzas Poblacionales desconocidas y estimadas a partir de las varianzas muestrales	Ho: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$	$t_0 = \frac{\overline{X}_1 - \overline{X}_2}{S_{\Delta\overline{X}}}$ Recordar que previamente debe calcularse la prueba de homogeneidad de varianzas para utilizar $S_{\Delta\overline{X}}$ que corresponda	$\lvert t_0 \rvert > t_{\frac{\alpha}{2};\nu}$ Bilateral
	Ho: $\mu_1 \geq \mu_2$ H1: $\mu_1 < \mu_2$		$t_0 < -t_{\alpha;\nu}$ Unilateral izquierda
	Ho: $\mu_1 \leq \mu_2$ H1: $\mu_1 > \mu_2$		$t_0 > t_{\alpha;\nu}$ Unilateral derecha
Prueba de Muestras apareadas	Ho: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$	$t_0 = \frac{\overline{X}_d - \mu_{\mu d}}{S_{\mu d}}$	$\lvert t_0 \rvert > t_{\frac{\alpha}{2};\nu}$ Bilateral
	Ho: $\mu_1 \geq \mu_2$ H1: $\mu_1 < \mu_2$		$t_0 < -t_{\alpha;\nu}$ Unilateral izquierda
	Ho: $\mu_1 \leq \mu_2$ H1: $\mu_1 > \mu_2$		$t_0 > t_{\alpha;\nu}$ Unilateral derecha
Prueba de diferencia Proporciones	Ho: $\pi_1 = \pi_2$ H1: $\pi_1 \neq \pi_2$	$Z_0 = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$	$\lvert z_0 \rvert > z_{\alpha/2}$ Bilateral
	Ho: $\pi_1 \geq \pi_2$ H1: $\pi_1 < \pi_2$		$z_0 < -z_\alpha$ Unilateral izquierda
	Ho: $\pi_1 \leq \pi_2$ H1: $\pi_1 > \pi_2$		$z_0 > z_\alpha$ Unilateral derecha

ANÁLISIS DE LA VARIANZA

COMPARACIONES MÚLTIPLES ENTRE MEDIAS MUESTRALES

Introducción

Se ha visto la importancia que tiene para los geólogos comparar dos medias muestrales. En el capítulo precedente se describieron dos procedimientos: estableciendo límites de confianza y a partir de la prueba de hipótesis *t*. Sin embargo a veces el problema que se investiga plantea el estudio y comparación de más de dos medias muestrales, por ejemplo el contenido de carbonatos en arenas de diferentes afloramientos o localidades ($\mu_1 = \mu_2 = \cdots = \mu_k$). Ante esta circunstancia se podrían intentar comparar, de a pares, todos las medias, utilizando test de hipótesis *t*, no obstante esto presenta dos problemas principales. El primer inconveniente surge pues a medida que el número de comparaciones aumenta, aumenta la probabilidad de cometer errores de Tipo I, rechazar la hipótesis nula cuando es verdadera (i.e. rechazar la hipótesis que dos medias poblacionales son iguales)[12]. Por ejemplo para cinco medias se deberían realizar diez pares de comparaciones (μ_1 vs. μ_2, μ_1 vs. μ_3, …, μ_4 vs. μ_5) usando un nivel de significación del $\alpha = 0{,}05$, la probabilidad de cometer al menos un error de Tipo I es 0,40. El segundo problema se relaciona con el tamaño de las muestras pues la mayoría de las veces se tienen muy pocas observaciones en cada muestra estadística como para tener una buena estimación de la varianza poblacional σ^2.

Para subsanar estas dificultades se han desarrollado un conjunto de estrategias que reciben el nombre general de Análisis de la Varianza, conocido por las siglas ANOVA (Analysis of variance). Estas metodologías normalmente se utilizan para el análisis de datos obtenidos a través de la aplicación de diferentes diseños de experimentos, no obstante, aunque los datos geológicos, debido a su duración y extensión areal son en su mayoría observacionales, también es posible usarlas.

Análisis de la Varianza de un factor

Definiciones

En esta oportunidad se utilizará un ejemplo hipotético muy simplificado para explicar conceptualmente las bases del análisis. Se sostiene que los sedimentos de fondo del arroyo Del Gato ubicado en el Partido de La Plata, provincia de Buenos Aires, se encuentran contaminados con PCB's (Bifenilos Policlorados). Se postula que la concentración de PCB´s no es la misma en diferentes sectores del arroyo debido a que el uso del territorio circundante varía. Para someter a prueba la hipótesis se dividió al arroyo en tres sectores bien diferentes. En términos estadísticos cada sector representa y es llamado un **tratamiento**. Se realiza un muestreo de sedimentos en cada sector, la selección de los puntos de muestreo en cada sector se efectua al azar.

En cabeceras el territorio se utiliza para el cultivo de hortalizas y flores (SHF), en el tramo medio es urbano (SU) y en el tramo inferior el arroyo se encuentra canalizado (SC). En cada sector se toman varias muestras de sedimentos, en la terminología de ANOVA las muestras de sedimento de cada sector se llaman **réplicas** (n_i). Se toman 6 muestras de sedimentos en SHF, 4 en SC y 4 SU, de modo que en total se tienen N = 14 datos (Fig. 1, Tabla 1).

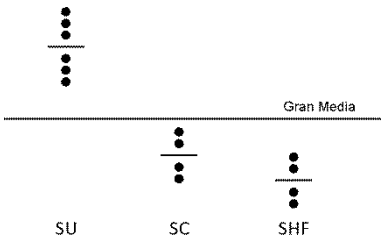


Figura 1. Representación de los datos de PCB´s del arroyo Del Gato. Los círculos representan los datos de cada tratamiento, las líneas cortas la media de cada tratamiento y la línea horizontal larga representa el valor medio de todos los datos.

En ANOVA lo más corriente es utilizar *X* para referirse a la variable dependiente o respuesta. En este caso la concentración de PCB´s en los sedimentos es la variable respuesta. La variable independiente, también es llamada tratamiento posee *k* categorías o condiciones en las cuales se toman los datos a comparar, en este caso corresponden a los tres sectores del arroyo. Ya se mencionó que n_i son las repeticiones o réplicas, es decir la cantidad de datos tomados en cada sector y *N* es el número total de datos. Cada dato es representado por la notación x_{ij} donde *j* es el *j*-ésimo dato en el *i*-ésimo tratamiento (Ver ejemplo).

El modelo

El modelo que utiliza el análisis de la varianza supone que si el proceso de muestreo ha sido **aleatorio**, los datos de PCB´s se definen como una muestra de n_i observaciones de la **población** correspondiente a **cada sector de la cuenca**. Esto es, hay una única muestra de n_i observaciones para

cada una de las tres **poblaciones normales** muestreadas. Cada media poblacional se designa con μ_i, y cada varianza poblacional con σ^2, **la varianza poblacional es la misma** para los tres sectores de la cuenca.

La **hipótesis nula** del ANOVA propone que las muestras son tomadas de la misma población y que además, esta población es normal, o bien que se han muestreado poblaciones idénticas, que tienen todas la misma media y varianza poblacional (Fig. 2).

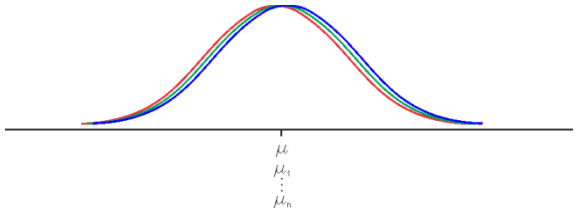


Figura 2.Hipótesis nula de ANOVA: Las k *muestras son tomadas de poblaciones idénticas.*

La **hipótesis alternativa** del ANOVA señala que las muestras son tomadas de diferentes poblaciones, o bien que al menos una muestra es tomada de una población diferente a las demás, pero todas estas poblaciones tienen la misma varianza (Fig. 3).

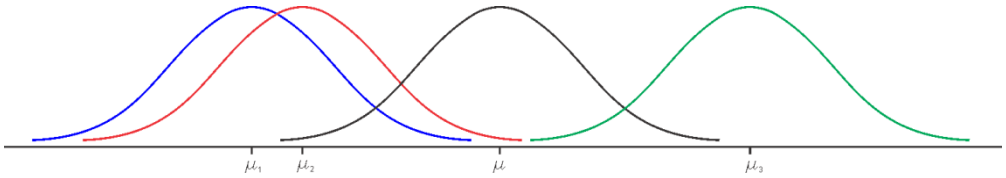


Figura 3.Hipótesis alternativa de ANOVA: Las k *muestras son tomadas de poblaciones normales diferentes con idénticas varianzas. Se muestra el punto de equilibrio de cada μ_i y el de todas ellas en μ (curva negra).*

Cuando el investigador acepta la hipótesis nula, en el ejemplo de las concentraciones de PBC´s en los sedimentos, se podrá concluir que este no varía en los diferentes sectores de la cuenca. En cambio, si rechaza la hipótesis nula, se concluirá que la concentración de PBC´s es diferente en al menos un sector. La interpretación depende del modelo matemático que se usa. Se trata de un modelo lineal que se define de la siguiente manera: cada observación x_{ij}, se puede representar con la distancia al punto de equilibrio del sistema, μ, más la distancia entre la media de la población de donde se extrajo, μ_i, al punto de equilibrio (μ_i - μ), mas una cantidad ε_{ij}, que representa la variación de la observación respecto a la media de su población. Se asume que el error épsilon tiene media cero y varianza igual a la varianza poblacional. La expresión para una observación es entonces:

$$x_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}\,, \qquad (7.1)$$

donde x_{ij} es la j-ésimo dato del i-ésimo factor, μ es la media general de los datos o el punto de equilibrio, (μ_i - μ) es el efecto del i-ésimo factor, ε_{ij} es una variable aleatoria normal, con esperanza 0 y varianza igual a la varianza poblacional σ^2 ($\mu_\varepsilon = 0;\ \sigma_\varepsilon^2 = \sigma^2$).

Si las medias de todas las poblaciones μ_i son iguales, el término del paréntesis (μ_i - μ) es igual a cero. Pero si las medias poblacionales no son iguales, una medida de estas diferencias es la varianza de la población de medias (Fig. 3).

Ahora bien, si las muestras son tomadas en forma aleatoria de una población común (**esta es la hipótesis nula**), es razonable esperar que las variaciones **entre** las muestras sea aproximadamente la misma que la variaciones **dentro** de las muestras y que ambas reflejen la variación de la población. Cualquier diferencia entre estas dos medidas de variación se debe puramente al azar y es producida por el procedimiento de muestreo. Sin embargo, si las muestras son tomadas de diferentes poblaciones (**la hipótesis alternativa**), no es razonable esperar esto, dado que la variación entre las muestras es el reflejo de la variación de la población de la cual es extraída. La lógica del ANOVA es encontrar, si es que existe, más variación Entre muestras diferentes o Dentro de una misma muestra. Hallar desviaciones entre muestras, en este caso, estarán mostrando la diferencia entre las poblaciones.

Entonces, el problema se circunscribe a encontrar la forma de estimar la varianza poblacional que es común a todas las distribuciones muestreadas.

Para **estimar** la varianza poblacional común σ^2, se calcula una varianza ponderada a partir de las varianzas muestrales de las distintas poblaciones. Esto es calculando lo que se conoce como el **Cuadrado Medio Dentro** (CMDentro) o **Cuadrado Medio del Error**.

$$CMDentro = \frac{SCDentro}{glDentro};\ SCDentro = \sum_{i=1}^{k}\left[\sum_{j=1}^{n}\left(x_{ij} - \overline{X}_i\right)^2\right];\ glDentro = \sum_{i=1}^{k}(n_i - 1) = N - k. \qquad (7.2)$$

Pero también se puede calcular una varianza ponderada a partir de las varianzas entre las medias muestrales de las distintas poblaciones y la gran media esto se conoce como **Cuadrado Medio Entre** (CMEntre):

$$CMEntre = \frac{SCEntre}{glEntre};\qquad SCEntre = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X});\qquad glEntre = k - 1. \qquad (7.3)$$

Además es posible calcular una **varianza total** (CMTotal):

$$CMTotal = \frac{SCTotal}{glTotal};\qquad SCTotal = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{X});\qquad glTotal = N - 1. \qquad (7.4)$$

Si $\sum_1^n x_{ij}$, es la suma de las observaciones de cada muestra, $\overline{X}i$ es al promedió de la *i*-ésima muestra y $\overline{X}$ el promedio de todos los datos, llamado la Gran Media, cada desviación de una observación a la gran media, se debe a la desviación de cada dato a la media grupal, más una desviación de la media de cada grupo a la gran media (Fig. 4).

$$\left(x_{ij} - \overline{X}\right) = \left(\overline{X}_i - \overline{X}\right) + \left(x_{ij} - \overline{X}_i\right) \qquad (7.5)$$

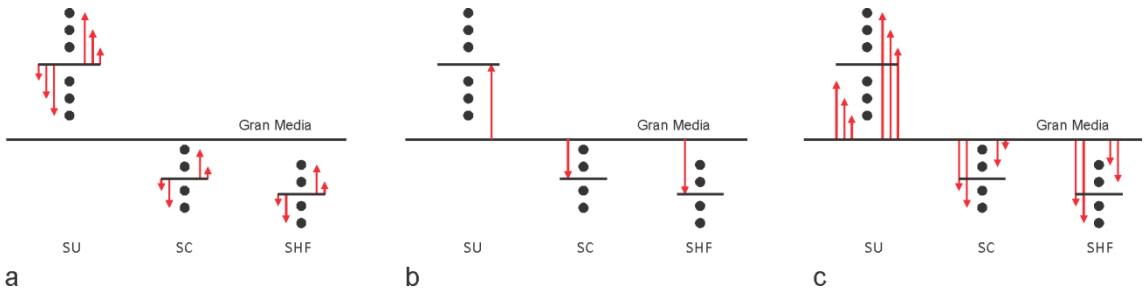


Figura 4. Fuentes de variación del ANOVA. a. Variación Dentro del tratamiento (las flechas indican la desviación de cada dato a la media de su respectivo tratamiento). b. Variación Entre tratamientos (las flechas indican la desviación entre la media de cada tratamiento y la Gran media). c. Variación Total (las flechas indican la desviación de los datos y la Gran Media).

La teoría estadística informa que la varianza total que presentan los datos puede descomponerse o ser atribuida a dos fuentes: la primera causada por la diferencia entre tratamientos, es lo que se llamó Cuadrado Medio Entre y la segunda producida por variaciones aleatorias llamada Dentro o Error. Entonces, si se consideran todas las observaciones, la expresión 7.5 se puede escribir como

$$SCTotal = SCEntre + SCDentro. \qquad (7.6)$$

Además, los grados de libertad Total son la suma de los grados de libertad Entre y los grados de libertad Dentro ($glTotal = glEntre + glDentro$). Pero no se cumple que el cuadrado medio total sea la suma de las cuadros medios entre y dentro ($CMTotal \neq CMEntre + CDentro$).

Retomando las hipótesis de ANOVA, como las poblaciones tienen todas la misma varianza, entonces las *k* varianzas muestrales estiman al mismo parámetro poblacional, y el promedio ponderado de estas varianzas es un buen estimador de esta varianza poblacional. Se puede probar que el **Cuadrado Medio Dentro** es un estimador insesgado de la varianza poblacional. El **Cuadrado Medio Entre**, en cambio, solamente estima a la varianza poblacional cuando la Hipótesis Nula del ANOVA es cierta. Es decir cuando las medias de las poblaciones muestreadas son iguales, ya que la componente de la varianza total producida por los factores de variación se anula y entonces Cuadrado Medio Entre es la varianza poblacional. Pero, si la Hipótesis nula no es verdadera, el **Cuadrado Medio Entre** estima a la varianza poblacional más una cantidad que representa una medida de la magnitud de los efectos del tratamiento. Luego, la comparación las varianzas calculadas, Cuadrado Medio Entre y Cuadrado Medio Dentro, permite cotejar medias poblacionales. El problema se reduce al cálculo de estas dos varianzas. Se trata entonces de descomponer la variabilidad total que presentan los datos de manera de distinguir estas dos fuentes de variación, entre tratamientos y dentro de cada tratamiento y luego compararlas.

Procedimiento para el cálculo

Recuerde que para facilitar los cálculos manuales y evitar errores al momento de calcular la varianza (Capítulo 2) se utilizó la Suma de Cuadrados (SC):

$$SC = \sum \left(x_i - \overline{X}\right)^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n} .$$

Se utilizaran diferencias análogas a esta para realizar los cálculos. Las varianzas o Cuadrados Medios se obtienen a partir del cociente entre suma de cuadrados y grados de libertad.

Para calcular la variabilidad de todos los datos, se usa lo que se conoce como la Suma de Cuadrados Total (***SCTotal***) que expresa la suma de las distancias de cada observación x_{ij} a la Media total de datos, la llamada Gran Media ($\overline{X}$)

$$SCTotal = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}^2 - \overline{X}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - C . \qquad (7.7)$$

Donde *C* es el Factor de corrección de la media:

$$C = \frac{\left(\sum_{i=1}^{K}\sum_{j=1}^{n} x_{ij}\right)^2}{N} . \qquad (7.8)$$

Los grados de libertad Total son el número total de datos menos uno (***glGLTotal = N-1***). El Cuadrado Medio Total (***CMTotal***) es entonces:

$$CMTotal = \frac{SCTotal}{N-1} . \qquad (7.9)$$

Por otra parte, la Suma de Cuadrados Entre (***SCEntre***) es el promedio de las distancias entre cada una de las medias muestrales a la gran Media:

$$SCEntre = \sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X}\right)^2 = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} x_{ij}\right)^2}{n_i} - C . \qquad (7.10)$$

Los grados de libertad entre tratamientos son el número de tratamientos menos uno (***glEntre = k - 1***). El Cuadrado Medio Entre (***CMEntre***) se calcula con la siguiente expresión:

$$CMEntre = \frac{SCEntre}{k-1} . \qquad (7.11)$$

Como se asumió que las varianzas poblacionales son iguales, la varianza dentro de los grupos o Error es la distancia que entre cada observación a la media de su propio grupo (tratamiento), es posible calcular la Suma de Cuadrados Dentro (***SCDentro***) como:

$$SCDentro = \sum_{i=1}^{K}\left(\sum_{j=1}^{n} x_{i\,j} - \overline{X}_i\right)^2 = SCtotal - SCEntre . \qquad (7.12)$$

Finalmente, los grados de libertad dentro surgen de la diferencia entre número total de datos y el número de tratamientos (***glDentro = glTotal – GLEntre = N - k***). El Cuadrado Medio Dentro es:

$$CMDentro = \frac{SCDentro}{N-k}. \tag{7.13}$$

Para probar la hipótesis nula que todas las medias son iguales frente a la alternativa que al menos una de las igualdades no se cumpla,

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$;

H_A: $\mu_i \neq \mu_j$, para al menos un par de (i, j)

se compara el CMEntre con el CMDentro ya que se trata de las dos varianzas que miden en forma independiente la varianza de la distribución de medias muestrales. La prueba de hipótesis para comparar si dos varianzas son iguales contrasta el cociente entre ambas con el estadístico *F* de Fisher**.** La hipótesis nula se rechaza cuando

$$\frac{CMEntre}{CMDentro} > F_{\alpha;k-1;N-k}. \tag{7.14}$$

Se realiza una prueba a una cola ya que se trata de detectar la variabilidad que tienda a aumentar la varianza Entre medias.

Los cálculos del análisis de la varianza se presentan generalmente en una tabla como la del Cuadro 1 que muestra, para el caso de un experimento aleatorizado que contiene *k* tratamientos, en la primera columna el origen de cada suma de cuadrados de las desviaciones; la segunda las Suma de Cuadrados, la tercera los Grados de Libertad, la cuarta los Cuadrados Medios y por último, la quinta presenta el valor calculado del estadístico de prueba *f*, comparando CMEntre y CMDentro.

Intuitivamente a mayor diferencia entre las medias observadas de los tratamientos, mayor es la evidencia que indica una diferencia entre las medias poblacionales correspondientes. Analizando la relación expresada en la SCEntre, es claro que a medida que las medias se alejan una de otras, las desviaciones aumentarán en valor absoluto y la SCEntre aumentará en magnitud. Por consiguiente a mayor valor de SCEntre mayor peso de la evidencia en rechazar la hipótesis nula.

Fuente de variación	**Suma de Cuadrados (SC)**	**Grados de libertad (gl)**	**Cuadrado Medio (CM)**	***F***
Entre Tratamientos	$\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} x_{ij}\right)^2}{n_i} - C$	*k-1*	$\frac{SCEntre}{k-1}$	$\frac{CMEntre}{CMDentro}$
Dentro de los Tratamientos	$SCTotal - SCEntre$	*N-k*	$\frac{SCDentro}{N-k}$	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - C$	*N-1*		

Cuadro 1. Tabla resumen de ANOVA. Para $C = \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n} Xij\right)^2}{N}$

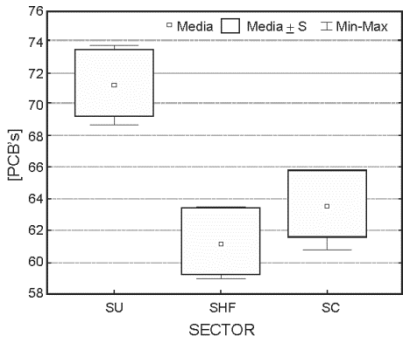
EJEMPLO 1

ANOVA de un factor con replicas diferentes

Se postula que la concentración de PCB´s en la cuenca del arroyo Del Gato no es la misma en diferentes sectores del mismo. En cabeceras el territorio se utiliza para el cultivo de hortalizas y flores (SHF), en el tramo medio es urbano (SU) y en el tramo inferior el arroyo se encuentra canalizado (SC). Se realizó un muestreo de sedimentos en cada sector, la selección de los puntos de muestreo en cada sector se efectuó al azar. Se toman 6 muestras de sedimentos en SHF, 4 en SU y 4 SC.

Sector / *Replica*	*SHF*	*SC*	*SU*	*Total*
1	59,1	60,8	69,6	
2	60,5	63,4	68,7	
3	63,5	65,5	72,3	
4	62,7	64,9	73,2	
5			70,4	
6			73,6	
n_i	4	4	6	14
$\sum_1^n x_{ij}$	245,8	254,6	427,8	928,2
$\bar{X}$	61,45	63,65	71,30	196,4
$\left(\sum_1^n x_{ij}\right)^2$	60417,64	64821,16	183012,84	308251,64
$\left(\sum_1^n x_{ij}\right)^2 \Big/ n_i$	15104,41	16205,29	30502,14	61811,84
$\sum_1^n x_{ij}^2$	15116,60	16218,46	30522,50	61857,56

Tabla 1. Datos de concentración de PCB's en sedimentos de fondo del arroyo Del Gato en tres sectores de la cuenca: Hortalizas y flores (SHF), urbano (SU) y canalizado (SC).



Box-plot de [PCB's] en sedimentos del arroyo del Gato.

Variable dependiente, *X*: [PCB's]
Tratamientos: sectores de la cuenca
k = nº tratamientos = 3 (varía de i = 1, …, k)
n_i = nº de replicas (varían según los sectores de la cuenca de j = 1, …, n)
N = nº total de observaciones = 14

H_0: las medias poblacional μ_i de [PCB's] de los diferentes sectores de la cuenca del arroyo son iguales
H_A: al menos una de las medias poblacional μ_i de [PCB's] de los diferentes sectores de la cuenca del arroyo es diferente

H_0: $\mu_{SHF} = \mu_{SU} = \mu_{SC}$; ó $\sigma^2_{ENTRE} = \sigma^2_{DENTRO}$
H_A: $\mu_i = \mu_j$ para al menos un par de (i, j); ó $\sigma^2_{ENTRE} > \sigma^2_{DENTRO}$
Riesgo de error de Tipo I: $\alpha = 0,05$

Factor de corrección: $C = \left(\sum_{1}^{k} \sum_{1}^{n} x_{ij} \right)^2 \Big/ N$

$$C = \frac{(245{,}8 + 254{,}6 + 427{,}8)^2}{14} = \frac{928{,}2^2}{14} = 61539{,}66$$

$SCTotal = \sum_{1}^{k} \sum_{1}^{n} x_{ij}^2 - C$ $\qquad SCTotal = 61857{,}65 - 61539{,}66 = 317{,}90$

$SCEntre = \sum_{1}^{k} \frac{\left(\sum_{1}^{n} x_{ij} \right)^2}{n_i} - C$ $\qquad SCEntre = 61811{,}84 - 61539{,}66 = 272{,}18$

$SCDentro = SCTotal - SCEntre$ $\qquad SCDentro = 317{,}90 - 272{,}18 = 45{,}72$

glT = N – 1 $\qquad$ *glE = k–1* $\qquad$ *glD = glT – glE = (N-k)*

glT = 14–1= 13 $\qquad$ *glE = 3 – 1 = 2* $\qquad$ *glD = 14 – 3 = 11*

CME = SCE/glE $\qquad$ *CME = 272,18/2 = 136,09*

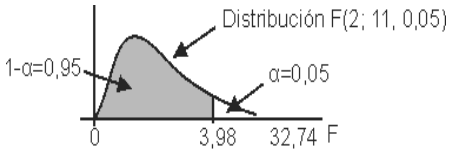
CMD = SCD/glD $\qquad$ *CMD = 45,72/11 = 4,16*

f = CME/CMD $\qquad$ *f = 136,09/4,16 = 32,74*

El valor crítico se encuentra en la Tabla 4 del Anexo, $F_{glE;\ glD;\ \alpha}$, $F_{2;\ 11;\ 0,05} = 3{,}98$

Tabla resumen de ANOVA para concentraciones de PCB's en sedimentos de fondo del arroyo Del Gato en tres sectores de la cuenca.

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Cuadrado Medio*	*f*	*F*
Entre sectores	272,18	2	136,09	32,74	3,98
Dentro de sectores	45,72	11	4,16		
Total	317,90	13			



Como $f > F_{2;\ 11;\ 0,05}$ (32,74 > 3,98), rechazo la Hipótesis nula. Es posible afirmar, con un error del 5%, que la concentración de PCB's en los sedimentos del arroyo Del Gato es diferente en los distintos sectores de la cuenca.

Comparaciones múltiples

Al hacer un análisis de varianza usando el modelo general lineal, cuando se rechaza la hipótesis nula del ANOVA se puede detectar si existen diferencias entre tratamientos, sin embargo, no se establecen dónde están las diferencias. Para detectar entre cuales tratamientos se encuentran estas diferencias se han propuesto algunas pruebas llamadas de Comparaciones Múltiples.

Aunque existen varios procedimientos, describiremos solamente la Prueba de Tukey por su sencillez.

La **prueba de Tukey**, también conocida como la prueba de Diferencia Significativa Honesta (**DSH**), es semejante a la de la prueba *t* donde se corrige el error estándar dado que, como se mencionó anteriormente, cuando se realizan comparaciones múltiples con la prueba de *t* aumenta la probabilidad de cometer error de tipo I.

Tukey propone contrastar la hipótesis nula, H_0: $\mu_A = \mu_B$ con la hipótesis alternativa, H_A: $\mu_A \neq \mu_B$, formulada para todos los pares posibles de comparaciones entre medias, donde A es la media más grande a comparar y B la más pequeña. Luego, se calcula el estadístico de prueba q_c como:

$$q_c = \frac{\overline{X_A} - \overline{X_B}}{SE}, \tag{7.15}$$

donde el error estándar, *SE*, es un valor asociado con el Cuadrado Medio Dentro (o Cuadrado Medio del Error) y al número de replicas del par de muestras que se comparan.

Para **tamaños de muestra iguales**, el error estándar se calcula como sigue:

$$SE = \sqrt{\frac{CMDentro}{ni}}\,. \tag{7.16}$$

Para **tamaños de muestra diferentes**, el error estándar se calcula con la expresión:

$$SE = \sqrt{\frac{CMDentro}{2}}\sqrt{\frac{1}{na} + \frac{1}{nb}}\,, \tag{7.17}$$

donde *na* es el tamaño de la muestra A y *nb* es el tamaño de la muestra B.

El estadístico q_c se aproxima a una distribución de $q_{k;\ N\text{-}k;\ \alpha}$ asociado al número de categorías del factor (*k*) y de los grados de libertad del CMDentro que se obtiene de la Tabla Rango Total Studentizado (Tabla 8 del Anexo). La hipótesis nula se rechaza cuando $q_c > q_{k;\ N\text{-}k;\ \alpha}$. Se suelen declarar diferencias significativas aquellas diferencias donde q_c es mayores que *q* para α **= 0,05** y diferencias altamente significativas a las que superan *q* para α **= 0,01**.

EJEMPLO 2

Prueba de Tukey para muestras de tamaño diferente

Debido a que se rechazó la hipótesis nula de la prueba de ANOVA del ejemplo de las concentraciones de PCB's en sedimentos del arroyo Del Gato, corresponde preguntarse entre que sectores de la cuenca se encuentran diferentes contenidos de PCB's en los sedimentos.

Sector / Replica	SHF	SC	SU
1	59,1	60,8	69,6
2	60,5	63,4	68,7
3	63,5	65,5	72,3
4	62,7	64,9	73,2
5			70,4
6			73,6
n_i	4	4	6
$\bar{X}$	61,45	63,65	71,30

CMDentro = 4,16

Para cada par posible de comparaciones contrastar las siguientes hipótesis:

H_0: $\mu_A = \mu_B$; Ha: $\mu_A \neq \mu_B$

$$q_c = \frac{\overline{X_A} - \overline{X_B}}{SE}$$

1°Calcular las diferencias de medias comenzando por las medias mayores $\overline{X}_A - \overline{X}_B$

2° Dado que se trata de comparaciones con tamaños de muestra diferentes $SE = \sqrt{\frac{CMDentro}{2}}\sqrt{\frac{1}{na} + \frac{1}{nb}}$

Para $n_A = 6$ y $n_B = 4$, $SE = \sqrt{\frac{4{,}16}{2}}\sqrt{\frac{1}{6} + \frac{1}{4}} = 0{,}931$

Para $n_A = 4$ y $n_B = 4$, $SE = \sqrt{\frac{0{,}056}{2}}\sqrt{\frac{1}{4} + \frac{1}{4}} = 1{,}020$

3° De la Tabla 8 del Anexo, $q_{(3;\ 11;\ 0{,}05)} = 3{,}82$ y $q_{(3;\ 11;\ 0{,}01)} = 5{,}15$

4° Armar la tabla, tomar la decisión estadística e interpretar los resultados.

Comparación (A vs. B)	*Diferencias* $\overline{X}_A - \overline{X}_B$	*SE*	q_c	$q_{(3;\ 11;\ 0{,}05)}$ / $q_{(3;\ 11;\ 0{,}01)}$	*Conclusión*
SU vs. SFH	9,85	0,931	10,58	3,82 / 5,15	Rechazar Ho: existen diferencias altamente significativas entre SU y SFH.
SU vs. SC	7,65	0,931	8,22	3,82 / 5,15	Rechazar Ho: existen diferencias altamente significativas entre SU y SC.
SC vs. SFH	2,20	1,02	2,16	3,82 / 5,15	Aceptar Ho: no existen diferencias entre SC y SHF.

SU SC SHF

Se puede afirmar, con un error de 1%, que la concentración de PCB's en sedimentos del sector urbano (SU) son diferentes a las del sector canalizado (SC) y a las del sector frutihorticola (SFH); además el sector canalizado y el frutihorticola no difieren entre sí.

Comprobación de supuestos

Discrepancias moderadas con el cumplimiento de los supuestos del ANOVA (aleatoriedad del muestreo, normalidad en las distribuciones y homogeneidad de varianzas) prácticamente no afectan las propiedades de la prueba. Sin embargo, si las diferencias son importantes se debe recurrir a otra estrategia de análisis.

Pruebas para comprobar el supuesto de normalidad

En cuanto a la normalidad, si bien es necesaria en las pruebas de hipótesis en algunos casos no es un supuesto crítico, a no ser que vaya acompañado de heterogeneidad de varianzas. Muchas pruebas de análisis de la varianza son robustas con relación a la falta de normalidad.

Para verificar el supuesto de normalidad se pueden utilizar pruebas de bondad de ajuste, como las de chi-cuadrado (χ^2), Kolmogorov-Smirnov y su modificación conocida como prueba de Lilllifors descriptas en el Capítulo 6. Además se puede examinar mediante el huso de histogramas, box plot y gráficos de probabilidad normal (normal probability plots).

Pruebas para comprobar el supuesto de homogeneidad de varianzas

Uno de los requerimientos para la aplicación de una prueba de ANOVA es el cumplimiento del supuesto de homogeneidad de varianzas entre las muestras. Esta característica es llamada homocedasticidad. Existen varias pruebas para chequear este supuesto, por su sencillez se describen la Prueba de Bartlett y la prueba F_{MAX} de Hartley.

Prueba de Bartlett

Introducida por Bartlett en 1937, la prueba se puede utilizar cuando los **tamaños de muestra** en cada tratamiento son **diferentes**. Para esta prueba se recomienda que cada tratamiento tenga al menos tres réplicas y preferentemente que éstas sean mayor a cinco. Si bien la prueba no requiere que los tamaños de las muestras sean iguales es muy sensible a alejamientos del supuesto de normalidad. Si existe evidencia fuerte de que los datos de cada población son normales, o casi normal, la prueba de Bartlett tiene un buen desempeño, pero si existe un fuerte apartamiento de la normalidad el desempeño es malo.

La hipótesis nula de la prueba postula que las varianzas de las k poblaciones muestreadas son iguales y la hipótesis alternativa que al menos una varianza poblacional es diferente ($H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$; $H_A: \sigma_i^2 \neq \sigma_j^2$ para al menos un par de (i, j)).

El estadístico de prueba se define como:

$$\chi^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(s_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)}, \tag{7.18}$$

para: k = número de tratamientos, n_i = tamaño del i-ésimo tratamiento, s_i^2 = varianza estimada para la i-ésima población a partir de una muestra de tamaño n_i, $N = n_1 + n_2 + \quad + n_k$ y

$$S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)s_i^2 \cdot \tag{7.19}$$

Este estadístico se aproxima a una distribución de χ^2 con $\nu = k - 1$ grados de libertad (χ^2_{k-1}). La hipótesis nula se rechaza cuando $\chi^2 > \chi^2_{k-1;\,\alpha}$ y se concluye que existen poblaciones con diferentes varianzas.

EJEMPLO 3

Calculo de la prueba de Homogeneidad de varianzas de Bartlett aplicado a los datos del ejemplo de las concentraciones de PCB's en sedimentos del arroyo del Gato.

	SHF	*SC*	*SU*	*Tolal*
	59,1	60,8	69,6	
	60,5	63,4	68,7	
	63,5	65,5	72,3	
	62,7	64,9	73,2	
			70,4	
			73,6	
N	4	4	6	*14*
n-1	3	3	5	
1/(n-1)	0,33	0,33	0,20	*0,87*
S^2	4,06	4,39	4,07	*12,53*
(n-1) S^2	16,25	17,56	24,43	*58,25*
lnS^2	1,40	1,48	1,40	*4,29*
(n-1) lnS^2	4,21	4,44	7,02	*15,66*

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$

$H_A : \sigma_i^2 \neq \sigma_j^2$ para al menos un par de (*i*, *j*)

$\alpha = 0{,}05$

χ^2_{k-1}, de la Tabla 2 del Anexo, $\chi^2_{3-1;0,05} = 5{,}99$

N = 14

k = 3

$$S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)s_i^2 \quad \rightarrow \quad S_p^2 = \frac{1}{14-3} \cdot 15{,}66 = 5{,}29$$

$$\ln S_p^2 = 1{,}67$$

$$\chi^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(s_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)} \quad \rightarrow \quad \chi^2 = \frac{(14-3)\,(1{,}67) - (15{,}66)}{1 + \frac{1}{3(3-1)}\left(0{,}87 - \frac{1}{14-3}\right)} = \frac{2{,}670}{5.655} = 0{,}472$$

Decisión estadística: 0,472 < 5,99 ($\chi^2 < \chi^2_{k-1,\alpha}$). Se acepta la hipótesis nula, las varianzas son iguales.

Prueba F_{max} de Hartley

Fue propuesta por Hartley (1940–1950). Se trata de una prueba muy sensible a alejamientos del supuesto de normalidad y requiere que los **tamaños de muestras** sean **iguales** ($n_1 = n_2 = \cdots = n$). Si los tamaños de muestras no son iguales, entonces la prueba ya no tiene soporte teórico fuerte y no es aplicable.

Las hipótesis que se contrastan son las mismas que las de la prueba de Bartlett ($H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$; $H_A: \sigma_i^2 \neq \sigma_j^2$ para al menos un par de (i, j)).

El estadístico de prueba *F*max es el cociente entre la mayor y la menor varianza muestral de los *k* tratamientos:

$$F\max = \frac{\max(s_i^2)}{\min(s_i^2)} \qquad (7.20)$$

donde: $i = 1, \ldots, k$, con *k* igual al número de muestras, $\max(S_i^2)$ la varianza mayor y $\min(S_i^2)$ la varianza menor de las *k* muestras.

Este estadístico calculado *F*max se aproxima a una distribución de F_{MAX} con *k* grados de libertad en el numerador y $\nu = n-1$ grados de libertad en el denominado. La hipótesis nula se rechaza cuando $F_{max} > F_{MAX}(k, n - 1, \alpha)$ (Tabla 9 del Anexo) y se concluye que las varianzas poblacionales son diferentes. Esta prueba si bien es una de las más fáciles de calcular requiere el uso de tablas especiales no siempre accesibles.

Incumplimiento de los supuestos

Los supuestos anteriores tienen por objetivo facilitar la interpretación de los resultados, tornar los métodos más simples y posibilitar la aplicación de las pruebas. Si bien la validez exacta de los supuestos es esencialmente teórica, en la práctica, cuando el alejamiento es poco el ANOVA puede realizarse igual.

Cuando los supuestos no se cumplen sus efectos varían dependiendo de la gravedad y de la situación. La homogeneidad de varianzas es el requisito necesario. Cuando se tiene heterogeneidad de varianzas se ven muy afectadas la prueba *F*, los métodos de comparaciones múltiples y los componentes de la varianza. En cuanto a la normalidad, si bien es necesaria en la pruebas de hipótesis, en algunos casos no es un supuesto critico a no ser que vaya acompañado de heterogeneidad de varianzas y muchas pruebas siguen siendo robustas ante la falta de normalidad.

La consecuencia del incumplimiento de los supuestos es que las conclusiones de los análisis realizados no sean validas puesto que los niveles de error pueden ser diferentes a los que se han elegido, pues por ejemplo los errores estándar alcanzan subestimar o sobreestimar los verdaderos errores poblacionales.

Si se observa un alejamiento grande de los supuestos se presentan dos alternativas: aplicar técnicas estadísticas que no requieran el cumplimiento de supuestos como las de la Estadística no Paramétrica, o utilizar alguna transformación en los datos que estabilice la varianza y realizar el análisis con estos datos transformados.

ANOVA Modelo I y ANOVA Modelo II

Para finalizar, se destaca que el Análisis de la Varianza es una metodología muy versátil que se aplica tanto cuando los datos se obtienen de un experimento planificado, hecho relativamente infrecuente en trabajos de investigación geológicos, como en datos obtenidos de muestreos aleatorios como los del ejemplo de concentración de PCB´s en sedimentos.

Este capítulo es sólo una introducción a los extensos métodos que comprende el Análisis de la Varianza. Se explicó cuál es el fundamento del análisis de la varianza de un factor que tiene varios niveles. En el ANOVA de un factor se distinguen dos modelos para la hipótesis alternativa. En el **Modelo I**, o de efectos fijos, la hipótesis alternativa supone que las *k* muestras son obtenidas de *k* poblaciones distintas y fijas. Es el caso que acabamos de ejemplificar donde el investigador está interesado en determinar las concentraciones de PCB' en tres sectores fijos de la cuenca del arroyo Del Gato. En el **Modelo II**, o de efectos aleatorios, la hipótesis alternativa supone que las *k* muestras, se han seleccionado aleatoriamente de un conjunto *m* mayor que *k* poblaciones. Por ejemplo existen numerosos afloramientos del Patagoniano y se muestrean solo algunos con el objeto de estudiar el tamaño que del bivalvo *Ostrea maxima*. Una manera sencilla de distinguir entre ambos modelos es pensar que, si se repitiera el estudio un tiempo después, en un Modelo I las muestras serían iguales (no los individuos que las forman) es decir corresponderían a la misma situación, mientras que en un Modelo II las muestras serían distintas. Aunque las asunciones iniciales y los propósitos de ambos modelos son diferentes, los cálculos y las pruebas de significación son los mismos y sólo difieren en la interpretación y en algunas pruebas de hipótesis suplementarias.

Además del ANOVA de un factor, el análisis de la varianza se puede aplicar para evaluar el efecto de dos factores sobre una variable dependiente. Este **ANOVA** se llama **bifactorial**. Cada factor tiene varios niveles. En el análisis bifactorial se formulan hipótesis sobre el efecto de cada factor por separado y de la interacción entre ambos factores. Por ejemplo puede plantearse la necesidad de conocer cuál es la variación de la concentración de algún ion en suelos diferentes (uno de los factores) y donde cada perfil está dividido en horizontes (el segundo factor de variación).

RELACIONES ENTRE DOS VARIABLES

Introducción

Algunos trabajos geológicos requieren conocer y modelar como son las relaciones entre variables que son medidas en una misma unidad experimental, es decir como es el patrón de variación conjunta. Son ejemplos transmisividad de un soluto y el espesor de la capa, distancia de transporte y la redondez de clastos, los volátiles y la viscosidad del magma, el caudal y las precipitaciones, la ley de Pb y la ley de Zn en una mena, y la ley de uranio y la radiactividad.

Cuando dos variables están relacionadas o asociadas varían en forma conjunta, si el valor de una aumenta el de la otra crece o decrece de manera permanente. Por otra parte si las dos variables están funcionalmente relacionadas no sólo varían simultáneamente, sino que el valor de una de ellas permite predecir el valor de la otra.

Para detectar si existe alguna relación entre dos variables ambas se miden en cada espécimen de la muestra, de este modo se obtiene una muestra de datos bivariados ($M = \{(x_1,y_1), (x_2,y_2), \ , (x_n,y_n)\}$). Con los datos se puede construir un diagrama de dispersión para visualizar si existe algún patrón de variación. Esta estrategia, sin embargo, no permite conocer si la relación (funcional o no según el caso) es significativa, pues puede suceder que por azar en la muestra las dos variables manifiesten una relación y en la población muestreada esa relación no exista. La aplicación de métodos estadísticos permite, al igual que en las situaciones semejantes ya vistas, dilucidar el problema. En este capítulo se abordarán los dos métodos principales que se utilizan para describir y cuantificar las relaciones entre dos variables: correlación y regresión lineal simple.

Correlación vs. Regresión

La correlación es una técnica exploratoria usada para examinar si los valores de las variables cambian simultáneamente en forma constante. El análisis de regresión, en cambio, describe la relación funcional entre las variables y permite predecir el valor que tomará una variable, llamada variable dependiente, conocido el valor de la otra variable, llamada independiente. En la relación funcional se

supone, de algún modo, que la variable dependiente puede ser afectada por el aumento o disminución de la variable independiente pero es imposible que se la situación inversa sea verdadera. Por ejemplo la meteorización química que sufre una roca depende de la humedad del ambiente pero es absurdo pensar que la relación inversa sea cierta. Por otra parte, la relación funcional no implica una relación causa-consecuencia sino que muchas veces los cambios son producidos por factores que inciden en la relación. Por ejemplo la viscosidad del magma se relaciona funcionalmente con la temperatura, sin embargo ambas son producidas por el contenido de sílice del magma.

Correlación

Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson[13], también conocido como coeficiente de correlación producto - momento, es una medida de de la relación lineal entre dos variables.

Las variables deben ser medidas en escalas de intervalo o de razón y deben provenir de poblaciones normalmente distribuidas. Esto significa que la distribución de frecuencia de ambas variables debe ser normal. Cuando el supuesto de distribución normal bivariada no se cumple repercute mucho en el coeficiente y el problema no se soluciona aumentando el tamaño de la muestra. Si las variables no se distribuyen normalmente se pueden transformar los datos para normalizarlos o utilizar el coeficiente de correlación no paramétrico de Spearman (Capítulo 9).

Es de uso corriente denominar a las dos variables *X* e *Y*, pero vale la pena aclarar que esta designación no implica que exista alguna relación de dependencia entre ambas. En los estudios de correlación se trata siempre de variables independientes, por ello es indistinto designar a una u otra con *X* o *Y*.

El coeficiente de correlación, *r* para la muestra y ρ para la población, se calcula con el cociente entre la covarianza de *X* e *Y* respecto al producto de sus respectivas desviaciones estándar

$$r = \frac{\sum_{i=1}^{n}(x-\bar{X})(y-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{X})^2 \cdot \sum_{i=1}^{n}(y-\bar{Y})^2}}. \qquad (8.1)$$

El numerador puede ser un numero positivo, cero o negativos en tanto el denominador es siempre un número positivo. Esto determina que *r* sea positivo, cero o negativo. Los valores de *r* positivos indican que el aumento de una de las variables va acompañado por el aumento de la otra. Valores de *r* negativos señalan que el aumento de una variable está seguido por la disminución de la otra variable. El valor 0 del coeficiente de correlación indica que no existe relación lineal pero puede haberla de otro tipo (Fig. 1).

Además el valor absoluto del numerador de la ecuación siempre es menor que el denominador, por lo que el coeficiente de correlación puede toma cualquier valor entre -1 y 1 ($-1 < r < 1$). El coeficiente no

tiene dimensiones pues las unidades se cancelan algebraicamente. Se recomienda calcular correlaciones con muestra de tamaño mayor a 12 porque cuando la cantidad de datos es demasiado pequeña los resultados son poco confiables.

El coeficiente de correlación mide la intensidad de la asociación entre las variables, no cuantifica el cambio de una respecto a la otra.

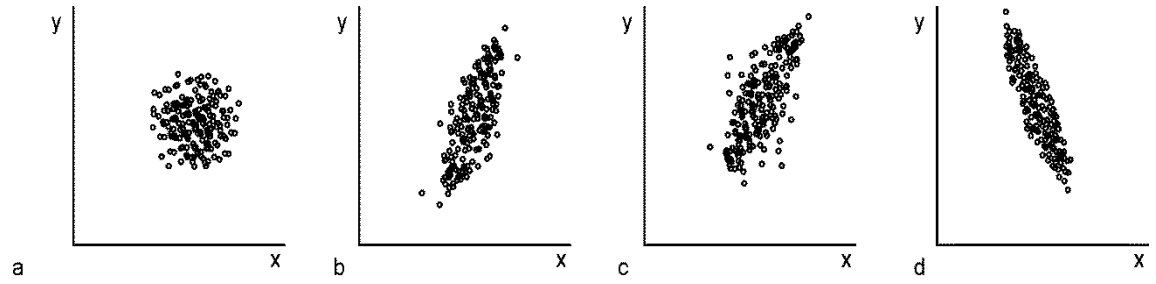


Figura 1. Gráficos de dos variables con diferente coeficiente de correlacionadas. a. ρ = 0. b. ρ = 0,95. c. ρ = 0,7. d. ρ = -0,95.

Coeficiente de Determinación ó Índice de Correlación

El Coeficiente de Determinación también llamado Índice de Correlación es el cuadrado del valor del coeficiente de correlación, r^2. Se puede interpretar como una medida de la intensidad de la relación lineal.

Pruebas de hipótesis sobre el coeficiente de correlación r

El coeficiente de correlación calculado a partir de una muestra es una estimación del coeficiente de correlación de la población (ρ). Es posible formular varias hipótesis sobre la correlación de las variables en la población.

Test de Hipótesis para ρ igual a cero

Es relevante conocer si en la población muestreada las variables están efectivamente correlacionadas. La hipótesis nula de la prueba postula que las variables no están correlacionadas y la hipótesis alternativa que sí lo están. Las hipótesis nula y alternativa se plantean respecto a cero (H_0: $\rho = 0$; H_A: $\rho \neq 0$). Recuerde que coeficientes de correlación 0 indica que las variables no están linealmente correlacionadas. Existen varias alternativas, con solo efectuar una prueba es suficiente.

Contrastar el valor absoluto de *r*, con unos valores tabulados críticos que se buscan en una tabla de valores críticos del coeficiente de Correlación *r* de Pearson con *n - 2* grados de libertad para 2 variables (Tabla 10 del Anexo).

Cuando el $r_c > r_{\alpha;\, n-2}$ se rechaza la hipótesis nula para el nivel de significación elegido para la prueba, se infiere que las variables están correlacionadas.

Contrastar la Hipótesis nula utilizando el estadístico *t* de Student que se obtiene a través de la siguiente expresión:

$$t_r = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{8.2}$$

Si $|t_r| \geq t_{(\alpha/2,\, n-2)}$, se rechaza la hipótesis nula para el nivel de significación elegido para la prueba (los valores de $t_{(\alpha/2,\, n-2)}$ se encuentran en la Tabla 3 del Anexo).

Contrastar la Hipótesis nula a través del estadístico *F* de Fisher que se calcula con la siguiente expresión:

$$F_r = \frac{1+|r|}{1-|r|} \tag{8.3}$$

Si $F_r \geq F_{(\alpha/2,\, n-1;\, n-1)}$, se rechaza la Hipótesis nula para el nivel de significación elegido para la prueba (los valores de $F_{(\alpha/2,\, n-1;\, n-1)}$ se encuentran en la Tabla 4 del Anexo).

Test de Hipótesis para cualquier otro ρ diferente de cero

Cuando se investiga si la muestra es extraída de una población donde las variables están correlacionadas de un modo especial se formula una prueba de hipótesis análoga a la prueba de *z* para una muestra (Capítulo 6). En este caso las hipótesis son H_0: $\rho = \rho_0$ y H_A: $\rho \neq \rho_0$.

La prueba requiere un paso previo, estandarizar el valor del coeficiente de correlación de la muestra (*r*) y de la población (ρ) utilizando la transformación de Fisher z_r,

$$z_r = 0{,}5\ \ln\frac{1+r}{1-r}, \tag{8.4}$$

cuya inversa es:

$$r = \frac{e^{2zr}-1}{e^{2zr}+1}. \tag{8.5}$$

Observe que aunque cuando el valor del coeficiente de correlación se encuentra entre 0 y 1, el correspondiente valor de z_r varía entre 0 e infinito, en tanto que cuando *r* se halla entre 0 y -1, el valor de z_r varía entre 0 y menos infinito.

Entonces para testear la hipótesis nula $\rho = \rho_0$, se utiliza el estadístico de prueba Z_0 normal estándar:

$$Z_0 = \frac{z_r - z_\rho}{\frac{1}{\sqrt{n-3}}} \tag{8.6}$$

Luego el Z_0 calculado se compara con $Z_{(\alpha/2)}$ (Tabla 1 del Anexo). Si $Z_0 \geq Z_{(\alpha/2)}$ se rechaza la hipótesis nula.

También se pueden realizar test a una cola (H_0: $\rho \leq \rho_0$, Ha: $\rho > \rho_0$ o H_0: $\rho \geq \rho_0$, Ha: $\rho < \rho_0$).

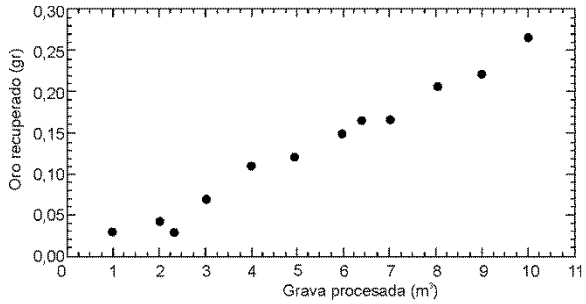
EJEMPLO 1

Coeficiente de correlación: cálculo, pruebas de hipótesis

Se realizó un muestreo de corriente para prospectar oro en la margen norte de Lago Fontana. Los datos muestran el volumen de grava procesada (m^3) y el peso de oro recuperado (gr). Interesa conocer la relación volumen procesado-gramos de oro recuperado.

	Grava procesada (m^3)	*Oro recuperado (gr)*
	1	0,025
	2	0,042
	2,5	0,038
	3	0,071
	4	0,103
	5	0,111
	6	0,142
	6,5	0,156
	7	0,164
	8	0,191
	9	0,220
	10	0,258
$\bar{X}$	*5,3*	*0,1*
S	*2,9*	*0,1*
S^2	*8,4*	*0,0*
$\sum x^2$	*433,5*	*0,3*
$\sum x$	*64,0*	*1,5*
$\left(\sum x\right)^2$	*4096,0*	*2,3*

1° Se realiza un gráfico bivariado con el objeto de explorar como es la relación entre las variables. La relación volumen procesado-gramos de oro recuperado es lineal.



2° Se calcula el coeficiente de correlación.

$$r = \frac{\sum_{i=1}^{n}(x-\bar{X})(y-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{X})^2 \cdot \sum_{i=1}^{n}(y-\bar{Y})^2}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

$$r = \frac{10{,}5 - \frac{64 \cdot 1{,}5}{12}}{\sqrt{\left(433{,}5 - \frac{4096}{12}\right)\left(0{,}3 - \frac{2{,}3}{12}\right)}} = 0{,}9942$$

3° Se realiza la prueba de hipótesis *t* de Student para saber si en la población de donde se extrajo la muestra las variables están correlacionadas.

r = 0,9942

H_0: $\rho = 0$
H_A: $\rho \neq 0$
$\alpha = 0{,}05$
$n = 12$
De la Tabla 3 del Anexo, $t_{(\alpha/2,\ n-2)} \Rightarrow t_{(0{,}025;10)} = 2{,}228$

$$t_r = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0{,}9942}{\sqrt{\frac{1-0{,}9942^2}{12-2}}} = 29{,}233$$

El $|t_r| \geq t_{(\alpha/2,\ n-2)}$ (29,333 ≥ 2,228) esto conduce al rechazo de la hipótesis nula. Se infiere que el volumen de grava procesado está correlacionado con el peso de oro recuperado.

4° El coeficiente de determinación es $r^2 = 0{,}9884$. El 98,84% de la variabilidad del peso de oro se puede explicar por el volumen de grava procesada.

5° Interesa conocer si la correlación entre el volumen de grava procesada y el peso de oro recuperado de la margen norte del Lago Fontana es mayor que la que se encuentra a nivel global que es de 0,90.
Se plantean las hipótesis de la prueba, se elige el nivel de significación y el valor crítico para tomar la decisión estadística:

$r = 0{,}9942$
H_0: $\rho < 0{,}90$
H_A: $\rho \geq 0{,}90$
$\alpha = 0{,}05$
De la Tabla 1 del Anexo, $Z_{0{,}05} = 1{,}645$

$$z_r = 0{,}5\ \ln\frac{1+r}{1-r} \qquad z_{0{,}9942} = 0{,}5\ \ln\frac{1+0{,}9942}{1-0{,}9942} = 2{,}92 \qquad z_{0{,}9} = 0{,}5\ \ln\frac{1+0{,}9}{1-0{,}9} = 1{,}47$$

$$Z_0 = \frac{z_r - z_\rho}{\frac{1}{\sqrt{n-3}}} = \frac{2{,}92-1{,}47}{\frac{1}{\sqrt{12-3}}} = 4{,}35$$

Se rechaza la Hipótesis nula pues $Z_0 \geq Z_{(\alpha)}$ (4,35 > 1,645). Se infiere que la correlación entre el volumen de grava y el peso de oro en la margen norte del Lago Fontana es mayor que la correlación entre ambas variables a nivel global.

Limites de confianza para el coeficiente de correlación poblacional

Para calcular los límites de confianza para el coeficiente de correlación poblacional ρ se procede de forma similar que con el cálculo de límites de otros parámetros. En este caso se calculan los límites de z_r cuyo parámetro poblacional ζ. Para la transformación de r a Z_r se utiliza la fórmula 8.4. De este modo se tiene

$$P(Z_r - Z_{(\alpha/2)} < \zeta < Z_r + Z_{(\alpha/2)}) = 1 - \alpha. \qquad (8.7)$$

Observe que se han calculado valores límites de Z_r, pero se buscan valores límites del coeficiente de correlación por lo que debe ahora transformar los valores de Z_r, en r utilizando la inversa de la transformación de Fisher (expresión 8.5).

Los límites inferior y superior de ρ serán $r_{inf} = \frac{e^{2zinf}-1}{e^{2zinf}+1}$ y $r_{sup} = \frac{e^{2zsup}-1}{e^{2zsup}+1}$.

EJEMPLO 2

Límites de confianza del coeficiente de correlación

Se utilizan los datos del ejemplo 1 del Lago Fontana para calcular los límites de confianza de *r*.

$\alpha = 0,1$
De la Tabla 1 del Anexo, $Z_{(\alpha/2)} = 1,645$

$z_r = 0,5 \ln\frac{1+r}{1-r}$ $\quad z_{0,9942} = 0,5 \ln\frac{1+0,9942}{1-0,9942} = 2,92$

$P(Z_r - Z_{(\alpha/2)} < \zeta < Z_r + Z_{(\alpha/2)}) = 1 - \alpha$
$P(2,92 - 1,645 < \zeta < 2,92 + 1,645) = 0,90$
$P(1,275 < \zeta < 4,565) = 0,90$

$r_{inf} = \frac{e^{2zinf}-1}{e^{2zinf}+1} = \frac{e^{2\cdot1,275}-1}{e^{2\cdot1,275}+1} = 0,8551$ $\qquad$ $r_{sup} = \frac{e^{2zsup}-1}{e^{2zsup}+1} = \frac{e^{2\cdot4,565}-1}{e^{2\cdot4,565}+1} = 0,9998$

$P(0,8551 < \rho < 0,9998) = 0,90$

Comparación de dos coeficientes de correlación

Se pueden testear hipótesis (a una o dos colas) sobre dos coeficientes de correlación. Para ello se recurre nuevamente a la transformación de Fisher (expresión 8.4). El estadístico de prueba es:

$$Z_0 = \frac{z_{r1}-z_{r2}}{\sqrt{\frac{1}{n1-3}+\frac{1}{n2-3}}}. \qquad (8.8)$$

En el caso de una prueba a dos colas las hipótesis son H_0: $\rho_1 = \rho_2$ y H_A: $\rho_1 \neq \rho_2$, luego el Z_0 calculado se compara con $Z_{(\alpha/2)}$. Si $Z_0 \geq Z_{(\alpha/2)}$ se rechaza la hipótesis nula.

Si la hipótesis nula se acepta se concluye que ambas muestras provienen de una población que tiene idéntico coeficiente de correlación. En estos casos se debe combinar la información de las dos muestras calculando un coeficiente de correlación único que estime mejor al coeficiente de correlación poblacional ρ. Se procede de la siguiente forma:

$$z_{r\ ponderado} = \frac{(n1-3)z_{r1}+(n2-3)z_{r2}}{(n1-3)+(n2-3)} \qquad (8.9)$$

Se ha calculado un valor de $Z_{r\ ponderado}$, por lo que para obtener el valor de ρ se debe aplicar la inversa de la transformación de Fisher (expresión 8.5).

EJEMPLO 3

Comparación de dos coeficientes de correlación

Continuando la prospección de oro en Lago Fontana se realizó un muestreo de corriente en la margen sur del lago. Se tomó una muestra de tamaño 20. El coeficiente de correlación del volumen de grava

procesado y el peso de oro recuperado fue r_S = 0,9687. Interesa conocer si ambas márgenes presentan características similares.

H_0: $\rho_N = \rho_S$
H_A: $\rho_N \neq \rho_S$
α = 0,1
$Z_{(\alpha/2)}$ = 1,645
n_N = 12
n_S = 20
r_N = 0,9942
r_S = 0,9687
$z_r = 0{,}5 \ln\frac{1+r}{1-r}$

$z_{0{,}9942} = 0{,}5 \ln\frac{1+0{,}9942}{1-0{,}9942} = 2{,}92$ $\qquad$ $z_{0{,}9687} = 0{,}5 \ln\frac{1+0{,}9687}{1-0{,}9687} = 2{,}07$

$Z_0 = \frac{z_{r1}-z_{r2}}{\sqrt{\frac{1}{n1-3}+\frac{1}{n2-3}}} = \frac{2{,}92-2{,}07}{\sqrt{\frac{1}{12-3}+\frac{1}{20-3}}} = 4{,}488$

El $Z_0 \geq Z_{(\alpha/2)}$ (4,488 > 1,645), se rechaza la hipótesis nula, se infiere que la correlación entre el volumen de grava y el peso de oro de ambas márgenes es igual. Corresponde hallar el coeficiente de correlación para el lago.

$$z_{r\ ponderado} = \frac{(n1-3)z_{r1} + (n2-3)z_{r2}}{(n1-3)+(n2-3)} = \frac{(12-3)2{,}92 + (20-3)2{,}07}{(12-3)+(20-3)} = 2{,}364$$

$$r_{ponderado} = \frac{e^{2zpon}-1}{e^{2zpon}+1} = \frac{e^{2\cdot 2{,}364}-1}{e^{2\cdot 2{,}364}+1} = 0{,}8551$$

El coeficiente de correlación entre el volumen de gravas y el peso de oro para el Lago Fontana es 0,8551.

Corrección del nivel de significación (α) para comparaciones múltiples

Es importante recordar que si se calculan todos los coeficientes de correlación de a pares cuando hay *m* variables, el número de correlaciones que se tienen es *m(m - 1)/2*. Por ejemplo, si *m* = 10, entonces se pueden calcular 45 correlaciones diferentes. En estos casos se espera que entre 5% y 10% (para el ejemplo, 2 a 5) de las correlaciones sean diferentes de 0 solamente por cuestiones de azar, es decir son falsos positivos.

Para evitar aceptar correlaciones debidas al azar, Bonferroni[14] propone reducir el nivel de significación de todas las pruebas de hipótesis sobre el coeficiente de correlación a α/m. Aunque usar la corrección Bonferroni controla la probabilidad de falsos positivos, lo hace a costa de incrementar los falsos negativos. A esto se suma que en ciertas situaciones se desea aceptar la hipótesis nula de la prueba, hecho que la corrección no contempla. Una alternativa para evitar ambos problemas es usar la corrección Šidák (o Dunn- Šidák) que es $1-(1-\propto)^{1/2}$.

Exactitud, Precisión y Correlación

Los diagramas de dispersión que muestran relaciones entre datos observados y valores estimados suministran información tanto de la exactitud como de la precisión. Por ejemplo, en un yacimiento es posible conocer la ley de los bloques explotados y las leyes estimadas por laboratorio a partir de un muestreo de los bloques. Si las estimaciones son exactas, se espera que los datos se alineen en torno a los 45° (equivalente a $Y = X$). En cambio, cuando las estimaciones no son perfectas los puntos no se disponen a 45°, existe un sesgo que se interpreta un problema de exactitud. Por otra parte si los puntos están dispersos, el ancho de la dispersión representa la variabilidad de los datos y va unido a la precisión del estimador o del laboratorio. La incertidumbre total es la suma de precisión más exactitud.

Correlación espuria en datos composicionales. El problema de la suma constante

Una de las prácticas habituales en el tratamiento de datos geoquímicos, mineralógicos y petrográficos (ígneos, sedimentarios y metamórficos) es su transformación de forma que para un espécimen o individuo la suma de todas las componentes individuales sea constante (i.e. 100%). Datos expresados como parte del todo (proporciones, porcentajes y partes por millón –ppm-) se conocen como **datos composicionales**. Este es un mecanismo sencillo que permite realizar comparaciones entre muestras, sin embargo puede conducir a resultados espurios e inducir a interpretaciones erróneas de los datos si no se analiza con los métodos estadísticos apropiados.

Los porcentajes y proporciones son razones numéricas complejas que contienen variables en su denominador que representan todos los constituyentes a ser examinados. Esto trae aparejado que los componentes de porcentajes no sean libres de variar independientemente. A medida que la proporción de un componente aumenta, la proporción de uno o más de los otros componentes debe decrecer. Por ejemplo si se analiza el quimismo de una roca y el contenido en sílice fuera 61,5%, entonces el contenido de alúmina no podrá ser cualquier valor, estará restringido a ser igual ó menor a (100 – 61,5) %. El siguiente óxido que forme parte de esta roca se verá también restringido por el contenido de los dos óxidos anteriores. Una de los problemas que se producen al analizar datos composicionales es que se introduce un sesgo negativo en las correlaciones. En el Capítulo 12 se abordarán los métodos estadísticos para analizar datos composicionales.

Regresión lineal simple

El análisis de regresión puede ser lineal o curvilíneo. A su vez el análisis lineal puede ser simple, parcial o múltiple y el curvilíneo simple o múltiple (potencial, exponencial o logarítmico). En el análisis de regresión se utiliza *X* para designar a la variable independiente e *Y* para referirse a la variable dependiente. La relación más simple que existe entre una variable dependiente y una independiente es la lineal (sólo se aborda esta en el libro). Solamente se necesitan dos estadísticos para describir la posición de la recta, el valor de *Y* cuando *X* es cero, llamado ordenada al origen y simbolizada con a y la pendiente de la recta, representada con b. Se puede ubicar cualquier punto sobre la recta con la ecuación:

$$Y_{ij} = a + b\,X_{ij} \tag{8.10}$$

El análisis de regresión consiste en hallar la recta que mejor describe la relación entre dos variables y realizar las pruebas estadísticas para corroborar que esa relación existe en la población muestreada (i.e. la pendiente es diferente de cero).

Cálculo de la ecuación de la recta

Suponga que se realiza un muestreo de una población donde las dos variables tienen una dependencia funcional que es lineal, cuando se grafican los datos de la muestra en un gráfico de dispersión es de esperar que los puntos se encuentren alineados (Fig. 2). Se puede calcular la recta que mejor ajusta a los datos con el método de mínimos cuadrados.

El método de ajuste de mínimos cuadrados considera la desviación vertical entre cada punto y la recta (i.e. desviación es $Y - \hat{Y}$, donde $\hat{Y}$ es el valor estimado por la recta de regresión) y define a la recta que mejor ajusta aquella donde el cuadrado de las desviaciones de todos los valores es mínimo, $[\sum_1^n (Y - \hat{Y})^2 = mínimo]$. La suma de cuadrados de estas desviaciones se llama Suma de Cuadrados de las desviaciones como se verá luego.

Antes de calcular los coeficientes de la recta conviene recordar que se define:

$$Suma\ de\ Cuadrados\ de\ X = SCx = \sum x^2 - \frac{(\sum x)^2}{n} \tag{8.11}$$

$$Suma\ de\ Cuadrados\ de\ Y = SCy = \sum y^2 - \frac{(\sum y)^2}{n} \tag{8.12}$$

$$Suma\ de\ Cuadrados\ de\ XY = SCxy = \sum xy - \frac{\sum x \sum y}{n} \tag{8.13}$$

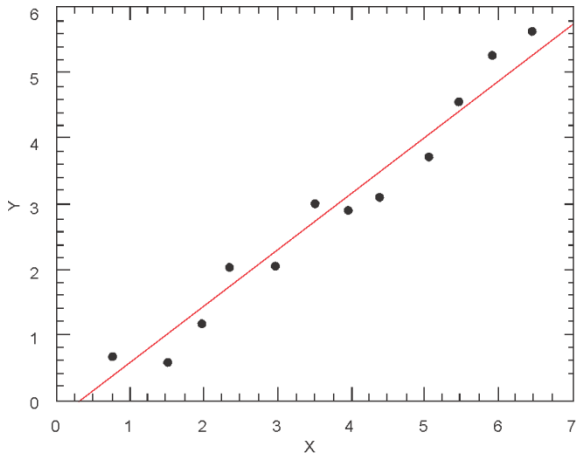


Figura 2. Los datos de esta figura muestran la relación lineal de Y en función de X.

Coeficiente de regresión

La expresión 8.14 se utiliza para calcular el coeficiente de regresión que es la pendiente de la recta.

$$b = \frac{SCXY}{SCX} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \qquad (8.14)$$

ó

$$b = r\frac{S_y}{S_x} \qquad (8.15)$$

El numerador de *b* puede ser positivo, cero o negativo, el denominador es siempre positivo, y el valor de la pendiente puede estar, teóricamente, entre menos infinito y más infinito. Pendientes positivas indican que al aumentar los valores de la variable también aumenta el valor de la variable dependiente. Pendientes negativas señalan que los incrementos de la variable independiente van acompañados por decrecimientos de la variable dependiente. Cuando la recta es paralela al eje de las X, la pendiente es cero lo que muestra que los cambios de la variable independiente no se relaciona con las variaciones de la variable dependiente. La pendiente se expresa en unidades de *X* sobre unidades de *Y*. El coeficiente de regresión expresa el cambio que se produce en *Y* con el cambio de una unidad de *X* (Fig. 3). Cuanto mayor es el valor absoluto de la pendiente, mayor es la magnitud del cambio de *Y* por cada unidad de cambio de *X*.

Ordenada al origen

La ordenada al origen es el valor de la variable dependiente correspondiente al valor cero de la variable independiente. Matemáticamente se prueba que el punto $\bar{X}$ e $\bar{Y}$ se encuentran en la recta de regresión que mejor se ajusta a los datos, entonces se puede sustituir ambos los valores $X_{i,j}$ e $Y_{i,j}$ en la

ecuación de la recta (expresión 8.10) por los valores medios de modo que se tiene $\bar{Y} = a + b\,\bar{X}$, y simplemente obtener la ordenada al origen como

$$a = \bar{Y} + b\,\bar{X}. \qquad (8.16)$$

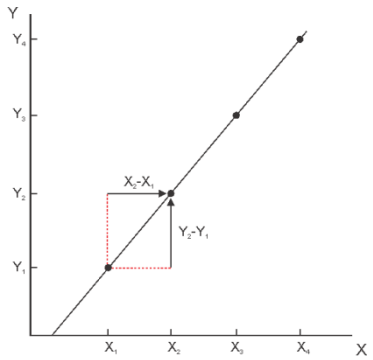


Figura 3. Por convención se utiliza la ordenada para la variable dependiente Y, y la abscisa para la variable independiente X. Cálculo de la pendiente de la recta que mejor ajusta a un conjunto de datos. La flecha vertical indica el cambio relativo de Y_1 a Y_2 que se produce con el incremento en X desde X_1 a X_2. Para cualquier par de puntos, el cociente $(Y_2 - Y_1)/(X_2 - X_1)$ es la pendiente de la recta, en este caso es positiva porque Y aumenta cuando X aumenta.

Estimación de la variable dependiente a partir de la variable independiente

Conocer la pendiente y la ordenada al origen de la recta que mejor ajusta permite estimar cualquier valor de la variable dependiente para cualquier valor de *X* usando la ecuación 8.10.

Para dibujar la recta de regresión solo se necesitan conocer dos valores de Y, se recomienda que esos dos valores sean la ordenada al origen y el que corresponde a $\bar{X}$.

Por otra parte, es aconsejable no extrapolar las estimaciones fuera del rango de la variable independiente que se utiliza para calcular la recta pues la función puede cambiar fuera de ese rango. Por ejemplo el gradiente geotérmico tiene un comportamiento lineal de 33 °C/km los primeros kilómetros, pero el gradiente cambia a medida que la profundidad se aproxima al manto.

EJEMPLO 4

Cálculo de la recta de regresión

El costo de producción de una empresa minera está integrado por varias variables (mano de obra directa, gastos de transformación, materiales consumibles como explosivos, etc.). En este ejemplo el problema se simplifica y se proponen un conjunto de datos hipotéticos de producción (X) y costo total (Y) mensual de una empresa minera. Producción y costo están multiplicados por contantes para facilitar los cálculos.

Producción t 10 (X)	*Costo $\$10^6$(Y)*
0,94	3,75
1,25	4,65
1,88	5,70
0,94	3,15
2,19	6,45
1,88	4,80
2,81	7,20

2,50	5,85
2,50	6,60
1,56	4,50
1,63	5,10
2,63	6,15

n = 12

$\bar{X} = 1{,}89$ $\bar{Y} = 5{,}33$

$\sum x = 22{,}69$ $\sum y = 63{,}90$

$S_X = 0{,}65$ $S_Y = 1{,}21$

$S_X^2 = 0{,}42$ $S_Y^2 = 1{,}47$

$\sum x^2 = 42{,}52$ $\sum y^2 = 356{,}45$

$(\sum x)^2 = 514{,}72$ $(\sum y)^2 = 4083{,}45$

$\sum xy = 128{,}92$

$$SCx = \sum x^2 - \frac{(\sum x)^2}{n} = 45{,}52 - \frac{514{,}72}{12} = 4{,}63$$

$$SCy = \sum y^2 - \frac{(\sum y)^2}{n} = 356{,}45 - \frac{4083{,}45}{12} = 16{,}18$$

$$SCxy = \sum xy - \frac{\sum x \sum y}{n} = 128{,}92 - \frac{22{,}69 \cdot 63{,}9}{12} = 8{,}10$$

a. Cálculo del coeficiente de regresión (pendiente *b* de la recta)

$$b = \frac{SCxy}{SCx} = \frac{8{,}10}{4{,}63} = 1{,}75\,\$/t$$

b. Cálculo de la ordenada al origen, *a*

$$a = \bar{Y} - b\,\bar{X} = 5{,}33 - 1{,}75 \cdot 1{,}89 = 2{,}01\,\$$$

c. La ecuación de la recta es

$$Y = 2{,}01\,\$ + 1{,}75\,\$/t \cdot X\ t$$

d. Estimación del costo de extracción de 2 t de material

x = 2t

$$\hat{y} = 2{,}01\,\$ + 1{,}75\,\$/t \cdot 2\,t$$

$$\hat{y} = 7{,}51\,\$$$

Supuestos de la regresión

Para validar las hipótesis sobre los parámetros poblacionales o para hallar los intervalos de confianza de la regresión se requiere el cumplimiento de los siguientes supuestos:

1° La variable dependiente es determinada por la variable independiente.

2° La relación entre las variables es lineal. Esto se corrobora con un gráfico de dispersión.

3° Para cada valor de *X* existe una población de valores de *Y* que están normalmente distribuidos.

4° Las varianzas de estas poblaciones de *Y* son iguales entre sí. Esto se puede comprobar graficando los residuos. El residuo es la diferencia entre el valor observado y el estimado por la recta ($Y - \hat{Y}$). En

un gráfico de dispersión se grafican los residuos para cada valor de *X*. Si hay homogeneidad de varianzas los residuos se distribuyen de manera uniforme en una línea horizontal en torno al cero pero cuando hay heterocedasticidad el ancho de la banda cambia a medida que *X* varía (Fig. 4).

5° El error de *Y* es aditivo.

6° Los valores de *Y* son independientes.

7° La variable independiente *X* se mide sin error. Si bien esto es impracticable, el error debe ser despreciable o al menos pequeño comparado con el error de medición de la variable dependiente.

Alejamientos pequeños de estas asunciones no invalidan las estimaciones. Si la relación entre las variables no es lineal se pueden transformar los datos antes de hacer el análisis de regresión.

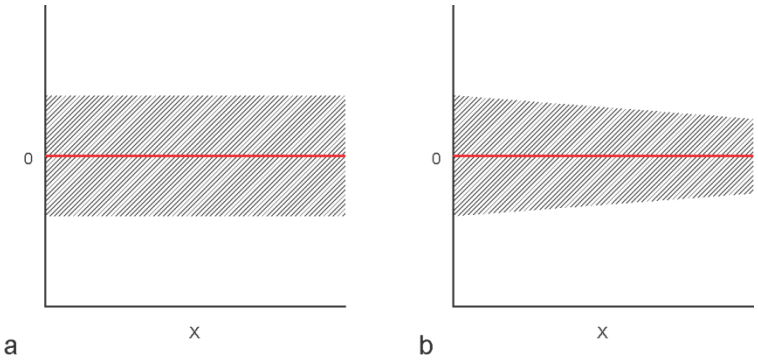


Figura 4. Gráfico de residuos: a) Homocedasticidad, b) Heterocedasticidad, en este caso, la varianza de los residuos decrece al aumentar X.

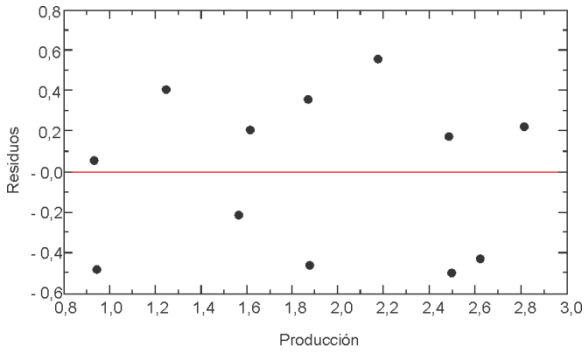
EJEMPLO 5

Comprobación del supuesto de homocedasticidad con el Método de los residuos

Se utilizan los datos del ejemplo 4 producción-costo de la empresa minera para las poblaciones de la variable dependiente que corresponde a cada valor observado.

El pronóstico de Costo se calcula con $y = 2{,}01\,\$ + 1{,}75\,\$/t \cdot x\,t$, para cada valor de producción observado. El residuo es $Y - \hat{Y}$.

Costo Y	*Costo estimado* $\hat{Y}$	*Residuos*
3,75	3,66	0,09
4,65	4,20	0,45
5,70	5,30	0,40
3,15	3,66	-0,51
6,45	5,85	0,60
4,80	5,30	-0,50
7,20	6,94	0,26
5,85	6,39	-0,54
6,60	6,39	0,21
4,50	4,75	-0,25
5,10	4,86	0,24
6,15	6,61	-0,46



Los puntos forman una franja horizontal con respecto a cero, entonces el modelo es aceptable.

Pruebas de Hipótesis sobre el coeficiente de regresión (pendiente)

Cuando la pendiente de la recta de regresión se calcula a partir de los datos de una muestra no informa sobre la relación entre las variables en la población. Puede suceder que aunque *b* sea diferente de cero se halla muestreado una población donde la relación de dependencia entre las variables no exista y que este hallazgo sea solamente una cuestión de azar (Fig. 5).

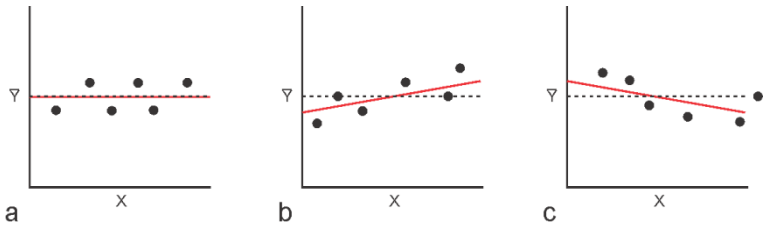


Figura 5. Tres situaciones donde X e Y no tienen relación de dependencia. Con línea roja recta de regresión de la muestra, línea cortada, regresión de la población. a) La recta de regresión de la muestra y de la población tienen pendiente 0. b) La recta de regresión de la muestra tiene suave pendiente positiva y la de la población tienen pendiente 0. c) La recta de regresión de la muestra tiene suave pendiente negativa y la de la población tienen pendiente 0.

Para conocer si en la población muestreada las variables tienen una dependencia lineal se debe realizar una prueba de hipótesis sobre el coeficiente de regresión poblacional (β). Se plantean la hipótesis nula H_0: $\beta = 0$ y la hipótesis alternativa H_A: $\beta \neq 0$. Fijado el nivel de significación α de la prueba, si la hipótesis nula se acepta existen evidencias para sostener, con probabilidad de error $1 - \alpha$, que la muestra es tomada de una población donde las variables no tienen una relación de dependencia. En tanto si la hipótesis nula no se rechaza, existen evidencias para sostener, con probabilidad de error $1 - \alpha$, que la muestra es tomada de una población donde las variables tienen una relación de dependencia. Estas hipótesis se pueden testear utilizando Análisis de la Variancia de la regresión o utilizando una prueba de *t* de Student.

ANOVA de la regresión

El análisis de la varianza de la regresión permite testear las hipótesis H_0: $\beta = 0$ y H_A: $\beta \neq 0$. Se trata de un análisis análogo al descripto en el Capítulo 7. En la figura 6 se representan las fuentes de variación del análisis. Para el dato *X-Y*, la **Variación Total** es la distancia vertical entre *Y* y la media de la variable dependiente $(Y - \bar{Y})$. La variación total se puede dividir en dos: a) la **Variación Explicada** por la recta de Regresión, está representada por la distancia entre el valor estimado por la recta y la media $(\hat{Y} - \bar{Y})$ y b) la variación no explicada por la regresión llamada Error o **Variación Residual**, es la distancia entre el valor observado y el estimado por la recta $(Y - \hat{Y})$.

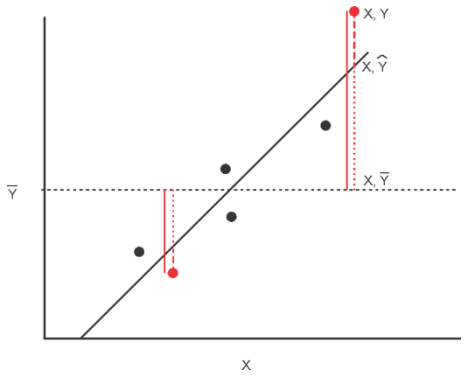


Figura 6. ANOVA de la regresión. La figura muestra la recta de regresión para seis datos con línea entera y la media de Y con línea de trazos. La línea roja entera representa la variación total, es la distancia del punto X,Y a la media $\bar{Y}$ $(Y - \bar{Y})$. *La línea roja de puntos representa la variación explicada por la recta de regresión* $(\hat{Y} - \bar{Y})$. *La línea roja entrecortada representa el Error, la variación no explicada por la recta de regresión* $(Y - \hat{Y})$.

Las sumas de cuadrados de las dos fuentes de variación del análisis se obtienen sumando para todos los pares de valores *X,Y* el cuadrado de esas diferencias. Los **grados de libertad** totales son *n - 1*, los del error, *n - 2* y el de la regresión[15] es 1. Los cuadrados medios se obtienen calculando los cocientes entre suma de cuadrados y grados de libertad respectivos. El estadístico de prueba *F* se consigue con el cociente entre el **Cuadrado medio de la regresión** (CMR) y el **Cuadrado medio del error** (CME). El Cuadro 1 muestra las fórmulas para realizar los cálculos.

Si la recta tiene pendiente cercana a cero o es cero, el CMR y el CME son muy semejantes, entonces $F \approx 1$. A medida que la pendiente aumenta la varianza explicada por la regresión y disminuye la varianza no explicada, $F > 1$. La decisión estadística se toma contrastando *F* calculado con $F_{(\alpha,,glR,\ glE)}$ (Tabla 4 del Anexo).

Los cuadrados medios, al igual que en el ANOVA se calculan como cocientes entre sumas de cuadrados y los grados de libertad.

La **Suma de Cuadrados Total** es $SCT = SCY = \sum y^2 - \frac{(\sum y)^2}{n}$.

Los grados de libertad total glTotal = *n* - 2

La **Suma de Cuadrados de la Regresión**:

$$SCRegresíon = \frac{(SCXY)^2}{SCX} = \frac{\left(\sum xy - \frac{\sum x \sum y}{n}\right)^2}{\sum xy - \frac{\sum x \sum y}{n}}.$$

Los grados de libertad total glRegresión = 1

La **Suma de Cuadrados del Error** o **Residual**: $SCError = SCT - SCRegresión.$

Los grados de libertad total glError = n - 2

Fuente de variación	**Suma de cuadrados**	**Grados de libertad (gl)**	**Cuadrados medios**	**F_c**
Regresión	$SCR = \frac{(SCXY)^2}{SCX} = b\,SCxy$ (8.17)	$glR = 1$	$CMR = SCR$ (8.19)	$\frac{CMR}{CME}$ (8.21)
Error o Residual	$SCE = SCT - SCR$ (8.18)	$glE = n - 2$	$CME = \frac{SCE}{glE}$ (8.20)	
Total	$SCT = SCy$	$glT = n - 2$	$CMT = \frac{SCT}{glT}$	

Cuadro 1. Tabla resumen de ANOVA de la regresión con fórmulas que se utilizan en el análisis.

EJEMPLO 6

Cálculo de ANOVA de la Regresión

Se utilizan los datos del ejemplo 4 de producción-costo de la empresa minera.

H_0: $\beta = 0$

H_A: $\beta \neq 0$

$n = 12$

$\alpha = 0{,}05$

De la Tabla 4 del Anexo, $F_{(\alpha;\ 1;\ 12\text{-}1)} = 4{,}84$

$SCy = 16{,}18$

$SCx = 4{,}63$

$SCxy = 8{,}10$

$SCT = SCy = 16{,}18$

$$SCR = \frac{(SCxy)^2}{SCx} = \frac{8{,}10^2}{4{,}63} = 14{,}20$$

$$SCError = SCT - SCR = 16{,}18 - 14{,}20 = 1{,}98$$

gl Total = 12 - 1 = 11; gl Regresión = 1 gl Error = 12 – 2 = 10

$$CMR = SCR = 14{,}20$$

$$CME = \frac{SCE}{glE} = \frac{1{,}98}{10} = 0{,}2$$

$$Fc = \frac{CMR}{CME} = \frac{14{,}2}{0{,}2} = 71{,}78$$

Fuentes de variación	*Suma de Cuadrados*	*Grados de libertad*	*Cuadrado medio*	*Fc*
Total	16,18	11		
Regresión	14,20	1	14,20	71,78
Residual	1,98	10	0,20	

Fc > $F_{(0{,}05;\ 1;\ 11)}$ (71,78 > 4,84) se rechaza la Hipótesis nula. En la población muestreada el coeficiente de regresión $\beta \neq 0$ (p > 0,05). Existe dependencia del costo con la producción de la empresa minera.

Error estándar de estimación

El Cuadrado Medio del Error o Residual a veces se indica como $S^2_{Y \cdot X}$, representa la varianza de *Y* después de tomar en cuenta la dependencia de *Y* sobre *X*. La raíz cuadrada de la varianza residual es el error estándar de estimación $S_{Y \cdot X}$.

El error estándar de estimación indica la precisión con la cual la recta de regresión predice la dependencia de *Y* por *X*. La ecuación de la recta se escribe entonces como

$$\hat{Y} = a + bX \mp S_{Y \cdot X}. \tag{8.22}$$

En el ejemplo producción-costo de la empresa minera $S_{Y \cdot X} = 1{,}4$. La ecuación de la recta es para este caso $\hat{y} = 2{,}01\,\$ + 1{,}75\,\$/t \cdot 2\,t \pm 1{,}4$.

Coeficiente de determinación

El coeficiente de determinación, r^2, como se indicó más arriba, sirve como medida de la linealidad de la regresión. También es un estadístico que indica la proporción o porcentaje de la variación total de los valores de la variable dependiente *Y* respecto a la media ($\bar{Y}$) que es explicada por la línea de regresión. r^2 se calcula como

$$r^2 = \frac{Suma\ de\ Cuadrados\ de\ la\ Regresión}{Suma\ de\ Cuadrados\ Total}, \tag{8.23}$$

y toma valores entre 0 y 1. Cuando todos los puntos se ubican sobre una recta de regresión cuya pendiente es diferente a cero, toda la variabilidad es explicada por la recta, numerador y denominador de la expresión 8.23 son iguales y r^2 es 1. Si la suma de cuadrados explicada es menor que la no explicada, r^2 es menor a uno. Cuando el coeficiente de determinación es pequeño se concluye que la relación lineal no es un buen ajuste para los datos, ya sea porque hay poca relación entre las variables o bien a que la relación subyacente no es lineal.

La contraparte del coeficiente de determinación es el **coeficiente de indeterminación**. Se trata de la proporción o porcentaje de varianza no explicada por la regresión, es $1 - r^2$.

En el ejemplo producción-costo de la empresa minera $r^2 = 0{,}8777$, la recta calculada ajusta bien a los datos, sólo el 12% de la variabilidad de los datos no se explica con el modelo.

Test de t sobre β (Coeficiente de regresión)

La prueba de *t* es una prueba alternativa que se puede realizar sobre el coeficiente de regresión para testear la hipótesis de β = 0. Sin embargo es la única para someter a prueba hipótesis sobre β diferente de cero (H_0: $\beta \leq 0$ y H_A: $\beta > 0$; H_0: $\beta \geq 0$ y H_A: $\beta < 0$).

Al igual que en otras pruebas de *t* el estadístico se calcula como la diferencia entre el estadístico y el parámetro hipotetizado sobre el error estándar del estadístico (Capítulo 6). Para este caso

$$t_b = \frac{b-\beta}{\sqrt{\frac{Cuadrado\ Medio\ Residual}{Suma\ de\ cuadrados\ de\ X}}} \tag{8.24}$$

Si $|t_b| \geq t_{(\alpha,\ n-2)}$, se rechaza la Hipótesis nula para el nivel de significación elegido para la prueba.

Límites de confianza en la regresión

Los intervalos de confianza de la regresión se calculan en forma análoga a otros intervalos de confianza sumando y restando al estadístico el producto del error estándar del estadístico y el valor critico de *t* para el error elegido para la prueba (Capítulo 6).

Límites de confianza para el coeficiente de regresión

Para el coeficiente de regresión los límites de confianza de β se calcula con la siguiente expresión:

$$P\left(b - t_{\frac{\alpha}{2};\nu}\sqrt{\frac{Cuadrado\ Medio\ Error}{Suma\ de\ Cuadrados\ de\ X}} < \beta < b + t_{\frac{\alpha}{2};\nu}\sqrt{\frac{Cuadrado\ Medio\ Error}{Suma\ de\ Cuadrados\ de\ X}}\right) = 1 - \alpha \tag{8.25}$$

para $\nu = n - 2$.

La figura 7 muestra los límites de confianza para la pendiente de la línea de regresión. Dentro de estos límites los posibles valores de *b* basculan alrededor del punto $(\bar{X}, \bar{Y})$.

EJEMPLO 7

Límites de confianza del coeficiente de regresión

Se utilizan los datos del ejemplo 4 de producción-costo de la empresa minera.

$n = 10$

$b = 1{,}75$

$SCx = 4{,}63$

$$CME = \frac{SCE}{glE} = \frac{1{,}98}{10} = 0{,}2$$

$\alpha = 0{,}05$

$\nu = n - 2 = 10$

De la Tabla 3 del Anexo, $t_{\alpha/2;\ \nu} = 2{,}228$

$$P\left(b - t_{\frac{\alpha}{2};\nu}\sqrt{\frac{Cuadrado\ Medio\ Error}{Suma\ de\ Cuadrados\ de\ X}} < \beta < b + t_{\frac{\alpha}{2};\nu}\sqrt{\frac{Cuadrado\ Medio\ Error}{Suma\ de\ Cuadrados\ de\ X}}\right) = 1 - \alpha$$

$$P\left(1{,}75 - 2{,}228\sqrt{\frac{0{,}2}{4{,}63}} < \beta < 1{,}75 + 2{,}228\sqrt{\frac{0{,}2}{4{,}63}}\right) = 1 - 0{,}05$$

$$P\left(1{,}29\frac{\$}{t} < \beta < 2{,}21\ \$/t\right) = 0{,}95$$

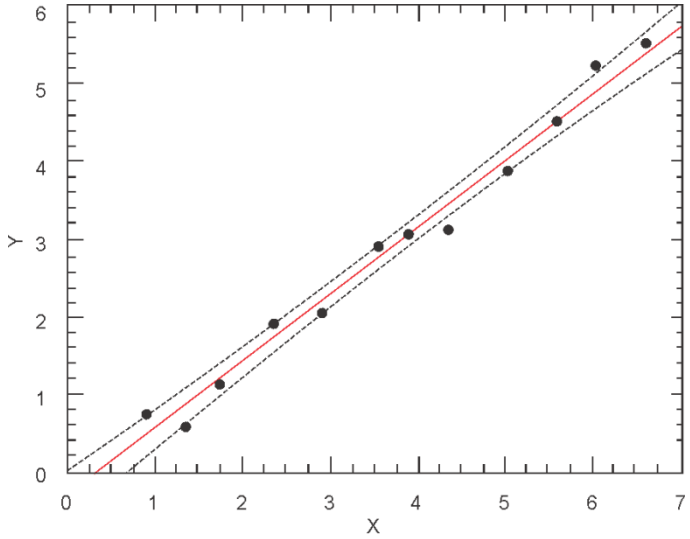


Figura 7. Límites de confianza del 95% de la línea de regresión. Línea roja, recta de regresión. Líneas negras cortadas límites de confianza.

Limites de confianza para el valor estimado de Y

Los límites de confianza de los valores estimados de *Y* generan una Banda de Confianza alrededor de la recta de regresión cuyo ancho es mínimo para $X = \bar{X}$ y se ensancha a medida que se aparta de la media (Fig. 8). La expresión 8.26 permite calcular los límites de confianza de un valor estimado

$$P\left(\hat{Y} - t_{\frac{\alpha}{2};\nu}\sqrt{S^2_{Y\cdot X}\left[\frac{1}{n} + \frac{(X-\bar{X})}{SCX}\right]} < \hat{Y_\mu} < \hat{Y} + t_{\frac{\alpha}{2};\nu}\sqrt{S^2_{Y\cdot X}\left[\frac{1}{n} + \frac{(X-\bar{X})}{SCX}\right]}\right) = 1 - \alpha \qquad (8.26)$$

para $\nu = n - 2$.

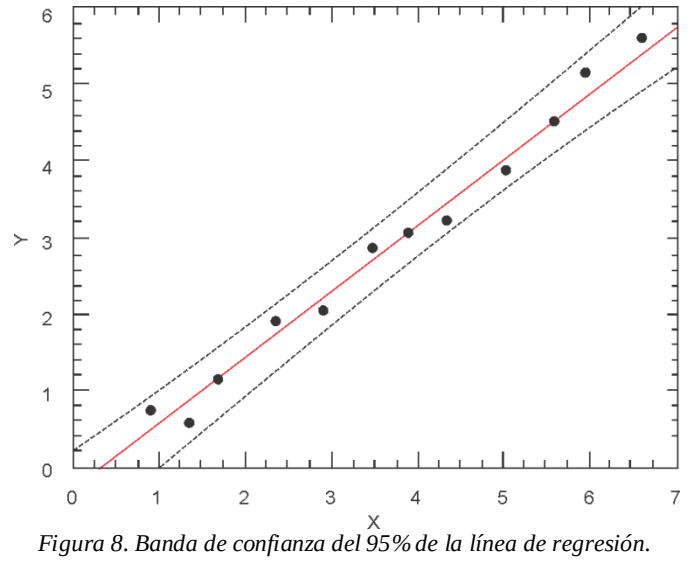


Figura 8. Banda de confianza del 95% de la línea de regresión.
Línea roja, recta de regresión. Líneas negras cortadas límites de confianza.

EJEMPLO 8

Límites de confianza del valor estimado con la recta de regresión
Se utilizan los datos del ejemplo 4 de producción-costo de la empresa minera.

$n = 12$
$\nu = 10$
$x = 2$
$\hat{y} = 2{,}01\,\$ + 1{,}75\,\$/t \cdot 2\,t$
$\hat{y} = 7{,}51\$$
$\alpha = 0{,}05$

De la Tabla 3 del Anexo, $t_{\alpha/2;10} = 2{,}228$

$\bar{X} = 1{,}89$
$SCx = 4{,}63$

$S^2_{Y \cdot X} = CME = 0{,}2$

$$P\left(\hat{Y} - t_{\frac{\alpha}{2};\nu}\sqrt{S^2_{Y \cdot X}\left[\frac{1}{n} + \frac{(X - \bar{X})}{SCX}\right]} < \hat{Y}_\mu < \hat{Y} + t_{\frac{\alpha}{2};\nu}\sqrt{S^2_{Y \cdot X}\left[\frac{1}{n} + \frac{(X - \bar{X})}{SCX}\right]}\right) = 1 - \alpha$$

$$P\left(7{,}51 - 2{,}228\sqrt{0{,}2\left[\frac{1}{12} + \frac{(2 - 1{,}89)}{4{,}63}\right]} < \hat{Y}_\mu < 7{,}51 + 2{,}228\sqrt{0{,}2\left[\frac{1}{12} + \frac{(2 - 1{,}89)}{4{,}63}\right]}\right) = 1 - 0{,}05$$

$$P\left(7{,}18\$/t < \hat{Y}_\mu < 7{,}83\$/t\right) = 0{,}95$$

Predicción inversa

Existen algunas pocas situaciones en las que es posible estimar el valor de la variable dependiente a partir de la variable independiente, esto se conoce como predicción inversa. Para el cálculo simplemente se acomodan los términos de la expresión 8.10 de modo que se tiene

$$\hat{X}_i = \frac{Y_i - a}{b} \tag{8.27}$$

Los límites de confianza del valor estimado de X no son simétricos, para calcularlos se usa la expresión

$$p\left(\bar{X} + \frac{b(Y_i - \bar{Y})}{K} \mp \frac{t_{\frac{\alpha}{2},n-2}}{K}\sqrt{S^2_{Y \cdot X}\left[\frac{(Y_i - \bar{Y})^2}{SCX} + K\left(1 + \frac{1}{n}\right)\right]}\right) = 1 - \alpha, \tag{8.28}$$

donde $K = b^2 - t^2_{\frac{\alpha}{2},n-2}S^2_b$.

Comparación de dos coeficientes de regresión (dos pendientes)

En los casos en los que se muestrean dos poblaciones y se obtienen dos rectas de regresión suele ser necesario conocer si las pendientes son significativamente diferentes entre sí o si son diferentes de cero. Para comparar las pendientes de dos rectas se utiliza una prueba de *t* de Student análoga a la que se utiliza para comparar dos medias (Capítulo 6) con un procedimiento que lleva varios pasos.

Prueba de igualdad de pendientes

Para analizar si las pendientes de las dos rectas son iguales (H_0: $\beta_1 = \beta_2$, H_A: $\beta_1 \neq \beta_2$) el primer paso es calcular el estadístico *t* con la expresión

$$t_b = \frac{b_1 - b_2}{S_{b_1 - b_2}} \tag{8.29}$$

donde el error estándar de la diferencia entre los dos coeficientes de regresión es

$$S_{b_1 - b_2} = \sqrt{\frac{(S^2_{Y \cdot X})_p}{(SCX)_1} + \frac{(S^2_{Y \cdot X})_p}{(SCX)_2}}. \tag{8.30}$$

El cuadrado medio residual ponderado se calcula con

$$(S^2_{Y \cdot X})_p = \frac{CMResidual_1 + CMResidual_2}{glResidual_1 + glResidual_2}. \tag{8.31}$$

El valor critico de *t* para esta prueba tiene grados de libertad $v = n_1 + n_2 - 4$ (la suma de los grados de libertad de los dos residuos). Como en otras pruebas, si $|t_b| \geq t_{(\alpha,\ n1+n2-4)}$, se rechaza la hipótesis nula para el nivel de significación elegido para la prueba.

Si **se rechaza la hipótesis nula** (H_0: $\beta_1 = \beta_2$) se asume que se han muestreado dos poblaciones diferentes, entonces conviene calcular el punto donde ambas rectas se cruzan utilizando

$$X_C = \frac{a_2 - a_1}{b_1 - b_2}. \tag{8.32}$$

Para estimar cualquier valor de *Y* se usa indistintamente las ecuaciones de las rectas:

$$\widehat{Y_C} = a_1 + b_1 X_C \qquad \text{o} \qquad \widehat{Y_C} = a_2 + b_2 X_C. \tag{8.33}$$

Cuando **la hipótesis nula no se rechaza** hay que estimar el parámetro poblacional β que corresponde a ambas rectas con

$$b_C = \frac{SCxy_1 + SCxy_2}{SCx_1 + SCx_2}. \tag{8.34}$$

Prueba para diferencia de pendientes igual, mayor o menor y diferente de cero

Es posible testear la hipótesis que la diferencia entre dos coeficientes de regresión es un valor específico (β_0). Las hipótesis pueden ser a dos colas (H_0: $\beta_1 - \beta_2 = \beta_0$, H_A: $\beta_1 - \beta_2 \neq \beta_0$) o a una sola cola ($H_0$: $\beta_1 - \beta_2 \geq \beta_0$, H_A: $\beta_1 - \beta_2 < \beta_0$ **o** H_0: $\beta_1 - \beta_2 \leq \beta_0$, H_A: $\beta_1 - \beta_2 > \beta_0$). El estadístico de prueba es

$$t = \frac{|b_1 - b_2| - \beta_0}{S_{b_1 - b_2}}. \qquad (8.35)$$

El valor critico de *t* para la prueba tiene grados de libertad $v = n_1 + n_2 - 4$ (la suma de los grados de libertad de los dos residuos). La hipótesis nula se rechaza si el *t* calculado es mayor que el *t* crítico.

EJEMPLO 9

Test de diferencia entre dos coeficientes de regresión

Se presentan las rectas de regresión que modelan los cambios del contenido de humedad en función de la profundidad de dos tipos de suelos diferentes que se desarrollan en una misma área de estudio.

H_0: $\beta_1 = \beta_2$; H_A: $\beta_1 \neq \beta_2$

Para la muestra 1

$n = 26$

$\sum x^2 = 1470{,}87$

$\sum y^2 = 13299{,}53$

$\sum xy = 4363{,}16$

$b = \frac{4363{,}16}{14670{,}87} = 2{,}97$

$SCError = 13299{,}53 - \frac{4363{,}16^2}{1470{,}87} = 356{,}73$

glError = 26 – 2 = 24

Para la muestra 2

$n = 30$

$\sum x^2 = 2272{,}47$

$\sum y^2 = 10964{,}09$

$\sum xy = 4928{,}81$

$b = \frac{4928{,}81}{2272{,}47} = 2{,}17$

$SCError = 10964{,}09 - \frac{4928{,}81^2}{2272{,}47} = 273{,}91$

glError = 30 – 2 = 28

$(S^2_{Y \cdot X})_p = \frac{356{,}73 + 273{,}91}{24 + 28} = 12{,}13$

$S_{b1-b2} = \sqrt{\frac{12{,}13}{1470{,}87} + \frac{12{,}13}{2272{,}47}} = 0{,}1165$

$t = \frac{2{,}97 - 2{,}17}{0{,}1165} = 6{,}867$

$\nu = 24 + 28 = 52$

$\alpha = 0{,}05$

De la Tabla 3 del Anexo, $t_{0{,}025;\ 52} = 2{,}007$

Se rechaza la Hipótesis nula, $t \geq t_{0{,}025;\ 52}$, $p < 0{,}001$ (6,867 > 2,007). Los coeficientes de regresión de los dos suelos son diferentes.

Interpretación de la función de regresión

Como se ha visto, la ecuación lineal calculada a partir de una muestra de datos bivariados (las constantes *a* y *b*) permite describir el ritmo de cambio de la variable dependiente (*Y*) con el cambio de la variable independiente (*X*). Sin embargo, aunque se asume que existe una dependencia matemática

de *Y* sobre *X*, no se debe asumir que exista un fenómeno geológico causa - efecto, ni que el fenómeno geológico pueda ser descripto en su totalidad por la recta. A pesar de estas limitaciones, la ecuación de regresión es útil para predecir los valores de *Y* para un *X* dado así como el error de la estima, valores muy útiles en sí mismos.

Regresión lineal en geocronología

Un uso importante de la regresión es en datación de muestras geológicas, rocas o fósiles usando el método radiométrico basado en la desintegración atómica. En la constitución de rocas y fósiles participan elementos inestables químicamente llamados isotopos radiactivos que se desintegran y se transforman en otros. El isotopo radiactivo se denomina padre (*P*) y el isotopo estable hijo (*D*). La ecuación 8.36 expresa la relación entre ambas ambos. *D* es la variable dependiente (equivalente a *Y*), *P* es la independiente (equivalente a *X*).

$$D = D_0 + \left(e^{\lambda t} - 1\right)P. \tag{8.36}$$

D es la cantidad de isotopo presente en la muestra, D_0 es la cantidad presente en la roca al momento de su formación, *e* es el exponencial, λ es la constante de decaimiento radiactivo (se mide en años), *t* es la edad de la muestra medido desde el presente (se expresa en Ma -10^6 años- o en Ga -10^9 años-) y *P* es la cantidad de isótopo padre de la muestra.

La desintegración o decaimiento se realiza a un ritmo constante que se puede medir en el tiempo geológico. El periodo de Semidesintegración (*t*/2) es el tiempo que tarda en desintegrarse la mitad de los núcleos del isotopo radiactivo (se mide en años). También se puede definir como el tiempo que tardan en transmutarse la mitad de los átomos radiactivos de una muestra. Un ejemplo es el Carbono 14 utilizado para datar restos fósiles no muy antiguos, su periodo es de 5760 años. Se conoce el periodo de semidesintegración de muchos isotopos que forman parte de la constitución de las rocas y se utilizan en geocronología, por ejemplo $^{14}C/^{14}N$, $^{40}K/^{40}Ar$, $^{87}Rb/^{87}Sr$ y $^{238}U/^{206}Pb$.

Un diagrama isócrono es un diagrama bivariado de la relación entre los isotopos padre - hijo para un conjunto de muestras de roca cogeneticas. Cuando la suite de muestras define un patrón lineal forman una **isócrona** y la pendiente de la recta es proporcional a la edad de la suite. La edad de la roca se calcula a partir de la pendiente de la recta con la ecuación,

$$t = \frac{1}{\lambda} \ln (pendiente + 1). \tag{8.37}$$

La recta que mejor ajusta la isocrona se obtiene con el método de mínimos cuadrados pesados que permite determinar, con precisión razonable, la pendiente y ordenada al origen de la isócrona. Utilizando este método es posible conocer la bondad del ajuste de la isócrona a partir del cuadrado

medio de las desviaciones y obtener los límites del error analítico de los datos. La **errocrona** es la recta que no ajusta a los datos dentro de los límites del error analítico.

ESTADÍSTICA NO PARAMÉTRICA

Introducción

Bajo el nombre de métodos no paramétricos se incluyen una variedad de pruebas que se emplean en el análisis de variables nominales (que se pueden clasificar en diferentes categorías) y de variables que se expresan en escalas ordinales. También se utilizan para datos medidos en escalas de intervalo o de razón donde la función de distribución de la variable aleatoria es inespecífica. Surgidos a fines de 1930, estos métodos llenan el vacío que dejan los métodos paramétricos descriptos en los capítulos precedentes.

Las pruebas no paramétricas tienen ventajas y desventajas respecto a las paramétricas. Se pueden utilizar cuando se desconoce la distribución de probabilidad la población, brindan respuestas rápidas con pocos cálculos, en los casos en que los datos son convertidos en rangos se elimina la incertidumbre asociada a la escala utilizada en las mediciones. Por otra parte, si los datos son continuos pero existen dudas del cumplimiento de los supuestos requeridos por las metodologías descriptas hasta ahora (normalidad y homogeneidad de varianzas) permiten hacer inferencias sobre los parámetros poblacionales. Sin embargo no usan toda la información disponible y al no haber parámetros es difícil hacer estimaciones cuantitativas, además son algo menos eficientes, para rechazar la hipótesis nula con el mismo nivel de confianza se necesitan muestras mayores.

Variable 1	***Variable 2***	
	Nominal	***Ordinal***
Nominal	Bondad de ajuste (χ^2) Tabla de Contingencia	Test de Kolmogorov – Smirnov (dos muestras)
Ordinal	Test de rachas* Test U de Mann – Whitney Test de Kruskal – Wallis	Coeficiente de correlación de Spearman

*Cuadro 1. Métodos no paramétricos. * El test de rachas se describe en el Capítulo 11 dado que se utiliza en el análisis de series.*

Dado que la estadística no paramétrica trabaja bien en suposiciones muy generales acerca de las características de cualquier distribución de probabilidades o parámetros involucrados en un problema inferencial, entonces si existen dudas sobre los supuestos que subyacen en los métodos paramétricos

es más seguro usar métodos no paramétricos, sin embargo cuando los supuestos se cumplen se deben aplicar métodos paramétricos.
Naturalmente el tipo de variable y el objetivo del análisis determinan cuál debe ser el método que se utilice. El cuadro 1 muestra los métodos que se abordarán en este capítulo.

Pruebas para datos nominales

Pruebas de bondad de ajuste χ^2

En los trabajos geológicos en los que se registra la presencia de alguna característica, por ejemplo se registra la presencia de estructuras en sedimentos cuando se levanta una columna estratigráfica o cuando muestras de agua se clasifican como potables o no, se utiliza la prueba χ^2. La prueba requiere datos nominales con categorías mutuamente excluyentes y su objetivo es inferir si la muestra fue tomada de una población cuyas componentes tienen proporciones establecidas. Aplicaciones geológicas se encuentran en la clasificación de rocas, problemas mineros y problemas paleontológicos entre otros.
Para calcular el estadístico de prueba se muestrea la población, los datos se expresan como **frecuencias**. El estadístico de prueba es el mismo que el de otras pruebas de bondad de ajuste (Capítulo 6, expresión 6.25),

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}, \qquad (9.1)$$

dónde *fo*: frecuencia observada, *fe*: frecuencia teórica o esperada y *k*: son las categorías de los datos.
Las hipótesis que se contrastan son:
H_0: *fo* = *fe* (las frecuencias observadas son iguales a las frecuencias esperadas) y
H_A: *fo* ≠ *fe* (las frecuencias observadas son diferentes a las frecuencias esperadas).
Se fija el nivel de significación de la prueba, α. La hipótesis nula se rechaza cuando $\chi_c^2 \geq \chi_{\alpha;\nu}^2$. Los grados de libertad, ν, se calculan como *k* (número de clases) menos 1 ($\boldsymbol{\nu = k - 1}$).

EJEMPLO 1

Prueba de bondad de ajuste de χ^2

El objetivo de la prueba es conocer si una muestra geológica de gneiss fue tomada de un afloramiento donde la relación entre los porfiroblastos de hornblenda (H), biotita (B) y granate (G) es 9:3:3. Los datos provienen de un análisis modal.

	H	*B*	*G*	*N*
fo	152	39	53	244
fe	146,4	48,8	48,8	

H_0: Los datos provienen de una población con relación H:B:G de 9:3:3.
H_A: Los datos provienen de una población donde la relación H:B:G no es 9:3:3.

N = 244
Grados de libertad $\nu = k - 1 = 3 - 1 = 2$
Nivel de significación $\alpha = 0,05$
De la Tabla 2 del Anexo, $\chi^2_{0,05;2} = 5,99$

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe} = \frac{(152 - 146,4)^2}{146,4} + \frac{(39 - 48,8)^2}{48,8} + \frac{(53 - 48,8)^2}{48,8} = 0,2142 + 1,9680 + 0,3615 = 2,544$$

Dado que 2,544 < 5,99 no existen evidencias para rechazar la hipótesis nula.
Se puede concluir que la muestra de gneiss proviene de un afloramiento donde hornblenda, biotita y granate tiene una relación 9:3:3.

Corrección de χ^2 por continuidad (Corrección de Yates)

Cuando los grados de libertad de la prueba es uno, $\nu = 1$ y el tamaño de la muestra es menor a doscientos ($n < 200$), el valor de χ^2 calculado con la expresión 9.1 no es exacto, está sobrestimado. Esto puede inducir a cometer **errores Tipo I**, rechazar la hipótesis nula cuando es verdadera, con probabilidad de error mayor a α, y el sesgo aumenta a medida que el tamaño de la muestra disminuye. En los casos en los que **$\nu = 1$** se debe **aplicar la Corrección por continuidad de Yates** donde el valor absoluto de las desviaciones entre la frecuencia observada y las esperadas se reduce 0,5, de modo que

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(|fo - fe| - 0,5)^2}{fe}, \qquad (9.2)$$

para χ_c^2 el valor de χ^2 corregido.

EJEMPLO 2
Corrección por continuidad de Yates
El sentido de enroscamiento de los caparazones del foraminífero *Globorotalia truncatulinoides* se usa como dato indirecto (*proxi data*) para estimar paleotemperatura del agua de mar. Las valvas dextrógiras ocurren en una relación 9:1 sobre las levógiras cuando las aguas son cálidas. El objetivo del estudio es determinar la paleotemperatura del agua en un un nivel de un testigo recogido a la latitud de Buenos Aires en la plataforma continental a partir de *G. truncatulinoides*.

H_0: Los datos provienen de una población con relación 9:1 de *G. truncatulinoides* dextrógiras-levógiras.
H_A: Los datos provienen de una población donde la relación *G. truncatulinoides* dextrógiras-levógiras no es 9:1.

	Categoría (sentido de enroscamiento)	
	Dextrógiras	*Levógiras*
fo	84	16
fe	90	10

N = 100

fe valvas dextrógiras = (0,9) 100 = 90
fe valvas levógiras = (0,1) 100 = 10
Grados de libertad $\nu = k - 1 = 2 - 1 = 1$
Nivel de significación $\alpha = 0,05$
De la Tabla 2 del Anexo, $\chi^2_{0,05;1} = 3,84$

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe} = \frac{(84-90)^2}{90} + \frac{(16-10)^2}{10} = 0,4000 + 3,6000 = 4,000$$

Utilizando la corrección por continuidad de Yates se obtiene

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(|fo - fe| - 0,5)^2}{fe} = \frac{(|84-90| - 0,5)^2}{90} + \frac{(|16-10| - 0,5)^2}{10} = 0,3361 + 3,0250 = 3,3611$$

Observe que si no se utiliza la corrección de Yates se rechaza la hipótesis nula dado que 4,00 > 3,84. Sin embargo, el resultado obtenido usando la corrección indica que no existen evidencias para rechazar la hipótesis nula puesto que 3,36 < 3,84.
Se puede concluir que los ejemplares provienen de una población donde la relación de *G. truncatulinoides* dextrógiras-levógiras es 9:1 lo que indicaría que se trata de aguas cálidas.

Pruebas de asociación. Tablas de contingencia

El estudio de la influencia de una variable, variable independiente, sobre la forma en que modifica otra, la variable dependiente, es un estudio bivariado. Las tablas de contingencia son tablas de doble entrada que constituyen una herramienta fundamental para el análisis de la dependencia o independencia entre variables categóricas. Las variables y sus categorías pueden ser nominales o estar codificadas. Las tablas están compuestas por *m* filas que presenta las categorías de una variable y *n* columnas para las categorías de la otra variable, en algunas ocasiones puede haber una tercera variable y en ese caso son tablas en tres dimensiones. La intersección de filas y columnas delimita celdas donde se vuelcan las **frecuencias absolutas** de los individuos que presentan las categorías simultáneamente. Es importante recalcar que deben ser tablas de frecuencias absolutas y nunca deben ser proporciones o porcentajes.
Por ejemplo, observar la frecuencia de una especie de bivalvo fósil en diferentes afloramientos de una Formación o la ocurrencia de un metal pesado por encima o debajo de un valor límite en diferentes sectores de un río o quizás la redondez y esfericidad de los clastos de una psefita. En cualquiera de estos casos y en otros muchos análisis de este tipo subyacen dos hipótesis sobre la existencia de alguna relación entre la frecuencia de ocurrencia simultánea de las dos variables. La primera afirma que existe una **relación** entre las variables estudiadas, ocasionalmente, si existe sustento teórico, es posible hablar de variable dependiente y de variable independiente. La otra hipótesis afirma que no existe tal relación y que ambas variables son totalmente independientes, esta es la hipótesis nula. Inspeccionar la tabla de contingencia no permite ser concluyente sobre cual hipótesis es válida y se

debe realizar una prueba, en este caso se trata de una prueba de hipótesis de χ^2. La hipótesis nula es sobre **ausencia de relación** de ocurrencia, sólo a veces de **independencia**. Si el supuesto de independencia se cumple, se espera encontrar frecuencias similares para la ocurrencia conjunta de las categorías de cada variable, pero cuando las frecuencias esperadas son diferentes de las observadas se deduce una relación de **dependencia**.

El estadístico de prueba es

$$\chi_c^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}, \tag{9.3}$$

dónde fo_{ij} es la frecuencia observada en la fila *i* y columna *j*, fe_{ij} la frecuencia esperada.

Las frecuencias esperadas se calculan con:

$$fe_{ij} = \frac{TF_i \cdot TC_j}{N}, \tag{9.4}$$

para TF_i = suma de las frecuencias de la fila *i*, TF_j = suma de las frecuencias de la columna *j*, N = cantidad de datos.

Las hipótesis que se contrastan son:

H_0: $fo = fe$ (no existe relación entre las variables) y

H_A: $fo \neq fe$ (las variables están relacionadas).

Para saber si la diferencia es significativa o no lo es, se compara el valor de χ^2 calculado con un $\chi^2_{\alpha;\nu}$ crítico para el nivel de significación, α, elegido para el test (Tabla 2 del Anexo). La hipótesis nula se rechaza cuando $\chi_c^2 \geq \chi^2_{\alpha;\nu}$. Los grados de libertad ν se calculan como cantidad de columnas (n) menos por cantidad de filas (m) menos 1 ($\nu = (m-1)(n-1)$).

EJEMPLO 3

Tablas de contingencia

En el noroeste de la provincia de Buenos Aires las aguas subterráneas suelen poseer elevados tenores de flúor. Concentraciones elevadas de este elemento en el agua para consumo resultan perjudiciales para la salud, la OMS estableció como límite 0,5 µg/l. Un hidrogeólogo estudia si los valores de F en el agua subterránea de 4 localidades del municipio de Carlos Tejedor se relacionan con su ubicación. Las muestras de agua de los pozos fueron clasificadas en dos categorías mutuamente excluyentes: por encima o por debajo del nivel establecido por la OMS.

H_0: $fo = fe$ (los niveles de F son independientes de la localidad)
H_1: $fo \neq fe$ (los niveles de F no son independientes de la localidad)
Nivel de significación de la prueba $\alpha = 0{,}05$
Grados de libertad $\nu = (m-1)(n-1) = (2-1)(4-1) = 3$
De la Tabla 2 del Anexo, $\chi^2_{0{,}05;3} = 7{,}815$

Frecuencias observadas

Nivel de F	*Localidad*				*Total Fila*
	A	*B*	*C*	*D*	
Debajo	32	43	16	9	100
Encima	55	65	64	16	200
Total Columna	87	108	80	25	300

A modo de ejemplo para la localidad A y valores debajo del nivel la $fe = \frac{(87)(100)}{300} = 29{,}000$

Frecuencias esperadas

Nivel de F	*Localidad*				*Total Fila*
	A	*B*	*C*	*D*	
Debajo	29,000	36,000	26,667	8,333	100
Encima	58,000	72,000	53,333	16,666	200
Total Columna	87	108	80	25	300

χ^2

Nivel de F	*Localidad*				*Total*
	A	*B*	*C*	*D*	
Debajo	*0,3103*	*1,3611*	*4,2667*	*0,0533*	
Encima	*0,1552*	*0,6806*	*2,1444*	*0,0267*	
Total					***8,987***

Dado que 8,987 > 7,815 existen evidencias para rechazar la hipótesis nula.
Se puede concluir que los niveles de F están relacionados con la localidad. Observe que en la localidad C hay menos pozos que los esperados con niveles bajos de F y más con niveles altos, por eso los valores de χ^2 subrayados en la tabla son elevados.

Prueba para datos ordinales y nominales

La **prueba de Kolmogorov-Smirnov** se puede usar para comparar dos distribuciones. Recuerde que en el capítulo 6 (expresión 6.28) se definió el estadístico de prueba *d* como la máxima diferencia entre la frecuencia relativa acumulada observada y esperada para el modelo teórico, sobre el número total de observaciones, la misma expresión se utiliza para comparar la distribución de dos poblaciones *A* y *B*,

$$d = \frac{\max|faA - faB|}{N} \tag{9.5}$$

dónde *faA*: máxima frecuencia relativa acumulada en la muestra *A*, *faB*: máxima frecuencia relativa acumulada en la muestra *B* y *N:* número total de observaciones.

Las hipótesis de la prueba son:

H_0: Las muestras provienen de la misma población.

H_1: Las muestras provienen de poblaciones diferentes.

H_0: $faA = faB$

H_1: $faB \neq faA$

La hipótesis nula se rechaza cuando $d \geq D_\alpha$. Los valores de *D* se encuentran en la Tabla 11 del Anexo.

EJEMPLO 4

Prueba de Kolmogorov-Smirnov para dos muestras

Frecuentemente se asevera que los procesos de desecación de los suelos arcillosos (F) que forman las grietas son similares a los que forman la disyunción columnar de los basaltos (B). Los barquillos de fango y las columnas de basalto tienen sección poligonal. Se plantea la siguiente hipótesis: si los procesos de formación son análogos, el número de lados de los barquillos de fango y de las columnas de

basalto serán iguales pues la contracción por enfriamiento o por desecación es equidistante desde un punto y tiende a formar estructuras hexagonales (tomado de Cheeny 1983).

X=N° lados	*f(B)*	*f(F)*	*fr(B)*	*fr(F)*	*fa (B)*	*fa (F)*	*d*
3	1	1	0.0303	0.0278	0.0303	0.0278	0.0025
4	3	7	0.0909	0.1944	0.1212	0.2222	0.1010
5	8	10	0.2424	0.2778	0.3636	0.5000	**0.1364**
6	15	8	0.4545	0.2222	0.8182	0.7222	0.0960
7	4	6	0.1212	0.1667	0.9394	0.8889	0.0505
8	1	4	0.0303	0.1111	0.9697	1.0000	0.0303
9	0	0	0.0000	0.0000	0.9697	1.0000	0.0303
10	1	0	0.0303	0.0000	1.0000	1.0000	0.0000

$n_B = 33$; $n_F = 36$

H_0: Las dos muestras son tomadas de poblaciones iguales (igual número de lados de los polígonos).
H_1: Las dos muestras son tomadas de poblaciones diferentes (diferente número de lados de los polígonos).
H_0: $faB = faF$
H_1: $faB \neq faF$

Nivel de significación, $\alpha = 0{,}05$
De la Tabla 11 del Anexo, $D_{(0,05)} = 0{,}241$

$$d = \max|faB - faF| = 0{,}1364$$

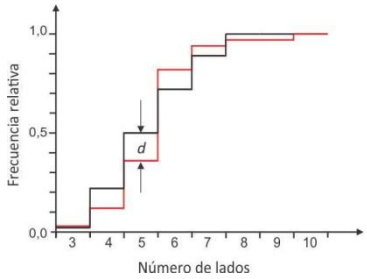


Figura ejemplo 4. En negro distribución de frecuencias relativas acumuladas de Fangos y en rojo de Basaltos.

Debido a que $d < D_{(0,05)}$ (0,1364 < 0,241), no existen evidencias para rechazar la hipótesis nula. Los procesos que originan las grietas de desecación y la disyunción columnar son similares.

Pruebas para datos ordinales

Muchas veces los trabajos requieren realizar pruebas pero los datos muestrales no cumplen con los requisitos de normalidad y homogeneidad de varianzas necesarios para aplicar las pruebas de *t*, ANOVA, coeficiente de correlación, entre otras. Las pruebas no paramétricas llamadas de distribución libre, resuelven el problema pues tienen exigencias menores. Estas pruebas, a diferencia de las paramétricas que utilizan los valores medios, comparan las distribuciones, por ello son menos potentes que sus equivalentes paramétricos. Las pruebas no paramétricas que se describen a continuación se utilizan con datos que originariamente fueron medidos en un escala ordinal o asignando rangos a datos medidos en escalas de razón o de intervalos (comúnmente de manera ascendente desde 1 al menor hasta *n* al mayor).

Aunque no tienen restricciones, se recomienda no usar pruebas no paramétricas cuando existen grandes diferencias entre la distribución de las muestras (incluyendo las varianzas). Como regla general la relación entre la mayor y menor varianza no debe ser superior a 4:1.

Test U de Mann - Whitney para comparar dos muestras independientes

Mann y Whitney en 1947 propusieron una prueba para comparar dos muestras independientes, cada individuo de una muestra se compara con los de la otra. La hipótesis nula postula que las dos muestras pertenecen a la misma población.

Entonces, se toman dos muestras aleatorias independientes (M_1 y M_2), con n_1 y n_2 observaciones cada una, los datos se combinan y se ordenan de **menor a mayor**, rango 1 al menor, 2 al siguiente, hasta n al mayor ($n_1 + n_2 = n$). Luego se suman los rangos de cada muestra por separado (R_1 y R_2 son la suma de rangos de la muestra 1 y 2 respectivamente).

Si las observaciones fueron seleccionadas de la misma población, esto es la hipótesis nula, la suma de los rangos de cada muestra (Ri) debería ser más o menos proporcionales al tamaño de cada una (n_1 y n_2). Es más, si n_1 y n_2 fueran iguales, las sumas de los rangos para las dos muestras son similares. Al contrario, si las observaciones fueron tomadas de diferentes poblaciones, esto es la hipótesis alternativa, los datos de una muestra serán diferentes a los de la otra (mayores o menores) y por lo tanto la suma de rangos, Ri, de cada muestra será diferente poniendo en evidencia que las muestras fueron tomadas de diferentes poblaciones.

Con los estadísticos de Mann-Whitney, U y U´, es posible discernir si los datos proceden de una misma población o no.

$$U = n_1 \cdot n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1, \tag{9.6}$$

$$U' = n_1 \cdot n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2. \tag{9.7}$$

Para confirmar $U + U' = n_1 \cdot n_2$.

Como se desprende de las expresiones 9.6 y 9.7, los valores de U y $U´$ dependen del tamaño de las muestras y de la suma de rangos. U es grande cuando R_1 es pequeño (la muestra tiene valores bajos) y viceversa. De igual forma $U´$ es grande cuando R_2 es pequeño (la muestra tiene valores bajos) y viceversa. Por consiguiente depende de las hipótesis del problema cual valor, U o $U´$, se contrastan con el valor crítico de tabla para el nivel de significación elegido para la prueba.

Si la **prueba es a dos colas**, las hipótesis son H_0: las muestras pertenecen a la misma población ($M_1 = M_2$) y H_A: las muestras son de poblaciones diferentes ($M_1 \neq M_2$), entonces el estadístico puede ser U o $U´$, **siempre el mayor**. La hipótesis nula se rechaza cuando U o $U´ > U_{Crítico\ \alpha/2}$.

Cuando la **prueba es unilateral de cola inferior**, con hipótesis H_0: $M_1 \geq M_2$ y H_A: $M_1 < M_2$ se debe calcular $\boldsymbol{U}$. La hipótesis nula se rechaza cuando $U > U_{Crítico\ \alpha}$. Por el contrario, cuando la **prueba es unilateral de cola superior**, con hipótesis H_0: $M_1 \leq M_2$ y H_A: $M_1 > M_2$ se debe calcular $\boldsymbol{U´}$. La hipótesis nula se rechaza cuando $U´ > U_{Crítico\alpha}$.

La Tabla 12 del Anexo tiene los valores críticos para muestras pequeñas (menor a 20). Cuando el tamaño de las muestras es mayor a 12, la distribución del estadístico *U* se aproxima a la distribución normal, con $\mu_U = \frac{n_1 n_2}{2}$ y $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$, esto permite calcular el estadístico de prueba $z = \frac{U-\mu_U}{\sigma_U}$ que se contrasta con *Z* (Tabla 1 del Anexo) para el nivel de significación elegido para la prueba. Para pruebas bilaterales la hipótesis nula se rechaza cuando $z > Z_{Crítico}$.

Por último, si hubiera valores iguales se producen **empates**, entonces se asigna a cada miembro del grupo empatado el promedio de los rangos que se les habría asignado a las observaciones. Por ejemplo suponga que hay empates entre el cuarto y quinto valor, en ese caso el rango de ambos será $\frac{4+5}{2} = 4{,}5$ y si el séptimo, octavo y noveno datos fueran iguales a los tres les correspondería el rango $\frac{7+8+9}{2} = 8$.

EJEMPLO 5

Prueba U de Mann-Whitney

El contenido de Fe^{+2} en los suelos es un importante proveedor de electrones y participa en la adsorción y mecanismos redox. Se diseña un procedimiento de muestreo en el horizonte A de dos suelos hidromórficos de importancia agrícola de la provincia de Buenos Aires: un suelo Natraquert (1) situado en la marisma y un Fluvaquert (2) situado en la llanura aluvial.

	Natraquert (1)		*Fluvaquert (2)*	
	Fe^{+2}	*Rango*	*Fe^{+2}*	*Rango*
	11,1	9	9,0	3,5
	12,3	12	10,7	7
	9,6	6	11,1	9
	9,0	3,5	11,1	9
	13,4	13	9,3	5
	11,8	11	8,7	2
	14,6	14	8,4	1
	15,2	15		
n_i	*8*		*7*	
R_i		*83,5*		*36,5*

a) Interesa saber si ambos suelos tienen igual contenido de Fe^{+2}.

H_0: El contenido de Fe^{+2} es el mismo en ambos suelos ($M_1 = M_2$)

H_A: El contenido de Fe^{+2} es diferente en ambos suelos ($M_1 \neq M_2$)

Prueba a dos colas, $\alpha = 0{,}05$

De la Tabla 12 del Anexo, $U_{7;\,8;\,0{,}05(1)} = 13$

$$U = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 7 \cdot 8 + \frac{8(8+1)}{2} - 83{,}5 = 8{,}5$$

$$U' = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 7 \cdot 8 + \frac{7(7+1)}{2} - 36{,}5 = 47{,}5$$

La prueba es a dos colas $U < U'$, entonces U' se usa para contrastar las hipótesis.

Debido a 47,5 > 13 ($U' > U_{8;\,7;\,0{,}05}$) existen evidencias para rechazar la hipótesis nula. El contenido de Fe^{+2} en el horizonte A del Natraquert es diferente al del Fluvaquert.

b) Interesa saber si el contenido de Fe^{+2} del Natraquert es mayor al del Fluvaquert.

H_0: el contenido de Fe^{+2} del Natraquert es menor o igual al del Fluvaquert ($M_1 \leq M_2$)

H_A: el contenido de Fe^{+2} del Natraquert es mayor al del Fluvaquert ($M_1 > M_2$)
Prueba de cola de cola derecha se debe calcular $U´$. $U´$= 47,5
α = 0,05
De la Tabla 12 del Anexo, $U_{8;\ 7;\ 0,05(1)} = 10$
Debido a 47,5 > 10 ($U´ > U_{5;\ 6;\ 0,05}$) existen evidencias para rechazar la hipótesis nula. El contenido de Fe^{+2} del Natraquert es mayor al del Fluvaquert.

Test de Kruskal-Wallis (ANOVA no paramétrico)

El test de Kruskal – Wallis, propuesto en 1952, se emplea para detectar diferencias entre más de dos poblaciones. Se aplica cuando el supuesto de normalidad de ANOVA no se cumple.
Los supuestos y el procedimiento de cálculo son similares en las etapas iniciales al descripto en la prueba de Mann – Whitney. Se obtienen *k* muestras aleatorias independientes (M_1, M_2, ..., M_k), con n_1, n_2 ..., n_k observaciones cada una, los datos se combinan y se ordenan de **menor a mayor**, rango 1 al menor, 2 al siguiente, hasta *n* el mayor ($n_1 + n_2 + ... + n_k = n$). Luego se suman los rangos de cada muestra por separado (R_i). Los empates se manejan como en la prueba de *U*, se asigna a cada miembro del grupo empatado el promedio de los rangos que se les habría asignado a cada observación.
La hipótesis nula de la prueba postula que las *k* distribuciones poblacionales de donde se obtuvieron las muestras son idénticas y la hipótesis alternativa que al menos una de las distribuciones poblacionales es diferente. Si la hipótesis nula es verdadera se espera que la suma de rangos de cada muestra, los R_i valores, sean prácticamente iguales, pero si existen diferencias entre las poblaciones las suma de rangos son diferentes. El estadístico *H* puede describir si existen o no esas diferencias.

$$H = \frac{12}{N(N+1)} \sum_{1=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \qquad (9.8)$$

Cuando todas las muestras tienen más de cinco datos ($n_i > 5$), el estadístico *H* aproxima a χ^2 con $\nu = k - 1$ grados de libertad. La hipótesis nula se rechaza cuando $H > \chi^2_{\nu;\alpha}$.
Si hay rangos empatados se debe corregir el valor de *H*.

$$H_c = \frac{H}{C} \qquad (9.10)$$

C se obtiene como,

$$C = 1 - \frac{\sum T}{n^3 - n} \qquad \text{y} \qquad \sum T = \sum_1^m (t_i^3 - t_i), \qquad (9.11 \text{ y } 9.12)$$

donde t_i es el número de empates en el *i*-ésimo grupo de empates y *m* es el número de rangos empatados.

EJEMPLO 6

Test de Kruskal – Wallis

Un problema limnológico cuyo objetivo es averiguar la calidad del agua en las lagunas llamadas "Encadenadas", ubicadas al suroeste de la provincia de Buenos Aires, se recogen datos de pH.

L. Epecuén		*L. del Monte*		*L. Cochicó*		*L. Alsina*	
7,68	(1)	7,71	(6*)	7,74	(13,5*)	7,71	(6*)
7,69	(2)	7,73	(10*)	7,75	(16)	7,71	(6*)
7,70	(3,5*)	7,74	(13,5*)	7,77	(18)	7,74	(13,5*)
7,70	(3,5*)	7,74	(13,5*)	7,78	(20*)	7,79	(22)
7,72	(8)	7,78	(20*)	7,80	(23,5*)	7,81	(26*)
7,73	(10*)	7,78	(20*)	7,81	(26*)	7,85	(29)
7,73	(10*)	7,80	(23,5*)	7,84	(28)	7,87	(30)
7,76	(17)	7,81	(26*)			7,91	(31)
n_i	*8*		*8*		*7*		*8*
R_i	*55*		*132,5*		*145*		*163,5*

*Entre paréntesis se encuentran los rangos.*Observaciones con rangos empatados.*

H_0: pH es el mismo en las 4 lagunas

H_A: el pH de al menos una es diferente

$\alpha = 0{,}05$

$k = 4$

$\nu = k - 1 = 3$

De la Tabla 2 del Anexo, $\chi^2_{(0{,}05;\ 3)} = 7{,}815$

$N = 8 + 8 + 7 + 8 = 31$

$$H = \frac{12}{N(N+1)} \sum_{1=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

$$H = \frac{12}{31(32)} \left(\frac{55^2}{8} + \frac{132{,}5^2}{8} + \frac{145^2}{7} + \frac{163^2}{8} \right) - 3(32) = 11{,}88$$

Como existen rangos empatados se debe corregir H.

Número de grupos con rangos empatados, $m = 7$.

$\sum T = \sum_1^m (t_i^3 - t_i)$

$\sum T = (2^3 - 2) + (3^3 - 3) + (3^3 - 3) + (4^3 - 4) + (3^3 - 3) + (2^3 - 2) + (3^3 - 3) = 168.$

$$C = 1 - \frac{\sum T}{n^3 - n} = 1 - \frac{168}{31^3 - 31} = 0{,}994$$

$$H_c = \frac{H}{C} = \frac{11{,}86}{0{,}9944} = 11{,}943$$

Dado que $H_c > \chi^2_{3;\ 0{,}05}$ (11,943 > 7,815) existen evidencias para rechazar la hipótesis nula. Se infiere que el pH de las lagunas es diferente.

Comparaciones múltiples no paramétricas

Cuando se rechaza la hipótesis nula de una prueba de Kruskal-Wallis corresponde preguntarse entre cuales muestras existen diferencias significativas. Para ello se realizan pruebas de hipótesis análogas a la prueba de Tukey (descripta en el capítulo 7) pero usando las sumas de rangos R_i en lugar de las medias muestrales. Esta prueba solo se puede realizar en **muestras de igual tamaño**.

Para cada par posible de comparaciones muestrales se formulan las hipótesis nula, las dos distribuciones son iguales, H_0: $\mu_A = \mu_B$ y la alternativa, las dos poblaciones son diferentes, H_A: $\mu_A \neq \mu_B$
El estadístico de contraste es ***q***,

$$q = \frac{R_B - R_A}{SE}, \tag{9.13}$$

donde el error estándar *SE* viene dado con la siguiente expresión:

$$SE = \sqrt{\frac{N\ (Nk)(Nk+1)}{12}}. \tag{9.14}$$

El estadístico q calculado se compara con el valor de Rangos estudentizados $q_{\alpha;\ \infty;\ k}$ (Tabla 8 del Anexo). La hipótesis nula se rechaza cuando $q > q_{\alpha;\ \infty;\ k}$.
Para facilitar las comparaciones se recomienda ordenarlas de mayor a menor de acuerdo a la suma de rangos y efectuar las diferencias entre la mayor y la menor de los posibles pares de comparaciones.
En el caso de **muestras de tamaño diferente** se calcula el estadístico ***Q***,

$$Q = \frac{\overline{R_A} - \overline{R_B}}{SE}, \tag{9.15}$$

donde $\bar{R}$ indica el valor medio de la suma de rangos (i. e. $\bar{R} = R/n$). Para muestras de tamaño diferente el error estándar, *SE*, se calcula como

$$SE = \sqrt{\frac{n(n+1)}{12} + \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}. \tag{9.16}$$

Si existen rangos empatados se utiliza,

$$SE = \sqrt{\frac{n(n+1)}{12} - \frac{\sum T}{12(n-1)}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}. \tag{9.17}$$

$\sum T$ es el calculado para la prueba del Kruskal-Wallis con rangos empatados.
La hipótesis nula se rechaza cuando $Q > Q_{\alpha;\ k}$ que se obtienen de la Tabla 13 del Anexo.

EJEMPLO 7

Comparaciones múltiples no paramétricas

Dado que existen diferencias en el pH de las lagunas Encadenadas (Ejemplo 6) pues se rechazó la hipótesis nula de la prueba de Kruskal–Wallis corresponde preguntar entre cuales lagunas se encuentran esas diferencias.

H_0: $\mu_A = \mu_B$

H_A: $\mu_A \neq \mu_B$

	L. Epecuén *(1)*	*L. del Monte* *(2)*	*L. Alsina* *(4)*	*L. Cochicó* *(3)*
R_i	55	132,5	163,5	145
n_i	8	8	8	7
$\bar{R}_i$	6,88	16,56	20,44	20,71

Las muestras tienen diferente tamaño y rangos empatados, entonces se utiliza

$$SE = \sqrt{\frac{N(N+1)}{12} - \frac{\sum T}{12(N-1)}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$\sum T = 168$
Para $n_A = 8$ y $n_B = 8$

$$SE = \sqrt{\frac{31(32)}{12} - \frac{168}{12(30)}\left(\frac{1}{8} + \frac{1}{8}\right)} = 4{,}53$$

Para $n_A = 7$ y $n_B = 8$

$$SE = \sqrt{\frac{31(32)}{12} - \frac{168}{12(30)}\left(\frac{1}{7} + \frac{1}{8}\right)} = 4{,}69$$

Comparación (B vs. A)	*Diferencias* $(\overline{RB} - \overline{RA})$	*SE*	*Q*	$Q_{0,05;\,4}$	*Conclusión*
3 vs. 1	20,71-6,88=13,83	4,69	2,95	2,639	Rechazar Ho: el pH del agua de las lagunas 1 y 3 es igual
3 vs. 2	20,71-16,56=4,15	4,69	0,88	2,639	Aceptar Ho: el pH del agua de las lagunas 2 y 3 es igual
3 vs. 4	20,71-20,44=0,27	4,53	0,596	2,639	Aceptar Ho: el pH del agua de las lagunas 3 y 4 es igual
4 vs.1	20,44-6,88=13,56	4,53	2,99	2,639	Rechazar Ho: el pH del agua de las lagunas 1 y 4 es igual
4 vs. 2	20,44-16,56=3,88	4,53	0,856	2,639	Aceptar Ho: el pH del agua de las lagunas 2 y 4 es igual
2 vs. 1	15,56-6,88=9,68	4,53	2,14	2,639	Aceptar Ho: el pH del agua de las lagunas 1 y 2 es igual

L. Epecuén L. del Monte L. Cochicó L. Alsina

Conclusión general: el pH de las lagunas Cochicó - del Monte, Cochicó – Alsina, Alsina – del Monte y del Monte – Epecuén no muestra diferencias significativas. La laguna de Epecuén tiene pH menor que las lagunas de Cochicó y Alsina.

Coeficiente de correlación de Spearman

El coeficiente de correlación de r_S Spearman se utiliza para saber si existe correlación entre dos variables cuando los supuestos para el cálculo del coeficiente de correlación *r* de Pearson no se cumplen.

El coeficiente de correlación r_S, al igual que *r*, varía entre +1 y –1; valores cercanos a +1 indican que las dos variables están positivamente correlacionadas, valores cercanos a –1 indican que la correlación entre ambas variables es inversa y valores cercanos a 0 indican ausencia de correlación. Si r_S es +1, entonces los rangos de las variables son idénticos: valores grandes de *x* corresponden a valores grandes de *y* y valores pequeños de *x* corresponden a valores pequeños de *y*. Sin embargo, a diferencia de *r* de Pearson, la relación entre las dos variables puede no ser necesariamente lineal. Por ejemplo si la relación es $Y = X^2$, r_S será cercano a 1 y el de *r* cercano a 0.

Comparar los dos coeficientes de correlación resulta una estrategia de análisis interesante. Cuando $r_S > r$ indica la existencia de pares de valores extremos, mientras que si $r > r_S$ indica la presencia de unos pocos valores extremos.

El procedimiento para el cálculo de r_S requiere **asignar rangos a *X* y rangos a *Y* de manera independiente** dado que se trata de variables independientes (en los procedimientos previos la variable es una sola y por lo tanto los rangos se asignan en un solo ranking).

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{9.18}$$

donde d_i es la diferencia de lós rangos de X y de Y.

Cuando existen **valores empatados** en X o en Y se emplea la expresión

$$r_S = \frac{(N^3-N)/6 - \sum_{i=1}^{n} d_i^2 - \sum T_x - \sum T_y}{\sqrt{\left[(N^3-N)/6 - 2\sum T_x\right]\left[(N^3-N)/6 - 2\sum T_y\right]}}, \tag{9.19}$$

donde $\sum T_x = \frac{\sum_1^m (t_i^3 - t_i)}{12}$ para t_i = número de empates en el i-ésimo grupo de empates de X y

$\sum T_y = \frac{\sum_1^m (t_i^3 - t_i)}{12}$ para t_i = número de empates en el i-ésimo grupo de empates de Y.

En la Tabla 14 del Anexo se encuentras los valores críticos de r_S para $\nu = n$ grados de libertad que permiten testear las hipótesis H_0: $\rho_S = 0$ y H_A: $\rho_S \neq 0$. La hipótesis nula se rechaza cuando $r_S > r_{S\,crítico}$.

EJEMPLO 8

Coeficiente de correlación de Spearman

En un estudio sedimentológico se desea investigar en los aglomerados que se encuentran en las costas del Nahuel Huapi como es la relación entre el ancho y el largo de los clastos. Se eligen 12 clastos al azar y se miden dos de sus ejes mayores ortogonales, A (largo) y B (ancho).

	Largo		Ancho		
x_i	*Rangos* x_i	y_i	*Rangos* y_i	d_i	d_i^2
5,20	4	3,70	5	-1	1
5,40	8,5^{++}	3,80	7	1,5	2,25
5,55	10	3,95	11	-1	1
5,10	1,5^{+}	3,60	2,5^{++}	-1	1
5,15	3	3,70	5	-2	4
5,10	1,5^{+}	3,55	1	0,5	0,25
5,35	7	3,70	5	2	4
5,25	5	3,60	2,5^{++}	2,5	6,25
5,40	8,5^{++}	3,90	9,5^{+}	-1	1
5,60	11	3,85	8	3	9
5,30	6	3,90	9,5^{+}	-3,5	12,25
5,70	12	4,15	12	0	0

$^{+}$ y $^{++}$ Rangos empatados.

$\sum d_i^2 = 42$,
$n = 12$

Hay 2 empates de 2 rangos en X y 2 empates de 2 rangos en Y.

$\sum T_x = \frac{(2^3-2)+(2^3-2)}{12} = 1$

$\sum T_y = \frac{(2^3-2)+(2^3-2)}{12} = 1$

$$r_S = \frac{(n^3-n)/6 - \sum_{i=1}^{n} d_i^2 - \sum T_x - \sum T_y}{\sqrt{\left[(n^3-n)/6 - 2\sum T_x\right]\left[(n^3-n)/6 - 2\sum T_y\right]}} = \frac{\left(12^3-12\right)/6 - 42 - 1 - 1}{\sqrt{\left[(12^3-12)/6 - 2\cdot 1\right]\left[(12^3-12)/6 - 2\cdot 1\right]}} = \frac{242}{284} = 0{,}852$$

H_0: $\rho_S = 0$
H_A: $\rho_S \neq 0$
$\alpha = 0{,}05$

De la Tabla 14 del Anexo, $r_{S\ (0,05;\ 2;\ 12)} = 0,587$

Existen evidencias para rechazar la H_0 ($r_S > r_{S(0,05;\ 2;\ 12)}$, 0,852 > 0,587). Se puede afirmar que existe una correlación positiva entre el largo y el ancho, a medida que aumenta el largo lo hace también el ancho, los clastos más largos son también los más anchos.

LA DISTRIBUCIÓN LOG-NORMAL

Introducción

La distribución log-normal se caracteriza por presentar una variable con mayor número de datos en torno a valores bajos y unos pocos datos hacia los valores más altos, de modo que es asimétrica (Fig. 1). Esta distribución describe adecuadamente datos originados por muchos procesos físicos, químicos, biológicos, toxicológicos, económicos (ingreso per capta, costos de inmuebles), epidemiológicos (tiempo de incubación de enfermedades, tiempo de supervivencia de enfermos de cáncer o HIV). Algunos de los procesos que se adecuan al modelo log-normal propios de las ciencias de la tierra son las leyes de muchos elementos traza, la trasmisividad de un acuífero, propiedades edáficas, dilución de sustancias en otro material, etcétera.

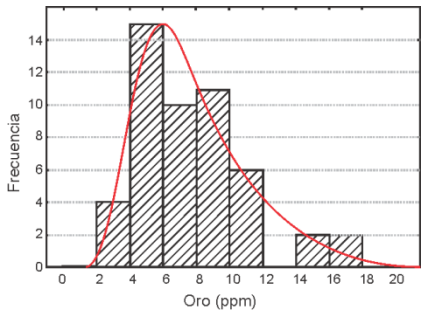


Figura 1. Histograma de concentraciones leyes de oro (ppm) en vetas. La ley promedio de estos datos es 7,6 ppm de Au.. Las leyes altas ejercen mucha influencia sobre el ley promedio.

Como se ha visto en capítulos anteriores muchas pruebas requieren normalidad y la transformación logarítmica suele ser la primera a que se recurre para lograrla. Operar con escalas logarítmicas no conlleva ninguna dificultad. Entre las escalas logarítmicas más usadas por los geólogos se encuentra el pH, la escala sismológica de Richter y la escala granulométrica Phi.

La distribución log-normal debe su nombre a que los logaritmos de los datos se distribuyen según el modelo normal. Una variable aleatoria *x* sigue un modelo log-normal si su logaritmo sigue una ley normal. Si se llama μ_x a la media y σ_x a la desviación estándar de *x* respectivamente, que *x* sea log-normal se expresa como $LN(\mu_x, \sigma_x^2)$. Por otra parte, si la variable *y* es igual al logaritmo natural de *x*:

$$y = \ln(x), \qquad (10.1)$$

entonces y está normalmente distribuida con media μ_y y desviación estándar σ_y y se expresa como $N(\mu_y, \sigma_y^2)$.

A partir de esta definición se puede demostrar que la función de densidad tiene por expresión:

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{\ln(x)-\mu_y}{\sigma_y}\right]^2}. \qquad (10.2)$$

Así como existen infinidad de distribuciones normales, dependiendo de los valores que tomen sus parámetros μ y σ^2, existen infinitas distribuciones log-normales dependiendo de los valores que tomen sus parámetros $LN(\mu_x, \sigma_x^2)$. Las distribuciones log-normales son siempre asimétricas con asimetría a la derecha, el grado de su asimetría depende de la varianza del logaritmo de las observaciones. Si el valor de la varianza es pequeño, entonces la asimetría y la distribución de frecuencias se aproxima a la normal (Fig.2.).

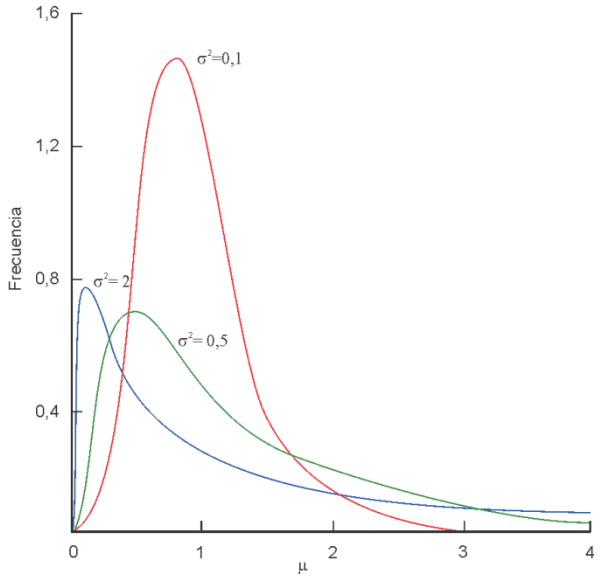


Figura 2. Modelos log-normal de la variable X. Observe como cambia la forma de la distribución a medida que cambia la varianza.

Estimación de parámetros

Aitchison y Brown (1957) notaron que cuando se conocen la media y desvío estándar poblacional de y ($y = \ln(x)$), μ_y, σ_y^2 se puede conocer la correspondiente media y varianza para x. En forma similar cuando se conoce μ_x, σ_x^2 para x (la variable sin transformar), se puede conocer la correspondiente media y varianza de y aplicando la expresión 10.1. En el cuadro 1 se presentan las expresiones que permiten realizar los cálculos.

Conocida *y* (ln(*x*))	Conocida *x* (la variable log-normal)
$\mu_x = e^{\mu_y + \frac{1}{2}\sigma_y^2}$ (10.3)	$\mu_y = \ln\left(\frac{\mu_x^2}{\sqrt{\mu_x^2 + \sigma_x^2}}\right)$ (10.5)
$\sigma_y^2 = e^{2\mu_y + \sigma_y^2}\left(e^{\sigma_y^2} - 1\right)$ (10.4)	$\sigma_y^2 = \ln\left(1 + \frac{\sigma_x^2}{\mu_x^2}\right)$ (10.6)

Cuadro 1. Expresiones que permiten calcular la media y desvío estándar poblacional de x *e* y.

Sin embargo, surgen problemas a la hora de estimar los parámetros μ_x, σ_x^2 a través de sus correspondientes estadísticos. Recuerde el teorema Central del límite, la media muestral $\bar{X}$ es un estimador insesgado de la media poblacional μ. De la misma forma la varianza muestral s^2 es un estimador insesgado de la varianza poblacional σ^2. Desafortunadamente cuando se trata de distribuciones log-normales estas estimaciones no son muy eficientes. Se ha descrito que las estimaciones más eficientes se obtienen primero estimando los parámetros μ_y, σ_y y luego estimando μ_x, σ_x, pues dado que los logaritmos de *x* están normalmente distribuidos, la media y varianza muestral de *y* resultan buenos estimadores de los parámetros poblacionales μ_y y σ_y respectivamente. Entonces, para estimar μ_y se calcula el promedio de los logaritmos de *x* (10.7) y para estimar σ_y^2 se calcula la varianza muestral de los logaritmos *x* (10.8).

$$\bar{X}_y = \frac{\sum_{i=1}^{n} \ln(x)}{n}, \qquad s_y^2 = \frac{\sum_{i=1}^{n} (\ln(x) - \bar{X}_y)^2}{n-1}. \qquad \text{(10.7 y 10.8)}$$

Para tener estimadores eficientes de la media y la varianza poblacional de *x* (μ_x, σ_x^2) a partir de *y* se utilizan las expresiones 10.9 y 10.10.

$$\bar{X}_x = e^{\bar{X}_y} \psi_n\left(\frac{1}{2} s_y^2\right), \qquad s_x^2 = e^{2\bar{X}_y} \phi_n(s_y^2) \qquad \text{(10.9 y 10.10)}$$

Los valores de $\psi_n = \left(\frac{1}{2} s_y^2\right)$ se buscan en la Tabla 17 del Anexo entrando con $T = \left(\frac{1}{2} s_y^2\right)$ y *n*, el tamaño de la muestra. Para hallar $\phi_n(s_y^2)$ se busca en la misma tabla entrando con $T = (s_y^2)$ y *n*, el tamaño de la muestra. Si los valores de *T* no se encuentran en la tabla se interpola linealmente con los valores más cercanos.

Sin embargo, si los datos no son exactamente log-normales las estimaciones de los parámetros poblacionales pueden conducir a un sesgo[16]. La media muestral $\overline{X_x}$ tiene una eficiencia de más del 90% en la estima de μ, en distribuciones con el coeficiente de variación es menor que 1,2 (corresponde a $s_y^2 \approx 0{,}9$). Por esta razón se recomienda usar $\overline{X_y}$ para estimar μ cuando el coeficiente

de variación de la distribución es menor a 1,2. El gráfico de la figura 3 permite calcular la eficiencia de la estimación usando el coeficiente de variación.

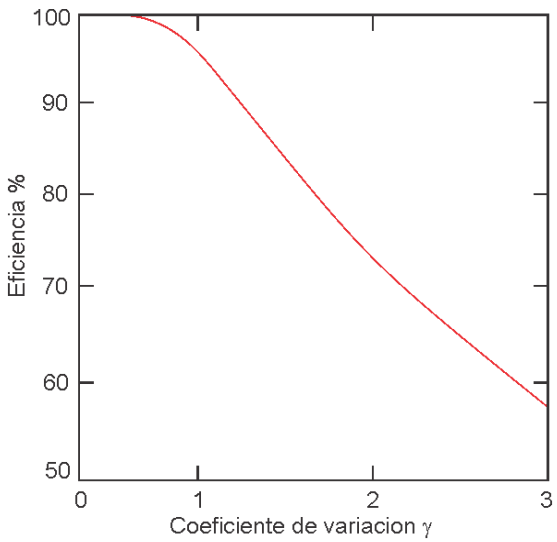


Figura 3. Gráfico que permite calcular la eficiencia de $\overline{X}_x$ para estimar μ conocido el coeficiente de variación.

El coeficiente de variación de la distribución adquiere un nuevo rol. El coeficiente de variación muestral, *CV*, es una estima del poblacional, γ, se puede hallar de varias maneras:

a) utilizando la varianza del ln(*x*) con la siguiente expresión

$$\gamma = \sqrt{e^{s_y^2} - 1}\,; \qquad (10.11)$$

b) a través del cociente entre el desvío estándar muestral y media muestral de la variable en escala logarítmica, *y*

$$\gamma = \frac{s_y}{\overline{X}_y}\,; \qquad (10.12)$$

c) utilizando el cociente entre el desvío estándar muestral y media muestral de la variable en escala aritmética, *x*

$$\gamma = \frac{s_x}{\overline{X}_x}\,. \qquad (10.13)$$

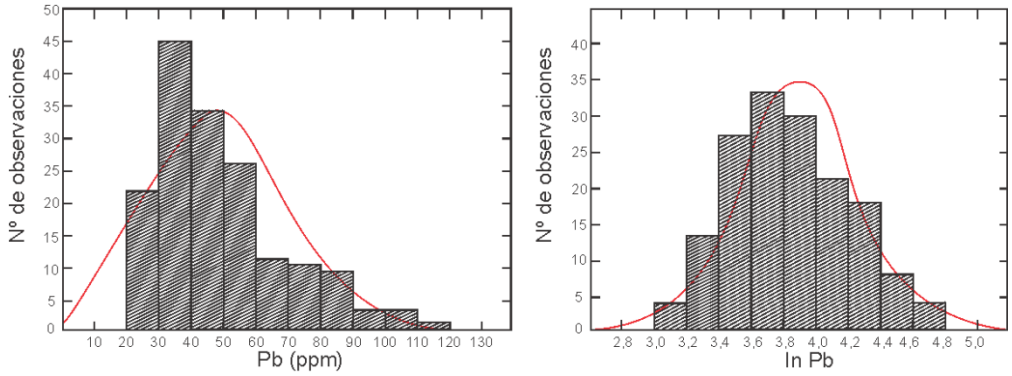
Cabe aclarar que las estimaciones obtenidas con estas expresiones difieren ligeramente entre sí.

EJEMPLO 1

Grado de eficiencia de la media en escala aritmética para estimar la media poblacional

La concentración de Plomo (ppm) en suelos de prados de un sector de los Jura suizos son log-normales (datos tomados de Goovaerts, 1997).

Se trata de 159 datos cuya distribución de frecuencias en escala aritmética y logarítmica se muestran en los histogramas de la figura de la izquierda y derecha respectivamente.



1° Se calculan los estadísticos en escala logarítmica

$$\overline{X_y} = \frac{\sum_{i=1}^{n} \ln(x)}{n} = 3{,}845$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (\ln(x) - \overline{X_y})^2}{n-1} = 0{,}133$$

2° Se obtienen los estimadores eficientes en escala aritmética,

$$\overline{X}_x = e^{\overline{X}_y} \psi_n\left(\frac{1}{2} s_y^2\right) = e^{3{,}845} 1{,}050 = 49{,}08 \quad \text{y} \quad s_x^2 = e^{2\overline{X}_y} \phi_n(s_y^2) = e^{2 \cdot 3{,}845} 1{,}128 = 2464{,}75 .$$

De la Tabla 17 del Anexo se obtienen $\psi_n = \left(\frac{1}{2} s_y^2\right)$ entrando con n = 50 y $T = \left(\frac{1}{2} s_y^2\right) = 0{,}066$ y $\phi_n(s_y^2)$ entrando con $T = (s_y^2) = 0{,}133$.
Ambas cantidades se obtuvieron por interpolación lineal entre los valores más cercanos, se uso n = 50 en lugar de n = 159 porque es el máximo valor tabulado.

3° Se calcula el coeficiente de variación poblacional,

$$\gamma = \sqrt{e^{s_y^2} - 1} = \sqrt{e^{0{,}133} - 1} = 0{,}377 .$$

La eficiencia de la media aritmética para estimar la media poblacional es aproximadamente 100% (Fig. 3). La media aritmética de la distribución de Pb es 50 ppm y la estimada es 49,08 ppm.

Inferencia

Pruebas de bondad de ajuste

Para evaluar la log-normalidad de un conjunto de datos se puede recurrir al método gráfico o las pruebas de bondad de ajuste de χ^2, Kolmogorov - Smirnov o Lillifords descriptos en el Capítulo 6. En todos los casos el procedimiento más simple es log transformar los datos y realizar las pruebas de normalidad. También se puede calcular el coeficiente de simetría y de curtosis y comparar los valores con los del modelo normal.

Límites de confianza

Para la media poblacional de X, conocida la varianza poblacional de Y

Cuando se conoce la varianza poblacional del logaritmo de la variable *y* (σ_Y^2) se pueden calcular los límites de confianza de μ_X siguiendo algunos pasos. Primero se calculan los límites en escala logarítmica (*y*)

$$p\left(\bar{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_Y^2}{n}} < \mu_Y < \bar{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_Y^2}{n}}\right) = 1 - \alpha. \quad (10.14)$$

Luego se transformar los límites a escala aritmética

$$antilogaritmo\ de\ \mu_{Yinf} = e^{\mu_{Yinf}}, \quad (10.15)$$

$$antilogaritmo\ de\ \mu_{Ysup} = e^{\mu_{Ysup}}. \quad (10.16)$$

Por último se expresar los límites en escala aritmética

$$p\left(e^{\mu_{Yinf}} - \frac{1}{2}\sigma_Y^2 < \mu_X < e^{\mu_{Ysup}} + \frac{1}{2}\sigma_Y^2\right) = 1 - \alpha. \quad (10.17)$$

Para la media poblacional de X, cuando no se conoce la varianza poblacional de Y

El cálculo de los límites de μ_X cuando se desconoce σ_Y^2 se hace también en etapas. Primero se calculan los límites de confianza de la varianza poblacional en escala logarítmica, σ_Y^2 (ver Capítulo 6)

$$p\left(\frac{(n-1)S_Y^2}{\chi^2_{1-\alpha/2}} < \sigma_Y^2 < \frac{(n-1)S_Y^2}{\chi^2_{\alpha/2}}\right) = 1 - \alpha \quad (10.18)$$

Los valores de χ^2 se encuentran en la Tabla 2 del Anexo.

Lugo se calcula los límites de μ_X utilizando los límites inferior y superior de σ_Y^2

$$p\left(\bar{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_{Yinf}}{n}} < \mu_Y < \bar{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_{Ysup}}{n}}\right) = 1 - \alpha. \quad (10.19)$$

Finalmente se expresan los límites en escala aritmética

$$p\left(e^{\mu_{Yinf}+0{,}5\sigma^2_{Yinf}} < \mu_X < e^{\mu_{Ysup}+0{,}5\sigma^2_{Ysup}}\right) = 1 - \alpha. \quad (10.20)$$

EJEMPLO 2

Límites de confianza de la media poblacional

Se utilizan los datos de concentración de Plomo (ppm) en suelos de prados de un sector de los Jura suizos del ejemplo 1.

Datos

n = 159 $\quad \overline{X_y} = 3{,}845 \quad s_y^2 = 0{,}133$

α = 0,05

De la Taba 1 del Anexo, $z_{\alpha/2} = 1{,}645$

$\chi^2_{158;\,0{,}975} = 194{,}6885$

$\chi^2_{158;\,0{,}025} = 125{,}0967$

Primer paso

$$P\left(\frac{(n-1)S_Y^2}{\chi^2_{1-\alpha/2}} < \sigma_Y^2 < \frac{(n-1)S_Y^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha$$

$$P\left(\frac{159 \cdot 0{,}133}{194{,}6885} < \sigma_Y^2 < \frac{159 \cdot 0{,}133}{125{,}0967}\right) = 1-0{,}05$$

$$P(0{,}1076 < \sigma_Y^2 < 0{,}1675) = 0{,}95$$

Segundo paso

$$P\left(\bar{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_{Yinf}}{n}} < \mu_Y < \bar{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_{Ysup}}{n}}\right) = 1-\alpha$$

$$P\left(3{,}845 - 1{,}645\sqrt{\frac{0{,}1076}{159}} < \mu_Y < 3{,}845 + 1{,}645\sqrt{\frac{0{,}1076}{159}}\right) = 1-0{,}05$$

$$p(3{,}7937 < \mu_Y < 3{,}9083) = 1-0{,}05$$

Tercer paso

$$P\left(e^{\mu_{Yinf}+0{,}5\sigma^2_{Yinf}} < \mu_X < e^{\mu_{Ysup}+0{,}5\sigma^2_{Ysup}}\right) = 1-\alpha$$

$$P\left(e^{3{,}7937+(0.5\cdot 0{,}1076)} < \mu_X < e^{3{,}9038+(0.5\cdot 0{,}1675)}\right) = 1-0{,}05$$

$$P(42{,}09 < \mu_X < 54{,}16) = 0{,}95$$

Para la varianza poblacional σ_X^2

Los límites de la media y varianza poblacional calculados con las expresiones 10.18 y 10.19 permiten encontrar los límites de confianza de la varianza poblacional en escala aritmética *x*.

$$P\left(\mu^2_{Xinf}\left(e^{\sigma^2_{Yinf}} - 1\right) < \sigma_X^2 < \mu^2_{Xsup}\left(e^{\sigma^2_{Ysup}} - 1\right)\right) = 1-\alpha \qquad (10.21)$$

Prueba de igualdad de contenido medio de dos poblaciones log-normales

Cuando se quieren comparar el contenido medio de dos poblaciones cuyos valores se comportan según el modelo log-normal se puede utilizar un test de *t* pero se deben utilizar los estadísticos de *y* (*y* = log *x*) para no llegar a falsas conclusiones. De modo que el estadístico de prueba que se construye debe tener la siguiente forma:

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$

$$t = \frac{\overline{Y_1} - \overline{Y_2} + \frac{S^2_{Y1} - S^2_{Y2}}{2}}{\sqrt{\frac{S^2_{Y1}}{n_1} + \frac{S^2_{Y2}}{n_2} + \left(\frac{S^4_{Y1}}{n_1 - 1} + \frac{S^4_{Y2}}{n_2 - 1}\right)^{\frac{1}{2}}}} \qquad (10.22)$$

La hipótesis nula se rechaza cuando $t > t_{\alpha/2;\ (n1+n2-2)}$. Al igual que otras pruebas de diferencias de medias también se pueden formular hipótesis a una sola cola, inferior o superior, dependiendo del problema.

EJEMPLO 3

Prueba de igualdad de contenidos medios

Se analizan datos de un estudio hidroquímico ([Cl^-] meq/l) en aguas y salmueras de la Salinas Grandes, Córdoba. Se realizó un muestreo de agua en el área de aporte de ríos y vertientes y en el Complejo salino (lagos y lagunas). La variable es log-normal.

H_0: $\mu_R = \mu_L$

H_A: $\mu_R \neq \mu_L$

$n_R = 30$ $\qquad$ $n_L = 35$

$\overline{X_R} = 7{,}13$ $\qquad$ $\overline{X_L} = 11{,}00$

$S_R^2 = 1{,}43$ $\qquad$ $S_L^2 = 0{,}45$

$\alpha = 0{,}05$

De la Tabla 3 del Anexo, $t_{(0{,}025;\ 63)} = 2{,}00$

$$t = \frac{7{,}13 - 11 + \frac{1{,}43 - 0{,}45}{2}}{\sqrt{\frac{1{,}43}{30} + \frac{0{,}45}{35} + \left(\frac{2{,}05}{29} + \frac{0{,}20}{34}\right)^{\frac{1}{2}}}} = -5{,}82$$

La hipótesis nula se rechaza 5,82 > 2,00 ($t > t_{(0{,}025;\ 63)}$). Se infiere que la concentración de cloruros de ríos y vertientes es diferente a la de lagos y lagunas.

Correlación y regresión

En el octavo capítulo se mencionó que el **coeficiente de correlación *r*** se debe obtener muestreando al azar una población normal bivariada. Cuando no se cumple el supuesto de normalidad es posible que el coeficiente indique que no hay correlación aún cuando existe asociación entre las variables con el consecuente error en la interpretación geológica del resultado. Una situación bastante habitual se plantea cuando se quiere investigar si dos variables log-normales que se encuentran correlacionadas, en estos casos el coeficiente de correlación se debe calcular con los **log-valores** de los datos de ambas variables ya que de esta manera se obtiene un *r* que da cuenta de la relación que existe entre ellas. La expresión que permite calcular *r* es igual a la 8.1, pero aquí $w = \ln(x)$ y $z = \ln(y)$.

$$r = \frac{\sum_{i=1}^{n}(w - \overline{W})(z - \overline{Z})}{\sqrt{\sum_{i=1}^{n}(w - \overline{W})^2 \cdot \sum_{i=1}^{n}(z - \overline{Z})^2}} \qquad (10.23)$$

Existen situaciones donde se detectan relaciones no lineales entre un par de variables, ya sea a partir de los valores de r^2 o analizando el gráfico de dispersión. En algunas de ellas, transformando la variable dependiente, la independiente o ambas usando logaritmos se logra obtener relaciones más o menos lineales que se pueden estudiar utilizando la metodología de la regresión lineal descripta en el capítulo 8. Por ejemplo si la regresión es exponencial, logarítmica o potencial, aplicando logaritmos se logran regresiones lineales y se evita recurrir a los métodos de ajuste no lineales más complejos.

Si la **regresión es exponencial**

$$y = a \cdot e^{bx} \quad \rightarrow \quad \ln y = \ln a + bx. \qquad \text{(10.24 y 10.25)}$$

Cuando la **regresión es logarítmica** $\ln y = a + b \ln x.$ (10.26)

Por último, **para regresión potencial**

$$y = a + x^b \quad \rightarrow \quad \ln y = \ln a + b \ln x. \qquad \text{(10.27 y 10.28)}$$

ANÁLISIS DE SERIES DE DATOS

SERIES DE TIEMPO, SERIES CRONOLÓGICAS Y OBSERVACIONES SECUENCIALES

Introducción

En este capítulo se abordará el estudio de fenómenos geológicos unidireccionales que se producen en el tiempo o en el espacio. Se describen algunos métodos de análisis de datos que se recolectan en un cierto orden o secuencia y donde la ubicación del dato en la secuencia se tiene en cuenta en el estudio. Existen numerosos datos de este tipo en geología, por ejemplo: datos que son tomados de secciones verticales como espesores, curvas de variación de isótopos ($\delta^{18}O$; $\delta^{13}C$), datos de perfiles eléctricos de perforaciones (resistividad eléctrica, Gamma ray, etc.), registros sísmicos, salidas o corridas de un registro continuo (difractogramas de rayos-X), datos climáticos (precipitaciones, temperaturas), datos hidrológicos e hidrogeológicos (caudal, altura de nivel freático), datos ambientales (calidad del agua), historia de producción de un pozo o yacimiento, etcétera.

La estrategia de estudio de las secuencias recibe el nombre general de Análisis de Series de Tiempo. Una serie temporal, cronológica, histórica o de tiempo es una sucesión de observaciones cuantitativas de un fenómeno ordenadas en el tiempo. El término serie de tiempo fue utilizado primero para referirse a observaciones realizadas en forma ordenada y en intervalos de tiempo dados, como el caso de las mediciones de temperatura diaria. Si bien la definición menciona solo el tiempo, en los estudios geológicos se puede ampliar para incluir a sucesiones de observaciones ordenadas en el espacio[17]. En algunas series la variable que se estudia es medida en una escala de intervalo (i.e. temperatura) o de razón (i.e. caudal) en tanto la escala de medida de la variable a lo largo de la cual se toman los datos suele ser una escala nominal (días, años, metros, etc.).

Por otro lado, existen otros tipos de series que requieren métodos de análisis diferentes al que comúnmente se entiende por análisis de series de tiempo. Son los casos de datos cuya escala de medida es nominal, como las que expresan las repeticiones rítmicas de los tipos litológicos en una columna estratigráfica o las de datos binomiales, como presencia/ausencia de estructuras o restos fósiles en una secuencia estratigráfica.

Naturalmente el tipo de variable y el objetivo del análisis determinan cuál debe ser el método que se utilice. El Cuadro 1 muestra los métodos que se abordarán en este capítulo.

VARIABLE MEDIDA EN ESCALA:	OBSERVACIONES CON ESPACIADO REGULAR	NO SE CONSIDERA EL ESPACIAMIENTO
DE RAZÓN O DE INTERVALOS	Análisis de Series de Tiempo	Auto-correlación
	Auto-correlación	Correlación cruzada
	Correlación cruzada	
ORDINAL O NOMINAL	Test de rachas	Test de rachas
	Auto-asociación y Asociación cruzada	Auto-asociación y Asociación cruzada
	Matrices de transición	Matrices de transición
	Cadenas de Markov	Cadenas de Markov

Cuadro 1. Métodos de análisis de series.

Series de tiempo

Se llama serie de tiempo o cronológica al conjunto de datos que surgen del registro de observaciones de una variable cuantitativa en función del tiempo. Las series pueden ser de dos tipos: evolutivas o estacionarias. En las series evolutivas el valor medio de la serie cambia a través del tiempo, en cambio, en las series estacionarias el valor medio permanece constante aunque se presenta variaciones en torno a ese valor.

La metodología de análisis que se describe en este apartado permite por una parte estimar los factores (o componentes) que producen el comportamiento general o patrón de la serie, y por otra usar las estimaciones para predecir el comportamiento futuro de la serie.

Componentes de una serie cronológica

En general se considera que las series cronológicas están formadas por cuatro componentes: la tendencia, la variación estacional, la variación cíclica y la variación irregular también llamada errática o residuo.

La **tendencia** (*T*) es el movimiento de una serie cronológica en un periodo largo, o en otras palabras, su comportamiento a largo plazo. El comportamiento puede ser de tipo estacionario o constante, lineal, parabólico, exponencial, logístico, etcétera. Por ejemplo la tendencia hacia el aumento de temperatura del planeta o la disminución de la porosidad con el aumento de la profundidad (Fig. 1).

Las **variaciones estacionales** (*VE*) representan las fluctuaciones que se repiten en periodos iguales o inferiores a un año. Su nombre proviene de las estaciones climatológicas (otoño, invierno, primavera y verano). Sin embargo aunque la periodicidad generalmente es el año, puede ser el mes, la semana o incluso el día dependiendo de los datos (Fig. 1).

El **componente cíclico** (*C*) representa el patrón de comportamiento que se repite en periodos de diferente duración, mayores a un año. Suelen ser más irregulares que las variaciones estacionales. La amplitud de los ciclos se mide en años. Por ejemplo periodos de crecidas o sequias en el caudal de un río (Fig. 1).

Las **variaciones irregulares** (*R*) pueden ser el producto de variaciones naturales aleatorias. Se producen en la serie de forma aislada y no permanente de modo que es prácticamente imposible preverlas. Se estima que su media es cero, su varianza es constante en el tiempo y se distribuyen según un modelo normal ($E \sim N(0, \sigma^2)$).

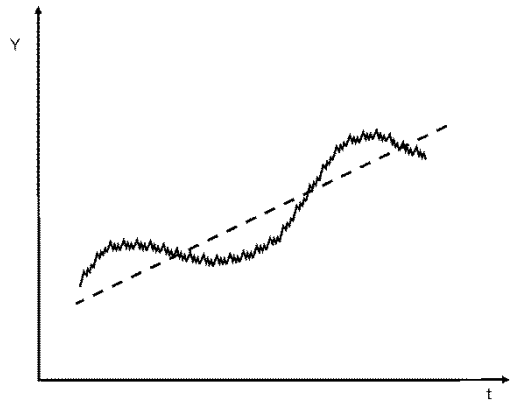


Figura 1. Componentes de una serie de tiempo. Serie con tendencia, ciclo y variaciones estacionales.

Modelos de la serie cronológica

Uno de los objetivos del estudio de una serie es evaluar que parte del valor de la variable dependiente y_t, para cualquier momento dado, puede atribuirse a la tendencia, a los factores estacionales, a los factores cíclicos o a una variación aleatoria. Para aislar las componentes correctamente se debe conocer la forma en que están relacionados en la población que se investiga. Existen algunos modelos series cronológicas.

Los modelos básicos son dos, según los componentes se expresen en términos de suma (modelo aditivo) o de producto (modelo multiplicativo).

En el **modelo aditivo** se considera que y_t es igual a la suma de los cuatro componentes:

$$y_t = T_t + VE_t + C_t + R_t. \qquad (11.1)$$

Al suponer que los componentes son aditivos se admite que son independientes entre sí. Entonces, por ejemplo, la tendencia no puede afectar ni a la variación estacional ni a la cíclica, ni estos componentes afectan a la tendencia.

En el **modelo multiplicativo** y_t se expresa en la forma de producto de componentes:

$$y_t = T_t \cdot VE_t \cdot C_t \cdot R_t. \qquad (11.2)$$

En este modelo se considera que las cuatro componentes están relacionadas entre sí, si bien se mantiene la posibilidad de que los componentes provengan de causas básicas diferentes.

Para conocer a cual modelo se ajusta una serie se calculan la Diferencia Estacional (*DE* = la diferencia entre un dato y el dato anterior) y el Cociente Estacional (*CE* = el cociente entre un dato y el anterior) para todos los pares de datos.

$DE = y_{t,i} - y_{t-1,i}$ $y_{i,t}$ = *y para el momento t* (11.3)

$CE = y_{t,i}/y_{t-1,i}$ y_{t-1} = *y para el dato equivalente del año siguiente* (11.4)

Luego se calculan los coeficientes de variación de *DE* y *CE*

$CV_{DE} = \frac{S_{DE}}{\bar{X}_{DE}}$, $CV_{CE} = \frac{S_{CE}}{\bar{X}_{CE}}$.

Cuando $CV_{CE} > CV_{DE}$ es indicio de un modelo aditivo. Mientras que si $CV_{CE} < CV_{DE}$ indicaría un modelo multiplicativo.

Análisis de la serie

El análisis de una serie cronológica consiste en estimar primero la tendencia y luego eliminar la variación que se debe a dicha tendencia. La variación restante se atribuye entonces a los factores *VE*, *C* y *R*. Luego se estiman y aíslan la variación de cada uno de estos componentes siguiendo con las *VE*, *C* y por último las variaciones *R*. Solamente después de haber estimado todos los componentes de una serie cronológica, es posible hacer predicciones del valor de la variable en algún punto del futuro comenzando a estimar primero a través de la tendencia y luego ajustando el valor con las otras componentes.

Determinación de la Tendencia (Trend Lineal)

El primer paso en el análisis de una serie consiste en graficar los datos en un sistema cartesiano en el que el tiempo siempre se pone en las abscisas y los valores de la serie, y_t, en las ordenadas. El diagrama permite conocer las características de la serie, tendencia, oscilaciones, valores anómalos, etcétera.

La Tabla 1 y contiene los datos de ancho de playa medido desde la línea de bajamar hasta la base espaldón durante un muestreo mensual efectuado en el periodo 2007-2011 en un perfil de playa de la provincia de Buenos Aires (Fig. 2) que se usaran para ejemplificar cada método.

	Ene.	*Feb.*	*Mar.*	*Abr.*	*May.*	*Jun.*	*Jul.*	*Ago.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
2007	225,6	226,2	222,7	218,1	212,3	209,1	200,9	198,9	204,0	207,3	209,3	207,5
2008	209,1	206,7	203,9	196,8	190,2	186,0	178,6	177,7	178,8	180,3	180,4	178,5
2009	180,4	178,3	175,7	170,8	164,9	161,2	155,1	155,4	157,0	159,1	162,3	161,3
2010	153,9	154,5	151,0	146,4	140,6	137,4	129,2	127,2	132,3	135,6	137,6	135,8
2011	137,4	135,0	132,2	125,1	118,5	114,3	106,9	106,0	107,1	108,6	108,7	106,8

Tabla 1. Ancho de playa (m) para el periodo 2007-2011.

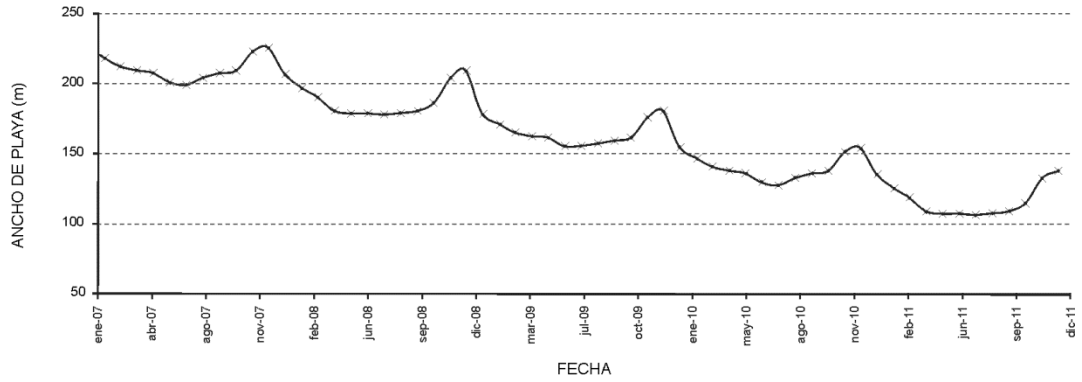


Figura 2. Variaciones del ancho de playa para el periodo 2006-2011.

El paso siguiente consiste en estimar la tendencia. Existen varios procedimientos que se pueden utilizar, se describen dos, el método analítico y el método de las medias móviles.

Método analítico

Este método de obtención de la tendencia tiene por objetivo seleccionar la función matemática que modele las variaciones a largo plazo de la serie. Como ya se mencionó la tendencia puede ser lineal, exponencial, logística, etc., se explicará solamente como obtener una función lineal.

Tal como ocurre en el análisis de regresión, primero se debe identificar si la tendencia es lineal. Cabe aclarar, sin embargo, que en el caso de series de tiempo el análisis de regresión no tiene por objetivo encontrar la capacidad explicativa del tiempo, sino modelar las variaciones que sufre la variable. El tiempo no explica nada, se trata solo del soporte en el que se mueve la serie. Dicho de otra forma, las variaciones de la variable en el tiempo no son producidas por el paso del tiempo sino que son causadas por otras variables que inciden en ella. Por ejemplo las variaciones en el caudal de un río no son función del tiempo sino que están reguladas por las precipitaciones y la temperatura del clima.

Si se acepta que se trata de una tendencia lineal ($y = a + b \cdot t$, dónde t es una medida del tiempo cronológico, a es la ordenada al origen y b la pendiente de la recta), se utiliza algún método para calcular la recta como el de Mínimos Cuadrados utilizado en el capítulo 8. Recuerde que para determinar los valores de a y b con éste método es necesario resolver las dos ecuaciones siguientes:

$$b = \frac{SCty}{SCt} \qquad SCt = \sum t^2 - \frac{\left(\sum t\right)^2}{n} \qquad SCty = \sum ty - \frac{\sum t \sum y}{n} \qquad a = \bar{y} - b\,\bar{t}$$

dónde y es la variable y n el número de datos de la serie.

Dado que el tiempo es continuo, no hay diferencias si se expresa en cualquier tipo de unidad, además, debido a que no tiene influencia en el análisis la asignación del valor $t = 0$ a cualquier punto hay libertad para elegir cuál es el origen de la serie.

Cuando se desconoce el punto de origen se suele asignar el 0 al punto central a través de una de las siguientes reglas:

a) Si el número de observaciones es **impar**, *t* se asigna al valor ubicado en el centro. Entonces los datos anteriores a *t = 0* se denotan por ... ,-3, -2, -1 y los posteriores por 1, 2 , 3, ...

b) Si el número de periodos es **par**, no hay punto medio, el valor *t* se toma entre la mitad de las dos medidas centrales, de manera que *t* = 0 se omite. Luego, los periodos anteriores a *t = 0* se asignan como ... ,-5, -3, -1 y los posteriores como 1, 3, 5, ...

Las ventajas de este procedimiento de codificación son que la media de los valores de *t* es igual a 0, además el punto central de cada periodo estará representado por un valor entero de *t*.

Obtener la tendencia mediante el método analítico permite tener una medida de la bondad del ajuste y determinar si la función elegida es correcta. Por otra parte, tener la función posibilita realizar predicciones. La medida de la bondad del ajuste se obtiene con el coeficiente de determinación r^2. En las series estacionarias la pendiente de la recta se aproxima a 0 y si además presentan poca variabilidad el estimador de la serie es simplemente la media.

Todos los resultados, así como la información previa se pueden expresar en forma resumida en los siguientes términos:

$$y_t^* = a + bt \qquad r^2 = \ldots \ (t_0 \text{ en } \underline{\text{fecha}} \text{ ; unidad temporal ...})$$

EJEMPLO 1

Estimación de la tendencia

Para los datos del ancho de playa la recta estimada por el método de mínimos cuadrados es:

$$y_t^* = 222{,}6\ m - 1{,}9t \qquad r^2 = 0{,}8791 \quad (t_0\text{: enero 2006; unidad temporal: mes})$$

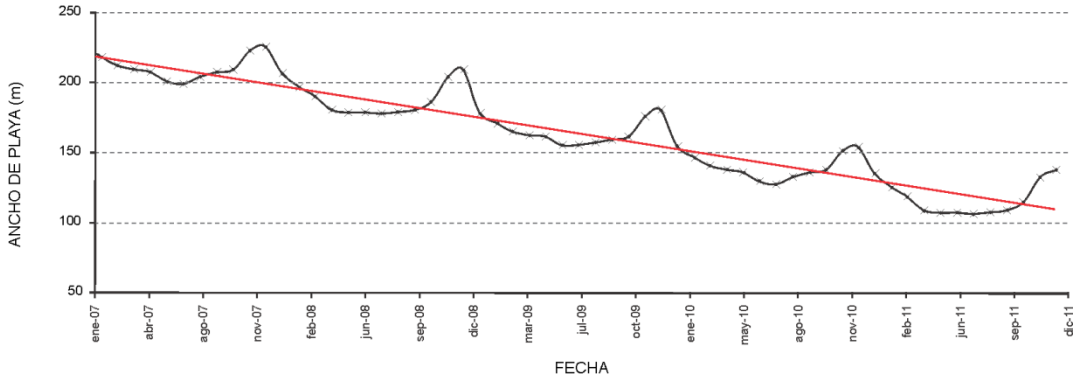


Figura 3. Tendencia del ancho de playa para el periodo 2006-2011. Recta calculada $y_t^* = 222{,}6\ m - 1{,}9t$.

En la playa monitoreada existe una tendencia lineal significativa con pendiente negativa (Fig. 3). Esto permite interpretar que el ancho de playa tiende a decrecer en el periodo considerado. Sin embargo, debido a que el lapso monitoreado es corto en términos de series temporales, no es posible asegurar que esta tendencia prosiga a futuro.

Eliminación de la tendencia

Para eliminar la tendencia de la serie partiendo de un modelo multiplicativo $y_t = T_t\ VE_t\ C_t\ R_t$, se calcula el cociente entre el valor observado en el momento t y el valor estimado con la recta de tendencia calculada. De este modo lo que se obtiene es:

$$\frac{T_t\ VE_t\ C_t\ R_t}{T_t} = VE_t\ C_t\ R_t = \frac{y_t}{y_t^*}, \quad y_t^* = y_t\ estimado\ con\ la\ recta\ de\ regresión\ calculada. \qquad (11.5)$$

Para modelos aditivos la tendencia se elimina por sustracción.

EJEMPLO 2

Eliminación de la tendencia

Para eliminar la tendencia de la serie de datos del ancho de playa bonaerense (Tabla 1) se calculan, para cada fecha de muestro, el ancho estimado con la ecuación de la recta de regresión calculada ($y_t^* = 222{,}6\ m - 1{,}9t$), (Tabla 2).

	Ene.	*Feb.*	*Mar.*	*Abr.*	*May.*	*Jun.*	*Jul.*	*Ago.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
2007	218,8	216,9	215,0	213,1	211,2	209,3	207,4	205,5	203,6	201,7	199,8	197,9
2008	196,0	194,1	192,2	190,3	188,4	186,5	184,6	182,7	180,8	178,9	177,0	175,1
2009	173,2	171,3	169,4	167,5	165,6	163,7	161,8	159,9	158,0	156,1	154,2	152,3
2010	150,4	148,5	146,6	144,7	142,8	140,9	139,0	137,1	135,2	133,3	131,4	129,5
2011	127,6	125,7	123,8	121,9	120,0	118,1	116,2	114,3	112,4	110,5	108,6	107,7

Tabla 2: Valores de ancho de playa estimados a partir de la recta de regresión calculada $y_t^* = 222{,}6\ m - 1{,}9t$

Luego, se calcula el cociente entre el ancho observado y el ancho estimado $\frac{y_t}{y_t^*} = VE_t\ C_t\ R_t$ (Tabla 3, Fig. 4). De este modo se obtiene una serie que solo posee las variaciones estacionales, las cíclicas y las residuales.

	Ene.	*Feb.*	*Mar.*	*Abr.*	*May.*	*Jun.*	*Jul.*	*Ago.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
2007	1,03	1,01	0,99	0,98	0,98	0,96	0,96	0,99	1,02	1,04	1,11	1,14
2008	1,05	1,01	0,99	0,95	0,95	0,96	0,96	0,98	1,00	1,04	1,15	1,19
2009	1,03	1,00	0,97	0,97	0,97	0,95	0,96	0,98	1,01	1,03	1,14	1,18
2010	1,03	0,99	0,96	0,95	0,95	0,92	0,92	0,97	1,00	1,03	1,15	1,19
2011	1,06	1,00	0,96	0,89	0,89	0,91	0,91	0,94	0,97	1,03	1,22	1,28

Tabla 3. Valores de ancho de playa sin tendencia (y_t/y_t^*)

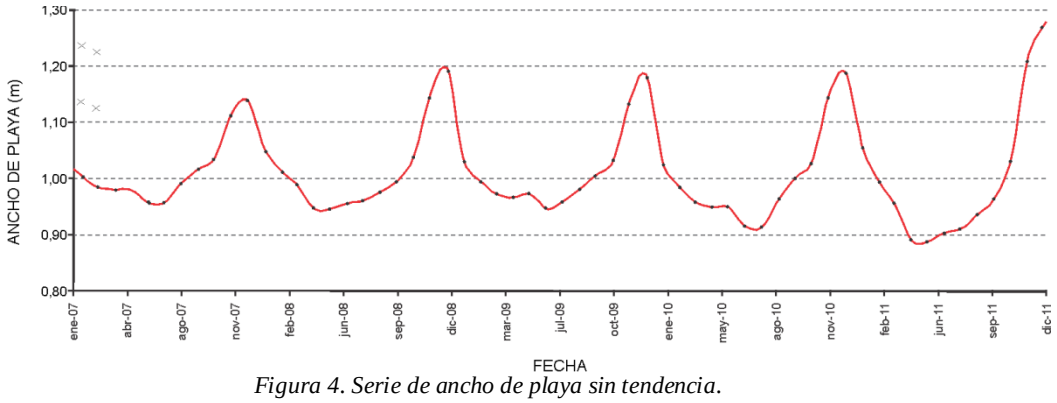


Figura 4. Serie de ancho de playa sin tendencia.

En la figura 4 se aprecia claramente un patrón que se repite año a año en el que el ancho de playa entre febrero y septiembre se encuentra debajo de los valores medios y entre octubre y enero por sobre los valores medios.

Método de las medias móviles

Este método sirve para suavizar la serie calculando promedios de valores de la misma para periodos de tiempo fijo pero que se desplazan a lo largo de toda la serie. En el vocabulario geofísico y de los ingenieros eléctricos se habla de eliminar el ruido (*noisy*) y de aplicar filtros. Es un método útil para analizar registros sísmicos y otros registros continuos como los difractogramas de rayos-X o los registros eléctricos de pozo. Con este procedimiento solo queda la tendencia ya que elimina los movimientos de corto y medio plazo y las anomalías debidas a causas impredecibles.

Un promedio móvil es la media aritmética de un conjunto de *k* valores consecutivos de la serie, el único requisito es que *k* sea menor que número total de datos.

Si el número ***k* de valores es impar** la media y_t^*, está centrada y se la hace corresponder con el valor del momento *t* que es el valor central de la suma. El promedio móvil es entonces:

$$y_t^* = \frac{\sum_{i=\frac{k-1}{2}}^{\frac{k-1}{2}} y_{t+i}}{k} = \frac{y_{t-\frac{k-1}{2}} + y_{t-\frac{k-1}{2}+1} + \cdots + y_t + \cdots + y_{t+\frac{k-1}{2}-1} + y_{t+\frac{k-1}{2}}}{k} \qquad (11.6)$$

Según esta expresión la primera media que se puede calcular es aquella correspondiente a un conjunto de *k* valores cuyo valor central coincide con el dato $t = (k - 1)/2$, y_t^*. La siguiente media se calcula para los *k* valores que tienen como dato central $t + 1$, y_{t+1}^*, y así sucesivamente. En otras palabras, a partir de la segunda media se elimina el primer dato y se agrega el siguiente de la serie.

Un ejemplo para $k = 3$, las sucesivas medias serán:

$$y_1^* = \frac{y_0 + y_1 + y_2}{3},$$

$$y_2^* = \frac{y_1 + y_2 + y_3}{3},$$

$$y_3^* = \frac{y_2 + y_3 + y_4}{3},$$

centrados en y_1, y_2 e y_3 respectivamente.

Si el número ***k* de valores es par** la media y_t^*, no está centrada y no corresponde a ningún valor observado de la serie original, sino a un punto medio entre dos valores. Por esa razón hay que promediar las medias calculadas de dos en dos sucesivamente para lograr una serie de medias centradas en un valor que se corresponda a un dato de la serie. Un ejemplo para $k = 4$ será de la forma:

$$y_{2|3}^* = \frac{y_1 + y_2 + y_3 + y_4}{4},$$

$$y_{3|4}^* = \frac{y_2 + y_3 + y_4 + y_5}{4},$$

$$y_3^* = \frac{y_{2|3}^* + y_{3|4}^*}{2}.$$

Un aspecto importante es la cantidad de datos que se deben incluir en la media móvil para lograr un buen resultado. Si bien cuanto mayor es el número de términos incluidos en el cálculo de las medias móviles mayor es el alisamiento, un número demasiado grande puede llevar a una pérdida de información. Por otra parte, si *k* es grande la serie resultante tiene pocos términos pues se pierden observaciones al principio y al final de la serie. Por el contrario si el número de datos es muy pequeño no se logra eliminar las variaciones que no son producidas por la tendencia y la suavización no cumple su cometido.

En series que tienen algún patrón de variación *k* es fácil de determinar. Por ejemplo en una serie de datos mensuales con estacionalidad, medias móviles centradas de 12 términos, eliminarían la estacionalidad y las variaciones accidentales.

Se recomienda usar el método de las medias móviles en series en las que el ritmo de crecimiento cambia con el tiempo o que presenten muchas variaciones irregulares ya que pueden proponer funciones que aproximen periodos cortos de tiempo. Sin embargo, aunque es útil en estos casos el método de la media móvil sólo suaviza la serie y no hay posibilidad algebraica de eliminar la tendencia.

EJEMPLO 3

Cálculo de medias móviles

Se suavizan los datos de ancho de playa calculando la media móvil de 3 términos (MM3) (Tabla 4). Las dos primeras medias son:

$$y^*_{Feb-07} = \frac{y_1 + y_2 + y_3}{3} = \frac{225{,}6 + 226{,}2 + 222{,}7}{3} = 224{,}8$$

$$y^*_{Mar-07} = \frac{y_2 + y_3 + y_4}{3} = \frac{226{,}2 + 222{,}7 + 218{,}1}{3} = 222{,}3$$

MM3	*Ene.*	*Feb.*	*Mar.*	*Abr.*	*May.*	*Jun.*	*Jul.*	*Ago.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
2007		224,8	222,3	217,7	213,2	207,4	203,0	201,3	203,4	206,9	208,0	208,6
2008	207,8	206,6	202,5	197,0	191,0	184,9	180,8	178,4	178,9	179,8	179,7	179,8
2009	179,1	178,1	174,9	170,5	165,6	160,4	157,2	155,8	157,2	159,5	160,9	159,2
2010	156,6	153,1	150,6	146,0	141,5	135,7	131,3	129,6	131,7	135,2	136,3	136,9
2011	136,1	134,9	130,8	125,3	119,3	113,2	109,1	106,7	107,2	108,1	108,0	

Tabla 4. Serie suavizada con promedios móviles de 3 términos.

También se suaviza la serie calculando la media móvil de 12 términos (MM12). Como el número de términos es par y cada media calculada no se corresponde con ningún dato observado se deben centrar calculando el promedio de dos MM12 sucesivas. Se obtienen así medias móviles de 12 términos centradas (MMC12) (Tabla 5, Fig. 5). La primera media móvil se calcula como sigue:

$$y^*_{Ene-07|Dic-07} = \frac{225,6 + 226{,}2 + \cdots + 209{,}3 + 207{,}5}{12} = 211{,}8$$

$$y^*_{Feb-07|Ene-08} = \frac{226{,}2 + 222{,}7 + \cdots + 207{,}5 + 209{,}1}{12} = 210{,}5$$

$$y^*_{Jul-07} = \frac{211{,}8 + 210{,}5}{2} = 211{,}1$$

MMC12	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.
2007							211,1	209,6	208,0	206,4	204,6	202,7
2008	200,8	199,0	197,0	194,9	192,5	190,1	187,7	185,3	183,0	180,7	178,6	176,5
2009	174,5	172,6	170,7	169,0	167,3	165,8	164,0	161,9	159,9	157,9	155,8	153,8
2010	151,8	149,5	147,3	145,3	143,3	141,2	139,4	137,9	136,3	134,7	132,9	131,0
2011	129,1	127,3	125,3	123,2	120,8	118,4						

Tabla 5. Serie suavizada con promedios móviles centrados de 12 términos.

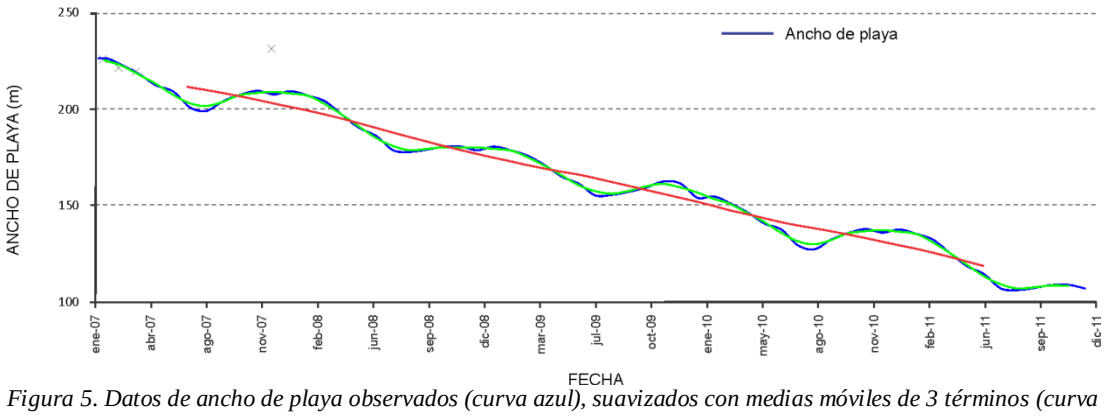


Figura 5. Datos de ancho de playa observados (curva azul), suavizados con medias móviles de 3 términos (curva verde) y con media móviles de 12 términos centradas (curva roja).

Métodos para aislar la estacionalidad

Inicialmente se definieron los movimientos estacionales de una serie como aquellos que se repiten sistemáticamente con una periodicidad menor al año. La estacionalidad produce que en determinados meses (u otros periodos de tiempo inferiores al año), sucedan movimientos ajenos a la tendencia que dificultan una educada comparación de esa serie en esos meses pues el nivel medio de la misma se va modificando.

Para evitar estas alteraciones en los niveles medios se realiza una corrección estacional o desestacionalización. El paso previo a esta corrección es aislar la componente estacional. Para ello se pueden usar varios métodos, aquí se explicará sólo el de la razón a la media móvil.

Método de la razón a la media móvil

Este método es el más difundido para estimar la variación estacional. Se sintetiza en los siguientes pasos:

1° **Estimación de la tendencia - ciclo** utilizando una media móvil cuyo orden permita eliminar la componente estacional. El orden de la media móvil depende de la periodicidad de los datos, si se trata de datos mensuales es de 12 términos, si los datos son trimestrales es de 4 términos. Las medias

móviles deben quedar siempre centradas. La nueva serie estará reducida pues se han perdido las primeras y las últimas observaciones.

2° **Obtención de las razones o porcentajes a las medias móviles** como cociente de la serie original y la estimación de la tendencia-ciclo calculada en el primer paso. Estos cocientes se denominan Índices Bruto Estacionales (*IBE*).

Si se trata de un modelo multiplicativo $y_t = T_t\, VE_t\, C_t R_t$, el cociente que se obtiene es:

$$\frac{y_t}{Media\ Móvil} = \frac{T_t\, VE_t\, C_t\, R_t}{T_t\, C_t} = VE_t\ R_t = IBE_t. \qquad (11.7)$$

Para modelos aditivos la *VE* se elimina por sustracción.

3° **Calcular la media de todos los *IBE*** correspondientes a un mismo mes o trimestre si los datos fueran trimestrales. De este modo se elimina gran parte de las variaciones irregulares.

En total se tienen *k* Índices Estacionales (*IE*) correspondientes a cada mes o a cada trimestre dependiendo de la periodicidad de los datos

$$IE = \frac{1}{n-1}\sum_{i=1}^{n-1} IBE_{t,i} \qquad i = 1,\ldots, k. \qquad (11.8)$$

Cuando no hay estacionalidad, el *IE* es 1, pero cuando existe estacionalidad los *IE* son diferentes a 1 y diferentes entre sí. Cabe aclarar que en modelos aditivos el *IE* es 0.

Algunos autores proponen realizar un ANOVA para detectar si existen diferencias significativas entre *IBE* y de existir diferencias, efectuar las pruebas *a posteriori* correspondientes.

La suma de todos los *IE* debería ser 12 si los datos son mensuales o 4 si son trimestrales. Cuando esto no sucede es necesario realizar un ajuste.

4° **Normalizar los índices estacionales** para expresarlos como proporción de su valor medio. Para obtener el Índice Estacional Normalizado (*IEN*) se calcula el cociente entre el *IE* por el promedio de los mismos.

. $$IEN_i = \frac{IE_i}{\sum_{i=1}^{k} IE}\, k. \qquad (11.9)$$

Los *IEN* se suelen expresar en porcentaje y son los que se utilizan para mostrar la desviación que se produce en cada mes del año. Para visualizar se realiza un grafico de barras marcando el nivel medio en 100.

EJEMPLO 4

Método de la razón a la media móvil

Para aislar la estacionalidad de la serie de ancho de playa del ejemplo se siguen los siguientes pasos:

1° Estimar la tendencia-ciclo de la serie de datos del ancho de playa. Los datos de la Tabla 4 muestra la serie obtenida con los promedios móviles de 12 términos centrados.

2° Obtener las razones a las medias móviles. El índice Bruto Estacional (*IBE*) se calcula realizando los cocientes de los datos de la Tabla 1 sobre los datos de la Tabla 5. La Tabla 6 muestra los cocientes; por ejemplo el IBE para julio de 2007 es:

$$\frac{y_t}{Media\ Móvil} = IBE_t \qquad \rightarrow \qquad IBE_{Jul-07} = \frac{246{,}4}{211{,}1} = 1{,}16$$

IBE	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.	
2007							1,167	1,161	1,145	1,128	1,118	1,118	Gran Media
2008	1,041	1,039	1,035	1,010	0,988	0,978	0,951	0,959	0,977	0,998	1,010	1,011	
2009	1,034	1,033	1,029	1,011	0,986	0,972	0,946	0,960	0,982	1,008	1,042	1,049	
2010	1,014	1,033	1,025	1,008	0,981	0,973	0,927	0,922	0,970	1,007	1,036	1,037	
2011	1,064	1,061	1,055	1,016	0,981	0,965							
IE	*1,038*	*1,042*	*1,036*	*1,011*	*0,984*	*0,972*	*0,998*	*1,000*	*1,019*	*1,035*	*1,051*	*1,054*	*1,020*

Tabla 6: Índices Bruto Estacional e Índices Estacionales.

3° El Índice Estacional de cada mes es el promedio de los *IBE* para cada mes. Los *IE* se presentan en la última fila de la Tabla 6.
4° Aunque el promedio de los IBE es aproximadamente 1 y para este caso no es necesario normalizar los índices se calculan igual a modo de ejemplo. La tabla 7 muestra los índices normalizados.

	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.
2007							1,144	1,138	1,123	1,106	1,096	1,096
2008	1,021	1,018	1,015	0,990	0,968	0,959	0,933	0,940	0,958	0,978	0,990	0,992
2009	1,014	1,013	1,009	0,991	0,966	0,953	0,927	0,941	0,963	0,988	1,021	1,028
2010	0,994	1,013	1,005	0,988	0,962	0,954	0,908	0,904	0,951	0,987	1,015	1,016
2011	1,044	1,040	1,034	0,996	0,961	0,946						
IEN	*1,018*	*1,021*	*1,016*	*0,991*	*0,965*	*0,953*	*0,978*	*0,981*	*0,999*	*1,015*	*1,031*	*1,033*

Tabla 7. Índice estacional normalizado.

La figura 6 representa el Índice Estacional normalizado para el ancho de playa. Los valores indican que existe un comportamiento estacional, durante los meses de otoño e invierno el ancho de playa disminuye y aumenta en los meses de primavera y verano respecto al valor medio.

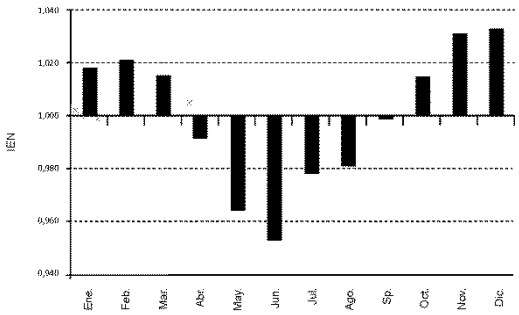


Figura 6. Índice estacional normalizado.

Desestacionalización

La desestacionalización consiste en eliminar de la serie la componente estacional para tener un diagnostico más fiel del fenómeno estudiado. Se parte del supuesto de que las fluctuaciones estacionales pueden medirse y aislarse de las variaciones producidas por la tendencia, las variaciones cíclicas e irregulares.

Para modelos multiplicativos se calcula el cociente entre cada dato de la serie original por su correspondiente *IEN*. Este cociente recoge el porcentaje de variación que sobre la tendencia provoca esta componente. Así

$$\frac{y_t}{IEN_t} = \frac{T_t\, VE_t\, R_t}{IEN_t} = T_t\, C_t\, R_t. \qquad (11.10)$$

Para modelos aditivos la componente estacional se elimina por sustracción.

Una vez que se elimina la componente estacional los valores de cualquier mes se pueden comparar entre sí porque perdieron la particularidad que le asignaba la estacionalidad.

EJEMPLO 5

Desestacionalización

La Tabla 8 muestra los datos desestacionalizados. Los valores surgen del cociente entre el dato original (Tabla 1) y el *IEN* (Tabla 7).

	Ene.	*Feb.*	*Mar.*	*Abr.*	*May.*	*Jun.*	*Jul.*	*Ago.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
2007	235,6	233,3	233,4	242,3	251,7	255,0	251,9	248,1	238,6	229,4	221,9	219,4
2008	221,6	221,5	219,3	220,0	220,1	219,4	205,4	202,8	204,3	204,3	203,1	200,9
2009	205,4	202,4	200,8	198,5	197,2	195,2	182,6	181,2	179,0	177,7	175,0	172,8
2010	177,2	174,6	173,0	172,3	171,0	169,1	158,6	158,4	157,2	156,8	157,5	156,1
2011	151,2	151,3	148,7	147,7	145,8	144,2	132,1	129,7	132,5	133,6	133,5	131,5

Tabla 8. Índice estacional normalizado.

La figura 7 muestra un patrón bastante definido: el ancho de playa disminuye paulatinamente durante el periodo analizado.

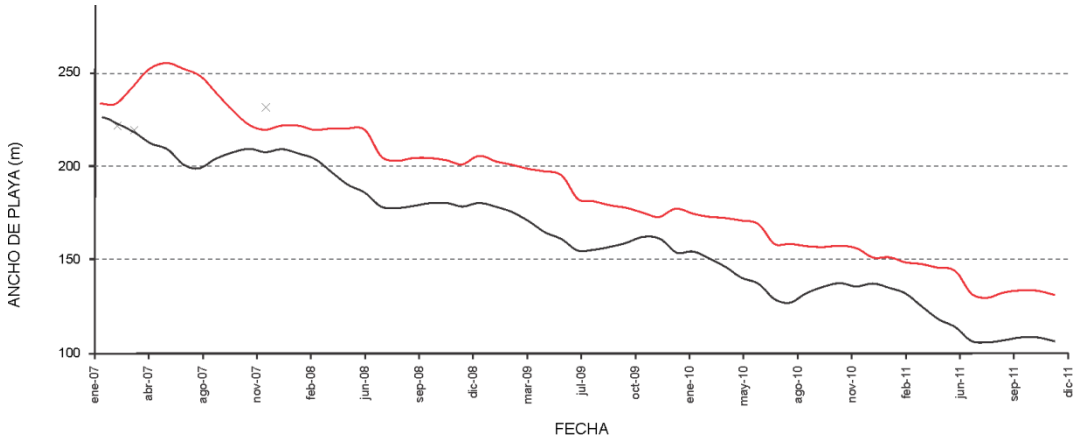


Figura 7. Ancho de playa desestacionalizado (curva negra), ancho de playa (curva negra).

Análisis del ciclo

La componente cíclica es la más difícil de detectar y estimar porque no siempre existe; cuando está presente sus oscilaciones suelen ser muy irregulares y no tienen periodo fijo, además, puede suceder que dos ciclos se superpongan. Sin embargo, a pesar de las dificultades, conviene aislar el

componente cíclico de una serie cronológica para estudiar los puntos de inflexión y los valores máximos y mínimos. Si la serie tiene un patrón cíclico estable, aislarla posibilita determinar las causas intrínsecas y la predicción de los movimientos futuros como por ejemplo los fenómenos del Niño.
La componente cíclica se puede aislar con el método de los residuos que se describe a continuación.

Método de los residuos

El método de los residuos consiste en identificar la Tendencia, las Variaciones Estacionales y las variaciones Irregulares y eliminarlas de la serie por división si el modelo es multiplicativo o por sustracción si el modelo es aditivo.
Para un modelo multiplicativo $y_t=T_t\ VE_t\ C_t\ R_t$, los pasos serán entonces primero desestacionalizar la serie $\left(\frac{y_t}{IEN}=\frac{T\ VE\ C\ R}{VE}=\boldsymbol{T\ C\ R}\right)$ y luego eliminar la tendencia $\left(\frac{T\ C\ R}{T}=\boldsymbol{C\ R}\right)$.
En el ejemplo del ancho de playa no se detectan ciclos probablemente porque el periodo analizado es muy corto.

Variaciones irregulares

Las variaciones irregulares o residuales son de poco interés. Podrían aislarse con una suavización con medias móviles. Sin embargo, cuanto mayor son las irregularidades que presente la serie mayor debe ser el número de términos que se incluyan en la media móvil. Luego se obtienen como cociente entre la serie sin tendencia ni estacionalidad ni ciclo como cociente.

Autocorrelación

Al analizar una serie cronológica a veces aparecen segmentos de curvas que aparentemente se repiten. Una forma de medir estas semejanzas es con un proceso de auto-comparación de la serie. Para medir estas semejanzas se utiliza el coeficiente de correlación lineal *r* de Pearson. El coeficiente se calcula entre la serie y la misma serie desplazada un paso o *lag* (*L*). Aunque la magnitud del paso puede variar teóricamente desde 1 hasta *n* - *L*, algunos autores recomiendan solo calcular *r* hasta desplazamientos igual a *n/4*, y algunos más conservadores hasta *n/10*.
Para el cálculo de r_L se utiliza una expresión análoga a la del coeficiente de correlación lineal *r* de Pearson. La autocorrelación para el paso *k* está dada por la siguiente expresión:

$$r_k = \frac{Cov.y_i y_{i+k}}{\sqrt{Var.y_i Var.y_{i+k}}} \quad \rightarrow \quad r_k = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y}_1)(y_{i+k} - \bar{y}_2)}{\sqrt{\left[\sum_{i=1}^{n-k}(y_i - \bar{y}_1)^2\right]\left[\sum_{i=1}^{n-k}(y_{i+k} - \bar{y}_2)^2\right]}}, \qquad (11.11)$$

donde $y_1, y_2, ..., y_n$ son los valores de la variable en los tiempos $t_1, t_2, ..., t_n$, $\bar{y}_1 = \sum_{i=1}^{n-k} y_i /(n-k)$

$\bar{y}_2 = \sum_{i=k+1}^{n-k} y_i /(n-k)$.

El coeficiente de autocorrelación toma valores entre 1 y -1, valores cercanos a 0 indican ausencia de autocorrelación. Se puede calcular la significancia de r_L con una prueba de hipótesis con el estadístico de prueba $z = r_L\sqrt{N-L}$. La prueba es a dos colas con hipótesis nula es ausencia de correlación (H_0: $\rho_L = 0$).

Si se calcula r_L para varios valores de L (i.e.: 1, 2, 3, ..., n - L) se obtienen los valores de la llamada función de autocorrelación cuya expresión gráfica es el **correlograma**. El correlograma es un gráfico bivariado que en abscisas tiene los valores de L y en ordenadas los valores de r_L. El correlograma facilita la detección de patrones de variación y se utiliza para describir la estructura de una serie cronológica.

Cuando el correlograma se calcula a partir de la serie cronológica original, se incluyen todas las componentes determinísticas (*T*, *C* y *VE*) y estocásticas (*R*). Valores de tendencia crecientes o decrecientes producen valores de autocorrelación cercanos a cero, a medida que aumenta el paso (Fig. 8). Si hay patrón cíclico o estacional el correlograma muestra una forma semejante a la función seno o coseno. En cambio, cuando el correlograma se calcula con los valores de la serie filtrados, por ejemplo con promedios móviles, se puede identificar o estimar la componente aleatoria.

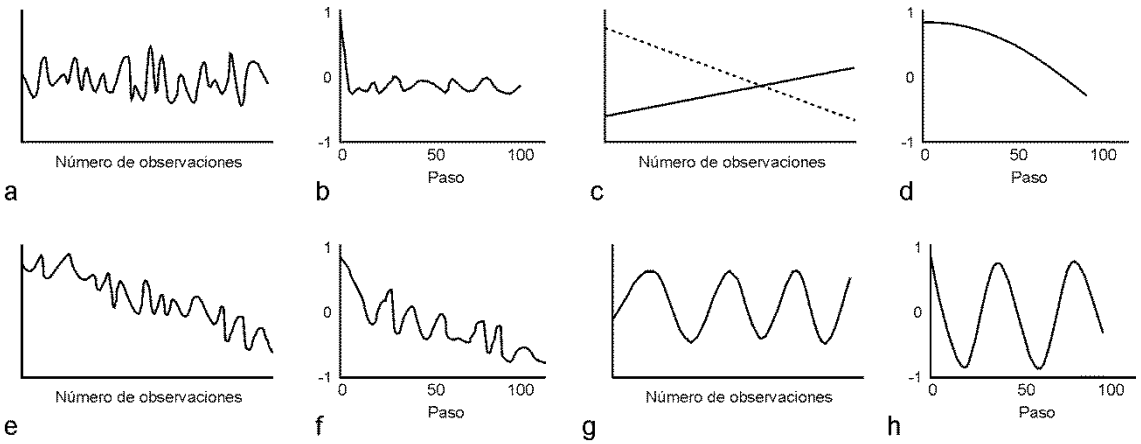


Figura 8. Ejemplos de secuencias de variables vs. tiempo y correlogramas (se plotea r *de Pearson vs. el numero de paso). a y b:Secuencia aleatoria estacionaria sin tendencia, el valor de* r *en el correlograma alcanza rapidammente el valor cero. c y d: Series con tendencia lineal positiva o negativa muestran el mismo correlograma con marcada correlacion positiva en los pasos bajos y marcadamente negativos a pasos distantes. e y f: Secuencia con tendencia negativa y componente aleatoria. g y h: Secuencia con componente estacionaria (puede ser ciclica dependiendo de la escala), el correlograma muestra el mismo patron (modificado de McKillup y Dyar 2010).*

EJEMPLO 6

Correlograma

El correglama obtenido con los datos del ancho de playa muestra claramente la componente cíclica del fenómeno que se repite, en este caso, con una periodicidad de un año (Fig. 9).

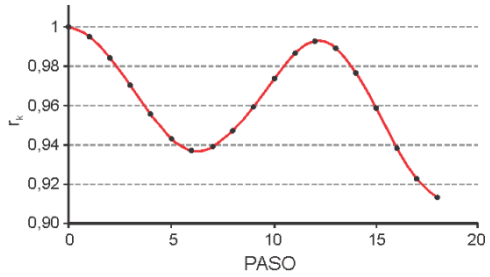


Figura 9. Correlograma del ancho de playa.

Correlación cruzada

El coeficiente de correlación cruzada se utiliza con el objeto de investigar si dos series de datos poseen patrones de variación semejantes. Si se designan a las dos series como Y_{1i} e Y_{2i}, se define n^* como el número de posiciones que se superponen entre las dos cadenas, el coeficiente de correlación cruzada para el desfasaje *m*, es:

$$r_m = \frac{n^* \sum Y_1 Y_2 - \sum Y_1 \sum Y_2}{\sqrt{[n^* \sum Y_1^2 - (\sum Y_1)^2][n^* \sum Y_2^2 - (\sum Y_2)^2]}}. \qquad (11.12)$$

La significancia del coeficiente de correlación cruzada en el desfasaje *m*, r_m, se calcula con una prueba de hipótesis de prueba *t*, donde el estadístico de prueba tiene la siguiente expresión:

$$t = r_m \sqrt{\frac{n^* - 2}{1 - r_m^2}}. \qquad (11.13)$$

La prueba es a dos colas con hipótesis nula es ausencia de correlación (H_0: $\rho_L = 0$), el valor crítico para tomar la decisión estadística se busca con ($n^* - 2$) grados de libertad en la Tabla 3 del Anexo.

Cabe aclarar que a diferencia del coeficiente de autocorrelación donde el desfasaje se produce siempre en el mimo sentido (L = 1, 2, ..., n) porque hay un inicio conocido de la serie, en el cálculo del coeficiente de correlación cruzada, las series se desfasan en los dos sentidos (L = -n, ..., -2, -1, 0, 1, 2, ..., n) porque no hay certeza de cuál es el punto o momento de coincidencia máxima de ambas series.

Por ejemplo, si se desea conocer cuando se producirá el aumento de caudal máximo de un río en el punto B de su cauce conociendo las precipitaciones caídas en el punto A ubicado aguas arriba se podrá inferir el tiempo que tarda en aumentar el caudal en B.

Test de rachas

El Test de Rachas de Wald-Wolfowitz (*Runs test*) se utiliza con series de datos nominales con solo dos categorías mutuamente excluyentes. Se trata de un test no paramétrico que se podría haber descrito en el Capítulo 11, pero dado que es una prueba que se aplica a las series más simples, se explica en este apartado. Imagine el siguiente ejemplo: en una playa de la provincia de Buenos Aires se realizan mediciones del ancho de playa una vez al mes durante 2 años, se define una fase de ampliación (A) como aquella en las que el ancho de playa aumenta respecto al mes anterior y una fase de reducción (R) como aquella en la que el ancho de playa disminuye respecto a la del mes anterior. Se registraron 10 fases de ampliación y 13 de reducción del ancho de playa.

Se define una racha *u* como la secuencia ininterrumpida de la misma categoría. La serie observada (S_O) tiene 8 rachas (cada racha se indica con una línea contíinua). Podría suceder que se hubiesen encontrado solo 2 rachas (serie S_1) o tal vez 20 (serie S_2) y ambas serían consistentes con las 10 fases de A y las 13 R.

S_O: A A A R R A A R R R R A A A R R R R A A R R R

S_1: A A A A A A A A A A R R R R R R R R R R R R R

S_2: A R A R A R A R A R A R A R A R A R A R A R R R R

El test de rachas permite evaluar la hipótesis que la secuencia de eventos es producto del azar. Para someter a prueba la hipótesis nula se usa el estadístico de prueba *Z*,

$$z_c = \frac{u-\mu_u}{\sigma_{\bar{u}}}, \qquad (11.14)$$

donde $\mu_u = \frac{2n_1n_2}{n_1+n_2} + 1$ y $\sigma_{\bar{u}} = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$, (11.15 y 11.16)

para n_1 es el número de elementos de la primera categoría (A para el ejemplo), n_2 el número de elementos de la segunda categoría (R, para el ejemplo) y *u* es el número de rachas de la serie.

La hipótesis nula se rechaza cuando $z_c > Z_{\alpha/2}$, los valores críticos del estadístico se encuentran en la Tabla 1 del Anexo.

EJEMPLO 7

Test de rachas

Serie de fases de ampliación-reducción del ancho de playa entre dos meses sucesivos observadas

S_0: A A A R R A A R R R R A A A R R R R A A R R R

H_0: $u = \mu_u$. La distribución de las fases de ampliación-reducción del ancho de playa es al azar

H_1: $u \neq \mu_u$. La distribución de las fases de ampliación-reducción del ancho de playa no es al azar

$\alpha = 0{,}05$ $\alpha/2 = 0{,}025$

De la Tabla 1 del Anexo, $Z_{0{,}025} = 1{,}96$

$n_A = 10$ $n_R = 13$ $u = 8$

$\mu_u = \frac{2\ 8\ 10}{8+10} + 1 = 9{,}9$ $\sigma_{\bar{u}} = \sqrt{\frac{2\ 8\ 10(2\ 8\ 10-8-10)}{(8+10)^2(8+10-1)}} = 2{,}03$

$z_c = \frac{8-9{,}9}{2{,}03} = -0{,}93$

Decisión estadística: $z_c > z_{\alpha/2}$ (2,03 > 1,96), existen evidencias para rechazar la hipótesis nula.
La serie de eventos de ampliación y reducción de la playa en el periodo analizado no ocurre al azar.

Rechazar esta hipótesis dará pie a investigar si la serie tiene: a) menos rachas que las que ocurren por azar, en ese caso la serie es mas agrupada o contagiosa que una aleatoria, b) si la serie tiene más rachas que las que ocurren por azar, indica una tendencia hacia una distribución uniforme. Para el ejemplo, si hay pocas rachas podría indicar que la acreción se produce solo en determinados meses del año y la reducción del ancho de playa en otros, por el contrario, si se encuentran muchas rachas existe alguna tendencia hacia la alternancia de faces de ampliación y reducción de la playa.

Para analizar se la serie es contagiosa o uniforme se realizan pruebas de hipótesis de una cola. La hipótesis nula de la situación contagiosa enuncia que los elementos de la población no están en forma agrupada (H_0: $u \geq \mu_u$) y la hipótesis alternativa que los elementos de la población están agrupados (H_A: $u < \mu_u$). La hipótesis nula se rechaza si $z_c \geq z_\alpha$ y $u \leq \mu_u$. Para el caso de S_1, con solo dos rachas (u = 2), el valor de z_c = -3,88 y el de $z_{0,05}$ = 1,645, además 2 es menor que la media poblacional (μ_u = 9,89), esto resultados llevan a rechazar la hipótesis nula y a concluir que esta serie presenta sus datos agrupados.

La hipótesis nula de la situación uniforme expresa que los elementos de la población se no se distribuyen uniformemente (H_0: $u \leq \mu_u$) y la hipótesis alternativa que los elementos de la población se distribuyen uniformemente (H_1: $u > \mu_u$). La hipótesis nula se rechaza si $z_c \geq z_\alpha$ y $u \geq \mu_u$. Para el caso de S_2, que posee un número elevado de rachas (u = 20), el valor de z_c = 9,85 y el de $z_{0,05}$ = 1,645, además 20 es mayor que la media poblacional (μ_u = 9,89), esto conduce a rechazar la hipótesis nula y a concluir que esta serie presenta sus datos uniformemente distribuidos.

Auto-asociación y Asociación cruzada

El método de auto-asociación y el de asociación cruzada tienen por objetivo calcular un índice para cuantificar las semejanzas entre dos series de datos nominales. Las categorías de clasificación deben ser mutuamente excluyentes. Por ejemplo series estratigráficas donde se suceden tipos litológicos (i.e. lutitas, areniscas, arcilitas, coquinas, etc.) o el registro en testigos de perforaciones de zonas mineralizadas (i.e. distintos tipo de ganga y mena), entre otros. Para medir las semejanzas cuando se comparan dos series se utiliza el **índice de coincidencias** (*ICo*) que es igual al número de

coincidencias (*C*) sobre el número de total de comparaciones (*N*), (*ICo = C/N*). El coeficiente se calcula desplazando las series una respecto a otra hasta que solo se compara una posición.

Los índices de coincidencia de cada comparación se pueden volcar en un gráfico bivariado similar al correloglama, poniendo en abscisas los valores de desfasaje y en ordenadas los valores del *ICo*.

La significación del *ICo* se determina con una prueba de hipótesis de Chi cuadrado (χ^2). Para realizar la prueba se necesita conocer el número de coincidencias y el de no coincidencias de la comparación. También se requiere conocer la probabilidad del número de coincidencias que tendrían dos series aleatorias con el mismo número de categorías y de la misma longitud (*Pc*) y la probabilidad del número de no coincidencias (*Pnc = 1 - Pc*). La probabilidad del número de coincidencias se calcula con la siguiente fórmula:

$$Pc = \frac{\sum_{k=1}^{m} X_{1k} X_{2k}}{n_1 n_2}, \tag{11.17}$$

donde, X_{1k} = número de observaciones en cada *k*-ésima categoría en la serie 1, X_{2k} = número de observaciones en cada *k*-ésima categoría en la serie 2, n_1 = es el número de observaciones de la serie 1, n_2 = es el número de observaciones de la serie 2.

El estadístico de prueba se calcula con la siguiente expresión:

$$\chi^2 = \frac{(|\mathrm{Co}-\mathrm{Ce}|-0{,}5)^2}{\mathrm{Ce}} + \frac{(|\mathrm{NCo}-\mathrm{NCe}|-0{,}5)^2}{\mathrm{NCe}}, \tag{11.18}$$

donde *Co* = número de coincidencias observadas, *Ce* = número de coincidencias esperadas (*Pc* x *N*), *NCo* = número de no coincidencias observadas, *NCe* = número de no coincidencias esperadas (*Pnc* x *N*).

La hipótesis nula de la prueba es que el número de coincidencias y de no coincidencias entre ambas series es producto del azar. La hipótesis nula se rechaza cuando $\chi^2 > \chi^2_{\alpha;1}$, note que el valor crítico de χ^2 tiene 1 grado de libertad (Tabla 2 del Anexo).

EJEMPLO 8

Autoasociación

Se relevaron dos perfiles en el Patagoniano aflorante en dos localidades distantes. La litologías halladas son: arena muy fina, (AMF), arena fina (AF), arena mediana (AM), arena gruesa (AG), sábulo (S). Los perfiles tienen las siguientes litologías de base a techo:

Perfil *1*: AG AM AF AM AF AG AF AMF AF S AF AG
Perfil *2*: AMF AF AM AG AF AG AM AF AM AF AMF AF
En la primera comparación *N* = 12, *C* = 2, *ICo* = 0,16

En la segunda comparación se desfasa una secuencia respecto a la otra una posición
AG AM AF AM AF AG AF AMF AF S AF AG
AMF AF AM AG AF AG AM AF AM AF AMF AF
N = 11, *C* = 4, *ICo* = 0,36

La prueba de significación de la comparación de las dos series sin desfasar se plantea como:

H_0: *Fo* = *Fe*. El número de coincidencias y de no coincidencias entre ambas series es producto del azar.
H_1: *Fo* ≠ *Fe*. El número de coincidencias y de no coincidencias entre ambas series no es producto del azar.

$\alpha = 0,05$
De la Tabla 2 del Anexo, $\chi^2_{0,05;\,1} = 3,84$

Categoría	*N° Perfil 1*	*N° Perfil 2*	*N°1 x N°2*
AMF	1	2	2
AF	5	5	25
AM	2	3	6
AG	3	2	6
S	1	0	0
Suma	*12*	*12*	*39*

$$Pc = \frac{\sum_{k=1}^{m} X_{1k}X_{2k}}{n_1 n_2} = Pc = \frac{39}{12\ x\ 12} = 0,27$$
$$Pnc = 1 - 0,27 = 0,73$$

Co = número de coincidencias observadas = 2
Ce = número de coincidencias esperadas (*Pc* x *N*) = 0,27 x 12 = 3,2
NCo = número de no coincidencias observadas = 10
NCe = número de no coincidencias esperadas (*Pnc* x *N*) = 0,73 x 12 = 8,8

$$\chi^2 = \frac{(|\mathrm{Co} - \mathrm{Ce}| - 0,5)^2}{\mathrm{Ce}} + \frac{(|\mathrm{NCo} - \mathrm{NCe}| - 0,5)^2}{\mathrm{NCe}} = \frac{(|2 - 3,2| - 0,5)^2}{3,2} + \frac{(|10 - 8,8| - 0,5)^2}{8,8} = 0,21$$

Decisión estadística: $\chi^2 < \chi^2_{0,05;1}$ (0,21 < 3,84), no existen evidencias para rechazar la hipótesis nula. Las coincidencias entre ambas series se producen por azar.

Matrices de transición

Las matrices de transición se utilizan para analizar series de datos nominales al igual que los métodos de auto asociación y asociación cruzada descriptos. Tiene por objetivo conocer la tendencia de paso de un estado o categoría a otro sin tener en cuenta el orden de los cambios de categoría en la secuencia.

Para el análisis de las transiciones de un estado a otro de una serie se puede construir la **matriz de frecuencias de transición** [*Tr*]. Esta matriz tiene un número de datos menos que la cantidad de datos de la serie (*n - 1*). La matriz se lee que el cambio ocurre **desde las categoría de las filas, hacia la las categorías de las columnas**.

Con el objeto de visualiza más claramente cuáles son las tendencias de transición, también se puede construir una **matriz de proporciones** (*proportion matrix*) dividiendo cada frecuencia observada en la matriz de transición por el total de transiciones registradas, $[Pp_{ij}] = (x_{ij}/(n - 1))$.

Para enfatizar la proporción de estados que preceden un estado dado, es útil calcular la **matriz de proporciones/probabilidades de transición** (*transition proportion matrix*). Esta matriz tiene el resultado del cociente de la frecuencia de transición de una categoría A en otra B, sobre el número total de A presentes en la secuencia, $[Pr_{ij}] = (x_{ij} / \textit{total de fila})$.

EJEMPLO 9

Matrices de transición

Los siguientes datos representan la sucesión de litologías en un perfil de detalle del Grupo Sierras Bayas. Se reconocieron 4 tipos litológicos: dolomías (D), calizas (C), cuarcitas (Q) y psamopelitas (P).

(Piso) D C D Q P Q P D C Q C D P Q P Q C D Q D C P D C Q P C D Q P D (Techo)

Matriz de transición Tr_{ij}		*HACIA* *D*	*C*	*Q*	*P*	*TOTAL DE FILA*
DESDE	*D*	0	4	3	1	*8*
	C	4	0	2	1	*7*
	Q	1	2	0	5	*8*
	P	3	1	2	0	*7*
TOTAL DE COLUMNA		*8*	*7*	*8*	*7*	*30*

Para Tr_{ij} por ejemplo, la frecuencia de transición desde D hacia C se registra en la celda $x_{1.2}$ y es igual 4(están subrayadas en la serie), mientras que la frecuencia de transición desde P hacia Q se registra en la celda $x_{4.3}$ y es igual a 2.

Matriz de proporciones Pp_{ij}		*HACIA* *D*	*C*	*Q*	*P*	*TOTAL DE FILA*
DESDE	*D*	0	0,13	0,10	0,03	*0,26*
	C	0,13	0	0,07	0,03	*0,23*
	Q	0,03	0,07	0	0,17	*0,27*
	P	0,10	0,03	0,10	0	*0,23*
TOTAL DE COLUMNA		*0,26*	*0,23*	*0,27*	*0,23*	*1*

En la matriz de proporciones se aprecia que la transición desde cuarcitas (Q) hacia psamopelitas (P) es la mayor, ocurre 17%. Por el contrario, la transición desde dolomitas (D) hacia psamopelitas (P), desde cuarcitas (Q) hacia psamopelitas (P), desde cuarcitas (Q) hacia dolomitas (D) y desde pasamopelitas (P) hacia calizas (C), son las menores, ocurren solo el 3% de las veces.

Matriz de proporciones de transición Pr_{ij}		*HACIA* *D*	*C*	*Q*	*P*	*TOTAL DE FILA*
DESDE	*D*	0	0,500	0,375	0,125	*1,00*
	C	0,571	0	0,286	0,143	*1,00*
	Q	0,125	0,250	0	0,625	*1,00*
	P	0,428	0,143	0,428	0	*≈1,00*

En la matriz de proporciones de transiciones se observa que el 50% de transiciones desde capas de dolomita (D) son hacia capas de caliza (C), hacia capas de cuarcitas (Q) es 37,5% y hacia psamopelitas (P) es de 12,5%.

Cadenas de Markov

Una cadena de Markov[18] es una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del estado del evento que le antecede. Las cadenas de este tipo **recuerdan** el estado del último evento y esto condiciona la probabilidad del estado de los eventos futuros, se dice que tienen **memoria**. Esta dependencia del evento anterior diferencia a las cadenas de Markov de las series de eventos independientes (i.e. tirar una moneda al aire).

Las series de eventos pueden ser totalmente determinísticas cuando la sucesión de estados es siempre la misma, por ejemplo si los estados son tres, A, B y C, a lo largo de la sucesión al estado A siempre le sucede B, a B siempre C y a C siempre A. La persistencia de la memoria es una función del

determinismo del sistema que, en este caso, recuerdar por siempre. En el extremo opuesto se encuentran las series totalmente aleatorias en las cuales la sucesión de eventos ocurre al azar como es el caso de una serie de lanzamiento de una moneda. Las series de este tipo olvidan el estado precedente por completo. Entre estos extremos se encuentran series parcialmente determinísticas, series de comportamiento semi-aleatorio, etcétera. Muchas de estas series tienen propiedades de las cadenas de Markov, esto es el estado en un punto cualquiera de la serie es dependiente, de manera estadística, del estado del suceso precedente. En estas series el número de pasos que persiste la memoria es un índice de orden de la serie.

Las aplicaciones de la teoría de Markov en problemas geológicos son para modelar o simular los procesos. Muchos procesos que operan en el interior o sobre la superficie de la tierra exhiben comportamiento de Markov, por ejemplo los procesos que condujeron a la depositación de secuencias hemipelágicas en el Jurásico superior de la Península antártica.

Para evaluar si una secuencia tiene propiedades de cadenas de Markov se recurre a las matrices de transición. Se ha visto que es posible usar la frecuencia relativa para estimar la probabilidad de ocurrencia de un suceso. Así la matriz de proporciones (Pp_{ij}) calculada con la matriz de frecuencia de transición de una secuencia (Tr_{ij}), es una estimación de las probabilidades de las transiciones de un estado a otro. De igual manera, la matriz de proporciones de transición (Pr_{ij}) es la matriz de probabilidad condicional, es decir la matriz tiene la probabilidad que ocurra un estado conociendo el estado precedente (P($A|B$)).

Con el objetivos de probar la significación de un proceso de Markov se realiza una prueba de hipótesis de χ^2. La prueba se desarrolla a partir de la matriz de transición (Tr_{ij}) y la matriz de probabilidad de transición (Pr_{ij}). El primer paso es calcular la probabilidad de que se llegue a un estado determinado desde cualquier estado anterior, es decir conocer las probabilidades marginales. La matriz de probabilidades marginales (Pr_j) se calcula sumando las frecuencias de cada columna de la matriz de transición *Tr*, y dividiendo por el número total de transiciones. Para una matriz de transición de *m x m* estados, Pr_j se calcula entonces con la siguiente expresión:

$$Pr_j = \frac{\sum_{i=1}^{m} Tr_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} Tr_{ij}}. \qquad (11.19)$$

El estadístico de prueba es:

$$\chi^2 = 2 \sum_{i=1}^{m}\sum_{j=1}^{m} Tr_{ij} \ln \frac{Pr_{ij}}{Pr_j}. \qquad (11.20)$$

Cada elemento de la matriz de probabilidad de transición (Pr_{ij}) se divide por la probabilidad marginal de la columna donde se encuentra el elemento (Pr_j) (Pr_{ij} / Pr_j). El número correspondiente de la matriz de transición (Tr_{ij}) se multiplica por el logaritmo natural de ese cociente. Por últimos se suman todos los valores m^2 y se multiplica por 2 para obtener el estadístico de prueba χ^2. El χ^2 calculado se contrasta con un valor crítico de tabla (Tabla 2 del Anexo) para el nivel de significación elegido y $\nu = (m - 1)^2$ grados de libertad.

La hipótesis nula de la prueba es que el estado de un evento en cualquier punto de la serie es independiente del estado del evento anterior. La hipótesis alternativa es el estado de un evento en un punto cualquiera de la serie depende del estado del evento anterior, o que la secuencia tiene propiedades de Markov.

EJEMPLO 10

Cadenas de Markov

Se desea investigar si la sucesión de sedimentos descriptos en un perfil de detalle del Grupo Sierras Bayas, presentado arriba, exhibe un comportamiento Markoviano que pueda asociarse a algún proceso.

Las hipótesis que se testean son:

H_0: el estado de un evento en cualquier punto de la serie es independiente del estado del evento precedente.

H_1: el estado de un evento en cualquier punto de la serie depende del estado del evento precedente.

$\alpha = 0{,}05$

$\nu = (m - 1)^2 = 9$

De la Tabla 2 del Anexo, $\chi^2_{0{,}05;\,9} = 16{,}92$

El estadístico de prueba es:

$$\chi^2 = 2 \sum_{i=1}^{m} \sum_{j=1}^{m} \mathrm{Tr_{ij}} \ln \frac{\mathrm{Pr_{ij}}}{\mathrm{Pr_j}}$$

Matriz de probabilidades marginales Pr_j	*D*	*C*	*Q*	*P*
	0,2 67	0,2 33	0,2 67	0,2 33

Pr_{ij}/Pr_j	*D*	*C*	*Q*	*P*
D	0,000	2,143	1,406	0,536
C	2,143	0,000	1,071	0,612
Q	0,469	1,071	0,000	2,679
P	1,607	0,612	1,071	0,000

Ln (Pr_{ij}/Pr_j)	*D*	*C*	*Q*	*P*
D		0,762	0,341	-0,624
C	0,762		0,069	-0,491
Q	-0,758	0,069		0,985
P	0,474	-0,491	0,069	

$Tr_{ij}\ ln\frac{Pr_{ij}}{Pr_j}$	*D*	*C*	*Q*	*P*
D		3,049	1,023	-0,624
C	3,049		0,138	-0,491
Q	-0,758	0,138		4,926
P	1,423	-0,491	0,138	

$$\sum_{i=1}^{m} \sum_{j=1}^{m} Tr_{ij}\ ln\frac{Pr_{ij}}{Pr_j} = 11{,}52$$

$$\chi^2 = 2 \cdot 11{,}52 = 22{,}04$$

Como $\chi^2 > \chi^2_{0,05;\,9}$ (22,04>16,92), se rechaza la hipótesis nula. Se infiere que esta serie tiene memoria de las litologías precedentes, se considera que existe un patrón de recurrencia en la secuencia. La figura muestra las transiciones más frecuentes con las probabilidades asociadas a cada transición.

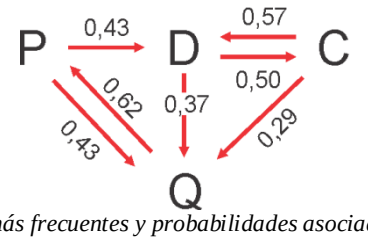


Figura 10. Transiciones litológicas más frecuentes y probabilidades asociadas en la secuencia de Sierras Bayas. P: psamopelitas, D: dolomías, C: calizas, Q: cuarcitas.

En la secuencia descripta las transiciones entre eventos sucesivos solamente pueden ser entre categorías diferentes, algunos autores las llaman transiciones verdaderas. Sin embargo, si se estudian secuencias litológicas, se puede registrar el tipo litológico que ocurre con un espaciamiento constante, por ejemplo relevar la litología cada metro. En estos casos pueden aparecer transiciones de la misma categoría y la matriz de transiciones Tr_{ij} no tendrá 0 en la diagonal. En estas secuencias el estadístico de la prueba de hipótesis χ^2 es el de las tablas de contingencia (Capítulo 9).

Análisis para series del mismo evento

Existen secuencias temporales o espaciales en donde el mismo evento se repite en el tiempo o distancia. Por ejemplo fechas de inundaciones o crecidas excepcionales de un río, flujos de barro, terremotos.

Si se requiere conocer si la **frecuencia de ocurrencia cambia** con el tiempo o la distancia se realiza un **análisis de regresión**. Primero se agrupan los registros de la secuencia en intervalos de la misma duración y se cuentan los eventos que ocurren en cada intervalo de modo de obtener la frecuencia en forma análoga a la construcción de un histograma. También en este caso es importante seleccionar adecuadamente la amplitud del intervalo de modo que no sean demasiado chicos para evitar que haya o sean muy pocos los intervalos sin datos, ni muy grandes como para que el número de intervalos sea bajo. Se recomienda contar al menos con 6 intervalos y que el 80% de ellos tengan al menos un dato. Luego se realiza el análisis de regresión con las frecuencias de ocurrencia. Se calcula la recta de regresión y se realiza la prueba de hipótesis sobre la pendiente (Capítulo 8). Si la pendiente es diferente que cero, existen evidencias de cambios en las frecuencias de ocurrencias. Dependiendo del patrón de cambio puede ser necesario efectuar un ajuste con una función diferente.

Por otra parte, cuando interesa saber si los eventos están **distribuidos al azar o agrupados o distanciados** en el tiempo, se realiza un **análisis de supervivencia**. Para ésta análisis, primero se obtiene información de la longitud del intervalo (tiempo o espacio) transcurrido entre eventos

sucesivos. Luego se construye la distribución de frecuencias para todas las longitudes de intervalo observadas. Se calcula además los porcentajes remanentes (o sobrevivientes) de casos cuyo intervalo **es mayor que** el intervalo analizado. Los porcentajes remanentes se utilizan para hacer un gráfico de supervivencia con la longitud del intervalo en abscisas y el logaritmo del porcentaje remanente en ordenadas. Para facilitar la interpretación del gráfico de supervivencia se suele realizar otro gráfico con el logaritmo del porcentaje de supervivientes. Este gráfico siempre tiene pendiente negativa. Cuando los eventos están distribuidos al azar la gráfica es lineal (Fig. 11a). Si los eventos están regularmente espaciados la pendiente es suave al inicio y luego se torna más pronunciada (Fig. 11b), mientras que si los eventos están agrupados la pendiente es abrupta en los intervalos cortos y decrece suavemente hacia los intervalos largos (Fig. 11c).

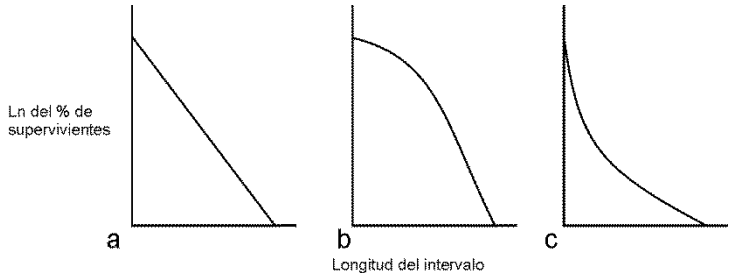


Figura 11. Logaritmo de la función de supervivencia de secuencias con patrón a: aleatorio, b: aproximadamente regular, c: agrupado.

EJEMPLO 11

Análisis de supervivencia

Los datos adjuntos muestran el año en que un volcán, suponga ubicado en la zona de subducción entre de la Placa Indoaustraliana bajo la Placa Euroasiática, mostró alguna actividad en el periodo 1850 - 2008. Interesa conocer si la frecuencia de ocurrencia en lapsos de 10 años cambia en el periodo analizado.

1850	1866	1887	1908	1934	1959	1977	1990
1851	1866	1890	1908	1934	1960	1978	1991
1852	1867	1892	1911	1938	1961	1978	1991
1852	1870	1894	1911	1939	1962	1979	1992
1853	1873	1895	1912	1942	1962	1980	1993
1854	1874	1895	1914	1943	1963	1983	1994
1854	1874	1896	1921	1944	1966	1983	1995
1854	1874	1900	1923	1946	1969	1984	1997
1856	1875	1903	1924	1950	1972	1984	2000
1857	1880	1904	1924	1952	1972	1986	2000
1859	1881	1904	1924	1953	1973	1987	2001
1860	1882	1905	1931	1955	1974	1987	2007
1861	1886	1907	1933	1956	1974	1988	2008

La Tabla 9 muestra el número de eventos volcánicos registrados en lapsos de 10 años. La figura 12 representa los datos de la tabla y la recta que mejor ajusta de ecuación $Y = 1{,}071 + 0{,}002\ X$. La prueba de hipótesis realizada sobre la pendiente confirma que la pendiente de la recta es 0 ($F_{1;14}$ = 0,044, No Significativa). Se interpreta que la frecuencia de ocurrencia de los eventos no cambia en el periodo 1850-2000.

Corresponde también estudiar el patrón de distribución temporal entre eventos sucesivos. La segunda columna de la tabla 10 muestra el número de eventos sucesivos cuya ocurrencia se produjo en periodos menores a 1 año, 1 año, 2 años, ..., 7 años que es la máxima separación entre un evento y el siguiente. Por ejemplo hay 21 eventos que se produjeron el mismo año, 37 que sucedieron con 1 año de diferencia (i.e. 1852-1853, 1856-1857), solo 1 que se distanció del precedente 7 años (2001-2007). La tercera columna de la tabla es el porcentaje de eventos respecto al total (104). Es evidente que el lapso entre eventos

sucesivos más frecuente es 1 año (41,1%). La cuarta columna es el porcentaje de eventos remanentes o "sobrevivientes" con intervalos "más largos que" 0 años, 1 año, ..., 7 años. Los porcentajes en este caso se calculan respecto al total de casos (90). Estos porcentajes se utilizan para hacer el gráfico de supervivencia (Fig. 13a). Los datos de la quinta columna (Ln % remanente) se usan para facilitar la interpretación del gráfico de supervivencia (Fig. 13b). En figura 13b los datos se disponen aproximadamente en una recta por lo que se infiere una distribución aleatoria.

Intervalo	*N° eventos*	*Intervalo*	*N° eventos*
1850-59	11	1930-39	6
1860-69	5	1940-49	4
1870-79	6	1950-59	6
1880-89	5	1960-69	7
1890-99	6	1970-79	9
1900-09	8	1980-89	9
1910-19	4	1990-99	8
1920-29	5	2000-09	5

Tabla 9. Número de eventos volcánicos registrados en intervalos de 10 años. Note que hay años en que ocurrió más de un evento, por ejemplo en 1924 se produjeron 3 eventos.

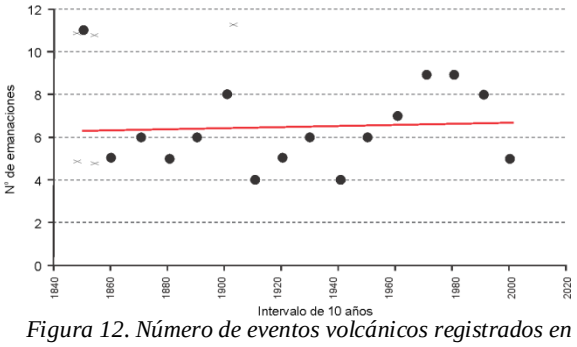


Figura 12. Número de eventos volcánicos registrados en lapsos de 10 años. La línea es la recta que mejor ajusta (Y = 1,071 + 0,002 X).

Eventos sucesivos en intervalos de longitud ...	*N° casos*	*% del total de casos*	*% de remanentes*	*Ln % remanentes*
0	21	23,3	76,7	4,34
1	37	41,1	35,6	3,57
2	12	13,3	22,2	3,10
3	12	13,3	8,9	2,19
4	4	4,4	4,4	1,48
5	2	2,2	2,2	0,79
6	0	0	2,2	0,79
7	1	2,2	0	0,00

Tabla 10. Número de años que transcurrieron entre eventos sucesivos.

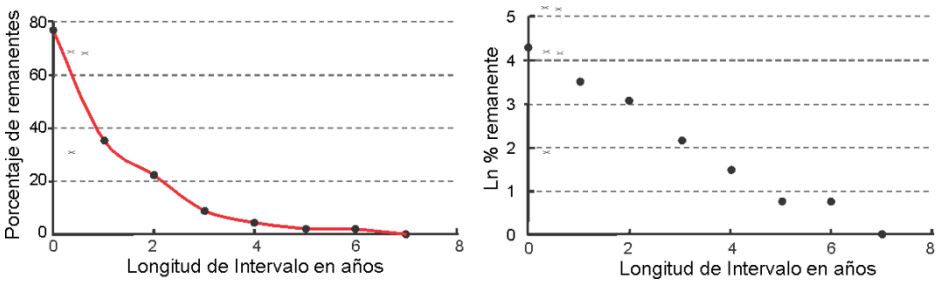


Figura 13. a) Gráfico de supervivencia. b) Gráfico del logaritmo de porcentaje de supervivientes.

INTRODUCCIÓN AL ANÁLISIS DE DATOS COMPOSICIONALES

Introducción

Los datos composicionales son aquellos que se expresan como proporciones, porcentajes o partes por millón y cuya suma es un valor constantes *k* (igual a 1, 100 o 10^6, respectivamente). Un sinnúmero de datos geológicos son composicionales: análisis químicos de elementos mayoritarios, de elementos minoritarios y traza, modas mineralógicas y sedimentológicas, análisis granulométricos de sedimentos, recuento de especies en asociaciones fósiles, hidroquímica de agua subterránea. También aparecen, frecuentemente en estudios ambientales de uso y distribución del suelo, distribución de contaminantes en agua, aire y suelo, etcétera. Además, en muchas otras oportunidades se utilizan subconjuntos de éstas composiciones como lo son los populares diagramas ternarios.

Aunque son numerosos los trabajos que presentan abundantes datos composicionales, al presente las conclusiones derivadas del análisis de estos datos numéricos son descriptivos, semicuantitativas y se limitan a caracterizar la muestra. Las estimaciones cuantitativas que tradicionalmente se utilizan se centran en describir las composiciones en términos de composición promedio y la variabilidad de cada componente en términos de poca, mucha, grande, mayor que, menor que, etcétera. Por otra parte a la hora de establecer relaciones entre componentes se indican en términos relativos y son del tipo "a medida que aumenta el contenido de A, decrece el de B".

Igualmente amplia es la utilización de diagramas ternarios, usados para visualizar datos numéricos que pueden ser expresados en términos de relaciones de tres componentes (porcentajes o proporciones ternarias). Ejemplos de diagramas ternarios se encuentran en geoquímica, sedimentología y paleontología pero son especialmente populares en las variadas ramas de la petrología de rocas ígneas, metamórficas y sedimentarias, aquí su rol se centra en esquemas de clasificación y de discriminación de ambientes geotectónicos y, otras veces, se indican líneas de tendencias. En algunas ocasiones se agregan a los datos puntuales los promedios aritméticos de los análisis, en un paso posterior, se suele proceder a la construcción de los llamados campos de variación composicional con el objeto de capturar la dispersión de los datos.

La primera referencia donde se identifican los problemas del análisis e interpretación de los datos composicionales se encuentra en el trabajo de Karl Pearson publicado en 1887. A mediados del siglo XX, Chayes (1960) identifica las dificultades a la suma constante. Krumbein (1962), Chayes y Kruskal (1966), Le Maitre (1982), Davis (1986) y Rock (1988) proponen algunas estrategias de análisis. Sin embargo no es hasta que en 1986 que Aitchison, en su monografía "*The statistical analysis of compositional data*", sienta las bases para el análisis moderno de los datos composicionales sorteando la restricción de la suma constante. Actualmente existen numerosas líneas de investigación en datos composicionales, el más importante es liderado por Vera Pawlowsky-Glhan en la Universidad de Girona, España.

Cabe mencionar, por último, que si bien los datos composicionales son esencialmente multivariantes y se analizan con métodos multivariantes (i.e. Análisis de Componentes Principales, Análisis de agrupamientos) que exceden los alcances de este libro, se presentan en este capítulo las bases del análisis y algunas estrategias exploratorias e inferenciales.

Datos Composicionales

Los datos composicionales, o cerrados, se caracterizan por que, para cada individuo, la suma de sus constituyentes o **partes** es aproximadamente un valor constante *k*, dentro de un cierto límite de incertidumbre. Por ejemplo la suma aproximada de concentraciones de ciertos elementos expresadas como gramos en 100 gramos de muestra es 100. La incertidumbre proviene, en el caso de datos geoquímicos, por errores en las mediciones o es causada al no considerar la presencia de los elementos traza. Además de estos datos, que se definen como intrínsecamente cerrados, existe otro conjunto en los que la suma constante se establece para presentar los datos en los diagramas ternarios.

Formalmente un dato composicional es un vector x ($x = (x_1, x_2, \cdots, x_D) \mid x_j >, j = 1, 2, ..., D; 1 + 2 + ...+ D = k$), cuyas D **partes** o componentes son **valores positivos** que sumados dan la unidad, o en forma general alguna constante fija k. Una matriz de datos composicionales **X** tiene como columnas las correspondientes d-partes composicionales y cada fila es un dato composicional.

$$\mathrm{X} = \begin{vmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2d} \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ x_{i1} & & & & x_{id} \end{vmatrix}$$

Los datos composicionales se enuncian en términos estadísticos como realizaciones (vector aleatorio) de una composición cuyo espacio muestral se llama simplex d-dimensional, S^D.

Para el caso particular en que $D = 3$, el simplex S^3 se puede representar con diagramas ternarios (Capítulo 2). Un ejemplo de tres componentes de una arena en el que las etiquetas de las tres componentes son cuarzo total (Qt), feldespatos (F) y líticos (L) [0,63; 0,35; 0,02] (Fig. 1).

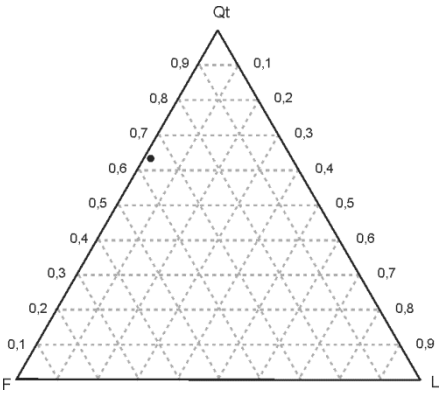


Figura 1. Diagrama ternario cuarzo (Qt), feldespatos (F), líticos (L), se muestra la ubicación de una muestra de composición [0,63; 0,35; 0,02].

Principales problemas que causa la restricción de la suma constante

La restricción impuesta por la suma constante impide la aplicación de los procedimientos estadísticos descriptos en los capítulos previos. Observe que el cambio de una de las partes provoca el cambio en al menos una de las otras partes de la composición. Ignorar o tratar inapropiadamente la restricción de la suma constante, tiene al menos tres inconvenientes graves que deben ser evitados para no incurrir en análisis erróneos y resultados irrelevantes: sesgo en las correlaciones, incoherencias subcomposicionales, problemas a la hora de establecer modelos lineales y en general aplicar las operaciones clásicas del espacio real vectorial a los datos composicionales (por ejemplo la distancia euclidiana).

Correlaciones espurias

Karl Pearson a fines del siglo XIX (1987) fue quien primero identifica el problema de las correlaciones espurias (falsas) entre proporciones. Posteriormente Chayes (1948, 1960) trabajando con geoquímica y mineralogía de rocas ígneas y metamórficas vincula el problema con la suma constante. La restricción de la suma constante determina que la matriz de correlaciones entre partes (*r* de Pearson, Capítulo 8) presente correlación negativas no nulas. Las correlaciones no son libres de tomar cualquier valor en el intervalo (-1,1). Se puede ver la trascendencia de este hecho pues, además de ser relevante por sí mismo, un gran número de pruebas estadísticas multivariantes parten del

cálculo de matrices de correlación o de varianza-covarianza (ej: Análisis de Componentes Principales; Análisis Cluster, Análisis Discriminante, etc.).

EJEMPLO 1

Correlaciones espurias

Se presentan los espesores de estratos presentes en dos secuencias A y B expresadas en metros y como porcentajes de la suma en ambas secuencias (tomado de Rollinson 1993).

	Espesor (m)		*Espesor (%)*	
	A	*B*	*A*	*B*
1	50	50	50,0	50,0
2	60	85	41,4	58,6
3	70	110	38,9	61,1
4	75	140	34,9	65,1
5	80	170	32,0	68,0
6	90	200	31,0	69,0
r	0,99		-1,00	

Claramente se ve que los espesores medidos (metros) en las secuencias expuestas en A y en B están positivamente correlacionados (r = 0,99), a medida que aumentan el espesor en A, aumenta en B. Sin embargo, al hacer la transformación de los datos a porcentajes se llega a la conclusión opuesta (r = -1), a medida que aumenta la proporción del espesor de A disminuye en B.

En el ejemplo 1 se cuenta con las mediciones, es decir con los datos absolutos, por lo que es fácil ver cual es la relación correcta, sin embargo, la gran mayoría de las veces se cuenta sólo con los valores relativos (proporciones o porcentajes) y el problema se traslada a composiciones con más de dos partes.

Incoherencias en las subcomposiciones

Una subcomposición es un subconjunto cualquiera de las partes de una composición en la que se reconstruye la condición de la suma constante. Intuitivamente se espera encontrar que una cierta correlación entre partes sea la misma en la subcomposición y en la composición original. Sin embargo las correlaciones varían desde la composición inicial a subcomposiciones cada vez más pequeñas.

EJEMPLO 2

Incoherencias en las subcomposiciones

Dos investigadores A y B estudian la composición de unas muestras de suelo. A está interesado en cuatro partes que clasifica como animal, vegetal, mineral y agua ($x_1, \ldots, x_4$). B seca las muestras y las analiza la composición animal, vegetal y mineral (s_1, s_2, s_3). De modo que los datos de B son una subcomposición de los de A (tomado de Aitchison 1997).

A	*B*
(x_1, x_2, x_3, x_4).	(s_1, s_2, s_3)
(0,1; 0,2; 0,1; 0,6)	(0,25; 0,50; 0,25)
(0,2; 0,1; 0,1; 0,6)	(0,50; 0,25; 0,25)
(0,3; 0,3; 0,2; 0,2)	(0,375; 0,375; 0,25)

La correlación entre las partes animal-vegetal registrada por el investigador A es corr(x_1, x_2) = 0,5 mientras que el investigador B obtiene corr(s_1, s_2) = -1.

Inconvenientes para establecer modelos lineales

Si se quiere establecer algún modelo lineal del tipo $y = a + bx$ fácilmente el valor calculado excede al del espacio muestral del simplex, es decir se excede el valor máximo de la composición total.

Bases del análisis de datos composicionales

Si bien los problemas de los datos composicionales se conocen desde hace más de cien años, es recién en 1982 que Aitchison propone una solución que evita la restricción de la suma constante. Las dificultades se sortean cuando la atención se centra en la **magnitud relativa de las partes**, es decir, en los cocientes x_i / x_j (i,j = 1,2, ..., D; $i \neq j$). Analizar las magnitudes absolutas de las partes x_1, x_2, ..., x_D de una composición carece de sentido. Este principio fundacional se denomina **invariancia por cambios de escala**.

Al trabajar con los cocientes entre partes desaparecen las correlaciones espurias y por otra parte, la magnitud relativa de las partes de una subcomposición no cambia en relación a la magnitud relativa entre las partes de la composición original ($s_i/s_j = x_i/x_j$). Por ejemplo, en la primera muestra de suelo del ejemplo 2, la relación animal/vegetal para la composición del investigador A y B son iguales, 0,1/0,2 = 0,25/0,50 = 0,5, la proporción de la componente animal es la mitad de la componente vegetal.

Sin embargo, la propuesta de Aitchison se enfrenta la dificultad para entender la geometría del espacio muestral de los datos composicionales, el simplex S^D, y su estructura algebraica ya que no es intuitiva como la geometría Euclidiana del espacio de los números Reales ($\Re$). Por ejemplo, considere la diferencia entre dos proporciones A y B en una muestra son 5% y 10% respectivamente, la distancia entre ellas es 5 unidades, pero esta diferencia no significa lo mismo cuando se tratara de 50% y 55%. En el primer caso la diferencia entre A y B es de 50%, en el segundo, en cambio es tan solo del 10%. La distancia Euclidiana es la misma, hay un incremento de 5 unidades en ambos casos pero el incremento relativo es sustancialmente diferente.

La metodología que propone Aitchison para sortear la compleja geometría del Simplex, se basa en tomar los logaritmos de los cocientes entre partes y llevando los datos al octante positivo de los reales $\Re^{D-1}$. Con esta simple estrategia es posible aplicar cualquier método de estadística clásica.

Operaciones básicas en el Simplex y transformaciones

Operación de Clausura

La operación de clausura (simbolizada con *C*) permite expresar un conjunto de datos como una composición, partiendo de la premisa que no existen datos ausentes o con valores negativos. Es una transformación que hace corresponder a cada vector $w = (w_1, w_2, \cdots, w_D)'$ de $\Re_+^D$ su dato composicional asociado

$$C(w) = \left(\frac{kw_1}{w_1+w_2+\cdots+w_D}, \frac{kw_2}{w_1+w_2+\cdots+w_D}, \cdots, \frac{kw_D}{w_1+w_2+\cdots+w_D}\right) \qquad (12.1)$$

en S^D con *k* la constante de clausura.

Esto es, la clausura se obtiene dividiendo cada componente w_i por la suma de todos todas las *d*-partes ($w_1 + \ldots + w_D$) y multiplicando el resultado por la constante *k*.

EJEMPLO 3

Clausura

Los siguientes datos corresponden al análisis modal de una arena. Se identificaron 444 granos de cuarzo (Qt), feldespato (F) y líticos (L).

Qt	*F*	*L*
302	136	6

Operando la clausura se obtiene la siguiente composición *x* = (68,02; 30,63; 1,35)

Operación de Perturbación

La operación de perturbación es equivalente a una traslación en el espacio de los reales. La perturbación es un cambio de escala. La perturbación (simbolizada ⊕) de una composición $x = (x_1, x_2, \ldots, x_D)$ en S^D por otro vector $y = (y_1, y_2, \ldots, y_D)$ en S^D, resulta en un nuevo vector *z*.

$$z = x \oplus y = \boldsymbol{C}(x_1 y_1, x_2 y_2, \ldots, x_d y_d) \qquad (12.2)$$

donde *C* significa la operación de clausura.

Perturbar *x* por su inversa $x^{-1} = (1/x_1, 1/x_2, \ldots, 1/x_D)$, resulta en el elemento neutro $e = C(1, 1, \ldots, 1) = (c/D, c/D, \ldots, c/D)$, que se ubica en el baricentro de la composición. Si se tratara de una composición ternaria $e = C(1, 1, 1) = (c/3, c/3, c/3)$ que es el centro del diagrama ternario[19].

Es común utilizar la inversa del centro de una composición ($g_{(X)}^{-1}$) como vector perturbador para lograr un re-escalamiento optimo enviando las observaciones a su baricentro (el concepto de centro se explica en el siguiente apartado). Es importante mencionar que la operación de perturbación puede ser

llevada a cabo, sobre valores muestrales, sobre líneas y límites de campos composicionales. Cuando se realiza una perturbación se aconseja entonces efectuarla sobre datos, líneas bordes de campo composicionales para no perder los marcos de referencia.

Cabe aclarar que, cuando se efectúa una perturbación lo único que queda invariante son, en el caso de los diagramas ternarios, los vértices V. Por convención para distinguir los vértices de aquellos diagramas en donde no se ha efectuado la operación de perturbación se los designa como "pV" (para V la etiqueta del vértice) o V^c con *c* como superíndice para indicar que se encuentra centrado.

Ciertos procesos geológicos como los cambios producidos durante el metamorfismo, la meteorización, la fragmentación que modifica la distribución del tamaño de las partículas se pueden modelar con perturbaciones.

EJEMPLO 4

Perturbación

Los datos corresponden a la composición mineralógica de 7 arenas cuyos granos son de cuarzo (Qt), feldespato (F) y líticos (L).

Qt	*F*	*L*
0,63	0,35	0,02
0,64	0,31	0,05
0,64	0,31	0,05
0,65	0,32	0,03
0,65	0,25	0,10
0,58	0,28	0,14
0,74	0,16	0,10

El vector perturbador es la inversa del centro $g_{(X)}$ = (0,66; 0,28; 0,06)

$g_{(X)}^{-1} = C\left(\frac{1}{0,66}, \frac{1}{0,28}, \frac{1}{0,06}\right) = (0,07; 0,15; 0,78)$

Para la primera muestra de arena

x = (0,63; 0,35; 0,02)

Perturbar *x* con $g_{(X)}^{-1}$

$x \oplus g_{(X)}^{-1}$ = C(0,63·0,07; 0,35·0,15; 0,02·0,78) =

= (0,37; 0,49; 0,14)

Operación de Potenciación

La potenciación, simbolizada con ⊗, es análoga a la multiplicación en los reales. Para cualquier composición $x \in S^D$ y cualquier número real $a \in \Re^+$, resulta una nueva composición *y*

$$y = x \otimes a = C(x_1^a, x_2^a, \cdots, x_D^a), \qquad (12.3)$$

donde *C* significa la operación de clausura.

La potenciación es útil para describir relaciones de regresión para composiciones.

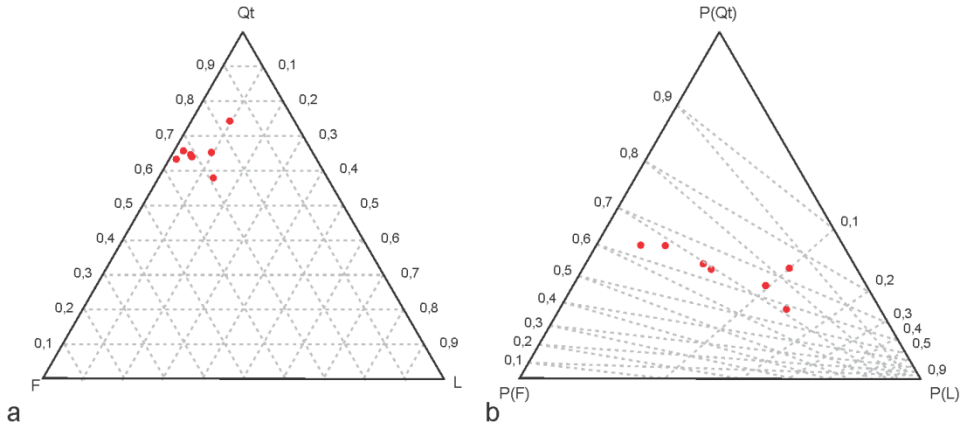


Figura 2. a. Diagrama ternario de la composición mineralógica de arenas.

b. Diagrama ternario de las mismas arenas centrado (perturbado por la inversa del centro).

EJEMPLO 5

Potenciación

Los datos del ejemplo 4 se potencian con $a = 3$.

Para el primer dato

$x = (0{,}63;\ 0{,}35;\ 0{,}02)$

$$y = x \otimes 3 = C(0{,}63^3;\ 0{,}35^3;\ 0{,}02^3)$$

$$y = (0{,}85;\ 0{,}15;\ 0{,}00)$$

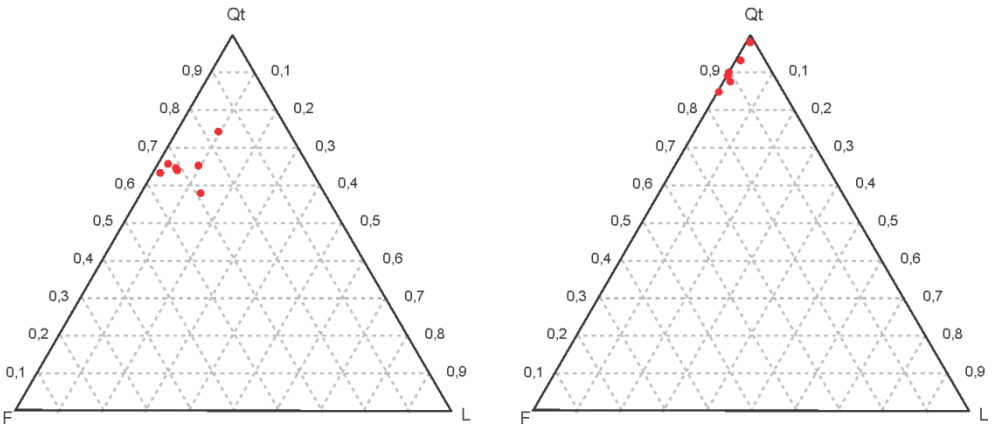


Figura 3. a. Diagrama ternario de la composición mineralógica de arenas.

b. Diagrama ternario de las mismas arenas después de efectuar la operación de potenciación con el escalar a = 3.

Subcomposiciones y amalgamas

Con la operación de clausura es posible hacer un recorte del espacio *d*-dimensional S^D, a un espacio de menores dimensiones. Esta disminución se puede lograr con una subcomposición o con una amalgama.

Como se indicó más arriba una **subcomposición** es un subconjunto de la composición, por ejemplo la subcomposición $s = (s_1, \cdots, s_C)$ de una composición $x = (x_1, \cdots, x_D)$ de *d*-partes. Se realiza cuando el interés es estudiar sólo parte de los componentes de un sistema, por ejemplo de todos los óxidos mayoritarios el interés está en las relaciones CNK (CaO, Na_2O, K_2O).

Una **amalgama** es un recorte del espacio *d*-dimensional que se realiza sumando dos o más partes, por ejemplo $a = (a_1, \cdots, a_{D-n})$ de una composición $x = (x_1, \cdots, x_D)$ de *d*-partes. El diagrama AFM de Irvine y Baragar (1971) que de los ocho óxidos mayoritarios de las rocas ígneas se centra en A: K_2O + Na_2O, F: Fe_2O_3 y M: MgO, es uno de los numerosos ejemplos geológicos. En estos casos se realiza primero la suma de algunos componentes y luego se procede a la operación de clausura.

EJEMPLO 6

Subcomposición y amalgama

La composición mineralógica de una arena es cuarzo monocristalino (Qm), feldespatos (F), cuarzo policristalino (Qp) y líticos (L) en la siguiente proporción:

(Qm, F, Qp, L) = (0,6; 0,2; 0,1; 0,1)

Se reduce una composición de 4 partes en 3, eliminando el cuarzo monocristalino

(s_F, s_{Qp}, s_L) = ***C***(0,2; 0,1; 0,1) = (0,50; 0,25; 0,25)

Se reduce la composición de 4 partes sumando cuarzo monocristalino y policristalino (Qt)

(a_{Qt}, a_F, a_L) = ***C***(0,7; 0,2; 0,1) = (0,7; 0,2; 0,1)

Transformaciones

Aitchison (1986) en su monografía propone una serie de transformaciones que tienen por objetivo llevar los datos desde el simplex S^D al espacio de los números reales $\Re^+$. Las transformaciones están basadas en el cálculo de logcocientes que permite, no solo describir adecuadamente los datos, sino que además utilizar los métodos estadísticos uni y multivariantes paramétricos y no paramétricos para analizar y modelar los datos composicionales.

La metodología que se aplica en el estudio de los datos composicionales consiste en: 1) transformar los datos para llevarlos al espacio sin restricciones, 2) analizar y modelar los datos en el espacio real y 3) transformar los datos para volverlos al simplex.

Transformación logcociente aditiva

La transformación logcociente aditiva, *alr*, de una composición $x \in S^D$ a $y \in \Re^{D-1}$, se define como:

$$y = alr(x) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, ..., \ln \frac{x_{D-1}}{x_D} \right) = (y_1, y_2, \cdots, y_{D-1}) \cdot \qquad (12.4)$$

Esto es, se elige una parte cualquiera como divisor del logcociente, siempre el mismo para todos los datos. Es importante aclarar que no existe pérdida de información al efectuar esta transformación.
La transformación logcociente aditiva es biyectiva[20] pero no es simétrica en las partes de *x* ya que la parte elegida como denominador tiene un protagonismo diferente respecto a las demás componentes.
La operación inversa de la transformación *alr* desde $\Re^{D-1}$ a S^D se realiza con:

$$agl(y) = C(e^{y_1}, e^{y_2}, \cdots, e^{y_{D-1}}) = x. \qquad (12.5)$$

Los datos *alr*-transformados sólo debe usarse cuando se requiere realizar pruebas de hipótesis (multinormalidad, contrastar hipótesis sobre el centro y estructura de covarianza) y regiones de confianza para el centro de una composición). Sin embargo, la falta de simetría impide que el cálculo de rectas de regresión cuando se opera con variables que surgen como cocientes con el mismo divisor y ser sumamente cuidadosos a la hora de interpretar diagramas de dispersión de logcocientes con el mismo divisor.

EJEMPLO 7

Transformación logcociente aditiva

Se realiza la transformación *alr* de la composición de una muestra de un gneiss tonalítico; se eligió la sílice como denominador del logcociente.

	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
x	62,74	0,75	18,42	4,71	0,05	1,62	4,03	5,59	1,90	0,19
alr(x)		-4,423	-1,226	-2,590	-7,131	-3,659	-2,746	-2,418	-3,499	-5,796

Transformación logcociente centrada

La transformación logcociente centrada *clr*, de una composición $x \in S^D$ a $z \in \Re^D$, se define como

$$clr(x) = \left(\ln \frac{x_1}{g(x)}, \ln \frac{x_2}{g(x)}, ..., \ln \frac{x_D}{g(x)} \right) = (z_1, z_2, \cdots, z_D) = z, \qquad (12.6)$$

con $g(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$, la media geométrica de cada dato composicional.
Cabe aclarar que en un conjunto de datos, cada composición (vector fila) tiene una media geométrica diferente. El resultado de la transformación de la matriz de datos (**X**), es una matriz con el mismo número de columnas que la matriz original (**Z**).
La inversa de esta transformación es

$$clr^{-1}(z) = C(e^{z_1}, e^{z_2}, \cdots, e^{z_D}) = x. \qquad (12.7)$$

La transformación *clr* es también biyectiva, pero a diferencia de la *alr* es simétrica entre las partes. Sin embargo, esta simetría en las componentes tiene el problema de una nueva restricción sobre los datos transformados cuya suma es cero y con una matriz de varianza-covarianza singular[21]. Esta

transformación se utiliza en pruebas que se basan en distancias tales como el análisis de agrupamiento (cluster) y para resolver problemas acerca del grado de correlación entre componentes y entre individuos (Análisis de Componentes Principales).

EJEMPLO 8
Transformación logcociente centrada
Se utilizan los datos de la composición de una muestra de un gneiss tonalítico del ejemplo 8.

$g(x) = (\prod_{i=1}^{D} x_i)^{1/D}$ =2,20

	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
x	62,74	0,75	18,42	4,71	0,05	1,62	4,03	5,59	1,90	0,19
clr(x)	3,349	-1,074	2,123	0,759	-3,782	-0,310	0,602	0,931	-0,150	-2,447

Análisis exploratorio de datos composicionales

Estadística composicional descriptiva

La estadística descriptiva estándar no se ajusta a la geometría del Simplex de ahí que se definen otros estadísticos que son el centro, la variancia total y la matriz de variación. Es útil también analizar la estructura de la matriz de varianza-covarianza de los logcocientes entre todas las partes.

Centro

La media o promedio de un conjunto de *n* datos composicionales **X** es un **vector de medias geométricas clausurado** llamado el **Centro** de la composición $\boldsymbol{g_m}$ (Aitchison, 1986, 1997)

$$\boldsymbol{g_m} = C(g_1, g_2, \ldots, g_D), \tag{12.8}$$

con $g_i = \left(\prod_{i=1}^{n} x_{ij}\right)^{1/n}$, $i = 1, 2, \ldots, D$.

Para obtener el centro se calcula la media geométrica, $g_1, g_2, \ldots, g_D$, de cada parte (columna de la matriz **X**) y luego se aplica la operación de clausura *C*. Note la diferencia con la media geométrica de la transformación *crl* (expresión 12.6) que utiliza la media geométrica de cada dato (fila de la matriz **X**).

Varianza total

Para medir la dispersión total se utiliza la varianza total, totvar[X] dada por

$$tot\,\mathrm{var}[X] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathrm{var}\left(\ln\frac{x_i}{x_j}\right). \tag{12.9}$$

La varianza total también se calcula como la suma de las varianzas de lo las partes *clr*-transformadas.

$$totvar[X] = \sum_{i=1}^{D} clr\ x_i \tag{12.10}$$

Matriz de variación

Otra medida apropiada para describir la variabilidad de un conjunto de datos composicionales es la matriz de variación que tiene en la mitad superior derecha las varianzas de los logcocientes entre partes y en la mitad inferior izquierda las medias aritméticas de los logcocientes entre partes.

	1	*2*	*3*	⋯	*d*	*D*
1		$var\left(ln\frac{x_1}{x_2}\right)$	$var\left(ln\frac{x_1}{x_3}\right)$			$var\left(ln\frac{x_1}{x_D}\right)$
2	$\bar{X}\left(ln\frac{x_1}{x_2}\right)$		$var\left(ln\frac{x_2}{x_3}\right)$			$var\left(ln\frac{x_1}{x_D}\right)$
3	$\bar{X}\left(ln\frac{x_1}{x_3}\right)$	$\bar{X}\left(ln\frac{x_1}{x_2}\right)$				
⋮						
d						$var\left(ln\frac{x_d}{x_D}\right)$
D	$\bar{X}\left(ln\frac{x_1}{x_D}\right)$	$\bar{X}\left(ln\frac{x_1}{x_2}\right)$			$\bar{X}\left(ln\frac{x_d}{x_D}\right)$	

Medias — *Varianzas*

(12.11)

Matriz de varianza-covarianza de los datos clr-*transformados*

La matriz de varianza-covarianza de los datos *clr*-transformados es útil para descifrar las relaciones entre las partes. La matriz es simétrica con las varianzas en la diagonal y las covarianzas en las demás celdas. El análisis de esta matriz se suele hacer gráficamente a partir de un *biplot* que resulta de un Análisis de Componentes Principales efectuado con los datos *clr*-transformados (Aitchison y Greenacre 2002).

	clr(1)	*clr(2)*	⋯	*clr(D)*
clr(1)	$var(1)$	$covar(1,2)$		$covar(1,D)$
clr(2)		$var(1)$		$var(2,D)$
⋮			⋱	
clr(D)				$var(D)$

(12.12)

EJEMPLO 9

Descripción estadística de datos composicionales

Se presentan datos de los óxidos mayoritarios de 17 muestras de lavas tolehiticas del Kilahuea Iki de Hawaii (tomado de Rollinson 1993, con datos de Richter y Moore 1966).

Muestra	SiO_2	TiO_2	Al_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
1	48,34	2,33	11,49	11,63	0,18	13,59	9,86	1,90	0,44	0,23
2	48,90	2,47	12,40	11,58	0,17	11,10	10,65	2,02	0,47	0,24
3	45,70	1,70	8,35	12,16	0,17	23,11	6,99	1,33	0,32	0,16
4	45,58	1,54	8,18	12,06	0,17	23,91	6,80	1,28	0,31	0,15
5	49,36	3,31	12,12	11,68	0,17	10,48	9,67	2,25	0,65	0,30
6	46,67	2,00	9,52	11,99	0,18	19,34	8,20	1,54	0,38	0,18
7	48,18	2,34	11,44	11,73	0,18	13,67	9,88	1,89	0,46	0,22
8	47,97	2,32	11,19	11,83	0,18	14,34	9,65	1,86	0,45	0,21
9	46,99	2,01	9,91	11,86	0,18	18,32	8,59	1,58	0,37	0,19
10	49,17	2,73	12,54	11,85	0,18	10,05	10,55	2,09	0,56	0,26
11	48,45	2,47	11,81	11,73	0,18	12,53	10,19	1,93	0,48	0,23
12	47,94	2,24	11,18	11,78	0,18	14,65	9,59	1,82	0,41	0,21
13	48,47	2,35	11,64	11,41	0,18	13,24	10,13	1,89	0,45	0,23
14	49,47	2,45	11,93	11,44	0,18	11,07	10,72	1,71	0,79	0,24
15	48,74	2,44	11,60	11,56	0,18	12,35	10,45	1,67	0,79	0,23
16	49,68	3,03	12,93	11,30	0,17	8,85	10,98	2,24	0,55	0,27
17	49,26	2,50	12,33	11,40	0,18	10,52	11,06	2,02	0,48	0,23
g_i	48,15	2,33	11,12	11,70	0,18	13,61	9,55	1,81	0,48	0,22
g_m	48,57	2,35	11,21	11,80	0,18	13,73	9,64	1,82	0,48	0,22
var clr	0,0121	0,0175	0,0134	0,0148	0,135	0,0715	0,0144	0,0160	0,0315	0,0165

La suma de las medias geométricas de los 10 óxidos es 99,15 por ello se opera la clausura y se obtiene el vector g_m.
La varianza total totvar = 0,2211

Matriz de variación

	SiO_2	TiO_2	Al_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
SiO_2		0,0244	0,0114	0,0019	0,0011	0,0931	0,0138	0,018	0,0573	0,0218
TiO_2	-3,0281		0,0062	0,0378	0,0328	0,2051	0,0094	0,0038	0,0292	0,0009
Al_2O_3	-1,4658	1,5623		0,0221	0,0165	0,1675	0,0008	0,0033	0,0347	0,0044
FeO	-1,4145	1,6136	0,0513		0,0011	0,0695	0,0253	0,0292	0,0748	0,0349
MnO	-5,6055	-2,5774	-4,1397	-4,191		0,0798	0,0183	0,0245	0,0665	0,0298
MgO	-1,2636	1,7645	0,2022	0,1509	4,3419		0,1728	0,1830	0,2593	0,1992
CaO	-1,6173	1,4107	-0,1515	-0,2028	3,9882	-0,3537		0,0066	0,0328	0,0074
Na_2O	-3,2836	-0,2555	-1,8178	-1,8691	2,3219	-2,02	-1,6663		0,0471	0,0036
K_2O	-4,6179	-1,5899	-3,1521	-3,2035	0,9876	-3,3544	-3,0006	-1,3343		0,0279
P_2O_5	-5,3918	-2,3637	-3,9259	-3,9773	0,2137	-4,1282	-3,7744	-2,1081	-0,7738	
	Medias $ln\frac{x_i}{x_j}$									

Varianzas $ln\frac{x_i}{x_j}$

Las características más destacadas de este conjunto de datos son:

La varianzas relativa más grandes aparece entre el K_2O y MgO $\left(var\left(ln\frac{K_2O}{MgO}\right) = 0{,}2593\right)$ además dado que la media es negativa $\left(\bar{X}\left(ln\frac{K_2O}{MgO}\right) = -3{,}3544\right)$ la proporción de K_2O tiende, en promedio, a ser menor que la de MgO (sombreado rosado). Por otra parte el hecho que $\bar{X}\left(ln\frac{K_2O}{MgO}\right) < \sqrt{var\left(ln\frac{K_2O}{MgO}\right)}$ sugiere que, para un número importante de observaciones, el $ln\frac{K_2O}{MgO}$ es negativo con el correspondiente porcentaje de MgO siempre mayor que el de K_2O.

En el otro extremo la varianza relativa más chicas aparece entre CaO y Al_2O_3 $\left(var\left(ln\frac{CaO}{Al_2O_3}\right) = 0{,}0008\right)$; además dado que la $\left(\bar{X}\left(ln\frac{CaO}{Al_2O_3}\right) = -0{,}1515\right)$es negativa la proporción de Al_2O es, en promedio, mayor que la de CaO y como $\bar{X}\left(ln\frac{CaO}{Al_2O_3}\right) < \sqrt{var\left(ln\frac{CaO}{Al_2O_3}\right)}$ en la mayoría de las muestras el porcentaje de Al_2O es mayor que CaO (sombreado celeste).

Las varianzas relativas mayores se encuentran entre el MgO (indicativo del olivino en la fase fraccionada) y los elementos que no intervienen en esta fase y que están concentrados en el magma (K, Na, Ca, P, Ti, y Al).

En las rocas ígneas, los valores altos de variabilidad en la matriz de variación indicarían relaciones entre la fase solida y la fundida y se produciría entre los elementos que forman parte de los cristales y la fase fundida o entre dos o más minerales cristalizados.

Matriz de varianza-covarianza de datos *clr*-transformados

	clr(SiO_2)	*clr*(TiO_2)	*clr*(Al_2O_3)	*clr*(FeO)	*clr*(MnO)	*clr*(MgO)	*clr*(CaO)	*clr*(Na_2O)	*clr*(K_2O)	*clr*(P_2O_5)
clr(SiO_2)	0,0022	-0,0047	-0,0023	0,0039	0,003	0,015	-0,0025	-0,003	-0,0071	-0,0044
clr(TiO_2)		0,0128	0,0056	-0,0087	-0,0075	-0,0357	0,005	0,0094	0,0122	0,0114
clr(Al_2O_3)			0,0046	-0,005	-0,0035	-0,021	0,0052	0,0055	0,0054	0,0055
clr(FeO)				0,0076	0,0057	0,0294	-0,0056	-0,0059	-0,0132	-0,0082
clr(MnO)					0,005	0,023	-0,0034	-0,0049	-0,0104	-0,007
clr(MgO)						0,1208	-0,0227	-0,0262	-0,0488	-0,0337
clr(CaO)							0,0066	0,0049	0,0073	0,005
clr(Na_2O)								0,0098	0,0018	0,0086
clr(K_2O)									0,0408	0,0119
clr(P_2O_5)										0,0109

Las covarianzas negativas entre Mg-K, Mg-Ti, Mg-P, Mg-Na y Mg-Ca, enfatizan la repulsión entre el Mg (en las olivinas) y los elementos K, Ti, P, Na y Ca que están concentradas en la fase fundida (sombreadas en color rosa). La covarianza positiva Fe-Mg refleja la asociación entre estos metales en las olivinas (sombreada en color celeste). Los otros elementos con covarianzas pequeñas o muestran fuertes asociaciones entre sí.

En el contexto de las rocas ígneas, altas covarianzas positivas indicarían fuerte asociación entre los elementos y esto se interpreta que coexisten en el mismo mineral.

Medidas de distancia entre composiciones

Para medir la distancia entre dos grupos de datos composicionales se utiliza la distancia que hay entre el centro de las dos composiciones. Sin embargo, dado que en el Simplex la distancia euclidiana no es apropiada se han propuesto dos medidas alternativas: la Distancia de Aitchison aplicada sobre la matriz de datos **X**, y la distancia de Mohalonobis aplicada a los datos *clr*-transformados.

$$d_A(X, X') = d_e(clr(X), clr(X')) \qquad (12.13)$$

X y X' son composiciones en S^D, d_E es la distancia Euclidiana en $\Re^D$ y *clr* la transformación logcociente centrada $clr(x) = \left(ln\frac{x_1}{g(x)}, \cdots, ln\frac{x_D}{g(x)}\right)$, con $g(x) = (\prod_{i=1}^{D} x_i)^{1/D}$.

Análisis estadístico inferencial de datos composicionales

Pruebas de hipótesis sobre normalidad multivariante de datos composicionales

Como se vio en los capítulos precedentes, establecer la distribución subyacente de un conjunto de datos es decisivo para la correcta implementación de algunos procedimientos estadísticos incluidos los datos composicionales.

Especialmente importante es chequear si un conjunto de datos composicionales siguen la distribución **lognormal aditiva** (Aitchison 1996). Se dice que una composición **X**, cuyo espacio muestral es el S^D, sigue una distribución log-normal aditiva (*additive logistic normality*), cuando el conjunto de los datos *alr* transformados (*alr*(X)) sigue una distribución normal multivariante.

Las hipótesis que se someten a prueba son H_0: las muestras proviene de una distribución log-normal aditiva o su equivalente, los datos *alr*-transformados provienen de una distribución normal multivariante, vs. H_A: las muestras no proviene de una distribución log-normal aditiva o su equivalente, los datos *alr*-transformados no provienen de una distribución normal multivariante.

Si bien los alcances de este libro no abarcan métodos multivariantes, solo se mencionara que Aitchison (1986) propone realizar tres pruebas de bondad de ajuste que se basan en comparar la distribución relativa acumulada observada con la distribución de probabilidades acumulada teórica: Anderson-Darling, Cramer-von Mises y Watson.

El método requiere calcular los estadísticos de prueba para: a) la distribución de cada uno de las partes, llamadas distribuciones marginales (*d*-pruebas marginales), b) cada una de las posibles distribuciones bivariadas, llamada distribución angular bivariada (½ $d(d - 1)$ pruebas bivariadas) y c) una prueba multivariante conjunta o global, la prueba Radius *d*-dimensional de la distribución *Radius*.

Todos los estadísticos calculados se contrastan con los valores críticos para el nivel de significación elegido para las pruebas. La hipótesis nula se acepta cuando el estadístico de prueba es menor que el valor crítico de tabla (Tabla 1). Todas estas pruebas se pueden realizar con el software CoDaPack v2.01.8 desarrollado por Marc Comas-Cufí y Santiago Thío-Henestrosa (Departamento de Ciencias de la Computación y Matemática Aplicada, Universidad de Girona, España).

Prueba de hipótesis para dos grupos de datos composicionales

Cuando se necesita conocer si existen diferencias entre dos conjuntos de datos composicionales las mismas pueden ubicarse entre los vectores de medias composicionales, entre la estructura de la matriz de varianza–covarianza o entre ambas. Cabe aclarar que las pruebas que permiten contrastar estas hipótesis sólo son válidas si existe normalidad composicional de cada grupo de datos (normalidad multivariante de los datos *alr*-transformados).

Considere dos muestras de **datos *alr*-transformados** de poblaciones normal multivariante N(μ_1, Σ_1) y N(μ_2, Σ_2) con tamaño n_1 y n_2 respectivamente (μ el vector de medias y Σ la matriz de varianza-covarianza). Se contrastan cuatro hipótesis en forma consecutiva:

	Nivel de significación (%)		10	5	2,5	1
Prueba de los Marginales	Anderson-Darling	$Q_A\left[1+\frac{4}{N}-\frac{25}{N^2}\right]$	0,656	0,787	0,918	1,092
	Cramer-von Mises	$Q_C\left[1+\frac{4}{2N}\right]$	0,104	0,126	0,148	0,178
	Watson	$Q_W\left[1+\frac{4}{2N}\right]$	0,096	0,116	0,136	0,163
Pruebas Bivariadas y Radius	Anderson-Darling	Q_A	1,933	2,492	3,070	3,857
	Cramer-von Mises	$\left[Q_C-\frac{0,4}{N}+\frac{0,6}{N^2}\right]\cdot\left[1+\frac{1}{N}\right]$	0,347	0,461	0,581	0,743
	Watson	$\left[Q_W-\frac{0,1}{N}+\frac{0,1}{N^2}\right]\cdot\left[1+\frac{0,8}{N}\right]$	0,152	0,187	0,221	0,267

Tabla 1. Valores críticos para las pruebas de log-normalidad aditiva (tomado de Aitchison 1986).

1) Los vectores de medias y las matrices de varianza-covarianza son iguales ($\mu_1 = \mu_1$, $\Sigma_1 = \Sigma_2$).
2) Las estructuras de covarianzas son las mismas pero los vectores de medias son diferentes ($\mu_1 \neq \mu_2$, $\Sigma_1 = \Sigma_2$).
3) Los vectores de medias y las estructuras de covarianzas son diferentes ($\mu_1 \neq \mu_2$, $\Sigma_1 \neq \Sigma_2$).

Si la primera hipótesis se acepta se acaba la prueba, si se rechaza se debe averiguar si las diferencias están en el vector de medias o en la estructura de la matriz de varianza-covarianza, por lo que se contrastan la 2º y 3º hipótesis, en este caso el orden es indistinto. Si se rechazan ambas se prosigue con la 4º hipótesis.

Para llevar adelante estas pruebas se utilizan estimadores de los vectores de medias y de la matriz de varianzas-covarianzas

$$\hat{\mu} = \bar{Y} = (\overline{y_1}, \overline{y_2}), \quad \text{con } \overline{y_j} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}\,, \;\; j = 1, 2 \tag{13.14}$$

$$\hat{\Sigma} = S = \begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}, \;\; \text{con} \begin{cases} s_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_{ij} - \overline{y_j}\right)^2, & j = 1,2, \\ s_{12} = \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - \overline{y_1})\,(y_{i2} - \overline{y_2}). \end{cases} \tag{13.15}$$

Observe la primera hipótesis $\mu_1 = \mu_1$, $\Sigma_1 = \Sigma_2$. Si esta hipótesis es verdadera, entonces las distribuciones son idénticas y solo es necesario estimar un vector esperado de medias y una matriz de varianza-covarianza que considere los datos de las dos muestras,

$$\overline{Y_c} = \frac{n_1\overline{Y_1} + n_2\overline{Y_2}}{n_1 + n_2} \tag{13.16}$$

y

$$S_c = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2} + \frac{n_1 n_2 (\overline{Y_1} - \overline{Y_2})'(\overline{Y_1} - \overline{Y_2})}{(n_1 + n_2)^2}, \tag{13.17}$$

$(\overline{Y_1} - \overline{Y_2})'(\overline{Y_1} - \overline{Y_2})$ es el producto de vector columna con dos componentes con el mismo vector traspuesto para una composición de tres partes en S^3. El estadístico para contrastar con la hipótesis alternativa $\mu_1 \neq \mu_1$, $\Sigma_1 \neq \Sigma_2$, es

$$n_1 ln \frac{|S_c|}{|S_1|} + n_2 ln \frac{|S_c|}{|S_2|}. \quad (13.18)$$

Este estadístico se compara con χ^2 con 1/2(3 – 1)(3 +2) grados de libertad.

Para la segunda hipótesis, $\mu_1 \neq \mu_2$, $\Sigma_1 = \Sigma_2$, solo se asumen iguales las matrices de covarianza. El estimador ponderado de ambas matrices de varianza covarianza es

$$S_p = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2}. \quad (13.19)$$

La hipótesis alternativa de esta prueba es de diferencia absoluta de los vectores de medias y de las matrices de varianza-covarianza, $\mu_1 \neq \mu_1$, $\Sigma_1 \neq \Sigma_2$ y el estadístico de prueba es

$$n_1 ln \frac{|S_p|}{|S_1|} + n_2 ln \frac{|S_p|}{|S_2|}. \quad (13.20)$$

Este estadístico se compara con χ^2 con 1/2(3 – 1)3 grados de libertad.

La tercera hipótesis, $\mu_1 \neq \mu_2$, $\Sigma_1 \neq \Sigma_2$, asume que el vector de medias es igual en ambas muestras. Existe aquí un problema, el estimador se calcula con un procedimiento iterativo cuyos pasos son los siguientes:

1. Sean los valores iniciales $S_{1h} = S_1$, $S_{2h} = S_2$,
2. Calcular $\bar{y}_h = \left(n_1 S_{1h}^{-1} + n_2 S_{2h}^{-1}\right)^{-1} \cdot \left(n_1 \overline{Y_1} S_{1h}^{-1} + n_2 \overline{Y_2} S_{2h}^{-1}\right)$,
3. Calcular $S_{1h} = S_1 + (\overline{Y_1} - \overline{Y_h})'(\overline{Y_1} - \overline{Y_h})$ y $S_{2h} = S_2 + (\overline{Y_2} - \overline{Y_h})'(\overline{Y_2} - \overline{Y_h})$,
4. Repetir pasos 2 y 3 hasta convergencia.

Para chequear la convergencia se calcula la máxima diferencia entre la estima de S_{1h} en dos iteraciones y también en S_{2h}. La máxima diferencia de ambas se compara con el valor fijo 0,01.

Nuevamente, la hipótesis alternativa de esta prueba es de diferencia absoluta de los vectores de medias y de las matrices de varianza-covarianza, $\mu_1 \neq \mu_1$, $\Sigma_1 \neq \Sigma_2$ y el estadístico de prueba es

$$n_1 ln \frac{|S_{1h}|}{|S_1|} + n_2 ln \frac{|S_{2h}|}{|S_2|} \quad (13.21)$$

Este estadístico se compara con χ^2 con $D - 1$ grados de libertad.

Todas estas pruebas se pueden realizar con el software CoDaPack v2.01.8.

El caso de los ceros

En el inicio del capítulo se mencionó que el vector composicional está restringido a valores positivos dado que toda la metodología propuesta se basa en el análisis de logcocientes. Sin embargo, en

muchos casos la matriz de datos presenta ceros. Estos ceros pueden ser por redondeo o ceros esenciales.

Los **ceros por redondeo** aparecen en una componente donde el cero se registra cuando no es detectado o es extremadamente chico (por ejemplo el valor es inferior al umbral de detección del método analítico). Los **ceros esenciales o estructurales** son los valores nulos que indican la ausencia de una de las partes en la composición. Dado que la presencia de ceros no admite ningún análisis composicional, se proponen diferentes estrategias para estudiarlos.

Para los ceros por redondeo, Martín-Fernández (2001) y Martín-Fernández *et al*. (2003) proponen la estrategia de **reemplazo multiplicativo**. Si δ_k es el valor de reemplazo derivado del umbral de detección para la *k*-esima parte y *x* es una observación que contiene ceros por redondeo, se construye la observación *r* (r_1, r_2, ..., r_D) sustituyendo los ceros de *x* mediante la siguiente expresión:

$$r_k = \begin{cases} \delta_k & si\ x_k = 0, \\ x_k(K - \sum_{\{x_l=0\}} \delta_k) & si\ x_k > 0, \end{cases} \qquad (12.22)$$

para *k* la constante de la composición y $\sum_{\{x_l=0\}} \delta_k$ la suma de los valores de reemplazo de cero.

Si los ceros son esenciales pueden indicar que las composiciones pertenecen a diferentes poblaciones o bien que el componente no tiene significación para el estudio. La estrategia de análisis composicional en estos casos puede ser dividir la muestra en submuestras y tratarlos de forma independiente o efectuar amalgamas de las diferentes partes y analizar la muestra entera (Aitchison 1986, Martín-Fernández *et al*. 2003).

EJEMPLO 10

Reemplazo multiplicativo de ceros

Los datos corresponden a una asociación de foraminíferos bentónicos.

$\delta_k = 0{,}005;$ $\quad \sum_{\{x_l=0\}} \delta_k = 0{,}015\,;$ $\quad 1 - \sum_{\{x_l=0\}} \delta_k = 0{,}985$

Composición	*Ammonia beccarii*	*Buccella peruviana*	*Bulimina patagonica*	*Cibicides dispars*	*Cibicides mckannai*	*Epistominella exigua*	*Globobulimina affinis*	*Nonionella auris*
Original (w_i)	0	0,25	0,15	0	0	0,25	0,22	0,13
Reemplazado el 0 (x_i)	0,005	0,246	0,148	0,005	0,005	0,246	0,217	0,128

INTRODUCCIÓN AL ANÁLISIS DE DATOS DIRECCIONALES

Introducción

En los trabajos geológicos es muy frecuente obtener datos de azimut/inclinación de planos (fallas, vetas, diaclasas, estratificación, planos axiales de pliegues, etc.) y azimut/buzamiento de líneas (ejes de pliegues, estrías, lineamientos minerales, fábrica deposicional o clástica, etc.). Estas orientaciones pueden estar en dos dimensiones (2D) en cuyo caso se llaman datos circulares o en tres dimensiones (3D) y se denominan datos esféricos.

La primera referencia de análisis de datos direccionales se encuentra en el trabajo de John Michell[22], quien, hacia finales del siglo XVIII, utiliza estos métodos para estudiar la separación angular entre las estrellas. Su análisis adquiere impulso a comienzos de 1950 con los estudios paleomagnéticos en rocas volcánicas de Islandia realizados por geólogos y geofísicos de la Universidad de Cambridge. Pero fue a partir de la publicación del trabajo de Fisher en 1953 que su utilización se afianza en las investigaciones paleomagnéticas y se expande a otras ramas de la geología.

Bien se trate de 2D o de 3D es posible distinguir dos categorías de orientaciones: las direcciones y los ejes. Una dirección es una línea que tiene orientación y sentido, se simboliza con una flecha, por ejemplo las direcciones paleomagnéticas, direcciones de sedimentos y estructuras sedimentarias clastos. Una dirección con el agregado de magnitud se convierte en un vector como la velocidad del viento. Un eje, en cambio, es una línea de dirección que no tiene sentido, por ejemplo los ejes cristalográficos y ópticos de minerales, el eje axial de un pliegue.

Una variable circular se puede definir como aquella que representa direcciones en el plano y que se cuantifican mediante ángulos que varían de 0° a 360° (2π radianes). Las direcciones se miden en escalas de intervalos que no tienen un punto cero verdadero. Son datos de este tipo los azimut e inclinaciones que se miden con una brújula donde el círculo está dividido en 360 intervalos iguales que son los grados. Los azimut podrían medirse desde el norte magnético o geográfico y la inclinación desde un plano horizontal o vertical ya que la posición del cero es, como se dijo, arbitraria.

Una de las diferencias más importantes respecto a las variables lineales es que las variables circulares toman valores cíclicos con periodos que se repiten en los 360° o en un rango más limitado como las inclinaciones (±90°). Debido a esta propiedad los datos temporales (diarios, mensuales o anuales) se

pueden analizar como datos direccionales. Es simple convertir los datos expresados en unidades de tiempo en direcciones, sólo se calcula el valor del ángulo que le corresponde a la unidad (α) dividiendo *360°/k* (*k* = número de unidades en el ciclo completo), luego se sitúa el 0 en el círculo y por último se calcula el ángulo correspondiente al dato que se quiere ubicar. Por ejemplo, para convertir las horas de un día en direcciones o ángulos, el ángulo unidad es α=360°/24 horas=15°/hora, si la hora 0 está en dirección Norte, la hora 3 se representa con un ángulo de 45° medido desde el Norte.

Todas estas características y razones de índole tanto teóricas como prácticas confluyen para dar a las variables direccionales un tratamiento estadístico distinto al de las variables lineales mediante estadísticos descriptivos, distribuciones específicas y métodos para realizar diferentes pruebas de hipótesis. En este capítulo se describen algunos estadísticos y pruebas de hipótesis para datos en 2D.

Para el cálculo de estadísticos y pruebas se requieren tres funciones trigonométricas básicas: seno, coseno y tangente. Recuerde que las funciones trigonométricas se definen a partir de las razones trigonométricas de los triángulos rectángulos. Si se considera el ángulo α del vértice A del triángulo rectángulo (Fig.1), los lados del triangulo se denominan: Hipotenusa (*h*) al lado opuesto al ángulo recto, o lado de mayor longitud del triángulo rectángulo; Cateto opuesto (*a*) al lado opuesto al ángulo α y Cateto adyacente (*b*) al lado adyacente al ángulo α.

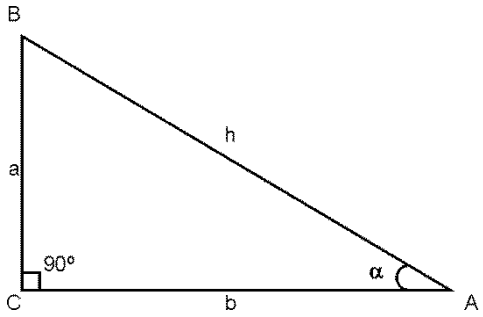


Figura 1. Triángulo rectángulo. a: cateto opuesto a α. b: cateto adyacente a α. h: hipotenusa, lado opuesto al ángulo recto.

El **seno** (sen) de un ángulo es la relación entre la longitud del cateto opuesto y la longitud de la hipotenusa (expresión 13.1).

$$sen\ \alpha = \frac{opuesto}{hipotenusa} = \frac{a}{h} \tag{13.1}$$

El **coseno** (cos) de un ángulo es la relación entre la longitud del cateto adyacente y la longitud de la hipotenusa (expresión 13.2).

$$cos\ \alpha = \frac{adyacente}{hipotenusa} = \frac{b}{h} \tag{13.2}$$

La **tangente** (tan) de un ángulo es la relación entre la longitud del cateto opuesto y la del adyacente (expresión 13.3).

$$tan\ \alpha = \frac{opuesto}{adyacente} = \frac{a}{b} \tag{13.3}$$

Suponga el caso de una circunferencia de radio uno (*r*) como la que se presenta en la figura 2, en el centro del círculo se encuentra el origen y se traza un eje vertical *X* y un eje horizontal *Y* (note la diferencia de las coordenadas cartesianas donde el eje horizontal es X y el vertical Y). Se acostumbra realizar las medidas desde el eje vertical X (Norte en la brújula) en sentido inverso a las agujas del reloj. En este círculo se puede ubicar un punto de dos modos diferentes: el primero es con sus **coordenadas polares** que son el ángulo α respecto al cero o punto de origen y la distancia *r* y, el segundo, es con sus **coordenadas rectangulares** (ver Capítulo 8) *X* e *Y* (Fig. 2). Las coordenadas polares *X* e *Y* son el seno y coseno del ángulo α respectivamente dado que

$$sen\ \alpha = \frac{opuesto}{hipotenusa} = \frac{X}{r} = \frac{X}{1}, \tag{13.4}$$

$$cos\ \alpha = \frac{adyacente}{hipotenusa} = \frac{Y}{r} = \frac{Y}{1}. \tag{13.5}$$

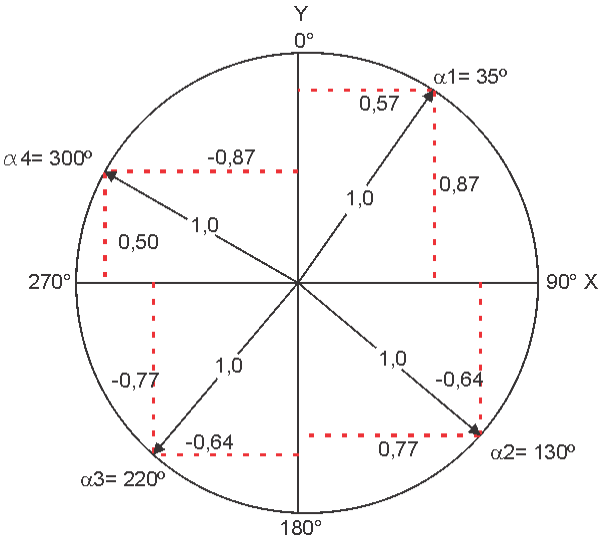


Figura 2. Circulo de radio 1. Para el punto 1, las coordenadas polares son α_1 = 35°, r = 1 y las coordenadas rectangulares X = 0,57 $\left(cos\ 35° = \frac{0,57}{1}\right)$ *e Y = 0,82* $\left(sen\ 35° = \frac{0,82}{1}\right)$*. Para el punto 2, las coordenadas polares son α_2 = 130°, r = 1 y las coordenadas rectangulares X = 0,77* $\left(cos\ 130° = \frac{0,77}{1}\right)$ *e Y = -0,64* $\left(sen\ 130° = \frac{-0,64}{1}\right)$*. Para el punto 3, las coordenadas polares son α_3 = 220°, r = 1 y las coordenadas rectangulares X = -0,77* $\left(cos\ 220° = \frac{-0,64}{1}\right)$ *e Y = -0,77* $\left(sen\ 220° = \frac{-0,64}{1}\right)$*. Para el punto 4, las coordenadas polares son α_4 = 300°, r = 1 y las coordenadas rectangulares X = -0 87*$\left(cos\ 300° = \frac{-0,87}{1}\right)$ *e Y = 0,50* $\left(sen\ 300° = \frac{0,50}{1}\right)$.

Estadística descriptiva

Representación gráfica

Un dato circular se puede simbolizar en el plano como un vector unitario o como un punto sobre un círculo unidad. La forma más simple de representar un conjunto de datos circulares es graficarlos en **diagramas de dispersión**, como puntos sobre un círculo unidad y cuando una dirección se repite los nuevo puntos se sitúan fuera del círculo sobre el radio correspondiente (Fig. 3a). El diagrama de dispersión permite visualizar las características más salientes de los datos tales como tendencia central, dispersión y el número de modas de la distribución.

Cuando los datos están agrupados la distribución se muestra con un **histograma circular** (Fig. 3b). Construir un histograma circular requiere calcular el área del sector del diagrama de modo que esta sea proporcional a la frecuencia de la clase que se representa en el perímetro. El radio del sector se obtiene con la expresión 13.6.

$$r^2 = \frac{2Af}{Nc}, \tag{13.6}$$

donde *r* es el radio del sector, *A* es el área total del histograma, *f* es la frecuencia de clase, *N* el tamaño de la muestra y *c* la amplitud del intervalo de clase expresada en radianes (1 radian = 180/π grados). Hay quienes indican que los histogramas de datos circulares no son adecuados para representar la distribución pues cambian su apariencia cuando se modifica el límite de los intervalos aun conservando la misma amplitud. En su reemplazo proponen graficar el histograma con círculos concéntricos que representen las frecuencias y un radios en cada intervalo de clase (Fig. 3c).

Los **diagramas de rosas** son muy difundidos cuando son muchos los datos a representar. Al igual que en el histograma los datos se agrupan en intervalos, en este caso la longitud de los radios que delimitan cada sector son proporcionales a la frecuencia relativa pero el área no lo es (Fig. 3d). El sector de la clase modal adquiere una proporción que no está acorde a su significado. Debido a esta distorsión no se recomienda este tipo de representación.

La representación gráfica de la distribución los datos esféricos se realizan casi siempre usando una proyección estereográfica como la red de Wulff o la red de Schmit donde la orientación se simboliza con un punto.

Estadísticos

Media angular

La media angular $\bar{\propto}$ de una muestra de *n* ángulos, α_1 hasta α_n, es una estima de la media angular de la población μ_α muestreada. La dirección de la media requiere conocer el ángulo medio $\bar{\propto}$ y la longitud.

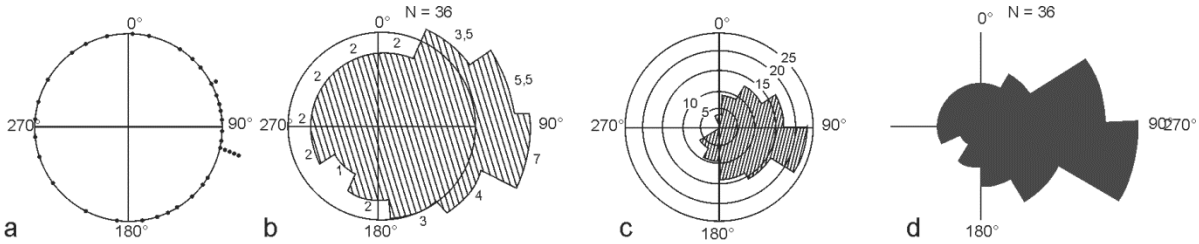


Figura 3: Gráficos de la distribución de frecuencias de la orientación del eje A de una muestra de 36 clastos psefiticos de un depósito de una planicie aluvial. a) diagrama de dispersión. b y c) histograma circular c) diagrama de rosas.

del vector medio *r*. Con las coordenadas polares de los datos se calcula

$$X = \frac{\sum_{i=1}^{n} cos\ \alpha_i}{n}, \tag{13.7}$$

$$Y = \frac{\sum_{i=1}^{n} sen\ \alpha_i}{n}, \tag{13.8}$$

$$r = \sqrt{X^2 + Y^2}. \tag{13.9}$$

El valor de $\bar{\propto}$ corresponde al ángulo cuyo coseno y seno se obtienen con

$$\cos \bar{\alpha} = \frac{X}{r}, \tag{13.10}$$

$$\text{sen}\, \bar{\alpha} = \frac{Y}{r}. \tag{13.11}$$

Otra alternativa para es a partir de la tangente de $\bar{\propto}$

$$\tan \bar{\alpha} = \frac{Y}{X}, \tag{13.12}$$

y luego calculando el arcotangente. En este caso se aplica

$$\bar{\propto} = \begin{cases} \arctan[Y/X] & si\ X > 0 \\ 180° + arctan[Y/X] & si\ X < 0 \end{cases}. \tag{13.13}$$

El vector medio *r* varía entre 0 y 1, cuando *r* es igual a cero, la media angular es indefinida y se concluye que no hay una dirección media, por el contrario cuando es 1, los datos están concentrados en una única dirección dominante. En el ejemplo 1 se muestra el cálculo para un conjunto de datos de azimut de 11 vetas del distrito minero de Manantial Espejo.

EJEMPLO 1

Ejemplos de cálculo de los estadísticos circulares: Media, Moda y Mediana angular, Dispersión media y Dispersión estándar

Los datos adjuntos son azimut de 11 vetas del Distrito Minero de Manantial Espejo, provincia de Santa Cruz.

AZIMUT	*Sen (Y)*	*Cos (X)*
55	0,8192	0,5736
81	0,9877	0,1564
92	0,9994	-0,0349
96	0,9945	-0,1045
109	0,9455	-0,3256
110	0,9397	-0,3420
111	0,9336	-0,3584
117	0,8910	-0,4540
132	0,7431	-0,6691
132	0,7431	-0,6691
154	0,4384	-0,8988

$\sum sen\ \alpha_i = 9{,}4352$ $\sum cos\ \alpha_i = -3{,}1264$

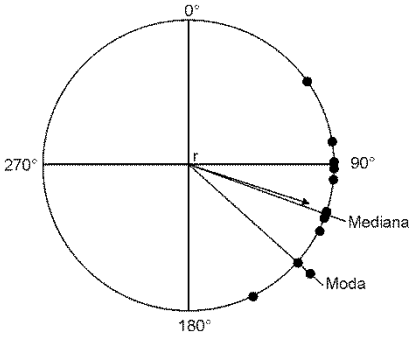
$n = 11;$ $\quad Y = \frac{9,4352}{11} = 0,8577;$ $\quad X = \frac{-3,1264}{11} = -0,2842$

$r = \sqrt{(-0,2842)^2 + 0,8577^2} = 0,9036$

$cos\,\bar{\alpha} = \frac{X}{r} = \frac{-0,2842}{0,9036} = -0,3145$ $\quad sen\,\bar{\alpha} = \frac{Y}{r} = \frac{0,8577}{0,9036} = 0,9492$

El ángulo con ese seno y coseno es $\bar{\propto} = 108°\ 19'$.
La mediana angular es 110°.
La moda angular es 132°.

Medidas de dispersión:
Rango = 154° - 55° = 99°
1 – *r* = 1 - 0,9036 = 0,0963
Desviación media $s = \frac{180°}{\pi}\sqrt{2(1-0,9036)} = 50,31°$
Desviación estándar $s' = \frac{180°}{\pi}\sqrt{-2\ln 0,9036} = 51,60°$



Media, Moda y Mediana angular, r de los datos de azimut de vetas del Distrito Minero de Manantial Espejo

Cuando los datos se presentan agrupados en arcos, la media angular se calcula con las siguientes fórmulas alternativas:

$$X = \frac{\sum_{i=1}^{n} f_i \cos \alpha_i}{n}, \qquad (13.14)$$

$$Y = \frac{\sum_{i=1}^{n} f_i sen\, \alpha_i}{n}, \qquad (13.15)$$

donde f_i es la frecuencia de la clase y α_i el punto medio del arco. El valor de *r* se subestima debido al agrupamiento y se debe corregir de modo que,

$$r_c = r\ c, \qquad (13.16)$$

$$c = \frac{\frac{d\pi}{360°}}{sen\left(\frac{2}{d}\right)} \quad \text{o} \quad c = \frac{\frac{d}{2}}{sen\left(\frac{d}{2}\right)}, \quad (d \text{ en radianes}) \qquad (13.17)$$

donde r_c es *r* corregido, *c* el factor de corrección y *d* la amplitud del arco.
El cálculo de $\bar{\propto}$ es válido sólo para muestras con una sola moda. Pero algunos datos geológicos pueden tener más de una moda y en estos casos el cálculo de la media angular difiere del descrito. Las muestras polimodales se consideran como extraídas de una distribución generada por el solapamiento de varias distribuciones unimodales por lo que se puede hablar de una mezcla de distribuciones.

Cuando las distancias entre modas son arbitraria no existen métodos para descomponer la muestra polimodal en varias muestras unimodales.

Sin embargo, si las **distribuciones son bimodales y diametralmente opuestas** (o axial) es posible reducirla a una sola muestra unimodal duplicando los ángulos ($2\alpha_i$). Si $2\alpha_i < 360°$ se registra $2\alpha_i$, pero si $2\alpha_i \geq 360°$ entonces se registra $360° - 2\alpha_i$. Luego se calcula el vector $\overline{\propto_2}$, r_2 utilizando las expresiones que permiten calcular la media angular (13.14, 13.15, 13.13 y 13.9). Para obtener el ángulo modal simétrico de la muestra original se debe cancelar el efecto de la duplicación de los ángulos, siendo:

$$\overline{\alpha_1} = \overline{\alpha_2}/2 \quad \text{o} \quad \overline{\alpha_1} = \overline{\alpha_2}/2 + 180°. \qquad (13.18)$$

Moda angular

La moda angular coincide con el dato que se repite más veces. Cuando los datos están agrupados se encuentra en la clase de mayor frecuencia (Fig. 3b y 3c). Pueden existir distribuciones con más de una moda, por ejemplo los sistemas de diaclasas conjugadas.

Mediana angular

La mediana angular es el valor del ángulo del diámetro, en el diagrama de dispersión, que divide a los puntos de la muestra de modo tal que deje en ambos semicírculos igual número de puntos (Fig. 3a). Si el número de puntos es impar la mediana puede coincidir con uno de los datos, si el número de datos es par se encuentra en medio de dos datos. Cuando los datos están uniformemente distribuidos alrededor del círculo, la mediana es indefinida. Para el caso de las vetas del distrito minero de Manantial Espejo la mediana coincide con el sexto dato y es 110°.

Dispersión angular

Para medir la dispersión angular existen medidas análogas a las medidas definidas en escala lineal.

El **rango** es el ángulo que corresponde al menor arco (porción de círculo de la circunferencia) que contiene a todos los datos de la distribución. Para los datos de azimut de vetas de Manantial Espejo el rango es 99° que corresponde al arco entre 55° y 154°.

La longitud del vector medio ***r*** de alguna forma se puede considerar una medida de concentración (*k*) de los datos y en contra partida *1 - r* es entonces una medida de dispersión. Recuerde que *r* no tiene

unidades y que varía entre 0, cuando los datos están dispersos y no se puede establecer una dirección media, y 1 cuando todos los datos están concentrados en la misma dirección (Fig. 4). De este modo, *1 – r = 0* indica entonces ausencia de dispersión y *1 – r = 1*, máxima dispersión.

Otra medida de dispersión es la **desviación media circular**, ***s***, que puede tener un valor comprendido entre un mínimo de 0 y un máximo de 81,03° (Fig. 4).

$$s = \frac{180°}{\pi}\sqrt{2(1-r)}. \qquad (13.19)$$

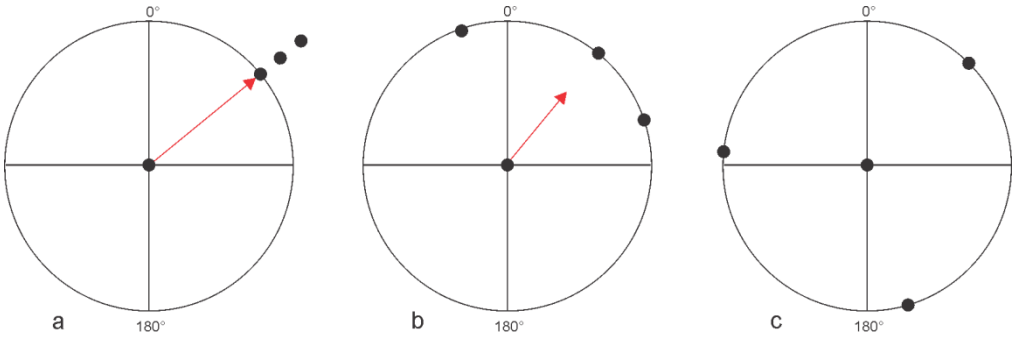


Figura 4. Medida de concentración y dispersión. a. r = 1, s = 0. b. r = 0,5, s = 57°. c. r = 0, s = 80°.

Por último, una medida análoga a la desviación estándar en escala lineal es la **desviación estándar circular**, ***s´***, cuyo valor varía entre cero e infinito.

$$s' = \frac{180°}{\pi}\sqrt{-2\ln r} \qquad (13.20)$$

Si los datos están agrupados se debe usar r_c en lugar de r. Ejemplos del cálculo se muestra para los datos de azimut de vetas de Manantial Espejo en el ejemplo 1.

Cuando las distribuciones son bimodales y diametralmente opuestas, el valor del desvío estándar medio, *s*, se calcula con el procedimiento descrito para la media angular, es decir duplicando los ángulos. El valor de *s* calculado corresponde a los valores angulares duplicados y es el doble del valor real, cuando debería ser la mitad (*s/2*).

EJEMPLO 2

Ejemplos de cálculo de los estadísticos circulares: Media, Moda y Mediana angular, Dispersión media y Dispersión estándar para una distribución bimodal diametralmente opuesta

Orientación de los fragmentos (céfalo, tórax y pigideo) del trilobites del Género Geragnostus hallados en una quebrada tributaria del río Huasamayo en la provincia de Jujuy (Cámbrico tardío-Tremadociano).

α_i	f_i	$2\alpha_i$	sen 2 α_i	f_i sen 2α_i	cos 2α_i	f_i cos 2α_i
0	2	0	0,0000	0,0000	1,0000	2,0000
15	3	30	0,5000	1,5000	0,8660	2,5980
30	5	60	0,8660	4,3300	0,5000	2,5000
45	10	90	1,0000	10,0000	0,0000	0,0000
60	6	120	0,8660	5,1960	-0,5000	-3,0000
75	3	150	0,5000	1,5000	-0,8660	-2,5980
90	1	180	0,0000	0,0000	-1,0000	-1,0000
180	1	0	0,0000	0,0000	1,0000	1,0000
195	3	30	0,5000	1,5000	0,8660	2,5980
210	5	60	0,8660	4,3300	0,5000	2,5000
225	10	90	1,0000	10,0000	0,0000	0,0000
240	6	120	0,8660	5,1960	-0,5000	-3,0000
255	3	150	0,5000	1,5000	-0,8660	-2,5980
270	2	180	0,0000	0,0000	-1,0000	-2,0000

$\sum sen\, 2\alpha_i = 45{,}052$ $\sum \cos 2\alpha_i = -1{,}000$

$n = 60;$ $Y = \frac{45{,}052}{60} = 0{,}7587;$ $X = \frac{-1{,}000}{60} = -0{,}0167$

$r = \sqrt{(-0{,}0167)^2 + 0{,}7587^2} = 0{,}7511$

$cos\, 2\bar{\alpha} = \frac{X}{r} = \frac{-0{,}0167}{0{,}7511} = -\,0{,}0222$ $sen\, 2\bar{\alpha} = \frac{Y}{r} = \frac{0{,}7587}{0{,}7511} = 0{,}9997$

El ángulo 2α es 91° 16′ y $\alpha \approx 45°$, significa que la distribución bimodal se ubica en el diámetro de la línea orientada a 45°.

Desviación media 2α $s = \sqrt{2(1 - 0{,}7511)} = 0{,}7056$

Desviación media α $\frac{s}{2} = 0{,}3528$

Modelos de distribuciones direccionales

Los modelos probabilísticos más comunes a los que se ajustan los datos direccionales son el uniforme y el circular normal.

Distribución circular uniforme

La distribución uniforme en el contexto de los datos direccionales describe la situación donde la probabilidad de ocurrencia de todos los puntos es la misma en todas direcciones, entonces la densidad de las direcciones es aproximadamente constante sobre la circunferencia del círculo (en el caso 2D) o sobre la superficie en la esfera (en el caso 3D). En contraposición una distribución circular no uniforme presenta uno o más grupos de datos que indican que tienen orientaciones preferenciales. La distribución circular uniforme es un buen modelo para procesos estocásticos y provee la densidad de probabilidad que se utiliza para testear la hipótesis nula de no existencia de una dirección preferencial.

Distribución circular normal o de Von Misses

La distribución de Von Misses es una de las distribuciones más usadas en el análisis direccional, su rol es análogo al de la distribución Normal en el análisis lineal razón por la cual es también conocida como Distribución Normal Circular. La función de densidad se muestra solo con fines ilustrativos. Para *n* muestras unimodales y simétricas está dada por

$$f(\alpha) = \frac{1}{2\pi I_0(k)} Exp[k \cos n\,(\alpha - \theta)], \qquad (13.21)$$

donde I_0 es la función de Bessel, *k* es el parámetro que de concentración que indica en qué medida se concentra la distribución alrededor de la dirección dominante θ. Ambos, θ y *k*, son los parámetros de la distribución. El parámetro de concentración *k* varía entre 0 y +∞ y en la medida que éste aumenta la función se aparta del círculo.

Si bien la distribución de muchos datos direccionales se aproxima a la distribución de Von Misses, los test estadísticos para probar el ajuste son complicados de calcular y están fuera del alcance de este libro. Por esta razón se asumirá que una distribución muestral de datos direccionales proviene de una distribución de Von Misses cuando su diagrama de dispersión es aproximadamente bilateral alrededor de la moda y muestra un decrecimiento paulatino en sus densidades desde la moda hacia la anti-moda.

Inferencia con datos direccionales

De igual forma que para datos lineales, las pruebas de hipótesis para los datos direccionales pueden ser no paramétricas o paramétricas, a dos colas o a una sola. Además se pueden calcular los intervalos de confianza para los parámetros poblacionales.

Límites de confianza para media angular

Los límites de confianza de la media angular se calculan con:

$$P(\bar{\alpha} - d < \mu_\alpha < \bar{\alpha} + d) = 1 - \alpha. \qquad (13.22)$$

La cantidad *d* se obtiene de las figura 1 del Anexo; se requiere conocer el tamaño de muestra *n* y el vector *r*.

Para el ejemplo de los azimut de las vetas de Manantial Espejo (n = 11, r = 0,9036, $\bar{\alpha} = 106°$), $d \approx 20$ para 95% de confianza. Los límites de la media poblacional angular son entonces

$$p(108° - 20° < \mu_\alpha < 108° + 20°) = 1 - 0{,}05$$
$$p(88° < \mu_\alpha < 128°) = 0{,}95.$$

Pruebas de bondad de ajuste

Existen varias pruebas de bondad de ajuste que se pueden aplicar con los datos direccionales para probar la hipótesis de uniformidad, el test de Chi cuadrado χ^2 (descripto en el Capítulo 6) y el test U^2 para una muestra de Watson (*Watson one-sample U^2 test*).

La prueba de **Chi cuadrado** χ^2 es la apropiada si los datos están agrupados. Requiere calcular las frecuencias esperadas, $\widehat{f_i}$, para cada frecuencia observada, f_i. Se recomienda agrupar los datos cuando la frecuencia esperada es menor a cuatro. En el ejemplo 3 se muestra una aplicación de la prueba.

EJEMPLO 3

Prueba de bondad de ajuste

Se pone a prueba la hipótesis que los datos de azimut de lineamientos del Distrito Manantial Espejo se distribuyen de manera uniforme.

H_0: $f_i = \widehat{f_i}$ (los datos de la muestra provienen de una población uniformemente distribuida alrededor de círculo)

H_1: $f_i \neq \widehat{f_i}$ (los datos de la muestra provienen de una población que no es uniformemente distribuida alrededor de círculo)

$\alpha = 0{,}01$

AZIMUT	f_i	$\widehat{f_i}$	χ_i^2
60-70	10	45,58	27,78
70-80	14	45,58	21,88
80-90	38	45,58	1,26
90-100	51	45,58	0,64
100-110	95	45,58	53,57
110-120	115	45,58	105,71
120-130	77	45,58	21,65
130-140	64	45,58	7,44
140-150	34	45,58	2,94
150-160	25	45,58	9,29
160-170	13	45,58	23,29
170-180	11	45,58	26,24

$$\widehat{f_i} = \frac{n}{k} = \frac{547}{12} = 45{,}58$$

$$\chi_c^2 = \sum_{i=1}^{k} \frac{\left(f_i - \widehat{f_i}\right)^2}{\widehat{f_i}}$$

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(10-45{,}58)^2}{45{,}58} + \cdots + \frac{(11-45{,}58)^2}{45{,}58} = 301{,}71$$

De la Tabla 2 del Anexo, $\chi^2_{(0{,}01;\ 11)} = 26{,}8$

La Hipótesis nula se rechaza P < 0,01. La población no está uniformemente distribuida.

El test U^2 **para una muestra de Watson** se usa cuando los datos no están agrupados. Para calcular el estadístico de prueba U^2, es necesario obtener u_i dividiendo cada dato por 360° ($u_i = \alpha_i / 360°$) y luego calcular las cantidades $\sum u_i$, $\sum u_i^2$, $\bar{u}$ y $\sum i u_i$,

$$U^2 = \sum u_i - \frac{(\sum u_i)^2}{n} - \frac{2}{n}\sum i\, u_i + (n+1)\bar{u} + \frac{n}{12}. \qquad (13.23)$$

Los valores críticos de esta prueba $U^2_{(\alpha;\ n;\ n)}$ se encuentran en la Tabla 16 del Anexo. En el ejemplo 4 se muestra una aplicación de la prueba.

EJEMPLO 4

Prueba de bondad de ajuste U^2 de Watson

Se utilizan los datos de azimut de vetas del Distrito Manantial Espejo del ejemplo 1.

H_0: Los datos de la muestra provienen de una población uniformemente distribuida alrededor de círculo.
H_A: Los datos de la muestra no provienen de una población uniformemente distribuida alrededor de círculo.
$\alpha = 0{,}05$

	i	*AZIMUT*	u_i	u_i^2	Iu_i
	1	55	0,1528	0,0233	0,1528
	2	81	0,2250	0,0506	0,4500
	3	92	0,2556	0,0653	0,7667
	4	96	0,2667	0,0711	1,0667
	5	109	0,3028	0,0917	1,5139
	6	110	0,3056	0,0934	1,8333
	7	111	0,3083	0,0951	2,1583
	8	117	0,3250	0,1056	2,6000
	9	132	0,3667	0,1344	3,3000
	10	132	0,3667	0,1344	3,6667
	11	154	0,4278	0,1830	4,7056
Suma	*n=11*		*3,3028*	*1,0480*	*22,2139*

$$\bar{u} = \frac{\sum u_i}{n} = \frac{3{,}3208}{11} = 0{,}3003$$

$$U^2 = \sum u_i - \frac{(\sum u_i)^2}{n} - \frac{2}{n}\sum i\, u_i + (n+1)\bar{u} + \frac{n}{12}$$

$$U^2 = 3{,}3028 - \frac{3{,}3028^2}{11} - \frac{2}{11}\ 22{,}2139 + 11\ (0{,}3003) + \frac{11}{12}$$

$$U^2 = 3{,}3028 - 0{,}9917 - \ 4{,}0389 + 3{,}3028 + 0{,}916 = 2{,}4917$$

De la Tabla 16 del Anexo, $U^2_{(0{,}05;\ 11;11)} = 0{,}182$

La Hipótesis nula se rechaza P < 0,01. La población no está uniformemente distribuida.

Significación de la media angular

La **prueba de Rayleigh** se utiliza para probar si una muestra se extrajo de una población que tiene distribución uniforme. Obviamente la estima de la media angular poblacional, μ_α, es mejor cuando la dispersión angular es baja que cuando es grande y cuando *r* es largo (≈1) que cuando es pequeño (≈0). La prueba se formula con el objetivo de testear cuán grande debe ser *r* para indicar que la población no es uniforme. Las hipótesis de la prueba son entonces: H_0: La población muestreada es uniforme (H_0: $\rho = 0$) lo que implica que no hay una dirección referencial y H_A: La población muestreada no es uniforme (H_A: $\rho \neq 0$), es decir que tiene una dirección preferencial. Las hipótesis se testean con el llamado *R de Rayleigh* y el *Z de Rayleigh*,

$$R = n{\cdot}r \tag{13.24}$$

$$z = \frac{R^2}{n} \quad \text{o} \quad z = n{\cdot}r^2. \tag{13.25}$$

Recuerde que *r* es la longitud del vector de la media angular. En la Tabla 17 del Anexo se encuentran los valores críticos de $Z_{\text{de Rayleigh } \alpha,n}$. Si la hipótesis nula se rechaza se puede concluir que la población tiene una dirección preferencial pero solo si la distribución es unimodal. Cuando la hipótesis nula no se rechaza se puede concluir que la distribución es uniforme solo si se asume que no tiene modas. El test falla si la distribución tiene más de una moda como en las distribuciones axiales que no son uniformes y tienen dos direcciones dominantes. Se muestra un ejemplo de cálculo para los datos del azimut de la vetas del Distrito Manantial Espejo.

EJEMPLO 5
Test de Rayleigh
Los datos corresponden al azimut de vetas de Manantial Espejo del ejemplo 1.

H_0: ρ = 0 (la población está uniformemente distribuida alrededor de círculo).
H_A: ρ ≠ 0 (la población no está uniformemente distribuida alrededor de círculo).

$n = 11$; $r = 0{,}9036$
$R = nr = (11)\,(0{,}9036) = 9{,}9397$

$$z = \frac{R^2}{n} = \frac{9{,}9397^2}{11} = 8{,}982$$

De la Tabla 17 del Anexo, $z_{\text{de Rayleigh } (0{,}05;\ 11)} = 2{,}92$
Se rechaza la hipótesis nula P < 0,01. La población no está uniformemente distribuida.

Pruebas de hipótesis para la media angular

Prueba para una muestra

El **test de Rayleigh modificado**, también llamado **test *V***, se utiliza para testear la hipótesis que supone, a priori, que la muestra fue tomada de una población que tiene una orientación particular. Por ejemplo si los vientos que soplan en el Río de La Plata un día de julio son predominantemente de dirección sudeste (N145°). Las hipótesis de la prueba son: H_0: La población muestreada es uniforme (H_0: ρ = 0) lo que implica que no hay una dirección preferencial y H_A: La población muestreada no es uniforme (H_1: ρ ≠ 0) y entonces existe una dirección particular. El estadístico *V* se calcula con la expresión

$$V = R\cos(\bar{\alpha} - \mu_0), \qquad (13.\ 26)$$

donde μ_0 es la dirección particular. En la Tabla 18 del Anexo se encuentran los valores críticos de $u_{(\alpha,\ n)}$.

$$u = V\sqrt{\frac{2}{n}}. \qquad (13.27)$$

Por otra parte se encontró que *V* aproxima a una cola de la distribución normal (*Z*). En estos casos si $n \geq 5$ se pueden usar $z_{0,05} = 1,645$ o $z_{0,01} = 2,236$, con desviaciones menores del 3% del valor nominal del error de tipo I (α).

Cuando los datos están agrupados *R* se debe determinar con r_c en lugar de con *r*.

EJEMPLO 6

Test de V

Se quiere testear la hipótesis que el Río de la Plata un día los vientos tuvieron dirección sudeste (N145°). Se dispone de datos de viento medidos cada tres horas.

H_0: $\rho = 0$ (la población está uniformemente distribuida alrededor de círculo)

H_A: $\rho \neq 0$ (la población no está uniformemente distribuida alrededor de círculo)

α_i	Sen α_i	Cos α_i
95	0,9962	-0,0872
110	0,9397	-0,3420
130	0,7660	-0,6428
145	0,5736	-0,8192
160	0,3420	-0,9397
150	0,5000	-0,8660
145	0,5736	-0,8192
150	0,5000	-0,8660

$n = 8$

$\sum sen\ \alpha_i = 5,1911\ ; \quad \sum cos\ \alpha_i = -5,3820$

$Y = \frac{5,1911}{8} = 0,6489\ ; \quad X = \frac{-5,3820}{8} = -0,6727$

$r = \sqrt{(-0,6727)^2 + 0,6489^2} = 0,9347$

$\cos \bar{\alpha} = \frac{X}{r} = \frac{-0,6727}{0,9347} = -0,7187; \quad \text{sen}\ \bar{\alpha} = \frac{Y}{r} = \frac{0,6489}{0,9347} = 0,6942$

El ángulo con ese seno y coseno es $\bar{\alpha} = 136°$.

R = 7,4775

V = R cos(136° - 145°)

= 7,4775 cos(4°) = 7,3855

$u = V\sqrt{\frac{2}{n}} \quad \rightarrow \quad u = 7,3855\sqrt{\frac{2}{8}} = 3,6927$

De la Tabla 18 del Anexo, $u_{(0,05;\ 8)} = 1,649$.

No existen evidencias para aceptar la hipótesis nula.

Otra alternativa para someter a prueba la hipótesis nula que los datos son muestreados de una población con dirección especifica (H_0: $\mu_\alpha = \mu_0$) y la hipótesis alternativa que la población muestreada no tiene esa dirección especifica (H_A: $\mu_\alpha \neq \mu_0$), es observar si la dirección especifica se encuentra comprendido entre los **límites del intervalo de confianza de la media angular**. En este caso el procedimiento es el mismo que el descripto para el cálculo de intervalo de confianza de la media

angular. La hipótesis nula se rechaza cuando la dirección específica está fuera de los límites del intervalo de confianza.

EJEMPLO 7

Test de la media angular para una muestra

Se utilizan los datos de viento del Río de La Plata del ejemplo 6.

H_0: $\mu_\alpha = 145°$ (la población tiene una media de 145°)
H_A: $\mu_\alpha \neq 145°$ (la población no tiene una media de 145°)

$$P(\bar{\alpha} - d < \mu_\alpha < \bar{\alpha} + d) = 1 - \alpha$$

La cantidad *d* se obtiene de la figura 1 del Anexo (*n* = 8, *r* = 0,9347, $\bar{\alpha} = 136°$), $d \approx 21$ para 95% de confianza.

$$P(136° - 21° < \mu_\alpha < 136° + 21°) = 1 - 0{,}05$$
$$P(115° < \mu_\alpha < 157°) = 0{,}095$$

Debido a que este intervalo de confianza contiene el valor hipotetizado de la media ($\mu_\alpha = 145°$), no se rechaza la hipótesis nula. Los vientos dominantes tuvieron dirección 145°.

Prueba para dos muestras

El **test de Watson-Williams** se utiliza para evaluar si dos muestras tienen la misma media angular. El estadístico de prueba es una modificación de la prueba lineal de *F* donde se reemplaza la media estimada por el *R* de Rayleigh obtenido de los datos de las muestras. El estadístico *F* se calcula con

$$F = K\frac{(N-2)(R_1+R_2-R)}{N-R_1-R_2}, \tag{13.28}$$

donde $N = n_1 + n_2$, R_1 y R_2 son los valores de *R* de Rayleigh (expresión 13.24) calculados para cada muestra en forma independiente, *R* es el valor de las dos muestras combinadas y *K* es un factor de corrección. Los valores de *K* se encuentran en la Tabla 18 del Anexo. Para usar el factor de corrección *K* hay que calcular el vector *r* promedio

$$\bar{r} = \frac{n_1 r_1 + n_2 r_2}{N} = \frac{R_1 + R_2}{N}. \tag{13.29}$$

El valor crítico de esta prueba es $F_{(\alpha;1;\,N-2)}$.

EJEMPLO

Test de Watson-Williams para dos muestras

Se pone a prueba la hipótesis que el rumbo de foliación S1 y S2 medidas a lo largo de un perfil en el río Catan Lil, provincia de Neuquén, son guales.

H_0: $\mu_{S1} = \mu_{S2}$
H_A: $\mu_{S1} \neq \mu_{S2}$

S1			S2		
$\alpha_i°$	sen α_i	cos α_i	$\alpha_i°$	sen α_i	cos α_i
144	0,5878	-0,8090	108	0,9511	-0,3090
140	0,6428	-0,7660	130	0,7660	-0,6428
122	0,8480	-0,5299	94	0,9976	-0,0698
88	0,9994	0,0349	104	0,9703	-0,2419
100	0,9848	-0,1736	76	0,9703	0,2419
70	0,9397	0,3420	94	0,9976	-0,0698
96	0,9945	-0,1045	146	0,5592	-0,8290
130	0,7660	-0,6428	164	0,2756	-0,9613
115	0,9063	-0,4226	178	0,0349	-0,9994
			174	0,1045	-0,9945

$\sum sen\ \alpha_{S1} = 7{,}6694 \qquad \sum cos\ \alpha_{S1} = -3{,}0716$

$n_{S1} = 9 \quad ; \qquad Y_{S1} = \frac{7{,}6694}{9} = 0{,}8522\ ; \quad X_{S1} = \frac{-3{,}0716}{9} = -0{,}3413$

$r_{S1} = \sqrt{(-0{,}3413)^2 + 0{,}8522^2} = 0{,}9180$

$\cos \bar{\alpha}_{S1} = \frac{X}{r} = \frac{-0{,}3413}{0{,}918} = -0{,}3718; \qquad \text{sen}\ \bar{\alpha}_{S1} = \frac{Y}{r} = \frac{0{,}8522}{0{,}918} = 0{,}9283$

α_{S1} = 126° 20´
R_{S1} = 8,22

$\sum sen\ \alpha_{S2} = 6{,}6271 \qquad \sum cos\ \alpha_{S2} = -4{,}8755$

$n_{S2} = 10 \quad ; \qquad Y_{S2} = \frac{6{,}6271}{10} = 0{,}6627; \quad X_{S2} = \frac{-4{,}8755}{10} = -0{,}4876$

$r_{S2} = \sqrt{(-0{,}4876)^2 + 0{,}6627^2} = 0{,}8227$

$\cos \bar{\alpha}_{S2} = \frac{X}{r} = \frac{-0{,}4876}{0{,}8227} = -0{,}5926 \qquad \text{sen}\ \bar{\alpha}_{S2} = \frac{Y}{r} = \frac{0{,}6627}{0{,}8227} = 0{,}8055$

α_{S2} = 111° 49´
R_{S2} = 8,2616

Al combinar los 19 datos de ambas muestras se obtiene:

$\sum sen\ \alpha_C = 7{,}6694 + 6{,}6271 = 14{,}2965$
$\sum cos\ \alpha_C = (-3{,}0176) + (-4{,}8755) = -7{,}9471$

$N = 19; \qquad Y_C = \frac{14{,}2965}{19} = 0{,}7524; \qquad X_C = \frac{-7{,}9471}{19} = -0{,}4183$
$r_C = 0{,}8609$

$$\bar{r} = \frac{R_1 + R_2}{N} = \frac{8{,}22 + 8{,}2616}{19} = 0{,}8678$$

R_C = 16,3568

De la Tabla 18 del Anexo, *K = 1,0916*

$$F = K\frac{(N-2)(R_1 + R_2 - R)}{N - R_1 - R_2} = 1{,}0916\frac{(19-2)(8{,}22 + 8{,}2616 - 16{,}3568)}{19 - 8{,}22 - 8{,}2616}$$

$F = 1{,}0916 \cdot 0{,}8944 = 0{,}9763$

De la Tabla 4 del Anexo, $F_{(0{,}05;\ 1;\ 17)} = 4{,}4513$

No existen evidencias para rechazar la hipótesis nula. Los rumbos de los planos de foliación S1 y S2 son iguales (p < 0,01).

Prueba para más de dos muestra

La **prueba de Watson-Williams** se generaliza a mas de dos muestras, en este caso la hipótesis nula es H_0: $\mu_{\alpha 1} = \mu_{\alpha 2} = \ldots = \mu_{\alpha k}$. El estadístico de prueba se calcula con

$$F = K\frac{(N-k)\left(\sum_{j=1}^{k} R_j - R\right)}{(k-1)\left(N-\sum_{j=1}^{k} R_j\right)}. \tag{13.30}$$

Acá *k* es el número de muestras, *R* es el R de Rayleigh combinado para las *k* muestras y *N* es la suma de los *k* tamaños de muestra. El factor de corrección *K* también se obtiene de la Tabla 18 del Anexo, en este caso

$$r_p = \frac{\sum_{j=1}^{k} n_j r_j}{N} = \frac{\sum_{j=1}^{k} R_j}{N}\ . \tag{13.31}$$

El valor crítico de esta prueba es $F_{(\alpha;\ k\text{-}1;\ N\text{-}k)}$. Este test se puede usar con datos agrupados sólo cuando los intervalos de agrupamiento no sean mayores a 10°. Por último, cabe aclarar que se trata de una prueba suficientemente robusta que es posible utilizar aunque no se cumplan los supuestos de normalidad circular (distribución de von Mises) e igualdad de dispersión de todas las poblaciones muestreadas.

INTRODUCCIÓN AL ANÁLISIS DE DATOS ESPACIALES

Introducción

En la mayoría de los trabajos geológicos se estudian fenómenos que se manifiestan en el espacio. Ya se trate de propiedades como la porosidad, permeabilidad, transmisividad, nivel piezométrico de un acuífero, de metales en un depósito mineral, de contaminantes en un sedimento, en el agua o en el aire, o del relieve, entre otros, siempre se utilizan mapas que muestran su distribución espacial.

Existen varias clases de mapas. Los más simples muestran sólo las coordenadas de los puntos en el plano. Otros, junto con las coordenadas, indican uno o dos valores de la variable (cota, espesores, valores de metales, etc.).

La distribución espacial de la variable se analiza con más facilidad en los mapas de curvas de isovalores (curvas de nivel, isopletas, isobaras, isohipsas, etc.) a partir del patrón que presentan las curvas (curvas más apretadas o más espaciadas). En otros mapas se utilizan colores o tonos de gris en lugar de curvas, un caso especial de este tipo son los mapas indicadores donde solo están presentes dos categorías.

Cuando se dibuja una curva o se colorean sectores de un mapa se están estimando valores de la variable en lugares que no han sido muestreados. La forma más corriente de realizar esas estimaciones es utilizando interpolación lineal, sin embargo existen otros métodos. En este capítulo se describen metodologías de análisis estadísticos que se utilizan para describir los patrones de variabilidad espacial.

Distribución de puntos

Algunos trabajos geológicos tienen por objetivo conocer la distribución en el espacio que presentan unos **puntos** sobre una superficie o en una línea. Los puntos pueden ser localidades, puntos de control geológico, trazas fósiles sobre una superficie, especímenes de plantas que crecen en suelos derivados de ciertas rocas (i.e. *Acacia burkittii*, *Acacia resinomarginea*) que se utilizan para prospección geobotánica o bien la presencia de algún atributo en una transecta o en un perfil. La distribución

espacial de los puntos puede ser totalmente al azar (Fig. 1a y 1b), seguir un patrón regular (Fig. 1c y 1d) o un patrón agrupado o contagioso (Fig. 1e y 1f).Cualquiera sea el caso si se quiere investigar como es el patrón de distribución de los puntos se puede recurrir a una prueba de hipótesis.

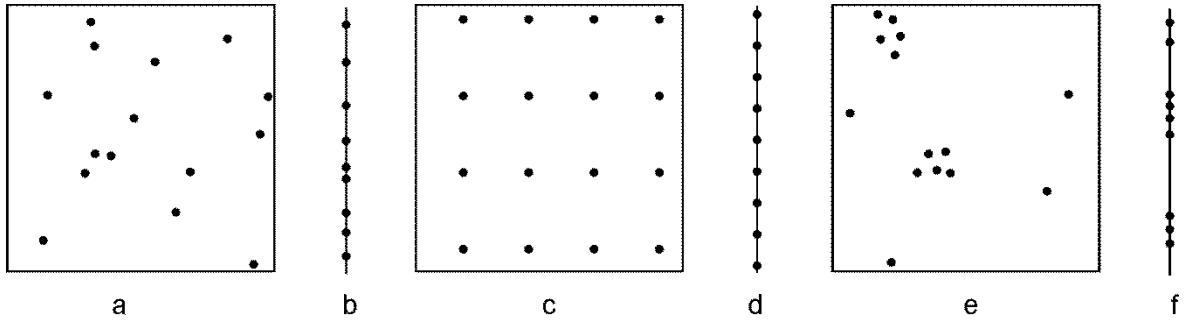


Figura 1. a y b. Distribución de puntos es al azar. c y d. Distribución regular. e y f. Distribución agrupada o contagiosa.

En el Capítulo 6 se describió la prueba de bondad de ajuste χ^2 cuyo objetivo es inferir si la población muestreada se ajusta a una cierta distribución teórica. Por otra parte, en el Capítulo 4 se mencionó que las variables Poisson describen la ocurrencia de sucesos independientes que ocurre en el espacio o en el tiempo. La variable aleatoria que se analiza en este apartado es *X*, el número de ocurrencias en una unidad especificada de superficie, o lineal. Recuerde que la función de probabilidad Poisson (expresión 4.3) permite calcular el número de ocurrencias de ese fenómeno aleatorio conociendo el parámetro λ que es el promedio de ocurrencia en la unidad especificada.

El procedimiento requiere dividir el área de trabajo en **unidades del mismo tamaño** de superficie o lineales y contar el número de puntos en cada una (Fig. 2). Si la distribución de los puntos es al azar, la frecuencia de puntos esperada en cada unidad tienen media y varianza iguales, la relación varianza media es igual a uno $\left(\frac{S^2}{\bar{X}} = 1\right)$, pues es una de las características del modelo Poisson (Capitulo 4) (Fig. 2a). En cambio, si los puntos están más uniformemente distribuidos que lo que determina el azar, la variación de la frecuencia de los puntos por unidad es pequeña porque todas las unidades tienen el mismo número de putos, entonces la relación varianza media es menor a uno $\left(\frac{S^2}{\bar{X}} < 1\right)$ (Fig. 2b). Por último, si los puntos están menos regularmente distribuidos que lo que determina el azar, la variación de la frecuencia de los puntos por unidad es grande porque hay unidades que contienen muchos puntos y otras con muy pocos o ninguno, de donde la relación varianza media es mayor que uno $\left(\frac{S^2}{\bar{X}} > 1\right)$ (Fig. 2c).

Entonces si se calculan las frecuencias esperadas para una distribución Poisson se puede utilizar la prueba de χ^2 para realizar la prueba de hipótesis que permita responder si la variable se distribuye al azar o no. La hipótesis nula de la prueba es que las frecuencias observadas y esperadas son iguales (H_0: *fo* = *fe*) y la hipótesis alternativa que las frecuencias observadas y esperadas son diferentes (H_1: *fo* ≠ *fe*). El estadístico de prueba es

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}, \qquad (14.1)$$

dónde *fo*: frecuencia observada, *fe*: frecuencia teórica o esperada y *k*: número de categorías.

La hipótesis nula se rechaza cuando $\chi_c^2 \geq \chi_{\alpha;\nu}^2$, de modo que se interpreta que la distribución de la variable en el espacio no es al azar. Los grados de libertad ,ν, se calculan como *k* (número de clases) menos el número de parámetros utilizados para estimar las frecuencias esperadas menos 1 *(k – nº de parámetros estimados - 1)*. Si el parámetro λ se desconoce se estima con el promedio de la muestra.

Cuando el patrón espacial no es al azar (se rechazó la hipótesis nula de la prueba de χ^2) conviene indagar si el patrón es agrupado o más o menos uniformemente distribuidos. Se realiza entonces una prueba de hipótesis que tiene en cuenta si la relación entre la varianza y la media es igual o diferente a uno, $H_0: \frac{S^2}{\bar{X}} = 1$ y $H_A: \frac{S^2}{\bar{X}} \neq 1$. El estadístico de prueba es una adaptación del la prueba de *t* para una muestra que tiene la expresión

$$t_{n-1} = \frac{\frac{S^2}{\bar{X}} - 1}{\sqrt{\frac{2}{n-1}}}. \qquad (14.2)$$

El análisis de la expresión 14.2 permite ver que cuando la relación entre la varianza y la media es 1, el numerador es 0 y también es 0 el valor de *t*. Un valor diferente a 0 indica que los datos están agrupados o están uniformemente distribuidos en el espacio. Si la varianza sobre la media es menor a uno, significa que los datos son uniformes. Cuando la varianza sobre la media es mayor que uno, los datos están agrupados. El valor crítico de *t* se busca en la Tabla 3 del Anexo para ν = *n* – 1 grados de libertad y α / 2 ($t_{n-1;\ \alpha/2}$).

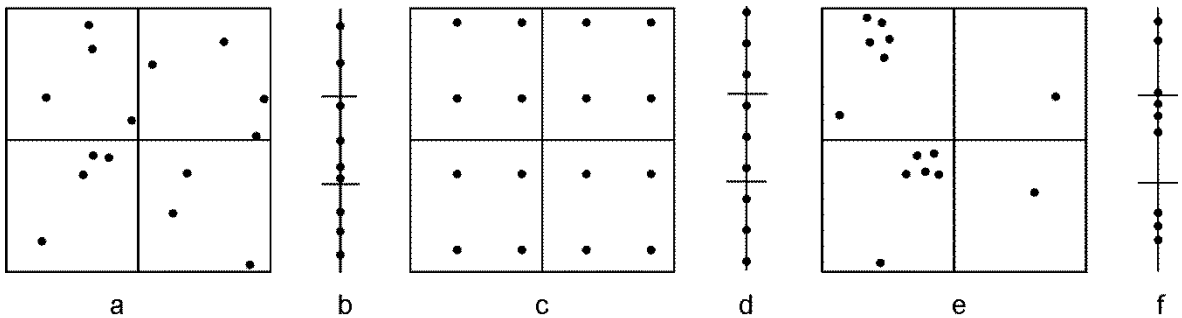


Figura 2. a y b. Distribución de puntos es al azar. c y d. Distribución regular. e y f. Distribución agrupada o contagiosa. Para estudiar que distribución espacial presentan los datos se divide el área en sectores todos del mimo tamaño.

EJEMPLO 1

Análisis espacial del patrón de distribución de puntos

Se realizan estudios paleoambientales en sedimentos marinos cretácicos de la cuenca Neuquina. Se desea averiguar si la distribución de los individuos de la especie de ostrea *Aetostreon latissimum* que aparecen en el techo del estrato transgresivo son el resultado del transporte por el oleaje y corrientes en un ambiente submareal somero o, por el contrario, refleja la distribución de los individuos en la comunidad en el momento de su muerte. Para responder sus interrogantes ubica sobre el techo del estrato una cuadricula de 100 celdas de 50 cm de lado.

	X = N° de individuos	*fo = N° de cuadriculas*	$P(x)=\frac{e^{-\lambda}\lambda^x}{x!}$	*fe = n P(x)*	χ^2
	0	44	0,3166	31,6637	4,8063
	1	24	0,3641	36,4132	4,2317
	2	17	0,2094	20,9376	0,7405
	3	7	0,0803	8,0261	1,4694
	4	5	0,0231	2,3075*	
	5	2	0,0053	0,5307*	
	6	1	0,0010	0,1017*	
	>7	0	0,0002	0,0167*	
Suma		100	1,0000	99,9973	11,2479

n = 100

m = número de ejemplares = 115

λ se estima con la media

$\bar{X}=\frac{m}{n}=\frac{115}{200}=1{,}15$; $S^2 = 1{,}89$

Las probabilidades se calculan con la expresión 4.3 $P(x)=\frac{e^{-\lambda}\lambda^x}{x!}$

H_0: *fo* = *fe*. Los individuos de la ostrea se disponen al azar.

H_A: *fo* ≠ *fe*. Los individuos de la ostrea no se disponen al azar.

α = 0,05

Como las frecuencias esperadas para x = 4, 5 , 6 y > 7 (señalados con *) son menores a 5 se suman a la frecuencia de x = 3. Las categorías k se reducen de 8 a 4.

ν = 4 - 1 - 1 = 2

De la Tabla 3 del Anexo, $\chi^2_{2;0,05}=5{,}99$

$$\chi_c^2=\sum_{i=1}^{k}\frac{(fo-fe)^2}{fe}$$

$\chi_c^2 = 11{,}25$

La hipótesis nula se rechaza, 11,25 > 5,99 ($\chi_c^2 > \chi^2_{2;0,05}$). Se interpreta que los individuos no fueron transportados por olas ni corrientes. ¿Cómo se disponían los individuos en la comunidad?

$H_0: \frac{S^2}{\bar{X}}=1$

$H_A: \frac{S^2}{\bar{X}}\neq 1.$

α = 0,05

ν = n – 1 = 99

De la Tabla 2 del Anexo, $t_{99,\ 0,05}$ = 1,984

$$t_{n-1}=\frac{\frac{S^2}{\bar{X}}-1}{\sqrt{\frac{2}{n-1}}}$$

$$t_{100-1}=\frac{\frac{1,89}{1,15}-1}{\sqrt{\frac{2}{100-1}}}=4{,}52$$

La hipótesis nula se rechaza, 4,52 > 1,984 ($t_c >\ t_{99,\ 0,05}$). Cuando se mira la relación entre la varianza y la media y se analiza el signo de t calculado, se puede afirmar que los individuos se disponen sobre el sustrato agrupados.

Predicción e interpolación de datos en 2D. Geoestadística

La **Geoestadística** es una rama de la estadística que estudia el comportamiento de las variables en el espacio conocidas como **variables regionalizadas**. Muchas ramas de la geociencias emplean métodos geoestadísticos para resolver sus problemas, la geología de minas, la del petróleo, hidrogeología, oceanografía, sensores remotos, edafología y meteorología entre otras. El objetivo de la Geoestadística es la caracterización del fenómeno natural para luego realizar estimaciones (inferencias) y obtener medidas de incertidumbre sobre las estimaciones realizadas.

Los orígenes de la Geoestadística se encuentran en los trabajos de Krige[23] (1951) quien estudió los problemas de estimación de las leyes de oro en las minas sudafricanas y propuso una metodología básica para su estudio. Matheron[24] y su grupo en la Escuela de Minas de París fueron quienes, diez años después, formulan en forma rigurosa y dan solución a los problemas de estimación. Actualmente los grupos de investigación más importantes son el francés y el de la Universidad de Stanford (California, USA) liderado por Andre Journel.

Una variable aleatoria regionalizada se define en todos los puntos del espacio, entonces cada realización de la función aleatoria es una función espacial. Cada realización tiene dos componentes, una estructurada y otra aleatoria. La componente estructurada es la que permite decir que dos mediciones cercanas se parecen, es probable que se encuentren valores altos próximos a otros valores altos. La componente aleatoria es la que impide predecir exactamente el valor de esas mediciones.

La estimación espacial que se realiza utilizando geoestadística utiliza modelos probabilísticos que consideran la incertidumbre de lo sucedido en la generación del fenómeno puesto que, como se indicó, los datos son considerados realizaciones de una función aleatoria espacial.

Es necesario mencionar que en este apartado se presentan los conceptos principales para efectuar los análisis geoestadísticos y que los mismos involucran gran cantidad de datos y la resolución de ecuaciones complejas razón por la cual se utilizan software. Existen muchos software para realizar los análisis algunos de acceso libre, *GEO-EAS* (1988), VARIOWIN 2.1 (1994), GSLIB (1997) y geoR entre otros.

La correlación espacial

Una metodología análoga a la que se utiliza para describir la correlación entre dos variable, calculando el coeficiente de correlación y a través de la construcción de un diagrama de dispersión (Capítulo 8) se puede usar para comenzar a describir la correlación espacial.

Antes de avanzar conviene aclarar algunos conceptos y símbolos que se utilizarán en este apartado. Los puntos en el espacio se identifican con las coordenadas, un punto $\boldsymbol{i}$ tiene coordenadas $\boldsymbol{x_i}$ e $\boldsymbol{y_i}$. La separación entre el punto $\boldsymbol{i}$ y el punto $\boldsymbol{j}$ es $(\boldsymbol{x_j - x_i}, \boldsymbol{y_j - x_i})$. Esta separación es un vector h que tiene módulo, dirección (θ) y sentido; el sentido se describe desde el origen hasta el extremo y puede ser de

i hacia ***j*** (h_{ij}) o desde ***j*** a ***i*** (h_{ji}). Por último las variables se denominarán *V* y *W* en lugar de *X* e *Y* para diferenciarlas de las coordenadas en el plano. Los valores que tomen la variable en un punto cualquiera del espacio se denota con letras minúsculas v_i o w_i.

Para comenzar a analizar la correlación espacial de una variable *V* se puede construir **diagrama de dispersión *h***. El diagrama muestra las relaciones entre los valores en un punto y en otro que se encuentre alejado un vector *h*. La figura 3 muestra dos modelos de adquisición de datos para construir el diagrama de dispersión.

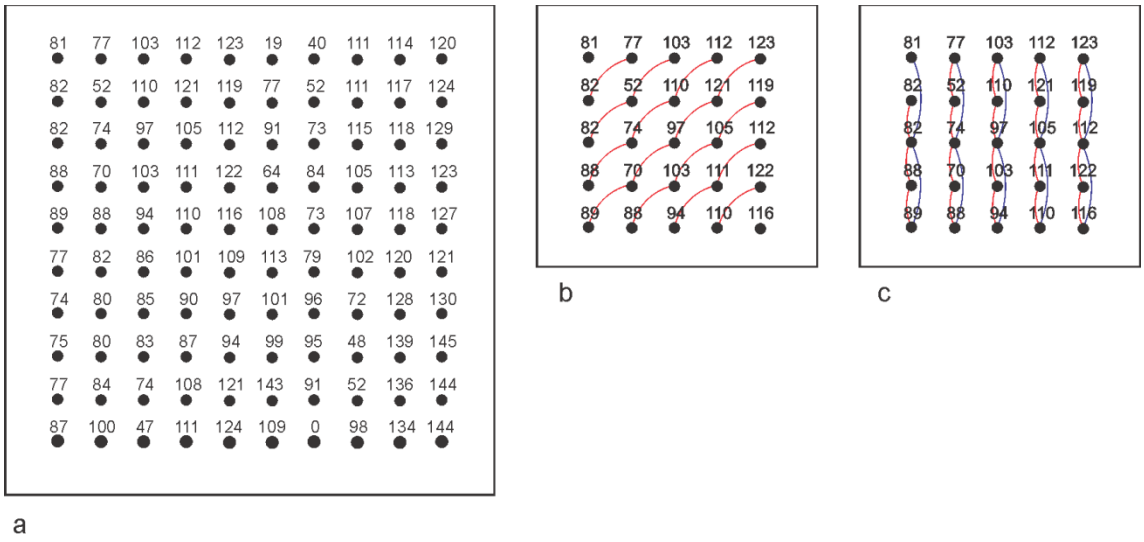


Figura 3. a) Datos de la variable V ubicados en los nodos de una cuadricula distanciada 1 metro (Tomado de Issaks y Srivastava 1989). Modelos de adquisición de datos (se muestra sólo para los 25 puntos ubicados en el noroeste). b) Pares de datos distanciados h = $\sqrt{2}$ en dirección Noreste. c) Pares de datos distanciados h = 1 y h =2 en dirección Norte. Los arcos definen los puntos de datos separados por cada vector.

En la figura 4 se muestran los diagramas de dispersión con los datos obtenidos de la figura 3, en abscisas se ubican los valores correspondientes al punto del origen del vector *h* (v_i) y en ordenadas los valores del extremo (v_j), cada punto en el diagrama de dispersión es entonces (v_i, v_j). El diagrama de dispersión generalmente cambia de forma par diferentes valores de *h* dado que depende tanto del módulo (largo) como de la orientación del vector utilizada para definir los pares. Cuando los valores de los puntos distantes *h* son semejantes los puntos en el diagrama de dispersión se encuentran alineados en una línea que pasa por el origen del diagrama e inclina 45°, en estos casos la correlación es alta y la variabilidad es baja (Fig. 4a). Por el contrario, cuando los valores de los puntos separados en *h* son muy diferentes forman una nube más o menos dispersa lo que indica baja correlación espacial y variabilidad alta (Fig. 4b). Se espera que la correlación disminuya al aumentar la distancia entre los puntos.

Por otra parte, el diagrama de dispersión *h* permite identificar datos anómalos que distorsionan la correlación espacial y se puede evaluar si se deben o no retener en el cálculo de medidas que describan la nube de puntos. Además sirve para determinar si existe más de una población en el

conjunto de datos puesto que el diagrama tendrá grupos de puntos correspondientes a los pares de datos dentro de cada población.

Correlograma y Variograma

Se ha visto que para explorar la correlación entre dos variables se utiliza sólo un diagrama de dispersión, pero no ocurre lo mismo con la correlación espacial puesto que se deben analizar una

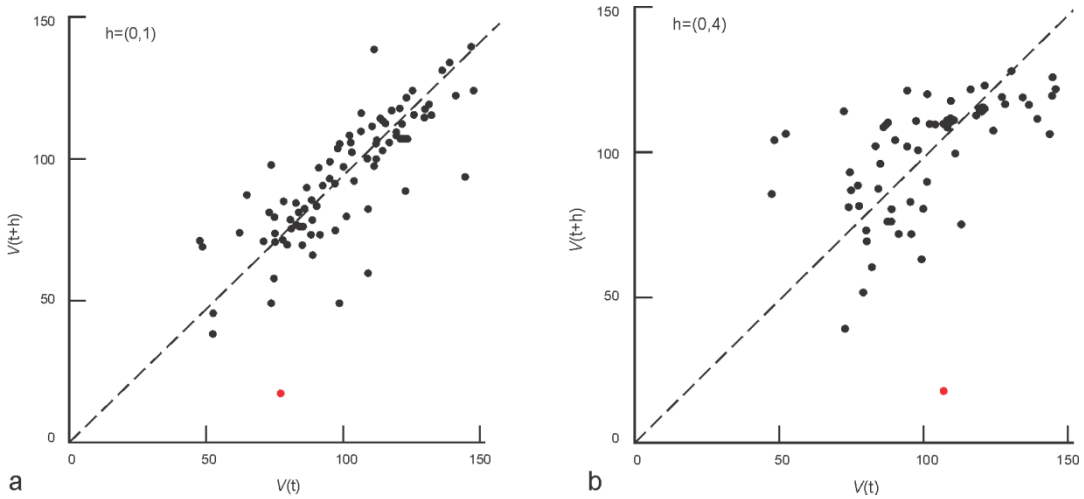


Figura 4.Diagramas de dispersión h de la variable V calculados con el software GEOEAS. a) h = 1 en dirección norte. b) h = 4 en dirección norte. Dato anómalo en rojo.

familia de diagramas de dispersión *h* para varias direcciones y distancia. Por otra parte si bien el coeficiente de correlación *r* sintetiza y cuantifica la información del diagrama de dispersión, analizar los coeficientes en forma aislada es complicado de ahí que se utiliza un correlograma. El **correlograma** es un gráfico de dispersión que muestra la relación entre *h* en las abscisas y el coeficiente de correlación de un gráfico de dispersión *h* en las ordenadas (Fig. 5a). Recuerde que el correlograma se utiliza también para detectar patrones temporales como se mencionó en el capítulo 11.

Otro índice para medir en este caso la discontinuidad espacial es la **semivarianza**,

$$\gamma = \frac{(v_i - v_j)^2}{2}, \qquad (14.3)$$

donde v_i y v_j son los valores de la variable *V* que se encuentran a una distancia h[25]. Para dos valores idénticos de la variable, la semivarianza es cero y a medida que la diferencia entre los valores aumenta también la semivarianza aumenta. La semivarianza de todos los puntos que están separadas por el vector *h* es entonces

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N} (v_i - v_j)^2, \qquad (14.4)$$

donde *N* es el número de pares de datos que se encuentran a la distancia *h* y v_i y v_j son los valores de *V* de el origen y extremo del vector *h*. El **variograma** es el gráfico de dispersión que muestra la relación entre *h* en las abscisas y la semivarianza de todos los puntos distantes *h* en las ordenadas (Fig. 5b). El nombre específico es en realidad función semi-variograma debido a que es la mitad del coeficiente que aparece en la expresión utilizada para definirlo. Con el tiempo el término semi-variograma pasó a variograma.

Tanto el coeficiente de correlación espacial como la semivarianza son muy sensibles a los valores extremos es por ello que se deben remover cuando se identificaron en el análisis de los diagramas de dispersión *h*. Además si existen diferentes poblaciones dentro del área se deben separar en conjuntos homogéneos antes de construir el variograma y el correlograma.

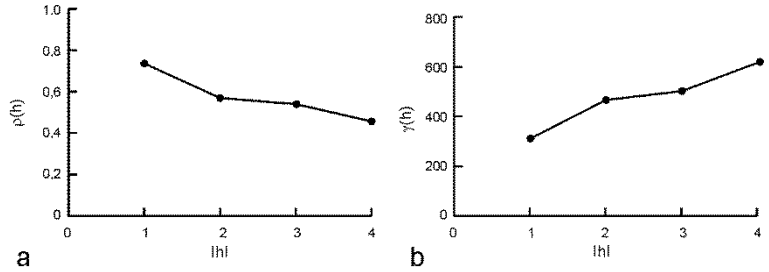


Figura 5. a) Correlograma y b) Variograma de la variable V calculados con el software GEOEAS

Para calcular estadísticos de correlación espacial significativos sobre pares de datos, los estadísticos de la variable (media y varianza) no deben cambiar dentro del área estudiada. Esta hipótesis acerca de la variable se define como **estacionaridad**. La hipótesis de estacionaridad no se prueba, se asume que se cumple, sin embargo es un requisito crítico, de ahí que antes de empezar el análisis espacial se debe explorar y calcular los estadísticos para el conjunto de datos y definir cuantas poblaciones están presentes en el área.

Variograma experimental omnidireccional y variogramas direccionales

Si el muestreo se realizó empleando una cuadricula más o menos regular, se utiliza el tamaño de la cuadricula para determinar la distancia entre los gráficos de dispersión *h*. Esta distancia se conoce como **paso** (*lag* en inglés). En cambio, si el muestreo fue aleatorio el espaciamiento se estima con el promedio de la distancia de muestreo entre dos puntos vecinos. Cuando a densidad de muestreo es más pequeña en una dirección que en otra, como ocurre con los datos obtenidos de perforaciones (la distancia vertical es mucho menor que la distancia entre pozos), los variogramas se calculan considerando las distancias de muestreo en ambas direcciones.

Los datos usados en el ejemplo 2 se tomaron en forma sistemática en los nodos de una cuadricula pero en la práctica no siempre es posible muestrear en los lugares planificados o, puede suceder, que se

cuente con información relevada con más intensidad en algún sector, de ahí que se permite cierta libertad. Debido a que cada diagrama de dispersión *h* depende de la dirección y distancia es necesario entonces tener **tolerancia** tanto en la distancia como la orientación del vector que define los pares. Esto significa que el variograma experimental se define en la distancia *h* mas/menos alguna distancia ($h \pm \Delta h$) y ángulo θ mas/menos algún ángulo ($\theta \pm \Delta\theta$). Generalmente se utiliza una tolerancia en la distancia igual a la mitad del desplazamiento (módulo del vector *h*). Cuando los variogramas son exploratorios la tolerancia angular suele ser chica (Fig. 6).

El análisis de la continuidad espacial se inicia con el cálculo de un variograma donde solamente se especifica una tolerancia de distancia, de aquí que todos los pares de datos separados por esa distancia ingresan al cálculo del variograma, no se considera ninguna orientación (la tolerancia angular es

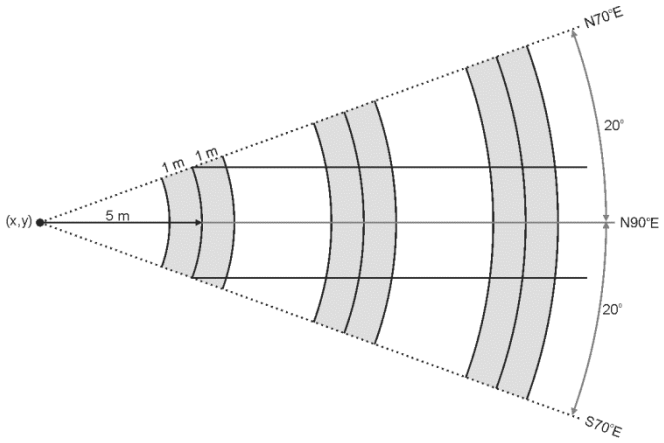


Figura 6. Tolerancia. Como al aumentar h aumenta el arco de búsqueda incluyendo puntos muy fuera de la dirección deseada se suele acotar la búsqueda definiendo una banda.

360°). Este variograma se denomina **omnidireccional** y se puede considerar como un variograma promedio del área de estudio (Fig. 7a). El variograma omnidireccional describe la estructura general de la correlación espacial y se utiliza para establecer los parámetros del variograma que se describen más adelante.

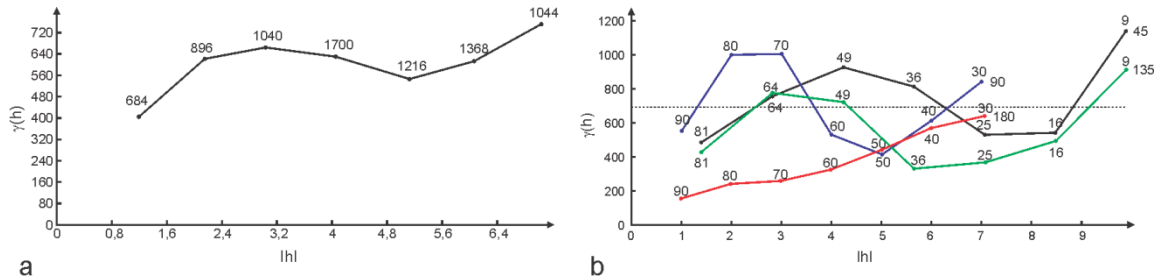


Figura 7. a) Variograma omnidireccional de la variable V (datos de la figura 3). b) Variogramas direccionales de la variable V (datos de la figura 3) obtenidos con el software VARIOWIN.

El siguiente paso tiene como objetivo detectar el patrón de variación espacial vinculado con algunas direcciones particulares, estas direcciones se conocen como **ejes de anisotropía**. Para ello es importante contar con información geológica y conocer la variable. Prácticamente no existen fenómenos donde la variabilidad espacial sea igual en todas direcciones, en general existen direcciones que presenten mayor continuidad que otras. Por ejemplo durante la formación de los depósitos minerales pueden existir controles estructurales o genéticos que determinen direcciones de máxima y mínima continuidad como ocurre con los depósitos de origen sedimentario donde la máxima continuidad es paralela al plano de estratificación. Si se trata de problemas hidrogeológicos asociados con la transmisividad (importante en el transporte de solutos) la variabilidad cambia en la dirección del flujo. Otros fenómenos que presentan variabilidad espacial son los vinculados con la dirección del viento como la contaminación aérea y los productos de emanaciones volcánicas que se espera que exista mayor continuidad en la dirección de los vientos dominantes.

Un panorama del patrón de variación espacial que permita detectar los ejes de anisotropía se obtiene con el **mapa de variograma** (Fig. 8). El mapa de variograma se dibujan con los valores de varios **variogramas direccionales** (Fig. 7b).

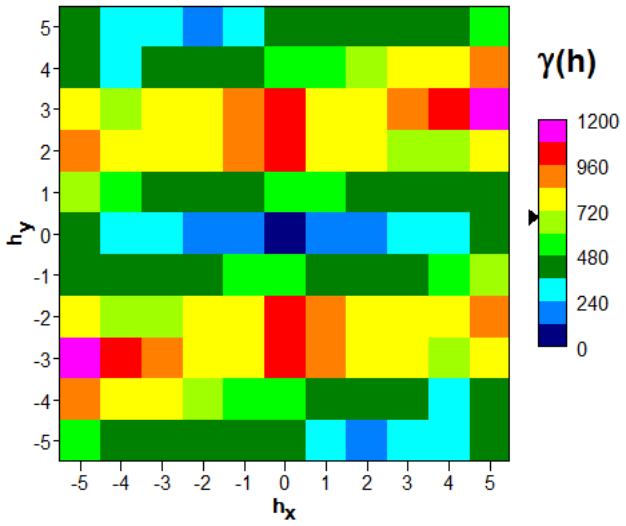


Figura 8. Mapa de variograma de los datos de la variable V (figura 3) obtenido con el software VARIOWIN.

Dependiendo del conocimiento geológico del fenómeno que se estudia y de a densidad de los datos se define el número, dirección y tolerancia espacial y angular de los variogramas direccionales. Si el muestreo no es regular se sugiere calcular variogramas en dirección N, N30E, N60E, E, S60E, S30E con tolerancia angular de 15°. De este modo se tendrá un panorama de la continuidad espacial de toda el área debido a que el variograma es igual en las distancias $+h$ y $-h$ (el variograma en la dirección N es igual al variograma en la dirección S, por ejemplo).

Partes del variograma

Si bien el valor del variograma a una distancia de separación cero, es cero (cada muestra es igual a sí misma), en el variograma experimental suele existir una discontinuidad en el origen que es llamada **efecto pepita**[26] (*nugget effect* en inglés) (Fig. 9). El efecto pepita representa la componente de variabilidad aleatoria que no está correlacionado espacialmente. Se origina por la suma de varios componentes: errores de laboratorio, errores de muestreo, errores de preparación de la muestra y las fluctuaciones aleatorias en la variable.

A medida que aumenta la distancia entre los puntos el valor del correspondiente variograma aumenta pero eventualmente, en una cierta distancia, se estabiliza y alcanza una **meseta** (*sill* en inglés). La meseta del variograma representa la variabilidad total, se relaciona con la varianza de los datos y es la parte del modelo que se correlaciona espacialmente. La meseta está integrada por una porción correlacionada (C) mas una componente aleatoria (CO) que corresponde al efecto pepita (Fig. 9).

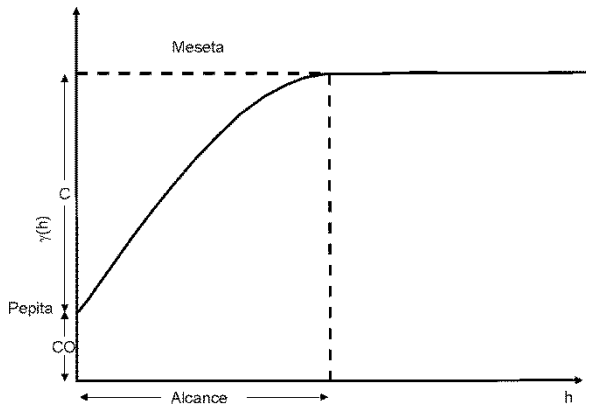


Figura 9. Partes del variograma.

El **alcance** del variograma (abreviatura del alcance de correlación, *range* en inglés) marca el límite de la distancia en que los datos tienen correlación espacial. Dos puntos separados por una distancia menor que el alcance están correlacionados, mientras que dos puntos separados por una distancia mayor no están correlacionados (son independientes). El alcance también se puede definir como la distancia a la que el variograma alcanza su valor de umbral (Fig. 9).

Modelado del variograma

Para estimar los valores que se encuentran entre dos puntos se suele usar interpolación lineal pues se asume que la pendiente entre los puntos es recta y cambia de manera uniforme entre la distancia que separa dos puntos vecinos. Sin embargo, el patrón de variación puede no ser lineal y es el variograma el que captura como cambia la variable en función de la distancia. Ahora bien, el variograma

experimental $\hat{\gamma}(h_k)$ se calcula para un número finito de vectores h (distancias y direcciones) y se necesita hallar la curva continua que mejor ajuste.

Por otra parte, como se ha visto en el Capítulo 5, para hacer inferencias y tener medidas de la incertidumbre asociada a ella es necesario contar con modelos. Los modelos teóricos básicos que se utilizan para ajustar el variograma experimental deben reunir algunas condiciones (definido positivo) para que sean útiles para estimar y predecir incertidumbre. Se utilizan seis modelos básicos: efecto pepita puro, esférico, exponencial, gaussiano, monomico y efecto agujero (Fig. 10).

El comportamiento del variograma en las cercanías del origen puede ser de tres tipos. En el modelo gaussiano es parabólico, lo que indica un fenómeno muy regular y que se encuentra muy correlacionado a distancias cortas como sucede con las cotas de un paisaje ligeramente ondulado o la potencia de una capa o de una veta. En los **modelos exponencial** y esférico el comportamiento es lineal, la pendiente de la recta indica la velocidad de cambio, a mayor pendiente se espera que los valores de puntos cercanos cambie más rápido que cuando la pendiente es baja. Por último el modelo **efecto pepita** es discontinuo en el origen que, como se indicó muestra la falta de continuidad espacial.

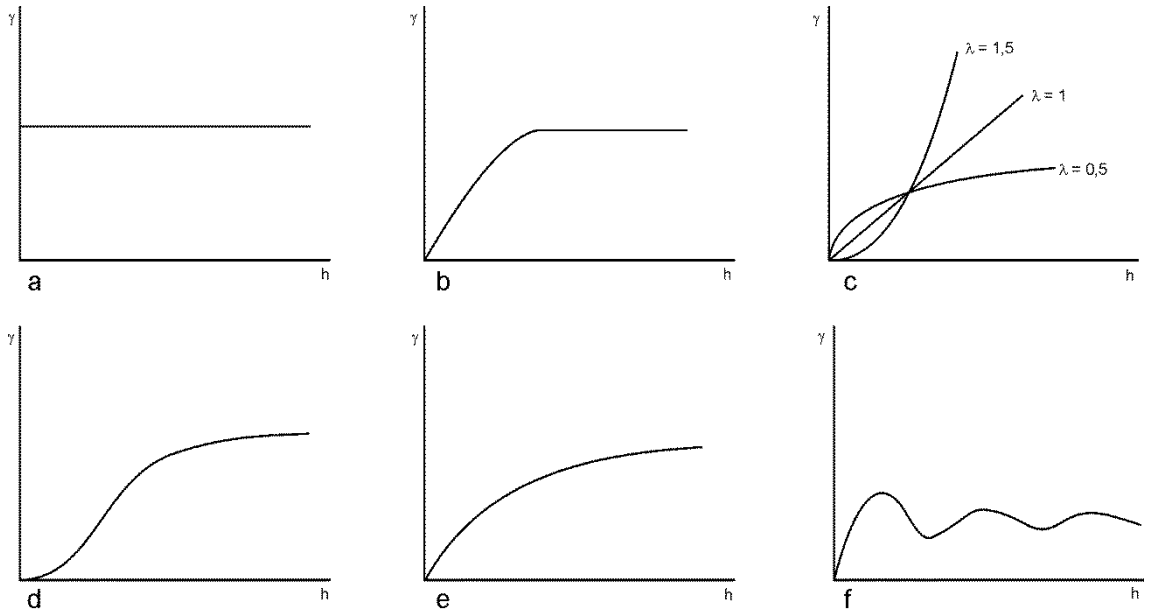


Figura 10. Modelos de variogramas. a) Efecto pepita. b) Esférico. c) Exponencial. d) Gaussiano. e) Monomico. f) Efecto agujero.

El **modelo efecto agujero** indica fenómenos con componentes periódicas o cuasi-periódicas. Se puede formar cuando se muestrean unidades plegadas o en yacimientos estratificados o sistemas de vetas paralelas cuando el variograma se construye en forma perpendicular al plano de las unidades estratificadas o de las vetas.

Para lograr un buen ajuste del variograma experimental se utiliza una combinación de modelos y se dice que los modelos están anidados. En la práctica, se suele usar el efecto pepita para modelar la

componente aleatoria y una forma lineal cerca del origen. El efecto pepita se estima prolongando una línea recta a través de los primeros puntos del variograma experimental hasta el origen.

Cuando la continuidad espacial cambia con la dirección se modela un variograma que combina los modelos en los ejes de máxima y mínima anisotropía. La anisotropía puede ser zonal o geométrica. La **anisotropía zonal** se presenta cuando la meseta de los variogramas depende de la dirección pero tienen el mismo alcance y en la **anisotropía geométrica** la meseta es la misma pero el alcance es diferente (Fig. 11).

Si bien existen varios métodos para ajustar el variograma experimental al modelo, mínimos cuadrados, jackknife, máxima verosimilitud, validación cruzada, el ajuste a sentimiento suele dar buenos resultados.

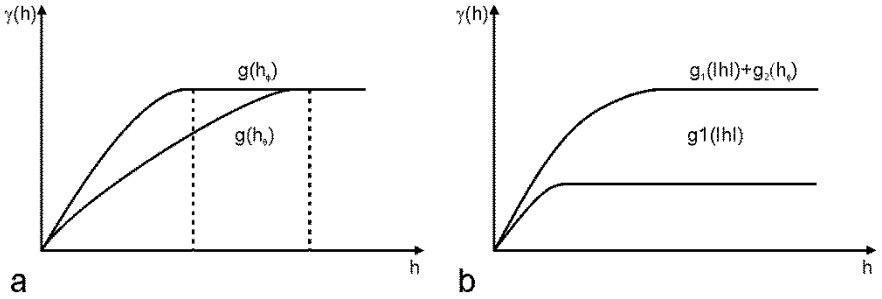


Figura 11. a) Anisotropía geométrica. b) Anisotropía zonal.

Krigeado y simulación

Una de las principales aplicaciones de la Geoestadística es la estimación de la variable en ciertos puntos o la estimación de los valores medios en una serie de zonas o bloques. Se ha desarrollado un estimador que minimiza la varianza de estimación que se denomina **krigeado**[27] (*krging* en inglés). El krigeado es un método de estimación lineal que calcula los estimados minimizando la varianza de los errores de estimación. Se suele usar la sigla BLUE que provienen de *Best Linear Unbiased Estimator* para sintetizar sus propiedaes. Además el krigeado proporciona un error de estimación conocido como **varianza de krigeado** que no dependen de los valores medidos de las variables sino de su posición y de los parámetros del modelado del variograma experimental, es decir que considera la variabilidad espacial de la variable que se estudia.

Los dos tipos principales de krigeado son el puntual y el de bloques. Los resultados del krigeado se vuelcan en dos mapas, uno con los valores estimados y el otro con las varianzas de la estimación (Fig. 12). El mapa de las varianzas de estimación se suele utilizar para planificar la ubicación de un futuro muestreo.

Con el krigeado se obtiene una imagen suavizada del fenómeno que se estudia que no considera las fluctuaciones. Pero si se enfoca el problema desde la perspectiva que la variable es una función

aleatoria se pueden tener innumerables realizaciones del fenómeno con las mismas características de continuidad espacial que los datos originales. Para estas realizaciones, llamadas **simulaciones**, se utiliza el variograma experimental ajustado con el modelo teórico. Los métodos de simulación más usados se reúnen bajo el nombre de **simulación condicional** (simulación secuencial gaussiana, bandas rotantes, booleana, etc.). En la simulación condicional todas las realizaciones proveen valores estimados y reproducen exactamente los datos (honran los datos) (Fig. 13).

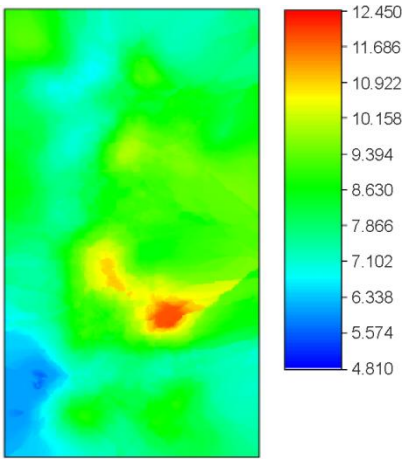


Figura 12. Mapa de porosidad efectuado con krigeado (tomado de Castaño Agudelo y Vergara Elorza 2004).

El krigeado y las simulaciones tienen diferentes objetivos, el del krigeado es la estimación, el de la simulación obtener escenarios diferentes de la variabilidad espacial. No hay un método mejor que otro, los dos se complementan.

Las ecuaciones de kriging y simulación están fuera del alcance de este libro. En Cressie (1980), Issaks y Srivastava (1989) y Goovaerts (1997) son algunos libros de referencia en el tema. Por otra parte, si bien se han mencionado sólo algunos software para realizar los análisis geoestadísticos, existen otros muchos de acceso libre

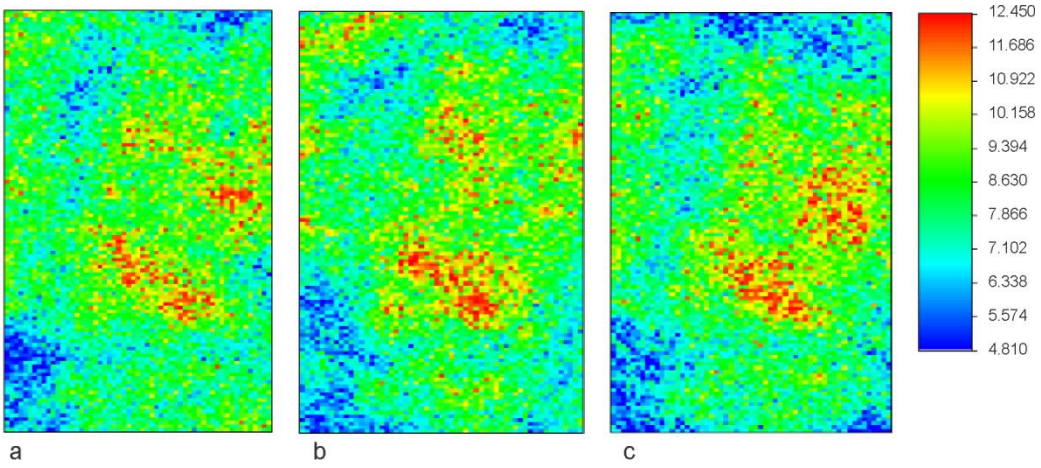


Figura 13. Simulaciones de porosidad (tomado de Castaño Agudelo y Vergara Elorza 2004).

Tabla 1. Valores de "*Z*" la distribución Normal estándar N(0,1) (en el cuerpo de la tabla se encuentra la probabilidad de hallar un valor de *Z* menor o igual a Z_0 (ej. $P(Z \leq 1{,}42) = 0{,}9222$).

Z	***0,09***	***0,08***	***0,07***	***0,06***	***0,05***	***0,04***	***0,03***	***0,02***	***0,01***	***0,00***
-3,5	0,00023	0,00017	0,00017	0,00018	0,00019	0,00019	0,00020	0,00021	0,00022	0,00022
-3,4	0,0003	0,0002	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003
-3,3	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0005	0,0005
-3,2	0,0007	0,0005	0,0005	0,0005	0,0006	0,0006	0,0006	0,0006	0,0006	0,0007
-3,1	0,0010	0,0007	0,0007	0,0008	0,0008	0,0008	0,0008	0,0009	0,0009	0,0009
-3,0	0,0010	0,0010	0,0011	0,0011	0,0011	0,0012	0,0012	0,0013	0,0013	0,0013
-2,9	0,0014	0,0014	0,0015	0,0015	0,0016	0,0016	0,0017	0,0018	0,0018	0,0019
-2,8	0,0019	0,0020	0,0021	0,0021	0,0022	0,0023	0,0023	0,0024	0,0025	0,0026
-2,7	0,0026	0,0027	0,0028	0,0029	0,0030	0,0031	0,0032	0,0033	0,0034	0,0035
-2,6	0,0036	0,0037	0,0038	0,0039	0,0040	0,0041	0,0043	0,0044	0,0045	0,0047
-2,5	0,0048	0,0049	0,0051	0,0052	0,0054	0,0055	0,0057	0,0059	0,0060	0,0062
-2,4	0,0064	0,0066	0,0068	0,0069	0,0071	0,0073	0,0075	0,0078	0,0080	0,0082
-2,3	0,0084	0,0087	0,0089	0,0091	0,0094	0,0096	0,0099	0,0102	0,0104	0,0107
-2,2	0,0110	0,0113	0,0116	0,0119	0,0122	0,0125	0,0129	0,0132	0,0136	0,0139
-2,1	0,0143	0,0146	0,0150	0,0154	0,0158	0,0162	0,0166	0,0170	0,0174	0,0179
-2,0	0,0183	0,0188	0,0192	0,0197	0,0202	0,0207	0,0212	0,0217	0,0222	0,0228
-1,9	0,0233	0,0239	0,0244	0,0250	0,0256	0,0262	0,0268	0,0274	0,0281	0,0287
-1,8	0,0294	0,0301	0,0307	0,0314	0,0322	0,0329	0,0336	0,0344	0,0351	0,0359
-1,7	0,0367	0,0375	0,0384	0,0392	0,0401	0,0409	0,0418	0,0427	0,0436	0,0446
-1,6	0,0455	0,0465	0,0475	0,0485	0,0495	0,0505	0,0516	0,0526	0,0537	0,0548
-1,5	0,0559	0,0571	0,0582	0,0594	0,0606	0,0618	0,0630	0,0643	0,0655	0,0668
-1,4	0,0681	0,0694	0,0708	0,0721	0,0735	0,0749	0,0764	0,0778	0,0793	0,0808
-1,3	0,0823	0,0838	0,0853	0,0869	0,0885	0,0901	0,0918	0,0934	0,0951	0,0968
-1,2	0,0985	0,1003	0,1020	0,1038	0,1056	0,1075	0,1093	0,1112	0,1131	0,1151
-1,1	0,1170	0,1190	0,1210	0,1230	0,1251	0,1271	0,1292	0,1314	0,1335	0,1357
-1,0	0,1379	0,1401	0,1423	0,1446	0,1469	0,1492	0,1515	0,1539	0,1562	0,1587
-0,9	0,1611	0,1635	0,1660	0,1685	0,1711	0,1736	0,1762	0,1788	0,1814	0,1841
-0,8	0,1867	0,1894	0,1922	0,1949	0,1977	0,2005	0,2033	0,2061	0,2090	0,2119
-0,7	0,2148	0,2177	0,2206	0,2236	0,2266	0,2296	0,2327	0,2358	0,2389	0,2420
-0,6	0,2451	0,2483	0,2514	0,2546	0,2578	0,2611	0,2643	0,2676	0,2709	0,2743
-0,5	0,2776	0,2810	0,2843	0,2877	0,2912	0,2946	0,2981	0,3015	0,3050	0,3085
-0,4	0,3121	0,3156	0,3192	0,3228	0,3264	0,3300	0,3336	0,3372	0,3409	0,3446
-0,3	0,3483	0,3520	0,3557	0,3594	0,3632	0,3669	0,3707	0,3745	0,3783	0,3821
-0,2	0,3859	0,3897	0,3936	0,3974	0,4013	0,4052	0,4090	0,4129	0,4168	0,4207
-0,1	0,4247	0,4286	0,4325	0,4364	0,4404	0,4443	0,4483	0,4522	0,4562	0,4602
0,0	0,4641	0,4681	0,4721	0,4761	0,4801	0,4840	0,4880	0,4920	0,4960	0,5000

Z	***0***	***0,01***	***0,02***	***0,03***	***0,04***	***0,05***	***0,06***	***0,07***	***0,08***	***0,09***
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Tabla 2. Valores críticos de la distribución χ^2 (el cuerpo de la tabla se encuentran los valores de χ^2 tales que la probabilidad sea mayor o igual a la especificada (ej. para $\nu = 10$, $P[\chi^2 \geq 15{,}99] = 0{,}10$).

Grados de	***Probabilidad***												
libertad ν	***0,995***	***0,99***	***0,975***	***0,95***	***0,90***	***0,75***	***0,5***	***0,25***	***0,10***	***0,05***	***0,025***	***0,01***	***0,005***
1	0,00	0,00	0,00	0,00	0,02	0,10	0,46	1,32	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	2,68	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	3,46	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	4,26	6,35	9,04	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	5,90	8,34	11,39	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	6,74	9,34	12,55	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	7,58	10,34	13,70	17,28	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	8,44	11,34	14,85	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	9,30	12,34	15,98	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	10,17	13,34	17,12	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	11,04	14,34	18,25	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	11,91	15,34	19,37	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	12,79	16,34	20,49	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	13,08	17,34	21,61	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	14,56	18,34	22,72	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	15,45	19,34	23,83	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	16,34	20,34	24,94	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	17,24	21,34	26,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	18,14	22,34	27,14	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	19,04	23,34	28,24	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	19,94	24,34	29,34	34,38	37,65	40,65	44,31	46,93
30	13,79	14,95	16,79	18,49	20,60	24,48	29,34	34,80	40,26	43,77	46,98	50,89	53,67
35	17,19	18,51	20,57	22,47	24,80	29,05	34,34	40,22	46,06	49,80	53,20	57,34	60,27
40	20,71	22,16	24,43	26,51	29,05	33,66	39,34	45,62	51,81	55,76	59,34	63,69	66,77
45	24,31	25,90	28,37	30,61	33,35	38,29	44,34	50,99	57,51	61,66	65,41	69,96	73,17
50	27,99	29,71	32,36	34,76	37,69	43,94	49,34	56,33	63,17	67,50	71,42	76,15	79,49
55	31,73	33,57	36,40	38,96	42,06	47,61	54,34	61,67	68,80	73,31	77,38	82,29	85,75
60	35,53	37,48	40,48	43,19	46,46	52,69	59,34	66,98	74,40	79,08	83,30	88,38	91,95
65	39,38	41,44	44,60	47,45	50,88	56,99	64,34	72,29	79,97	84,82	89,18	94,42	98,10
70	43,28	45,44	48,76	51,74	55,33	61,70	69,33	77,58	85,53	90,53	95,02	100,4	104,2
75	47,21	49,48	52,94	56,05	59,79	66,42	74,33	82,86	91,06	96,22	100,8	106,4	110,3
80	51,17	53,54	57,15	60,39	64,28	71,15	79,33	88,13	96,58	101,9	106,6	112,3	116,3
85	55,17	57,63	61,39	64,75	68,78	75,88	84,33	93,39	102,1	107,5	112,4	118,2	122,3
90	59,20	61,75	65,65	69,13	73,29	80,63	89,33	98,65	107,6	113,2	118,1	124,1	128,3
95	63,25	65,90	69,92	73,52	77,82	85,38	94,33	103,90	113,0	118,8	123,9	130,0	134,3
100	67,33	70,06	74,22	77,93	82,39	90,13	99,33	109,14	118,49	124,34	129,56	135,81	140,17
140	100,65	104,14	109,14	113,68	119,03	128,38	139,33	150,89	161,83	168,61	174,65	181,84	186,85

Tabla 3. Valores críticos de la distribución "*t*" de Student (el cuerpo de la tabla se encuentran los valores de *t* tales que la probabilidad sea mayor o igual a la especificada, ej. para $\nu = 10$, $P[t \geq |1{,}812|] = 0{,}05$)

Grados libertad ν	***Probabilidad***									
	0,40	***0,30***	***0,20***	***0,15***	***0,10***	***0,075***	***0,05***	***0,025***	***0,01***	***0,005***
1	0,325	0,727	1,376	1,963	3,078	4,165	6,314	12,706	31,821	63,656
2	0,289	0,617	1,061	1,386	1,886	2,282	2,920	4,303	6,965	9,925
3	0,277	0,584	0,978	1,250	1,638	1,924	2,353	3,182	4,541	5,841
4	0,271	0,569	0,941	1,190	1,533	1,778	2,132	2,776	3,747	4,604
5	0,267	0,559	0,920	1,156	1,476	1,699	2,015	2,571	3,365	4,032
6	0,265	0,553	0,906	1,134	1,440	1,650	1,943	2,447	3,143	3,707
7	0,263	0,549	0,896	1,119	1,415	1,617	1,895	2,365	2,998	3,499
8	0,262	0,546	0,889	1,108	1,397	1,592	1,860	2,306	2,896	3,355
9	0,261	0,543	0,883	1,100	1,383	1,574	1,833	2,262	2,821	3,250
10	0,260	0,542	0,879	1,093	1,372	1,559	1,812	2,228	2,764	3,169
11	0,260	0,540	0,876	1,088	1,363	1,548	1,796	2,201	2,718	3,106
12	0,259	0,539	0,873	1,083	1,356	1,538	1,782	2,179	2,681	3,055
13	0,259	0,538	0,870	1,079	1,350	1,530	1,771	2,160	2,650	3,012
14	0,258	0,537	0,868	1,076	1,345	1,523	1,761	2,145	2,624	2,977
15	0,258	0,536	0,866	1,074	1,341	1,517	1,753	2,131	2,602	2,947
16	0,258	0,535	0,865	1,071	1,337	1,512	1,746	2,120	2,583	2,921
17	0,257	0,534	0,863	1,069	1,333	1,508	1,740	2,110	2,567	2,898
18	0,257	0,534	0,862	1,067	1,330	1,504	1,734	2,101	2,552	2,878
19	0,257	0,533	0,861	1,066	1,328	1,500	1,729	2,093	2,539	2,861
20	0,257	0,533	0,860	1,064	1,325	1,497	1,725	2,086	2,528	2,845
21	0,257	0,532	0,859	1,063	1,323	1,494	1,721	2,080	2,518	2,831
22	0,256	0,532	0,858	1,061	1,321	1,492	1,717	2,074	2,508	2,819
23	0,256	0,532	0,858	1,060	1,319	1,489	1,714	2,069	2,500	2,807
24	0,256	0,531	0,857	1,059	1,318	1,487	1,711	2,064	2,492	2,797
25	0,256	0,531	0,856	1,058	1,316	1,485	1,708	2,060	2,485	2,787
26	0,256	0,531	0,856	1,058	1,315	1,483	1,706	2,056	2,479	2,779
27	0,256	0,531	0,855	1,057	1,314	1,482	1,703	2,052	2,473	2,771
28	0,256	0,530	0,855	1,056	1,313	1,480	1,701	2,048	2,467	2,763
29	0,256	0,530	0,854	1,055	1,311	1,479	1,699	2,045	2,462	2,756
30	0,256	0,530	0,854	1,055	1,310	1,477	1,697	2,042	2,457	2,750
> 30	0,253	0,524	0,842	1,036	1,282	1,440	1,645	1,960	2,326	2,576

Tabla 4. Distribución F para α **= 0,10**; ($P[F > F_0] = \alpha$)

Grados de libertad ν2	***Grados de libertad del numerador ν1***																				
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***12***	***15***	***20***	***25***	***30***	***40***	***50***	***60***	***80***	***100***	***120***
1	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,8	60,1	62,7	61,2	61,7	62,0	62,2	62,5	52,7	62,8	62,9	63,00	
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,47	9,48	9,48	9,48
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,17	5,17	5,16	5,15	5,15	5,15	5,14	5,14
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,80	3,79	3,78	3,78	3,78
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,15	3,14	3,13	3,13	3,12
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,81	2,80	2,78	2,77	2,76	2,75	2,75	2,74
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,57	2,56	2,54	2,52	2,51	2,50	2,50	2,49
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,35	2,34	2,33	2,32	2,32
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,27	2,25	2,23	2,22	2,21	2,20	2,19	2,18
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,17	2,16	2,13	2,12	2,11	2,09	2,09	2,08
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,03	2,01	1,99	1,97	1,96	1,95	1,94	1,93
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,89	1,87	1,85	1,83	1,82	1,80	1,79	1,79
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,76	1,74	1,71	1,69	1,68	1,66	1,65	1,64
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,68	1,66	1,63	1,61	1,59	1,58	1,56	1,56
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,63	1,61	1,57	1,55	1,54	1,52	1,51	1,50
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,48	1,47	1,45	1,43	1,42
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73	1,68	1,63	1,57	1,53	1,50	1,46	1,44	1,42	1,40	1,39	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,50	1,48	1,44	1,41	1,40	1,37	1,36	1,35
80	2,77	2,37	2,15	2,02	1,92	1,85	1,79	1,75	1,71	1,68	1,63	1,57	1,51	1,47	1,44	1,40	1,38	1,36	1,33	1,32	1,31
100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,69	1,66	1,61	1,56	1,49	1,45	1,42	1,38	1,35	1,34	1,31	1,29	1,28
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,44	1,41	1,37	1,34	1,32	1,29	1,28	1,26

Tabla 4 (continuación). Distribución F para α **= 0,05**; ($P[F > F_0] = \alpha$)

Grados de libertad v2	***Grados de libertad del numerador v1***																				
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***12***	***15***	***20***	***25***	***30***	***40***	***50***	***60***	***80***	***100***	***120***
1	161	199	216	224	230	234	237	239	240	242	244	246	248	249	250	251	252	252	253	253	
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,46	19,46	19,47	19,48	19,48	19,48	19,49	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,63	8,62	8,59	8,58	8,57	8,56	8,55	8,55
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,70	5,69	5,67	5,66	5,66
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,52	4,50	4,46	4,44	4,43	4,41	4,41	4,40
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,83	3,81	3,77	3,75	3,74	3,72	3,71	3,70
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,40	3,38	3,34	3,32	3,30	3,29	3,27	3,27
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,11	3,08	3,04	3,02	3,01	2,99	2,97	2,97
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,89	2,86	2,83	2,80	2,79	2,77	2,76	2,75
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,73	2,70	2,66	2,64	2,62	2,60	2,59	2,58
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,50	2,47	2,43	2,40	2,38	2,36	2,35	2,34
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,28	2,25	2,20	2,18	2,16	2,14	2,12	2,11
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,07	2,04	1,99	1,97	1,95	1,92	1,91	1,90
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,84	1,82	1,80	1,78	1,77
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,88	1,84	1,79	1,76	1,74	1,71	1,70	1,68
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,78	1,74	1,69	1,66	1,64	1,61	1,59	1,58
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,95	1,87	1,78	1,73	1,69	1,63	1,60	1,58	1,54	1,52	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,69	1,65	1,59	1,56	1,53	1,50	1,48	1,47
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,88	1,79	1,70	1,64	1,60	1,54	1,51	1,48	1,45	1,43	1,41
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,85	1,77	1,68	1,62	1,57	1,52	1,48	1,45	1,41	1,39	1,38
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,60	1,55	1,50	1,46	1,43	1,39	1,37	1,35

Tabla 4 (continuación). Distribución F para α **= 0,025**; ($P[F > F_0] = \alpha$)

Grados de libertad ν2	***Grados de libertad del numerador ν1***																				
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***12***	***15***	***20***	***25***	***30***	***40***	***50***	***60***	***80***	***100***	***120***
1	648	799	864	899	922	937	948	957	963	969	977	985	993	998	1001	1005	1008	1009	1011	1013	
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,48	39,49	39,49	39,49
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	14,01	13,99	13,97	13,96	13,95
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,50	8,46	8,41	8,38	8,36	8,33	8,32	8,31
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,27	6,23	6,18	6,14	6,12	6,10	6,08	6,07
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,11	5,07	5,01	4,98	4,96	4,93	4,92	4,90
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,40	4,36	4,31	4,28	4,25	4,23	4,21	4,20
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,94	3,89	3,84	3,81	3,78	3,76	3,74	3,73
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,60	3,56	3,51	3,47	3,45	3,42	3,40	3,39
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,35	3,31	3,26	3,22	3,20	3,17	3,15	3,14
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,01	2,96	2,91	2,87	2,85	2,82	2,80	2,79
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,69	2,64	2,59	2,55	2,52	2,49	2,47	2,46
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,40	2,35	2,29	2,25	2,22	2,19	2,17	2,16
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,23	2,18	2,12	2,08	2,05	2,02	2,00	1,98
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,12	2,07	2,01	1,97	1,94	1,90	1,88	1,87
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	1,99	1,94	1,88	1,83	1,80	1,76	1,74	1,72
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32	2,22	2,11	1,99	1,92	1,87	1,80	1,75	1,72	1,68	1,66	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,87	1,82	1,74	1,70	1,67	1,63	1,60	1,58
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,35	2,28	2,21	2,11	2,00	1,88	1,81	1,75	1,68	1,63	1,60	1,55	1,53	1,51
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	2,08	1,97	1,85	1,77	1,71	1,64	1,59	1,56	1,51	1,48	1,46
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,75	1,69	1,61	1,56	1,53	1,48	1,45	1,43

Tabla 4 (continuación). Distribución F para α **= 0,01**; ($P[F > F_0] = \alpha$)

Grados de libertad ν2	***Grados de libertad del numerador ν1***																				
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***12***	***15***	***20***	***25***	***30***	***40***	***50***	***60***	***80***	***100***	***120***
1	4052	4999	54035	5624	5764	5859	5928	5981	6022	6056	6106	5157	6208	6239	6260	6286	6302	6313	6326	6334	
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,48	99,48	99,49	99,49
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,87	26,69	26,58	26,50	26,41	26,35	26,32	26,27	26,24	26,22
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,91	13,84	13,75	13,69	13,65	13,61	13,58	13,56
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,45	9,38	9,29	9,24	9,20	9,16	9,13	9,11
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,30	7,23	7,14	7,09	7,06	7,01	6,99	6,97
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72º	6,62	6,47	6,31	6,16	6,06	5,99	5,91	5,86	5,82	5,78	5,75	5,74
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,26	5,20	5,12	5,07	5,03	4,99	4,96	4,95
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,71	4,65	4,57	4,52	4,48	4,44	4,41	4,40
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,31	4,25	4,17	4,12	4,08	4,04	4,01	4,00
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,76	3,70	3,62	3,57	3,54	3,49	3,47	3,45
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,28	3,21	3,13	3,08	3,05	3,00	2,98	2,96
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,84	2,78	2,69	2,64	2,61	2,56	2,54	2,52
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,60	2,54	2,45	2,40	2,36	2,32	2,29	2,27
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,45	2,39	2,30	2,25	2,21	2,16	2,13	2,11
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,27	2,20	2,11	2,06	2,02	1,97	1,94	1,92
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,56	2,42	2,27	2,17	2,10	2,01	1,95	1,91	1,86	1,82	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,10	2,03	1,94	1,88	1,84	1,78	1,75	1,73
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55	2,42	2,27	2,12	2,01	1,94	1,85	1,79	1,75	1,69	1,65	1,63
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,37	2,22	2,07	1,97	1,89	1,80	1,74	1,69	1,63	1,60	1,57
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,93	1,86	1,76	1,70	1,66	1,60	1,56	1,53

Tabla 5. Valores críticos "*d*" de la prueba Kolmogorov-Smirnov para datos continuos

N	***α***		
	0,10	***0,05***	***0,01***
1	,950	,975	,995
2	,776	,842	,929
3	,642	,708	,828
4	,564	,624	,733
5	,510	,565	,669
6	,470	,521	,618
7	,438	,486	,577
8	,411	,457	,543
9	,388	,432	,514
10	,368	,410	,490
11	,352	,391	,463
12	,338	,375	,450
13	,325	,361	,133
14	,314	,349	,418
15	,304	,338	,404
16	,295	,328	,329
17	,286	,318	,381
18	,278	,309	,371
19	,272	,301	,363
20	,264	,294	,356
25	,24	,27	,32
30	,22	,24	,29
35	,21	,23	,27
> 35	$\frac{1,22}{\sqrt{N}}$	$\frac{1,36}{\sqrt{N}}$	$\frac{1,63}{\sqrt{N}}$

Tabla 6. Valores críticos "*d*" de la prueba Kolmogorov-Smirnov para datos continuos corregido por Lillifords

N	***α***		
	0,10	***0,05***	***0,01***
4	,352	,381	,417
5	,315	,337	,405
6	,294	,319	,364
7	,276	,300	,348
8	,261	,285	,331
9	,249	,271	,311
10	,239	,258	,294
11	,230	,219	,281
12	,223	,242	,275
13	,214	,234	,268
14	,207	,227	,261
15	,201	,220	,257
16	,195	,213	,250
17	,189	,206	,245
18	,184	,200	,239
19	,179	,195	,235
20	,174	,190	,231
25	,165	,180	,203
30	,144	,161	,187
> 30	$\frac{,805}{\sqrt{N}}$	$\frac{,886}{\sqrt{N}}$	$\frac{1,031}{\sqrt{N}}$

Tabla 7. Valores críticos "*d*" de la prueba Kolmogorov-Smirnov para datos discretos o agrupados

k	*n*	0.1 (alfa 2) 0.05 (alfa 1)	0.05 0.025	0.02 0.01	0.01 0.005
	3	2	3	3	3
	6	3	3	4	4
	9	4	4	4	5
	12	5	4	6	6
	15	4	5	6	6
	18	5	5	6	6
	21	5	6	6	7
	24	5	6	7	1
	21	6	6	7	8
	30	6	7	8	8
	33	6	7	8	8
	35	6	7	8	9
	39	7	7	8	9
	42	7	8	8	9
	45	7	8	9	10
3	***48***	7	8	9	10
	51	7	8	10	10
	54	8	9	10	11
	57	8	9	10	11
	60	8	9	10	11
	63	8	9	11	12
	66	8	9	10	12
	69	8	9	11	12
	72	8	9	11	12
	75	8	10	11	12
	78	9	10	11	12
	81	9	10	11	13
	84	9	10	12	13
	87	9	10	12	13
	90	9	10	12	13
	99	9	10	12	13
	4	3	3	3	3
	8	4	4	4	5
	12	4	5	5	6
	15	5	5	6	6
	20	5	6	6	7
	24	6	6	7	8
	28	6	7	7	8
	32	6	7	8	9
	36	7	7	8	9
	40	7	8	9	9
	44	7	8	9	10
	48	7	8	10	10
4	***52***	8	9	10	11
	56	8	9	10	11
	60	8	9	10	11
	64	8	9	11	12
	68	9	10	11	12
	72	9	10	11	12
	76	9	10	11	12
	80	9	10	11	12
	84	9	10	12	13
	88	9	10	12	13
	92	9	10	12	13
	96	9	10	12	13
	100	9	11	12	13

k	*n*	α (2)0.1 (alfa 2) 0.05 (alfa 1)	0.05 0.025	0.02 0.01	0.01 0.005
	5	3	3	4	4
	10	4	4	5	5
	15	4	5	5	6
	20	5	5	6	7
	25	5	6	6	7
	30	5	6	7	8
	35	6	7	7	8
	40	6	7	8	9
	45	6	7	8	9
5	***50***	7	8	9	10
	55	7	8	9	10
	60	7	8	9	11
	65	7	9	10	11
	70	8	9	10	11
	75	8	9	10	12
	80	8	9	11	12
	85	8	9	11	12
	90	8	9	11	12
	95	8	9	11	12
	100	8	9	11	12
	6	3	4	4	4
	12	4	5	5	6
	18	5	6	6	7
	24	6	6	7	8
	30	6	7	8	9
	36	7	8	9	9
	42	7	8	9	10
6	***48***	8	9	10	11
	54	8	9	10	11
	60	9	10	11	12
	66	9	10	11	12
	72	9	10	12	13
	78	9	11	12	13
	84	9	11	12	13
	90	10	11	13	14
	96	10	11	13	14
	7	3	4	4	4
	14	4	5	5	6
	21	5	6	6	7
	28	5	6	7	8
	35	6	7	8	9
	42	6	7	8	9
7	***49***	7	8	9	10
	56	7	8	9	11
	63	8	9	10	11
	70	8	9	10	12
	77	8	9	11	12
	84	8	10	12	12
	91	8	10	11	13
	98	8	10	11	13

Tabla 7 (continuación). Valores críticos "*d*" de la prueba Kolmogorov-Smirnov para datos discretos o agrupados

		0.1 (alfa 2)	0.05	0.02	0.01
k	***n***	0.05 (alfa 1)	0.025	0.01	0.005
8	***8***	3	4	4	5
8	***16***	4	5	6	6
8	***24***	5	6	7	7
8	***32***	6	7	7	8
8	***40***	6	7	8	9
8	***48***	7	8	9	10
8	***56***	7	9	10	11
8	***64***	8	9	10	11
8	***72***	8	9	11	12
8	***80***	8	10	11	12
8	***88***	8	10	11	13
8	***95***	9	10	11	13
9	***9***	4	4	4	5
9	***18***	5	5	6	7
9	***27***	6	6	7	8
9	***36***	6	7	8	9
9	***45***	7	8	9	10
9	***54***	7	9	10	11
9	***63***	8	9	10	11
9	***72***	8	10	11	12
9	***81***	8	10	11	13
9	***90***	9	10	11	13
9	***99***	9	10	12	13
10	***10***	4	4	5	5
10	***20***	5	6	6	7
10	***30***	6	7	7	8
10	***40***	7	8	8	9
10	***50***	7	8	9	10
10	***60***	8	9	10	11
10	***70***	8	10	11	12
10	***80***	9	10	11	13
10	***90***	9	10	12	13
10	***100***	9	10	12	13

Tabla 8. Valores críticos de "*q*" para la prueba de Tukey (**α = 0,05**)

v	*k: 2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99	7,17	7,32	7,47	7,60	7,72	7,83	7,93	8,03	8,12	8,21
6	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49	6,65	6,79	6,92	7,03	7,14	7,24	7,34	7,43	7,51	7,59
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16	6,30	6,43	6,55	6,66	6,76	6,85	6,94	7,02	7,10	7,17
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05	6,18	6,29	6,39	6,48	6,57	6,65	6,73	6,80	6,87
9	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74	5,87	5,98	6,09	6,19	6,28	6,36	6,44	6,51	6,58	6,64
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60	5,72	5,83	5,93	6,03	6,11	6,19	6,27	6,34	6,40	6,47
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	5,61	5,71	5,81	5,90	5,98	6,06	6,13	6,20	6,27	6,33
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39	5,51	5,61	5,71	5,80	5,88	5,95	6,02	6,09	6,15	6,21
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43	5,53	5,63	5,71	5,79	5,86	5,93	5,99	6,05	6,11
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,55	5,64	5,71	5,79	5,85	5,91	5,97	6,03
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20	5,31	5,40	5,49	5,57	5,65	5,72	5,78	5,85	5,90	5,96
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,35	5,44	5,52	5,59	5,66	5,73	5,79	5,84	5,90
17	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11	5,21	5,31	5,39	5,47	5,54	5,61	5,67	5,73	5,79	5,84
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27	5,35	5,43	5,50	5,57	5,63	5,69	5,74	5,79
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23	5,31	5,39	5,46	5,53	5,59	5,65	5,70	5,75
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11	5,20	5,28	5,36	5,43	5,49	5,55	5,61	5,66	5,71
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,10	5,18	5,25	5,32	5,38	5,44	5,49	5,55	5,59
30	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82	4,92	5,00	5,08	5,15	5,21	5,27	5,33	5,38	5,43	5,47
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73	4,82	4,90	4,98	5,04	5,11	5,16	5,22	5,27	5,31	5,36
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81	4,88	4,94	5,00	5,06	5,11	5,15	5,20	5,24
120	2,80	3,36	3,68	3,92	4,10	4,24	4,36	4,47	4,56	4,64	4,71	4,78	4,84	4,90	4,95	5,00	5,04	5,09	5,13
121	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62	4,68	4,74	4,80	4,85	4,89	4,93	4,97	5,01

Tabla 8 (continuación). Valores críticos de "*q*" para la prueba de Tukey (α = **0,01**)

v	*k = 2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
5	5,70	6,98	7,80	8,42	8,91	9,32	9,67	9,97	10,24	10,48	10,70	10,89	11,08	11,24	11,40	11,55	11,68	11,81	11,93
6	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,10	9,30	9,48	9,65	9,81	9,95	10,08	10,21	10,32	10,43	10,54
7	4,95	5,92	6,54	7,01	7,37	7,68	7,94	8,17	8,37	8,55	8,71	8,86	9,00	9,12	9,24	9,35	9,46	9,55	9,65
8	4,75	5,64	6,20	6,62	6,96	7,24	7,47	7,68	7,86	8,03	8,18	8,31	8,44	8,55	8,66	8,76	8,85	8,94	9,03
9	4,60	5,43	5,96	6,35	6,66	6,91	7,13	7,33	7,49	7,65	7,78	7,91	8,03	8,13	8,23	8,33	8,41	8,49	8,57
10	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21	7,36	7,49	7,60	7,71	7,81	7,91	7,99	8,08	8,15	8,23
11	4,39	5,15	5,62	5,97	6,25	6,48	6,67	6,84	6,99	7,13	7,25	7,36	7,46	7,56	7,65	7,73	7,81	7,88	7,95
12	4,32	5,05	5,50	5,84	6,10	6,32	6,51	6,67	6,81	6,94	7,06	7,17	7,26	7,36	7,44	7,52	7,59	7,66	7,73
13	4,26	4,96	5,40	5,73	5,98	6,19	6,37	6,53	6,67	6,79	6,90	7,01	7,10	7,19	7,27	7,35	7,42	7,48	7,55
14	4,21	4,89	5,32	5,63	5,88	6,08	6,26	6,41	6,54	6,66	6,77	6,87	6,96	7,05	7,13	7,20	7,27	7,33	7,39
15	4,17	4,84	5,25	5,56	5,80	5,99	6,16	6,31	6,44	6,55	6,66	6,76	6,84	6,93	7,00	7,07	7,14	7,20	7,26
16	4,13	4,79	5,19	5,49	5,72	5,92	6,08	6,22	6,35	6,46	6,56	6,66	6,74	6,82	6,90	6,97	7,03	7,09	7,15
17	4,10	4,74	5,14	5,43	5,66	5,85	6,01	6,15	6,27	6,38	6,48	6,57	6,66	6,73	6,81	6,87	6,94	7,00	7,05
18	4,07	4,70	5,09	5,38	5,60	5,79	5,94	6,08	6,20	6,31	6,41	6,50	6,58	6,65	6,73	6,79	6,85	6,91	6,97
19	4,05	4,67	5,05	5,33	5,55	5,73	5,89	6,02	6,14	6,25	6,34	6,43	6,51	6,58	6,65	6,72	6,78	6,84	6,89
20	4,02	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09	6,19	6,28	6,37	6,45	6,52	6,59	6,65	6,71	6,77	6,82
24	3,96	4,55	4,91	5,17	5,37	5,54	5,69	5,81	5,92	6,02	6,11	6,19	6,26	6,33	6,39	6,45	6,51	6,56	6,61
30	3,89	4,45	4,80	5,05	5,24	5,40	5,54	5,65	5,76	5,85	5,93	6,01	6,08	6,14	6,20	6,26	6,31	6,36	6,41
40	3,82	4,37	4,70	4,93	5,11	5,26	5,39	5,50	5,60	5,69	5,76	5,83	5,90	5,96	6,02	6,07	6,12	6,16	6,21
60	3,76	4,28	4,59	4,82	4,99	5,13	5,25	5,36	5,45	5,53	5,60	5,67	5,73	5,78	5,84	5,89	5,93	5,97	6,01
120	3,70	4,20	4,50	4,71	4,87	5,01	5,12	5,21	5,30	5,37	5,44	5,50	5,56	5,61	5,66	5,71	5,75	5,79	5,83
121	3,64	4,12	4,40	4,60	4,76	4,88	4,99	5,08	5,16	5,23	5,29	5,35	5,40	5,45	5,49	5,54	5,57	5,61	5,65

Tabla 9. Valores críticos para la prueba F_{MAX}

Grados de libertad	***Número de tratramientos k***										
v	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
2	39,0	87,5	142	202	288	333	403	475	550	828	704
3	15,4	27,8	39,2	50,7	62,0	72,9	83,5	93,9	104	114	124
4	9,6	15,5	20,6	25,2	29,5	33,6	37,5	41,1	44,6	48,0	51,4
5	7,2	10,8	13,7	16,3	18,7	20,8	22,9	24,7	26,5	28,2	29,9
6	5,82	8,38	10,4	12,1	13,7	15,0	16,3	17,5	18,6	19,7	20,7
7	4,99	6,94	8,44	9,70	10,8	11,8	12,7	13,5	14,3	15,1	15,8
8	4,43	6,00	7,18	8,12	9,03	9,78	10,5	11,1	11,7	12,2	12,7
9	4,03	5,34	6,31	7,11	7,80	8,41	8,95	9,45	9,91	10,3	10,7
10	3,72	4,85	5,87	6,34	6,92	7,42	7,87	8,28	8,66	9,01	9,34
12	3,28	4,16	4,75	5,30	5,72	6,09	6,42	6,72	7,00	7,25	7,43
15	2,86	3,54	4,01	4,37	4,68	4,95	5,19	5,40	5,59	5,77	5,95
20	2,46	2,95	3,29	3,54	3,76	3,94	4,10	4,24	4,37	4,49	4,59
30	2,07	2,40	2,61	2,78	2,91	3,02	3,12	3,21	3,29	3,36	3,39
60	1,67	1,85	1,96	2,04	2,11	2,17	2,22	2,26	2,30	2,33	2,36
∞	1	1	1	1	1	1	1	1	1	1	1

Tabla 10. Valores críticos del coeficiente de correlación "*r*" de Pearson

N – 2	***α***		***N-2***	***α***	
	0,05	***0,01***		***0,05***	***0,01***
1	0,997	1	***24***	0,388	0,496
2	0,950	0,990	***25***	0,381	0,487
3	0,878	0,959	***26***	0,374	0,478
4	0,811	0,917	***27***	0,367	0,47
5	0,764	0,874	***28***	0,361	0,463
6	0,707	0,834	***29***	0,355	0,456
7	0,666	0,798	***30***	0,349	0,449
8	0,672	0,765	***35***	0,325	0,418
9	0,602	0,735	***40***	0,304	0,393
10	0,576	0,708	***45***	0,288	0,372
11	0,553	0,864	***50***	0,273	0,354
12	0,532	0,661	***60***	0,250	0,325
13	0,514	0,641	***70***	0,232	0,302
14	0,497	0,623	***80***	0,217	0,283
15	0,482	0,606	***90***	0,205	0,267
16	0,456	0,590	***100***	0,195	0,254
17	0,456	0,575	***125***	0,174	0,228
18	0,444	0,561	***150***	0,159	0,208
19	0,433	0,549	***200***	0,138	0,181
20	0,423	0,537	***300***	0,113	0,148
21	0,413	0,526	***400***	0,098	0,128
22	0,404	0,526	***500***	0,088	0,115
23	0,396	0,505	***1000***	0,062	0,081

Tabla 11. Valores críticos de "*d*" para la prueba de Kologorov-Smirnov para dos muestras

k	n	0,05	0,025	0,01	0,005
3	3	2	3	3	3
	6	3	3	4	4
	9	4	4	4	5
	12	5	4	6	6
	15	4	5	6	6
	18	5	5	6	6
	21	5	6	6	7
	24	5	6	7	1
	21	6	6	7	8
	30	6	7	8	8
	33	6	7	8	8
	35	6	7	8	9
	39	7	7	8	9
	42	7	8	9	9
	45	7	8	9	10
	48	7	8	9	10
	51	7	8	10	10
	54	8	9	10	11
	57	8	9	10	11
	60	8	9	10	11
	63	8	9	11	12
	66	8	9	10	12
	69	8	9	11	12
	72	8	9	11	12
	75	8	10	11	12
	78	9	10	11	12
	81	9	10	11	13
	84	9	10	12	13
	87	9	10	12	13
	90	9	10	12	13
	99	9	10	12	13
4	4	3	3	3	3
	8	4	4	4	5
	12	4	5	5	6
	15	5	5	6	6
	20	5	6	6	7
	24	6	6	7	8
	28	6	7	7	8
	32	6	7	8	9
	36	7	7	8	9
	40	7	8	9	9
	44	7	8	9	10
	48	7	8	10	10
	52	8	9	10	11
	56	8	9	10	11
	60	8	9	10	11
	64	8	9	11	12
	68	9	10	11	12
	72	9	10	11	12
	76	9	10	11	12
	80	9	10	11	12
	84	9	10	12	13
	88	9	10	12	13
	92	9	10	12	13
	96	9	10	12	13
	100	9	11	12	13

k	n	0,05	0,025	0,01	0,005
5	5	3	3	4	4
	10	4	4	5	5
	15	4	5	5	6
	20	5	5	6	7
	25	5	6	6	7
	30	5	6	7	8
	35	6	7	7	8
	40	6	7	8	9
	45	6	7	8	9
	50	7	8	9	10
	55	7	8	9	10
	60	7	8	9	11
	65	7	9	10	11
	70	8	9	10	11
	75	8	9	10	12
	80	8	9	11	12
	85	8	9	11	12
	90	8	9	11	12
	95	8	9	11	12
	100	8	9	11	12
6	6	3	4	4	4
	12	4	5	5	6
	18	5	6	6	7
	24	6	6	7	8
	30	6	7	8	9
	36	7	8	9	9
	42	7	8	9	10
	48	8	9	10	11
	54	8	9	10	11
	60	9	10	11	12
	66	9	10	11	12
	72	9	10	12	13
	78	9	11	12	13
	84	9	11	12	13
	90	10	11	13	14
	96	10	11	13	14
7	7	3	4	4	4
	14	4	5	5	6
	21	5	6	6	7
	28	5	6	7	8
	35	6	7	8	9
	42	6	7	8	9
	49	7	8	9	10
	56	7	8	9	11
	63	8	9	10	11
	70	8	9	10	12
	77	8	9	11	12
	84	8	10	12	12
	91	8	10	11	13
	98	8	10	11	13

Tabla 11 (continuación). Valores críticos de "*d*" para la prueba de Kologorov-Smirnov para dos muestras

k	*n*	*0,05*	*0,025*	*0,01*	*0,005*
8	*8*	3	4	4	5
	16	4	5	6	6
	24	5	6	7	7
	32	6	7	7	8
	40	6	7	8	9
	48	7	8	9	10
	56	7	9	10	11
	64	8	9	10	11
	72	8	9	11	12
	80	8	10	11	12
	88	8	10	11	13
	95	9	10	11	13
9	*9*	4	4	4	5
	18	5	5	6	7
	27	6	6	7	8
	36	6	7	8	9
	45	7	8	9	10
	54	7	9	10	11
	63	8	9	10	11
	72	8	10	11	12
	81	8	10	11	13
	90	9	10	11	13
	99	9	10	12	13
10	*10*	4	4	5	5
	20	5	6	6	7
	30	6	7	7	8
	40	7	8	8	9
	50	7	8	9	10
	60	8	9	10	11
	70	8	10	11	12
	80	9	10	11	13
	90	9	10	12	13
	100	9	10	12	13

Tabla 12. Valores críticos de la prueba *U* de Mann-Whitney

N2 / N1	3		4		5		6		7		8		9		10	
3	0 -	- -														
4	0 -	- -	1 -	0 -												
5	1 -	0 -	2 0	1 -	4 1	2 0										
6	2 -	1 -	3 1	2 0	5 2	3 1	7 3	5 2								
7	2 0	1 -	4 1	3 0	6 3	5 2	8 4	6 3	11 6	8 4						
8	3 0	2 -	5 2	4 1	8 4	6 3	10 6	8 4	13 8	10 6	15 10	13 8				
9	3 1	2 0	6 3	4 1	9 5	7 3	12 7	10 5	15 9	12 7	18 11	15 9	21 14	17 11		
10	4 1	3 0	7 3	5 2	11 6	8 4	14 8	11 6	17 11	14 9	20 13	17 11	24 16	20 13	27 19	23 16
11	5 1	3 0	8 4	6 2	12 7	9 5	16 9	13 7	19 12	16 10	23 15	19 13	27 18	23 16	31 22	26 18
12	5 2	4 1	9 5	7 3	13 8	11 6	17 11	14 9	21 14	18 12	26 17	22 15	30 21	26 18	34 24	29 21
13	6 2	4 1	10 5	8 3	15 9	12 7	19 12	16 10	24 16	20 13	28 30	24 17	33 23	28 20	37 27	33 24
14	7 2	5 1	11 6	9 4	16 10	13 7	21 13	17 11	26 17	22 15	31 22	26 18	36 26	31 22	41 30	36 26
15	7 3	5 2	12 7	10 5	18 11	14 8	23 15	19 12	28 19	24 16	33 24	29 20	39 28	34 24	44 33	39 29
16	8 3	6 2	14 7	11 5	19 12	15 9	25 16	21 13	30 21	26 18	36 26	31 22	42 31	37 27	48 36	42 31
17	9 4	6 2	15 8	11 6	20 13	17 10	26 18	22 15	33 23	28 19	39 28	34 24	45 33	39 29	51 38	45 34
18	9 4	7 2	16 9	12 6	22 14	18 11	28 19	24 16	35 24	30 21	41 30	36 26	48 30	42 31	55 41	48 37
19	10 4	7 3	17 9	13 7	23 15	19 12	30 20	25 17	37 26	32 22	44 32	38 28	51 38	45 33	58 44	52 39

Tabla 12 (continuación). Valores críticos de la prueba *U* de Mann-Whitney

N2 / N1	11		12		13		14		15		16		17		18		19	
11	34 25	30 21																
12	38 28	33 24	42 31	37 27														
13	42 31	37 27	47 35	41 31	51 39	45 34												
14	46 34	40 30	51 38	45 34	56 43	50 38	61 47	55 42										
15	50 37	44 33	55 42	49 37	61 47	54 42	66 51	59 46	72 56	64 51								
16	54 41	47 36	60 46	53 41	65 51	59 45	71 56	64 50	77 61	70 55	83 66	75 60						
17	57 44	51 39	64 49	57 44	70 55	63 49	77 60	67 54	83 66	75 60	89 71	81 65	96 77	87 70				
18	61 47	55 42	68 53	61 47	75 59	67 53	82 65	74 58	88 70	80 64	95 76	86 70	102 82	93 75	109 88	99 81		
19	65 50	58 45	72 56	65 51	80 63	72 57	87 69	78 63	94 75	85 69	101 82	92 74	109 88	99 81	116 94	106 87	123 101	113 93

La tabla está dividida en celdas. Cada columna de celdas corresponde al número de datos en la muestra 1 (N1) y cada fila al número de datos en la muestra 2 (N2). En la prueba los datos se arreglan de forma que N1 es menor que N2. Dentro de cada celda hay 4 valores críticos de U arreglados del siguiente modo:

		Número de colas	
		una	dos
α	0,05		
	0,01		

La región critica contiene valores de *U* igual o menor el valor crítico.

Tabla 13. Valores críticos de "*Q*" para Comparaciones Múltiples No Paramétricas para tamaño de muestra diferentes

k	**0,10**	**0,05**	**0,01**	**0,005**	**0,001**
2	1,645	1,960	2576	2,807	3,291
3	2,128	2,394	2,936	3,144	3,588
4	2,394	2,639	3,144	3,342	3,765
5	2,576	2,807	3,291	3,481	3,891
6	2,713	2,936	3,403	3,558	3,988
7	2,823	3,038	3,494	3,675	4,067
8	2,914	3,124	3570	3,748	4,134
9	2,992	3,197	3,635	3,810	4,191
10	3,059	3,261	3,692	3,865	4,241
11	3,119	3,317	3,743	3,914	4,286
12	3,172	3,368	3,789	3,957	4,363
13	3,220	3,414	3,830	3,997	4,363
14	3264	3,456	3,868	4,034	4,397
15	3,304	3,494	3,902	4,067	4,428
16	3,342	3,529	3,935	4,098	4,456
17	3,376	3,562	3,965	4,127	4,483
18	3,409	3,593	3,993	4,154	4,508
19	3,439	3,622	4,019	4,179	4,532
20	3,467	3,649	4,044	4,203	4,554
21	3,494	3,675	4,067	4,22o	4,575
22	3,519	3,699	4,089	4,247	4,595
23	3,543	3,722	4,110	4,268	4,614
24	3,566	3,744	4,130	4,287	4,632
25	3,5 8	3,765	4,149	4,305	4,649

Tabla 14. Valores críticos para el coeficiente de correlación "*R*" de Spearman

n	***0,05***	***0,01***
5	0,900	
6	0,829	0,943
7	0,714	0,893
8	0,643	0,833
9	0,600	0,783
10	0,564	0,745
11	0,523	0,736
12	0,497	0,703
13	0,475	0,673
14	0,457	0,646
15	0,441	0,623
16	0,425	0,601
17	0,417	0,582
18	0,399	0,564
19	0,388	0,549
20	0,377	0,534
21	0,368	0,521
22	0,359	0,534
23	0,351	0,521
24	0,343	0,508
25	0,336	0,496
26	0,329	0,485
27	0,323	0,475
28	0,317	0,465
29	0,311	0,456
30	0,305	0,448
35	0,283	0,394
40	0,264	0,368
45	0,248	0,347
50	0,235	0,329
60	0,214	0,300
70	0,198	0,278
80	0,185	0,260
90	0,174	0,245
100	0,165	0,233

Tabla 15. Factor de multiplicación para la media geométrica

Valores de $\psi_n\left(\frac{1}{2}s_y^2\right)$ y $\phi_n(s_y^2)$ para distribuciones de frecuencias log-normal

T	*Tamaño de la muestra (n)*							
	2	***5***	***8***	***10***	***13***	***15***	***25***	***50***
0,05	1,025	1,041	1,045	1,046	1,047	1,048	1,049	1,050
0,10	1,050	1,082	1,091	1,093	1,096	1,097	1,100	1,103
0,15	1,076	1,125	1,138	1,143	1,147	1,149	1,154	1,158
0,20	1,102	1,169	1,187	1,194	1,200	1,203	1,210	1,216
0,25	1,128	1,214	1,238	1,247	1,255	1,259	1,268	1,276
0,30	1,154	1,260	1,291	1,302	1,312	1,317	1,330	1,340
0,35	1,180	1,307	1,345	1,359	1,372	1,378	1,393	1,406
0,40	1,207	1,356	1,401	1,418	1,433	1,441	1,460	1,476
0,45	1,234	1,406	1,459	1,479	1,498	1,506	1,530	1,548
0,50	1,261	1,457	1,519	1,542	1,564	1,574	1,602	1,625
0,55	1,288	1,509	1,581	1,608	1,633	1,645	1,618	1,705
0,60	1,315	1,563	1,645	1,675	1,705	1,719	1,757	1,789
0,65	1,343	1,618	1,711	1,746	1,780	1,796	1,840	1,876
0,70	1,371	1,875	1,779	1,818	1,857	1,876	1,926	1,968
0,75	1,399	1,733	1,849	1,894	1,938	1,958	2,016	2,064
0,80	1,427	1,792	1,922	1,971	2,021	2,045	2,110	2,165
0,85	1,456	1,853	1,996	2,052	2,108	2,134	2,208	2,270
0,90	1,485	1,915	2,074	2,135	2,197	2,227	2,310	2,381
0,95	1,514	1,979	2,153	2,221	2,291	2,323	2,417	2,496
1,00	1,543	2,044	2,235	2,310	2,387	2,424	2,528	2,617
1,05	1,573	2,111	2,320	2,403	2,487	2,528	2,644	2,744
1,10	1,602	2,180	2,407	2,498	2,591	2,636	2,765	2,876
1,15	1,632	2,250	2,497	2,596	2,698	2,748	2,891	3,014
1,20	1,662	2,321	2,589	2,698	2,810	2,864	3,022	3,159
1,25	1,693	2,395	2,685	2,803	2,926	2,985	3,159	3,311
1,30	1,724	2,470	2,783	2,911	3,045	3,111	3,301	3,470
1,35	1,754	2,547	2,884	3,023	3,169	3,241	3,450	3,636
1,40	1,786	2,626	2,988	3,139	3,298	3,376	3,604	3,809
1,45	1,817	2,706	3,096	3,259	3,431	3,515	3,766	3,991
1,50	1,849	2,788	3,206	3,382	3,569	3,661	3,933	4,181
1,55	1,880	2,873	3,320	3,510	3,711	3,811	4,108	4,379
1,60	1,913	2,959	3,437	3,642	3,859	3,967	4,291	4,587
1,65	1,945	3,047	3,558	3,777	4,012	4,129	4,480	4,804
1,70	1,977	3,137	3,682	3,918	4,171	4,297	4,678	5,031
1,75	2,010	3,229	3,810	4,062	4,334	4,471	4,883	5,269
1,80	2,043	3,323	3,942	4,212	4,504	4,651	5,097	5,517
1,85	2,077	3,420	4,077	4,366	4,680	4,838	5,320	5,776
1,90	2,110	3,518	4,216	4,525	4,861	5,031	5,552	6,048
1,95	2,144	3,619	4,359	4,688	5,049	5,232	5,794	6,331
2,00	2,178	3,721	4,506	4,857	5,243	5,439	6,045	6,628

Los valores de $\psi_n\left(\frac{1}{2}s_y^2\right)$ se buscan en la Tabla entrando con $T=\left(\frac{1}{2}s_y^2\right)$ y *n*= tamaño de la muestra. Para hallar $\phi_n(s_y^2)$ se busca en Tabla entrando con $T=(s_y^2)$ y *n*= tamaño de la muestra. Valores intermedios se obtiene interpolando linealmente con los valores más cercanos.

Tabla 16. Valores críticos U^2 para la prueba de Watson para una muestra

n	*α: 0,10*	*0,05*	*0,01*
4	0,144	0,171	0,230
5	0,146	0,175	0,238
6	0,147	0,177	0,243
7	0,148	0,179	0,247
8	0,149	0,180	0,250
9	0,149	0,181	0,252
10	0,150	0,182	0,253
11	0,150	0,182	0,255
12	0,150	0,183	0,256
13	0,150	0,183	0,257

Tabla 17. Valores críticos para la prueba Z de Rayleigh

n \ *α*	*0,10*	*0,05*	*0,02*	*0,01*	*0,001*
6	2,74	2,86	3,57	4,06	5,29
10	2,30	2,92	3,72	4,29	5,99
20	2,30	2,96	3,82	4,45	6,47
30	2,30	2,97	3,85	4,50	6,62
40	2,30	2,98	3,86	4,53	6,69
50	2,30	2,98	3,87	4,54	6,74
100	2,30	2,99	3,89	4,57	6,82
200	2,30	2,99	3,90	4,59	6,87
500	2,30	2,99	3,91	4,60	6,89
∞	2,30	3,00	3,91	4,61	6,91

Tabla 18. Valores críticos de *"u"* para el test *V* de uniformidad circular

n \ *α*	*0,10*	*0,05*	*0,025*	*0,01*	*0,001*
8	1,296	1,649	1,947	2,280	2,916
9	1,294	1,649	1,948	2,286	2,937
10	1,293	1,648	1,950	2,290	2,954
20	1,287	1,646	1,955	2,308	3,025
30	1,285	1,646	1,957	2,315	3,047
40	1,284	1,646	1,957	2,317	3,058
50	1,284	1,645	1,958	2,319	3,065
75	1,283	1,645	1,959	2,322	3,073
100	1,283	1,645	1,959	2,333	3,077
200	1,282	1,645	1,959	2,325	3,084
300	1,282	1,645	1,960	2,325	3,086
∞	1,2818	1,6449	1,9598	2,3256	3,0877

Tabla 19. Factor de corrección "*K*" para el test de Watson

r	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0		188,49	94,747	63,5015	47,8749	38,4992	32,2498	27,7851	24,4367	21,8325
0,01	19,7489	18,0444	16,6239	15,4219	14,3916	13,4986	12,7173	12,0278	11,4150	10,8667
0,02	10,3731	9,9266	9,5206	9,1500	8,8103	8,4976	8,2091	7,6941	7,6938	7,4628
0,03	7,2472	7,0455	6,8564	6,6787	6,5115	6,3539	6,2050	6,0641	5,9306	5,8040
0,04	5,6837	5,5693	5,4603	5,3564	5,2572	5,1625	5,0718	4,9850	4,9102	4,8219
0,05	4,7453	4,6717	4,6009	4,5322	4,4672	4,4039	4,3430	4,2841	4,2273	4,1724
0,06	4,1194	4,0680	4,0184	3,9703	3,9237	3,8785	3,8347	3,7922	3,7510	3,7109
0,07	3,6720	3,6342	3,5974	3,5616	3,5268	3,4930	3,4600	3,4278	3,3965	3,3666
0,08	3,3362	3,3072	3,2789	3,2512	3,2243	3,1979	3,1722	3,1470	3,1224	3,0984
0,09	3,0749	3,0519	3,0294	3,0074	2,9858	2,9648	2,9441	2,9239	2,9041	2,8846
0,1	2,8656	2,8469	2,8286	2,8107	2,7931	2,7758	2,7589	2,7423	2,7259	2,7099
0,11	2,,6942	2,6787	2,6636	2,6487	2,6340	2,6196	2,6055	2,5915	2,5779	2,5644
0,1 2	2,5512	2,5382	2,5254	2,5128	2,5004	2,4882	2,4762	2,4644	2,4528	2,4413
0,1 3	2,4301	2,4189	2,4080	2,3972	2,3866	2,3762	2,3658	2,3557	2,3457	2,3358
0,14	2,3261	2,3165	2,3070	2,2977	2,2885	2,2794	2,2705	2,2616	2,2529	2,2443
0,1 5	2,2358	2,2275	2,2192	2,2110	2,2030	2,1950	2,1872	2,1794	2,1718	2,1642
0,16	2,1567	2,1494	2,1421	2,1349	2,1278	2,1208	2,1138	2,1070	2,1002	2,0935
0,17	2,0868	2,0803	2,0738	2,0764	2,0611	2,0549	2,0487	2,0426	2,0365	2,0305
0,18	2,0246	2,0188	2,0130	2,0072	2,0016	1,9960	1,9904	1,9849	1,9795	1,9741
0,19	1,9688	1,9635	1,9583	1,9532	1,9481	1,9430	1,9380	1,9331	1,9282	1,9233
0,2	1,9185	1,9137	1,9090	1,9043	1,8997	1,8951	1,8906	1,8861	1,8817	1,8772
0,21	1,8729	1,8685	1,8643	1,8600	1,8558	1,8516	1,8475	1,8434	1,8393	1,8353
0,22	1,8313	1,8274	1,8234	1,8195	1,8157	1,8119	1,8081	1,8043	1,8006	1,7969
0,23	1,7933	1,7896	1,7860	1,7825	1,7789	1,7754	1,7719	1,7685	1,7651	1,7617
0,24	1,7583	1,7550	1,7516	1,7484	1,7451	1,7419	1,7386	1,7355	1,7323	1,7292
0,25	1,7261	1,7230	1,7199	1,7169	1,7138	1,7080	1,7079	1,7049	1,7020	1,6991
0,26	1,6962	1,6933	1,6905	1,6877	1,6849	1,6821	1,6793	1,6766	1,6739	1,6712
0,27	1,6685	1,6658	1,6632	1,6606	1,6579	1,6554	1,6528	1,6502	1,6477	1,6452
0,28	1,6427	1,6402	1,6377	1,6353	1,6328	1,6304	1,6280	1,6256	1,6233	1,6209
0,29	1,6186	1,6162	1,6139	1,6116	1,6094	1,6071	1,6048	1,6026	1,6004	1,5982
0,3	1,5960	1,5938	1,5916	1,5895	1,5873	1,5852	1,5831	1,5810	1,5789	1,5768
0,31	1,5748	1,5727	1,5707	1,5687	1,5667	1,5647	1,5627	1,5607	1,5587	1,5568
0,32	1,5548	1,5529	1,5510	1,5491	1,5472	1,5453	1,5434	1,5416	1,5397	1,5379
0,33	1,5360	1,5342	1,5324	1,5306	1,5288	1,5270	1,5253	1,5235	1,5217	1,5200
0,34	1,5183	1,5165	1,5148	1,5131	1,5114	1,5097	1,5081	1,5064	1,5047	1,5031
0,35	1,5014	1,4998	1,4982	1,4986	1,4950	1,4934	1,4918	1,4902	1,4886	1,4871
0,36	1,4885	1,4839	1,4824	1,4809	1,4793	1,4778	1,4763	1,4748	1,4733	1,4718
0,37	1,4703	1,4689	1,4674	1,4659	1,4645	1,4630	1,4616	1,4602	1,4587	1,4573
0,38	1,4559	1,4545	1,4531	1,4517	1,4503	1,4490	1,4476	1,4462	1,4449	1,4435
0,39	1,4422	1,4408	1,4395	1,4382	1,4368	1,4355	1,4342	1,4329	1,4316	1,4303

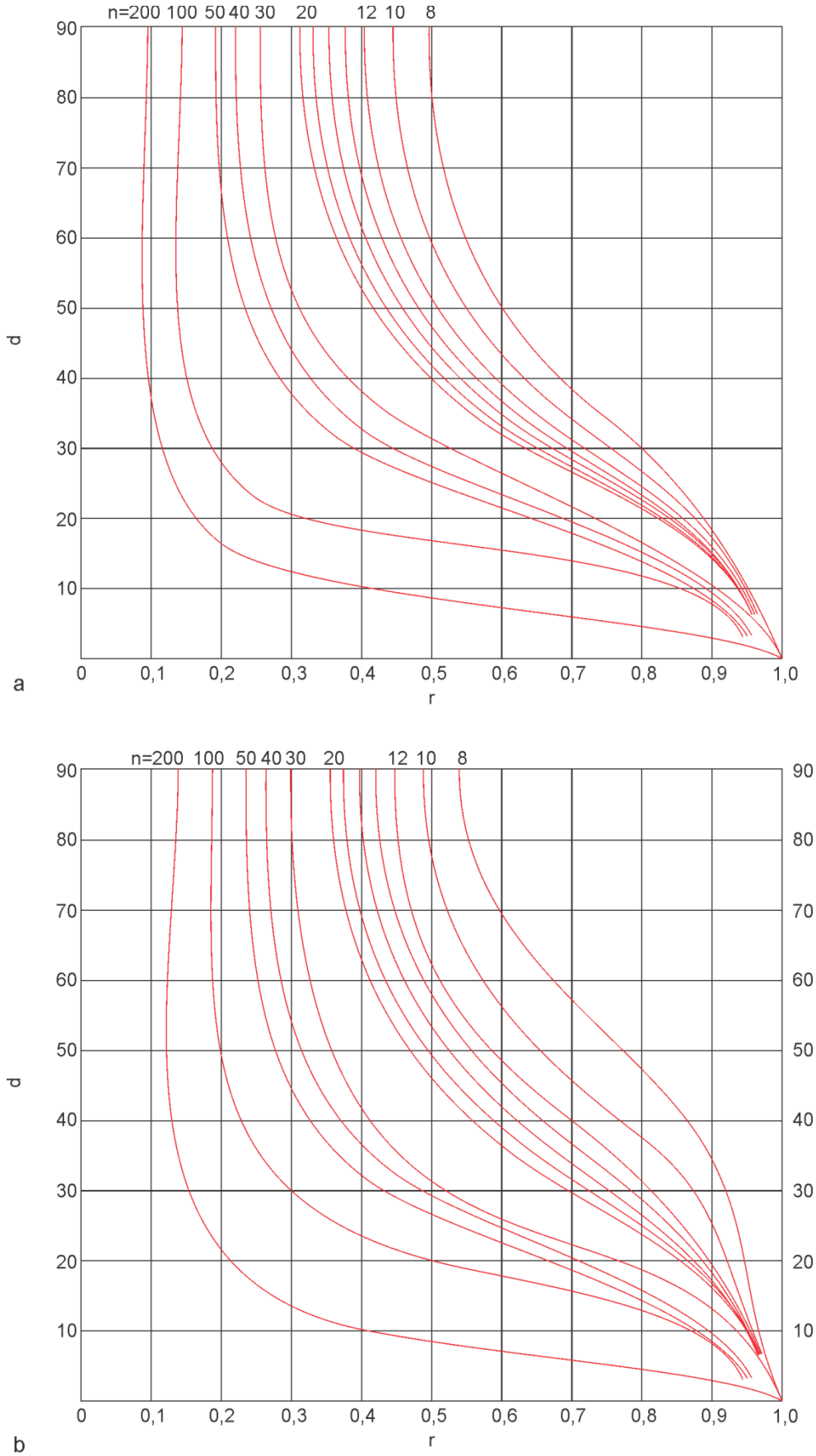
Tabla 19 (continuación). Factor de corrección "*K*" para el test de Watson

r	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0,4	1,4290	1,4277	1,4205	1,4252	1,4239	1,4227	1,4214	1,4202	1,4189	1,4177
0,41	1,4165	1,4152	1,4140	1,4128	1,4116	1,4104	1 ,4092	1,4092	1,4068	1,4056
0,42	1,4044	1,4033	1,4021	1,4009	1,3998	1 ,3986	1 ,3975	1,3963	1,3952	1,3940
0,43	1,3929	1,3918	1,3907	1,3895	1,3844	1,3873	1 ,3862	1,3851	1,3840	1,3829
0,44	1,3818	1,3808	1,3797	1,3786	1,3775	1,3765	1,1754	1,3744	1,3733	1,3723
0,45	1,3712	1,3702	1,3691	1,36SI	1,3671	1,3660	1,3650	1,3640	1,3630	1,3620
0,46	1,3610	1,3600	1,3590	1,3580	1,3570	1,3560	1,3550	1,3540	1,3530	1,3521
0,47	1,3511	1,3501	1,3492	1,3482	1,3472	1,3463	1,3453	1,3444	1,3434	1,3425
0,48	1,3416	1,3406	1,3397	1,3388	1,3378	1,1369	1,3360	1,3351	1,3342	1,3333
0,49	1,3324	1,3315	1,3306	1,3237	1,3281	1,3279	1,3270	1,3261	1,3252	1,3243
0,5	1,3235	1,3226	1,3217	1,3209	1,3200	1,3191	1,3183	1,3174	1,3166	1,3157
0,51	1,3148	1,3140	1,3132	1,3123	1,3115	1,3106	1,3098	1,3090	1,3081	1,3073
0,52	1,3065	1,3057	1,1049	1,3040	1,3032	1,3024	1,3016	1,3008	1,3000	1,2992
0,53	1,2984	1,2976	1,2968	1,2960	1,2952	1,2944	1,2936	1,2929	1,2921	1,2913
0,54	1,2905	1,2897	1,2890	1,2882	1,2874	1,2867	1,2859	1,2851	1,2844	1,2836
0,55	1,2829	1,2821	1,2814	1,2806	1,2799	1,2791	1,2784	1,2776	1,2769	1,2762
0,56	1,2754	1,2747	1,2740	1,2732	1,2725	1,2718	1,2710	1,2703	1,2692	1,2689
0,57	1,2682	1,2674	1,2667	1,2660	1,2653	1,2646	1,2639	1,2632	1,2625	1,2618
0,58	1,2611	1,2604	1,2597	1,2590	1,2583	1,2576	1,2569	1,2562	1,2555	1,2548
0,59	1,2542	1,2535	1,2528	1,2521	1,2514	1,2509	1,2501	1,2494	1,2487	1,2481
0,6	1,2474	1,2467	1,2461	1,2454	1,2447	1,2441	1,2434	1,2428	1,2421	1,2414
0,61	1,2408	1,2401	1,2395	1,2388	1,2382	1,2315	1,2369	1,2362	1,2356	1,2350
0,62	1,2341	1,2337	1,2330	1,2324	1,2318	1,2311	1,2305	1,2298	1,2292	1,2286
0,63	1,2280	1,2273	1,2267	1,2261	1,2254	1,2248	1,2242	1,2236	1,2230	1,2223
0,64	1,2217	1,2211	1,2205	1,2199	1,2193	1,2186	1,2180	1,2174	1,2168	1,2162
0,65	1,2156	1,2150	1,2144	1,2138	1,2132	1,2126	1,2120	1,2114	1,2108	1,2102
0,66	1,2096	1,2090	1,2084	1,2078	1,2072	1,2066	1,2060	1,2054	1,2048	1,2042
0,67	1,2036	1,2030	1,2024	1,2018	1,2013	1,2007	1,2001	1,1995	1,1989	1,1983
0,68	1,1977	1,1972	1 1966	1,1960	1,1190	1,1948	1,1948	1,1937	1,1931	1,1925
0,69	1,1920	1,1914	1,1908	1,1902	1,1897	1,1897	1,1885	1,1879	1,1874	1,1868
0,7	1,1862	1,1857	1,1851	1,1845	1,1840	1,1834	1,1828	1,1823	1,1817	1,1811
0,71	1,1806	1,1800	1,1794	1,1789	1,1783	1,1777	1,1772	1,1766	1,1761	1,1755
0,72	1,1749	1,1744	1,1738	1,1733	1,1727	1,1721	1,1716	1,1710	1,1705	1,1699
0,73	1,1694	1,1688	1,1682	1,1677	1,1671	1,1666	1,1660	1,1655	1,1649	1,1644
0,74	1,1638	1,1633	1,1627	1,1621	1,1616	1,1610	1,1605	1,1599	1,1594	1,1588
0,75	1,1583	1,1517	1,1572	1,1566	1,1561	1,1555	1,1550	1,1544	1,1539	1,1533
0,76	1 ,1528	1,1522	1,1517	1,1511	1,1505	1,1500	1,1494	1,1489	1,1483	1,1478
0,77	1,1472	1,1467	1,1461	1,1456	1,1450	1,1445	1,1439	1,1434	1,1428	1,1423
0,78	1,1417	1,1412	1,1406	1,1401	1,1395	1,1389	1,1384	1,1378	1,1373	1,1367
0,79	1,1362	1,1356	1,1351	1,1345	1,1340	1,1334	1,1328	1,1323	1,1317	1,1312
0,8	1,1306	1,1300	1,1295	1,1289	1,1284	1,1278	1,1272	1,1267	1,1261	1,1256

Tabla 19 (continuación). Factor de corrección "*K*" para el test de Watson

r	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0,81	1,1250	1,1244	1,1239	1,1233	1,1227	1,1222	1,1216	1,1210	1,1205	1,1199
0,82	1,1193	1,1188	1,1182	1,1176	1,1170	1,1165	1,1159	1,1153	1,1147	1,1142
0,83	1,1136	1,1130	1,1124	1,1119	1,1113	1,1107	1,1101	1,1095	1,1090	1,1084
0,84	1,1078	1,1072	1,1060	1,1060	1,1054	1,1049	1,1043	1,1037	1,1031	1,1025
0,85	1,1019	1,1013	1,1007	1,1001	1,0995	1,0989	1,0983	1,0977	1,0971	1,0965
0,86	1,0959	1,0953	1,0947	1,0941	1,0935	1,0928	1,0922	1,0916	1,0910	1,0904
0,87	1,0898	1,0892	1,0885	1,0879	1,0873	1,0867	1,0861	1,0854	1,0848	1,0842
0,88	1,0815	1,0829	1,0823	1,0816	1,0810	1,0804	1,0797	1,0791	1,0785	1,0778
0,89	1,0772	1,0765	1,0759	1,0752	1,0746	1,0740	1,0733	1,0727	1,0720	1,0713
0,9	1,0707	1,0700	1,0694	1,0687	1,0681	1,0674	1,0667	1,0661	1,0654	1,0647
0,91	1,0641	1,0634	1,0627	1,0621	1,0614	1,0607	1,0601	1,0594	1,0587	1,0580
0,92	1,0573	1,0567	1,0560	1,0553	1,0546	1,0539	1,0533	1,0526	1,0519	1,0512
0,93	1,0505	1,0498	1,0491	1,0484	1,0477	1,0470	1,0463	1,0456	1,0449	1,0443
0,94	1,0436	1,0429	1,0422	1,0414	1,0407	1,0400	1,0393	1,0386	1,0379	1,0372
0,95	1,0365	1,0358	1,0351	1,0344	1,0337	1,0330	1,0322	1,0315	1,0308	1,0301
0,96	1,0294	1,0287	1,0279	1,0272	1,0265	1,0258	1,0251	1,0243	1,0236	1,0229
0,97	1,0222	1,0214	1,0207	1,0200	1,0192	1,0185	1,0178	1,0170	1,0163	1,0156
0,98	1,0148	1,0141	1,0134	1,0126	1,0119	1,0119	1,0104	1,0097	1,0089	1,0082
0,99	1,0075	1,0067	1,0060	1,0052	1,0015	1,0037	1,0030	1,0022	1,0000	1,0000

Figura 1. Valores *d* para Límites de confianza para la media angular. a. $\alpha = 0,05$. b. $\alpha = 0,01$



NOTAS

1. La introducción de la cantidad (***n* - 1**) requiere explicación. Dado que la varianza se define como el cuadrado del promedio de las desviaciones respecto al promedio, cuando se procede al muestreo se desconoce la media poblacional μ pero la estimamos con la media muestral, $\overline{X}$. La media muestral, $\overline{X}$, se calcula de manera que minimiza los cuadrados de las desviaciones con respecto a ella (ver 3º propiedad de la media). Esta propiedad de la media muestral tiende a subestimar la varianza cuando se usa la ecuación para calcular σ^2. Esto es S^2 es un estimador sesgado de σ^2, y para corregir este sesgo se reduce el denominador de la varianza muestral a ***n* -1** lo que produce un estimador mayor de σ^2.

2. Recuerde que el promedio de estas desviaciones es cero, en tanto el cuadrado de las desviaciones, la varianza, es siempre positivo.

3. Los modelos son aproximaciones de la realidad que pueden expresarse verbalmente o como relaciones matemáticas. Los modelos deterministicos y los probabilísticos son dos tipos de modelos matemáticos. En los modelo determinsticos se expresa la relación funcional entre las variables (dependiente-independiente). Muchos modelos geológicos son determinísticos, por ejemplo la "ley de Stokes" que describe la velocidad de caída de una partícula en un fluido.

4. Pierre-Simon Laplace (1749-1827) fue un astrónomo, físico y matemático francés que inventó y desarrolló la transformada y la ecuación que llevan su nombre. Además, demostró también la estabilidad del sistema solar, sentó las bases científicas de la teoría matemática de probabilidades, formuló el método de los mínimos cuadrados (fundamental para la teoría de errores). Fue creyente del determinismo (el estado actual "determina" en algún sentido el futuro).

5. Charles Lyell (1797-1875) fue un abogado y geólogo británico. En su obra "*Priciples of Geology*" publicado entre1830 y 1833 presentó principios fundamentales para la historia de la Geología moderna. Lyell fue uno de los representantes más destacados del uniformismo y el gradualismo geológico. El uniformismo o actualismo, es el principio según el cual los procesos naturales que actuaron en el pasado son los mismos que actúan en el presente.

6. Axioma: proposición matemática tan clara y evidente que se acepta sin demostración. Sinónimo: postulado.

7. Andréi Nikoláyevich Kolmogórov (1903 – 1987) fue un matemático ruso que hizo progresos importantes en los campos de la Teoría de la Probabilidad y aplicaciones, Teoría de Funciones, Lógica Matemática, Problemas de Estacionalidad, Educación e Historia de las matemáticas. En particular, estructuró el sistema axiomático de la Teoría de la Probabilidad a partir de la Teoría de Conjuntos, donde los elementos son eventos. Trabajó al principio de su carrera en lógica constructivista.

8. Los fenómenos que tienen esta propiedad: "la frecuencia relativa con que ocurren en una gran serie de observaciones es estable" reciben el nombre de fenómenos aleatorios o estocásticos.

9. Thomas Bayes (1702 – 1761) fue un matemático británico. Su obra más conocida es el Teorema que lleva su nombre. Bayes fue uno de los primeros en utilizar la probabilidad inductivamente y establecer una base matemática para la inferencia probabilística. Actualmente, con base en su obra, se ha desarrollado una poderosa teoría que se aplica en áreas del conocimiento muy diversas.

10. La distribución normal fue descubierta por De Moivre (matemático francés, 1667 – 1774) quien en 1773 enuncia la ecuación de la distribución. En 1774 Laplace (físico-matemático francés, 1749 – 1827) la redescubre. Sin embargo fue Gauss (matemático y físico alemán, 1777 – 1855) quien en 1809 la desarrolla. Durante los siglos XVIII y XIX se realizaron estudios que intentaban demostrar que el modelo normal es el que poseen la mayoría de las variables continuas (de ahí el nombre normal). Hoy se sabe que los intentos fracasaron.

11. La tabla de números aleatorios contiene 5000 dígitos. Idealmente, estos números son generados por un mecanismo tal que cada dígito es el resultado de un ensayo que consiste en una extracción de un número de 0,1... ,9 con una probabilidad igual a 1/10; los dígitos en posiciones diferentes son los resultados de repeticiones independientes de tales ensayos. El modelo en que está basada la tabla de números aleatorios asegura que todos los dígitos simples tienen la misma probabilidad de ocurrencia de 1/10, que todos los pares de dígitos 00, 01, ...,99 tienen una probabilidad de ocurrencia igual a 1/100, y así sucesivamente.

12. La probabilidad de error de tipo *I* aumenta con el número de pares de comparaciones. Cada vez que se formula una hipótesis nula, el riesgo de error de tipo *I* es α. Para $\alpha = 0,05$, la probabilidad de no cometer error de tipo *I* es $1 - \alpha = 0,95$. Para tres medias existen tres pares de comparaciones posibles, la probabilidad de no cometer error de tipo *I* es $0,95^3 = 0,86$ y la probabilidad de cometer error de tipo *I* es 0,14. Para cuatro medias existen seis pares de comparaciones posibles, la probabilidad de no cometer error de tipo *I* es $0,95^6 = 0,74$ y la probabilidad de cometer error de tipo *I* es 0,26.

13. Karl Pearson (1857-1936), científico, matemático y pensador británico, que estableció la disciplina de la estadística matemática, fue el fundador de la bioestadística.

14. Carlo Emilio Bonferroni (1892-1960) matemático italiano que trabajó en teoría de probabilidades.

15. Los grados de libertad de la regresión son el número de parámetros estimados menos uno. En este caso se estiman a partir de la muestra la pendiente (β) y la ordenada al origen (α) poblacional, de ahí que los grados de libertad son $2 – 1 = 1$.

16. Se llama sesgo de un estimador a la diferencia entre su esperanza matemática y el valor del parámetro que estima. Un estimador cuyo sesgo es nulo se llama insesgado o centrado.

17. Estas metodologías de análisis de series no se pueden usar con mediciones realizadas en transectas, que si bien pueden ser unidireccionales, generalmente ocurre que los datos están vinculados con sus vecinos de ambos lados.

18. Andrei Andreevitch Markov (1856-1922) fue un matemático ruso conocido por sus trabajos en la teoría de los números y la teoría de probabilidades. Su trabajo teórico en el campo de los procesos en los que están involucrados componentes aleatorios originó lo actualmente se conoce como cadena de Márkov consideradas una herramienta esencial en disciplinas como la economía, la ingeniería, la investigación de operaciones y muchas otras.

19. $\mathbf{x} \oplus \mathbf{y}$ es un elemento del S^d y puede mostrarse como $(S^d, \oplus)$. La operación de Perturbación satisface la condición de ser asociativa, conmutativa y tener un elemento neutro (e) lo que le confiere una estructura de grupo conmutativo. Como consecuencia dado dos vectores cualesquiera existe siempre un vector perturbador $\boldsymbol{x}$ que transforma el primero en el segundo (von Eynatten, *et al*., 2002).

20. Una función es biyectiva cuando todos los elementos del conjunto de salida tienen una imagen distinta en el conjunto de llegada, y a cada elemento del conjunto de llegada le corresponde un elemento del conjunto de salida.

21. Matriz singular es aquella tiene determinante cero. Esta dificultad suele ser superada dado que muchos paquetes estadísticos aplican inversas generalizadas para superar el problema cuando se requiere utilizar la matriz de varianza-covarianza como es el caso de Análisis de Componentes Principales.

22. John Michell (1724-1793) filosofo naturalista inglés su trabajo abarca temas teóricos y experimentales de la astronomía, geología, óptica y gravitación. Los principales aportes a la geología se vinculan con los terremotos. Además fue el primero en definir la estratigrafía del mesozoico del Reino Unido.

23. Danie Gerhardus Krige, geólogo de minas sudafricano nació en el Estado Libre de Orange en 1919 es pionero en el campo de la geoestadística.

24. Georges François Paul Marie Matheron (1930 – 2000) fue un matemático y geólogo francés fundador de la Geoestadística y co-fundador junto con Jean Serra de la morfología matemática.

25. La semivarianza es la misma varianza muestral $S^2 = \frac{\sum_{i=1}^{n}(x-\bar{X})^2}{n-1}$, dado que se trata de sólo dos valores $S^2 = \frac{(x_1-\bar{X})^2-(x_2-\bar{X})^2}{2-1} = \frac{(x_1-x_2)^2}{2}$.

26. El nombre efecto pepita proviene de la minería de oro, la distribución de las pepitas de oro en un yacimiento ocurren al azar, entonces se encuentran **pepitas** en una muestra y no en otra.

27. Krigeado y kriging derivan del nombre de D. G. Krige precursor del método.

BIBLIOGRAFÍA

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, London, 416p.

Aitchison, J. (2003). *A concise guide to compositional data analysis*. In CDA Workshop, Girona, 134p.

Aitchison, J. y Brown, J.A.C. (1957). *The Logormal Distribution*. Cambridge University Press, Cambridge.

Aitchison, J y Greenacre, M. (2002). *Biplots of compositional data*. Appl. Statist. 51, Part 4 :375-392.

Alfaro Sironvalle, M. A., (2000). *Estadística*. Universidad de Chile y Universidad de Santiago de Chile, 114p. http://www.marcoalfaro.cl/archivos/estadisticas.pdf

Alfaro Sironvalle, M. A. (2002). *Introducción al Muestreo Minero*. Instituto de Ingenieros de Minas de Chile. 83p. http://www.marcoalfaro.cl/archivos/MUESTREOV-2.pdf

Batschelet, E. (1972). *Recent statistical method for orientation data*. En: S.R. Galler, K. Scmidt-Koening, G.J. Jacobs, and R.E. Belleview (eds.), Animal Orientation and Navigation, U. S. Governement Printing Office, Washington, D. C.

Borradaile, G. (2003). *Statistics of Earth Science Data: their distribution in time, space, and orientation*. Springer. Berlin. 351 p.

Brown, C.E. (1998). *Applied Multivariate Statstics in Geohydrology and Related Sciences*. Springer, Berlin. 248p.

Castaño Agudelo, A.F. y Vergara Elorza, F. (2004). *Simulación geoestadística aplicada al modelamiento de yacimientos de petróleo.* Tesis de pregrado. Universidad Nacional de Colombia sede Medellín, Facultad de minas. http://www.bdigital.unal.edu.co/944/1/3474407_15272213_2004.pdf

Chalmers, A.F. (1990). *¿Qué es esa cosa llamada ciencia? Una valoración de la naturaleza y el estatuto de la ciencia y sus métodos*. Siglo veintiuno editores. México. 245p.

Chayes, F. (1949). *On ratio correlation in petrography*. Jour. Geology. 57(3): 239-254.

Chayes, F. (1960). *On correlation between variables of constant sum*. Jour. Geophys. Research. 65(2): 4185-4193.

Chayes, F. (1962). *Numerical correlation and petrographic variation*. Jour. Geology. 70(4): 440-552.

Chayes, F. and Kruskal, W. (1966). *An approximate rest for correlation between proportions*. Mathematical Geology, 74(5): 692-702.

Cheeny, R.F. (1983). *Statistical Methods in Geology for field and lab decisions*. George Allen & Unwin (Publishers) Ltd. London, UK. 169 p.

Chou, Y.L. (1977). *Análisis estadístico*. Traducción al español por Vincent Agut Ammer. 2ª edición. Iberoamericana. Buenos Aires. 808p.

Cressie, N.A.C. (1980). *Statistics for Spatial data*. John Wiley & Sons. 900p.

Conover, W.J. (1999). *Practical Nonparametric Statistics.* John Wiley & Sons, Inc., New York

Davis, J.C. (2002). *Statistics and data analysis in Geology*. Third Edition. John Wiley & Sons, Inc., New York. 548p.

Deutsch, C. and Journel, A.G. (1997). *GSLIB: Geostatistical Software Library and user´s guide (Applied Geostatistics)*. Oxford University Press Inc., 2nd Revised edition, 384p.

Diggle, P.J. & Ribeiro Jr., P.J. 2007. *Model Based Geostatistics*. Springer, New York.

Durrand, D. and Greenwod, J.A. (1958). *Modifications of Rayleigh test for the uniformity in analysis of two-dimensional orientation data*. Journal of Geology 66: 229-238.

Englud, E. and Sparks, A. (1988). *GEO-EAS (Geostatistical environmental assessment software).* Environmental Protection Agency United States Publications.

Fisher, R.A. (1953). *Dispersion on a sphere*. Proceeding of Royal Socity of London. Series A, 217: 295-305.

Fisher, N.I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press, Great Britain.

Goovaerts, P. (1997). *Geostatistics for natural resoureces evaluation*. Oxford University Press. New York, Oxford. 483p.

Isaaks, E.H. and Srivastava, R.M. (1989). *An introduction to applied Geostatistics*. Oxford University Press, New York. 592p.

Kietzmann, D.A., Cuitiño, J.I. Medina, R. A. y Scasso, R.A. (2009). *Análisis de cadenas de Markov y series de Fourier en una secuencia hemipelágica del Jurásico superior de la Península Antártica*. Latin American Journal of Sedimentology and Basin Analysis. 16 (1): 45-56.

Koch, G.S. and Link, R.E. (1980). *Statistical analysis of Geological Data*. Dos volúmenes en un tomo. Dover Publication, Inc. New York. Vol. I, 375 p, Vol. II, 438p.

Koch y Link el artículo

Krige, D.G. (1951). *A statistical approach to some basic mine valuation problems on the Witwatersrand*. J. of the Chem., Metal. and Mining Soc. of South Africa 52 (6): 119–139.

Krumbein, C. (1962). *Open and closed number systems: stratigraphic mapping*. Bull. Amer. Assoc. Petrol. Geologists, 46: 322-337.

Krumbein W.C. and Graybill F.A. (1965). *An introduction to statistical models in geology.* McGraw-Hill, New York. 475p.

Le Maitre, R. (1982). *Numerical petrography*. Elsevier. Amsterdam. 281p.

Mahan, R.P. (1991). *Applications in Spatial and Temporal Performance Analysis*. United States Army Research Institute for the Behavioral and Social Sciences, Special Report 16. 53p.

Mardia, K.V. (1981). *Statistics of Directional Data*. Academic Press, New York. 357p.

Martín-Fernández, J.A. (2000). *Medidas de diferencias y clasificación automática no paramétrica de datos composicionales.* Tesis doctoral del Programa de doctorado Matemática Aplicada. U.P.C. Girona. 243p.

Matheron, G. (1962). *Traité de géostatistique appliquée*. Tome I: Mémoires du Bureau de Recherches Géologiques et Minières. Pairs: Editions Technip, 14.

McKillup, S. and Dyar, M. D. (2010). *Geostatistics Explain An introduction for earth scientists*. Cambridge University Press, New York. 396p.

Miall, A.D. 1973. *Markov chain analysis applied to an ancient alluvial plain secession*. Sedimentology 20: 347-364.

Mendenhall, W., Wackerly, D. y Scheaffer, R. (1994). *Estadística matemática con Aplicaciones*. 2ª Edición. Edit. Grupo Editorial Iberoamericano. 772p.

Merodio, J. C. (1985). *Métodos estadísticos en Geología*. Asociación Geológica Argentina. Serie B Didáctica y Complementaria N°13. Bs. As. Argentina. 230p.

Muñoz Vicente, M.D. (2008). *Aportaciones al Estudio de la Anisotropía y Modelado Espacial de Información*. Tesis Doctoral de la Universidad de Salamanca. 168p.

Orche García, E. (1999). *Manual de evaluación de yacimientos minerales.* Entorno Gráfico. 300p.

Popper, K. (1968). *The logic of Scientific Discovery*. Taylor & Francis e-Library, 2005. 513p.

Riccardi, A.C. (1977). *Geología: ¿Protociencia, Especulación o Ciencia*? Revista de la Asociación Geológica Argentina 32(1): 52-69.

Rock, N.M. (1988). *Numerical petrology*. Springer-Verlag, Berlin. 427p.

Rodríguez Morilla, C. (2000). *Análisis de series temporales.* La Muralla S.A. / Hespérides S.L. Madrid. 166p.

Rollinson, H.R. (1993). *Using geochemical data: evaluation, presentation, interpretation*. Longman Scientific & Technical. New York. 352p.

Sánchez Fernández, J. (2004). *Capítulo 4.- Series temporales. Introducción a la Estadística Empresarial*. 46p. www.eumed.net/cursecon/libreria/2004/jsf/4.pdf

Santaló, L. A. (1980). *Probabilidades e Inferencia Estadística*. Colección de Monografías Científicas. Serie de Matemática Nº 1. Edit. OEA. 137p.

Scasso, R.A., Kiessling, W. y Santisteban, M. (1998). *Ciclos Markovianos de Tobas-Radiolaritas con dependencia de transición simple en el Jurásico superior de Antártida*. 10° Congreso Latinoamericano de Geología y 6° Congreso Nacional de Geología Económica. Buenos Aires. Actas 1: 84.

Sironvalle, M.A. (2002). *Introducción al muestreo minero*. Instituto de Ingenieros de Minas de Chile. Santiago, Chile. 83p. http://www.marcoalfaro.cl/archivos/MUESTREOV-2.pdf

Sullivan, J. (1998). *Curso de Geoestadística para minería I*. Corporación Nacional del Cobre de Chile. Casa Matriz. 45p.

Thomopoulos N.T. and Johnson A. C. (2003). Tables and Characteristics of the Satandardized Lognormal Distribution. Proceedings of the Descision Sciences Institute, 103: 1-6.

Tulcanaza, E. (1992). *Técnicas geoestadisticas y criterios técnico-económicos para la estimación y evaluación de yacimientos mineros*. Estudios mineros, Santiago, Chile. 256p.

Pananatier, Y. 1994. VARIOWIN 2.1 Softweare. Institute of Mineralogy University of Lausanne, Switzerland.

Zar, H.J. (2009). *Biostatistical analysis*. 5 th ed. Prentice Hall, New Jersey. 945p.

AUTORA

Marta Alperin. Licenciada en Geología (1982) y Doctora en Ciencias Naturales (1988) de la Facultad de Ciencias Naturales y Museo (FCNyM) de la UNLP. Profesora Adjunta de la cátedra de Estadística de la FCNyM desde 2005. En investigación sus principales aportes se realizan en el análisis estadístico de datos geológicos y paleontológicos (Geoestadística y Análisis de datos Composicionales) en el marco de más de doce proyectos de investigación. Cuenta con trabajos publicados en revistas nacionales e internacionales como la Revista de la Asociación Geológica Argentina, Ameghinana, Journal of Foraminiferal Research y Revista de Geología aplicada a la Ingeniería y el Ambiente entre otras. En el campo profesional se desempeñó principalmente en el área de Medio ambiente y en el área de Educación y Geología. Por otra parte participó de la actividad académica como Coordinadora del Departamento de Posgrado (1995-2001) y Consejera Académica (2004-2007) de FCNyM.

Alperin, M., Echeveste, H., Fernández, R, y Bellieni, G. 2007. Análisis estadístico de datos composicionales en volcanitas jurásicas del Macizo del Deseado, provincia de Santa Cruz, Argentina. Revista de la Asociación Geológica Argentina. 62(2): 200-209. 2007. ISSN 0004-4822.

Alperin, M., Cusminsky. G. and Bernasconi. E. 2011. *Benthic foraminiferal morphogroups on the Argentine Continental shelf.* Journal of Foraminiferal Research 41(2): 155-166. ISSN 0096-1191. http://jfr.geoscienceworld.org/cgi/reprint/41/2/155

del Río, J.L., Aperin, M., Bó, M.J., López de Armentia, A., Álvarez, J., Camino, M. y S. Bazzini. 2011. *Cambios inducidos por obras portuarias en ambientes de playa, Quequén, provincia de Buenos Aires, Argentina*. Revista de Geología aplicada a la Ingeniería y el Ambiente. 26, 53-62. ISSN Nº 1851-7838.