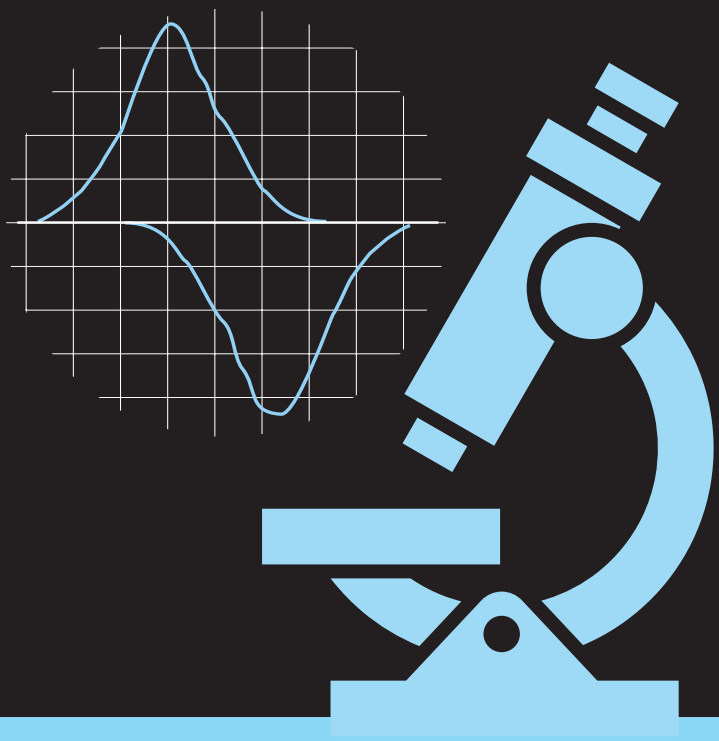
Estadística y salud

Principios metodológicos en las guías de publicación



Erik Cobo
José Antonio González

UPCGRAU 73

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Estadística y salud

Principios metodológicos en las guías de publicación

Erik Cobo
José Antonio González

UPCGRAU 73

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Primera edición: octubre de 2023

Oficina de Publicacions Acadèmiques Digitals de la UPC
Edificio K2M, Planta S1, Despacho S103-S104
Jordi Girona 1-3, 08034 Barcelona
Tel.: 934 015 885
www.upc.edu/idp
E-mail: info.idp@upc.edu

Producción: Service Point
Pau Casals, 161-163
08820 El Prat de Llobregat (Barcelona)

ISBN:978-84-10008-00-7
ISBN digital: 978-84-10008-01-4
DL: B 18425-2023
DOI: 10.5821/ebook-9788410008014

Presentación

Este texto afronta el principal reto de la estadística para la salud: alcanzar un lenguaje, un léxico compartido por teóricos y aplicados. En consecuencia, explicamos conceptos básicos a toda la comunidad académica, tanto aplicada (medicina, enfermería, farmacia, psicología, odontología, fisioterapia, etc.) como teórica (estadística, matemática, informática, ingeniería, etc.).

Abordamos solo conceptos que disponen de consenso científico. Hemos encontrado este acuerdo en las guías de publicación impulsadas por la red EQUATOR y refrendadas por la junta internacional del consejo editor de revistas médicas (ICJME). Este acuerdo se basa en especialistas en el tema, expertos en campos complementarios, formando grupos eclécticos de clínicos, estadísticos, metodólogos, reguladores, financiadores, editores y, cuando ha sido posible, representantes de los pacientes. Con la ayuda del método Delphi, han consensuado el conjunto mínimo de puntos clave que, en su opinión, debe incluir el informe de resultados de cada tipo de estudio. A ser posible, descansan en estudios empíricos que identifican las características que facilitan replicar los resultados, impulsando así la *metainvestigación*, o investigación sobre la investigación.

Pretendemos ser concisos: el texto es denso. No pretenda leerlo de un tirón. Lo hemos estructurado en pequeñas unidades docentes o *píldoras* que pueden abordarse *independientemente*. Al terminar cada píldora, imagine una pregunta de investigación, o un artículo, y pregunte: "Este concepto, ¿cómo le afecta?" Abórdelo en grupos de trabajo que le ayudarán a 1) resolver sus propias dudas; 2) aprender de las de sus compañeros, y 3) ampliar su perspectiva.

Hemos agrupado las píldoras en 9 capítulos. El primero expone las bases metodológicas (replicación, medida, variable, causalidad, etc.); el segundo, las estadísticas (varianza, error de estimación, sesgo, regresión a la media, etc.); y el tercero, las de inferencia estadística (confianza, credibilidad, decisión de Neyman-Pearson, equivalencia, metaanálisis, etc.). En el cuarto continuamos

hablando del ensayo clínico ya que, contrariamente a la tradición pero en línea con las guías, sostenemos que el ensayo clínico facilita el aprendizaje de los restantes tipos de estudios. Además, sus resultados tienen mayor implicación para pacientes y disponen de guías más maduras (CONSORT, SPIRIT). En el quinto, explicamos las medidas del efecto de las intervenciones; en el sexto, los métodos para condicionar o bloquear terceras variables; en el séptimo explicamos el metaanálisis, que acumula la información de varios estudios; en el octavo, las extensiones de CONSORT a ensayos distintos del habitual (2 muestras *paralelas* con 2 tratamientos).

Y dejamos para el final los estudios observacionales, más habituales. Los etiológicos, núcleo de los epidemiológicos clásicos, avanzan en el conocimiento de las causas de las enfermedades (STROBE). El segundo gran grupo de observacionales aborda los indicadores, ya sean diagnósticos o pronósticos, distinguiendo si el objetivo es construir (TRIPOD) o validar (STARD) el indicador.

Enfocamos este texto introductorio a la interpretación. En consecuencia, no abordamos ningún paquete estadístico. Si necesita utilizar alguno, le recomendamos R porque: 1) es libre y gratuito; 2) es transparente; 3) es el favorito del mundo académico, y 4) dispone de ayudas, también libres y gratuitas.

La bibliografía importante está en EQUATOR, LIGHTS y Wikipedia. Enlazamos a unas pocas referencias. También a otros recursos docentes nuestros: aplicativos informáticos para practicar conceptos, vídeos explicativos y artículos divulgativos.

Agradecimientos: Este texto se basa en nuestra experiencia docente en asignaturas como Ensayos Clínicos, del Máster en Estadística e Investigación Operativa (UPC-UB); Estadística Médica, del Grado en Estadística (UB-UPC), y Probabilidad y Estadística, del Grado en Informática (UPC), así como en otras asignaturas oficiales que les precedieron. Citemos también nuestros cursos de divulgación científica, sobre todo para las industrias de fármacos (en especial, Novartis) y de dispositivos médicos (Medtronic). Estamos muy agradecidos a todas las generaciones de estudiantes que desde 1991 nos han discutido cada concepto. A los equipos de AptaTargets, empresa a la cual asesoramos, y a los de Ferrer y Almirall, con quienes hemos trabajado en el pasado.

Debemos mencionar a quienes nos ayudan a madurar semana a semana nuestros puntos de vista: los miembros del consejo editorial de Medicina Clínica, los compañeros del grupo de investigación GRBIO y los profesores de todas las asignaturas de nuestro departamento.

Claudia Agüero nos ha ayudado con las figuras y las tablas. Enrique Ventura dibujó las ilustraciones para dos monografías de la Fundación Esteve.

Contribuciones: En general, de la primera versión de los textos, EC, y de los aplicativos, JAG. Ambos somos responsables de la versión final.

Erik Cobo y José Antonio González
bioestadística.fme@upc.edu

Índice

1
Conceptos metodológicos

Empecemos con los aspectos metodológicos básicos para toda investigación en salud. Veremos, por ejemplo, las propiedades de la medida y el papel de las variables en un estudio. Comenzaremos por la característica esencial de la ciencia: la reproducibilidad de los resultados. Y terminaremos explicando cómo las guías de publicación pretenden facilitarla.

1.1. Reproducibilidad

El primer requisito científico es la reproducibilidad de los resultados: aquello que no logramos reproducir no tiene valor científico.

Ni clínico, ya que necesitamos reproducir estos resultados en sus futuros destinatarios.

Opinión: Los milagros existen: eventos poco probables pasan. Aunque, si no son reproducibles, no sirven para mejorar el futuro.

Un científico está obligado a reportar sus resultados de forma **reproducible**.

Así, si su investigación tiene éxito, sus destinatarios podrán beneficiarse de ella. Y, si no lo tiene, facilitará el trabajo de otros investigadores, que no reproducirán los mismos estudios, sin introducir ningún cambio.

En 2009, Paul Glasziou, junto con Iain Chalmers, fundador en 1993 de la colaboración Cochrane y en 2004 de la alianza James Lind, postularon que el 84% de la inversión en investigación en salud se pierde (literalmente, "*waste*"), ya que no termina en resultados reproducibles.

En efecto, dijeron que tiramos a la basura el 84% de la inversión en investigación sobre salud.

La reproducibilidad puede ser valorada por los mismos autores o por otros distintos, y en los mismos datos o en diferentes.

En el caso de los mismos autores y datos, puede demostrarse si los autores proporcionan acceso a los datos anonimizados, a los informes de monitorización y a los códigos de los programas de análisis. Diversas herramientas informáticas lo facilitan. Además, compartir los códigos permitirá a otros avanzar más rápido en sus propias investigaciones.

El **azar** y la **cobertura** de los intervalos de incertidumbre facilitan la reproducibilidad: sabemos cuantificar la incertidumbre introducida por el **azar y** podemos comprobar si los intervalos tienen la **cobertura** deseada.

La reproducibilidad implica otros requisitos, como la transparencia, la legibilidad y la concisión.

Transparencia: el informe de resultados debe incluir todos los detalles que permitan reproducirlos.

Legibilidad: se refiere a la facilidad para leer y comprender el informe.

Concisión: no debe incluir más detalles de los necesarios.

A todo ello ayudan las recomendaciones de las guías de publicación.

El gran requisito científico es la reproducibilidad.

1.2. Unidades

Denominamos individuos o **unidades** a los objetos de la investigación en que determinamos las **medidas** de interés.

Las unidades no tienen que ser necesariamente personas.

Pueden coexistir diferentes unidades, estructuradas de forma jerárquica o anidadas en niveles.

- Por **ejemplo**, las células dentro de los tejidos, los órganos dentro de los voluntarios, los pacientes dentro de los hospitales, etc. Para aportar información completamente nueva, las unidades deben ser **independientes** entre sí.

Contraejemplo: las medidas repetidas en la misma unidad comparten información. El análisis de la incertidumbre será sofisticado.

Contraejemplo: las grandes bases de datos (*big data*) presentan retos muy interesantes para su procesamiento, que la informática va afrontando. El reto es más exigente desde el punto de vista estadístico, ya que suelen contener abundante información repetida, irrelevante.

Broma. Y podrían no ser tan grandes.

Valora la posibilidad de recoger las unidades de manera independiente.

1.3. Medida

Definimos dos propiedades de la **medida**: la fiabilidad y la validez.

La **fiabilidad** valora el grado de repetibilidad de la determinación, es decir, la concordancia entre los valores obtenidos en el mismo individuo, en las mismas condiciones.

La **validez** valora la pertinencia de la medida, es decir, si realmente estamos valorando el objetivo que buscamos con la medida.

- Por **ejemplo**, la evaluación de un examen de problemas puede ser muy fiable, en el sentido de que distintos profesores pueden discrepar, supongamos, en muy pocas décimas. En cambio, puede argumentarse que no es una medida válida de la capacidad de un estudiante.

- Veamos otro **ejemplo**: la puntuación de un trabajo final de carrera podría ser menos fiable, con una discrepancia de puntos en lugar de décimas entre los evaluadores, aunque muchos podrían argumentar que es más válida, ya que informa mejor del rendimiento futuro del estudiante.

Hablamos en condicional (*podría*) porque siempre necesitamos datos: por ejemplo, medir la fiabilidad.

Como **ejercicio**, podríamos preguntar por la fiabilidad y la validez de los dos componentes de la evaluación de la selectividad: el examen y el rendimiento escolar previo.

Razona sobre la validez y la fiabilidad de tus determinaciones.

1.4. Papel de las variables

Denominamos **variables** a todas las determinaciones que recogemos en las unidades y que se caracterizan por tomar distintos **valores** en diferentes unidades.

Nota técnica. Matemáticamente, definimos las funciones de probabilidad acumulada con sumatorios para las variables **discretas** y con integrales para las **continuas**. Las segundas pueden tomar cualquier valor dentro de su rango, pero las primeras dan saltos entre los posibles valores.

Ejemplos de variables continuas son determinaciones físicas como el peso o la altura, que tienen unidad de medida, como el kilogramo o el metro. Y de las discretas, los recuentos, como el número de hermanos.

Nota técnica. Otra clasificación habla de las escalas de medida: **nominal**, **ordinal**, de **intervalo** y de **razón**, según si van añadiendo las propiedades de equivalencia, orden, unidad de medida y definición de cero absoluto.

Sus **ejemplos** respectivos serían: el género, la clase social, la temperatura en grados centígrados y el peso en kilogramos.

También cabe distinguir entre determinaciones físicas con unidad de medida, como el peso, y valoraciones clínicas, como una escala (Glasgow, Rankin, etc.).

Opinión. No nos gusta la etiqueta de variables objetivas y subjetivas, pues la consideramos pobre.

Más importante es tener claro el papel de cada variable en una investigación.

1. Denominamos variable respuesta Y (*outcome, output, endpoint*) el criterio de valoración del desenlace.

   Por **ejemplo**, valoramos el desenlace de un accidente cerebral (ictus) mediante la escala de Rankin modificada (mRS), que va de 0 (= sin síntomas residuales) a 6 (= deceso).

   Además, tenemos dos tipos de variables previas, que podrían ayudar a determinar esta respuesta:

2. Denominamos **covariables Z** aquellas condiciones cuyo valor no depende de nuestra voluntad (auténticas *variables aleatorias*), como la edad o el género.

3. Denominamos **intervención X** aquella variable cuyo valor depende de nuestra voluntad.

   En un diseño experimental, asignamos el valor de **X** a cada unidad. En un estudio observacional, el valor de todas las variables no depende de nosotros: todas son **Z**. Incluso potenciales **X** como las exposiciones, que no hemos asignado.

   Dawid y Senn utilizan el término *decisiones* para las **X**. En este texto, hablamos de *intervenciones* en los estudios experimentales y de *exposiciones*, en los observacionales.

Clasifica tus variables en: respuesta **Y**, intervención **X** y covariables **Z**.

1.5. Intervenciones

Los ensayos clínicos tradicionales estudian intervenciones farmacológicas. Representaron el 25% de los estudios publicados en 2000.

Además, tenemos otros tipos de intervenciones.

- **Ejemplos:** quirúrgicas, psicológicas, de acupuntura, de gestión, de fisioterapia, etc.

El principal reto **clínico** de las intervenciones no farmacológicas consiste en describirlas de forma reproducible. La guía TIDieR enumera los puntos clave para replicar una intervención sofisticada.

El reto **metodológico** consiste en enmascarar a los distintos investigadores. El intervencionista y el proveedor de los cuidados conocen el tratamiento que

deben aplicar en cada caso. Conviene incomunicarlos de otros investigadores, como el evaluador de la respuesta, para mantener oculta la intervención.

Y dos retos **estadísticos**: 1) modelar la variabilidad del efecto asociada a unos intervencionistas y centros que serán distintos en el futuro y 2) considerar la posible existencia de un efecto compartido (*clúster*) o contagiado, transmitido entre unidades, ya que resultados similares para pacientes vecinos podrían implicar un efecto de clúster y la necesidad de adaptar la muestra o los métodos estadísticos.

La extensión CONSORT para las intervenciones no farmacológicas contiene recomendaciones útiles:

1. Considerar a los intervencionistas como variables y reportar: i) qué criterios de selección tenían; ii) cómo fueron asignados a cada grupo en una nueva caja del flujograma, y iii) cómo eran los finalmente reclutados.

2. Reportar la adherencia de los intervencionistas y los participantes a la intervención.

3. Minimizar los lapsos de tiempo entre aleatorización e intervención, que podrían propiciar cambios del grupo asignado, el no cumplimiento del protocolo y pérdidas.

Consultar tanto la extensión CONSORT a las intervenciones no farmacológicas como la guía TIDieR.

1.6. Relación causal

Podemos hablar de relación causal o de causalidad. Asociación y relación son lo mismo. Pero causalidad es otro concepto: obtenemos cambios en la respuesta **Y** (efectos) mediante manipulaciones en **X** (intervenciones).

- Como primer **ejemplo** de **causalidad,** utilizamos el interruptor para apagar y encender la luz. Introducimos cambios voluntarios en el interruptor (*intervención* **X**) para inducir cambios en el estado de la luz (*respuesta* **Y**). Si el sistema funciona sin margen de error, la relación es *determinista* o matemática, en un sentido clásico del término.

- En un segundo **ejemplo** de causalidad, invirtiendo más horas de estudio (*intervención* **X**) podemos influir en la nota (*respuesta* **Y**), aunque ahora

existe cierta variabilidad restante en la nota, lo cual obliga a recurrir a la estadística para explicarel efecto de las horas de estudio en la nota.

Utilizaremos las relaciones causales para *cambiar* el futuro. Y cuantificamos las consecuencias de la intervención **X** mediante medidas del *efecto* en **Y**.

- En un primer **ejemplo** de **relación** no causal, en un centro de enseñanza primaria un maestro puede utilizar la altura (*covariable* **Z**) de una niña para aproximarse al curso al que va (*respuesta* **Y**). Decimos *aproximarse* porque la relación no es determinista: existe una relación entre el curso y la altura, aunque también mucha variabilidad dentro de un mismo curso. Y, por supuesto, la relación no es causal: alargando las piernas con prótesis no alcanzaremos el nivel que permita un cambio de curso.

Ejercicio. Valora los requisitos para considerar la *variable* **Z** altura como una *intervención* **X**.

- En un segundo **ejemplo**, un observador hábil puede usar el consumo de internet (*variable* **Z**) de un país como chivato de la mortalidad por cáncer (*respuesta* **Y**) en esa zona. Recuerda que internet y cáncer son más comunes en los países desarrollados. Y quizás sea un buen chivato, con una buena capacidad predictiva. Pero, una vez más, prohibir internet no tendría efecto alguno sobre el cáncer.

- En un tercer **ejemplo**, otro observador hábil puede utilizar el hecho de llevar mechero como chivato de la morbimortalidad (*respuesta* **Y**). Llevar mechero es un buen indicador de fumador. Y también podría ser un buen chivato sobre el futuro de esa persona: una vez más, conviene valorar la capacidad predictiva de llevar mechero para anticipar los años de vida que le quedan a nuestro sujeto.

Hernán y Robbins explican la diferencia entre relación y causalidad.

Hemos hablado de *chivatos* para dejar claro que no pretendemos interpretar causalmente: el objetivo no es intervenir.

Nota técnica. Si hay relación, los pacientes de la **población** tratada evolucionan de forma diferente de los de la población de referencia. Si la relación es causal, el conjunto de la **población** responde de forma diferente cuando se le asigna una u otra intervención.

Si hay asociación, siempre puedes utilizar una variable para anticipar el valor de otra (ver píldora siguiente). Sin embargo, intervenir sobre la primera solo cambiará el valor de la segunda si hay relación causal.

1.7. Predicción

Un objetivo distinto de investigación analiza si es posible anticipar los valores de la respuesta.

En meteorología, por **ejemplo**, deseamos anticipar el tiempo; no modificarlo.

Para seleccionar un método de previsión del tiempo, conviene conocer con qué precisión podemos anticiparlo: qué combinación (algoritmo o fórmula) de las variables reduce más la incertidumbre previa sobre el tiempo.

Avancemos un poco lo que veremos en el capítulo 9: cómo valorar la calidad de la predicción.

► Por **ejemplo**, deseo anticipar el peso de la próxima persona que se sentará en la silla de mi despacho. La puerta translúcida me informa de que esa persona mide 190 cm. Para cuantificar cómo esta información reduce mi incertidumbre sobre el peso, puedo comparar el de (A) **todas** las posibles personas que entrarán con el de (B) **las que miden 190 cm**. Supongamos que, en ese entorno, el peso corporal en el escenario A tiene una media de 70 kg y una desviación típica de 10 kg, y en el B una media de 90 kg, con una desviación típica de 7 kg. Entonces, cuando actualizo la predicción general de 70 kg, según A, a 90 kg, según B, se reduce el error (cuadrado) de predicción de 10^2 a 7^2 kg^2, o desciende la incertidumbre de 100 kg^2 a 49 kg^2: una reducción del 51%. [Recordaremos la definición de desviación típica en la píldora 2.1.]

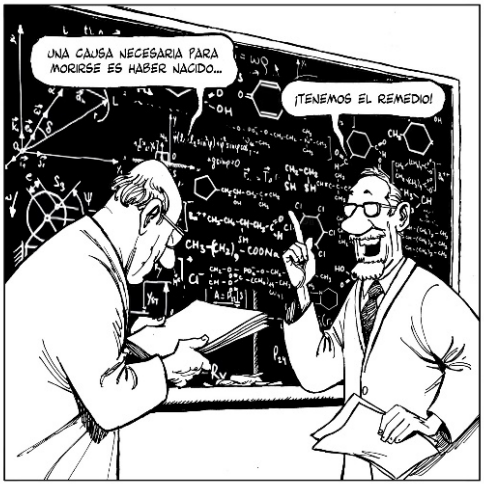
La predicción será más precisa cuanto más intensa sea la relación entre la variable predictora **Z** y la respuesta **Y**.

En la píldora 9.5, conoceremos cómo cuantificar la capacidad para reducir la incertidumbre. En la 9.4, veremos que la respuesta **Y** puede ser futura (pronóstico) o presente (diagnóstico).

Si el interés es predecir, valora la capacidad para reducir la incertidumbre.

1.8. Causas frente a efectos

Distinguimos entre los efectos de las causas y las causas de los efectos.



- Por **ejemplo**, no es lo mismo decir "¿Se irá el dolor si me tomo un analgésico?" que "¿Se fue el dolor porque me tomé un analgésico?"

Las preguntas sobre los efectos parten de una causa bien definida.

- Por **ejemplo**, el prospecto de una intervención farmacológica define la causa (algo así como "adminístrese oralmente en ayunas 1 vez al día").

En cambio, las preguntas sobre las causas parten de un efecto bien delimitado.

- Por **ejemplo**, bronquitis crónica, quizás definida como "tos productiva durante más de 3 meses al año".

Las causas reciben nombres distintos: 1) en farmacología, cirugía o fisioterapia, se llaman **intervenciones**, y 2) en epidemiología, **exposiciones**.

Las intervenciones pueden ser decisiones **X** del investigador; las exposiciones siempre son variables observadas **Z**.

Los pretendidos efectos de las intervenciones son positivos, por lo que podremos asignar la causa a los voluntarios respetando los principios éticos, y accediendo a las ventajas del diseño de experimentos. Como los efectos de las exposiciones suelen ser negativos, debemos fundamentarnos en estudios observacionales.

Las preguntas sobre los efectos miran al futuro, hacia adelante, mientras que las preguntas sobre las causas miran al pasado, hacia atrás. Ello marca una diferencia fundamental en sus métodos.

En este texto, siguiendo el consejo de STROBE, evitaremos el uso de los términos prospectivo y retrospectivo con más de un significado.

El estudio de las causas (ver píldora 9.1) siempre es *tentativo*; lanza *conjeturas*, **postula** nuevas ideas y requiere una redacción prudente, del tipo: "estos resultados sugieren".

Opinión. La etiología está muy asociada a nuestra cultura judeocristiana, cuya pregunta favorita es: "¿De quién es la culpa?"

Nota técnica. Interpretar como causal una relación observada requiere condiciones sobre las terceras variables. Las dos más simples son: 1) la suficiencia del modelo, o creer que hemos medido sin error y especificado bien en el modelo

todas las covariables **Z** con capacidad predictiva; y 2) la asignación al azar. Esta segunda la podemos garantizar con un buen diseño, pero la primera siempre quedará en mera suposición, premisa o asunción.

Si deseas interpretar causalmente una relación observada, lee atentamente la bibliografía sobre DAG (ver píldora 6.8) y sobre la inferencia causal. Y sé muy cauto: recuerda que es una interpretación.

Postula las causas y estima los efectos.

1.9. Hacer (*Do*) frente a ver (*See*)

Distingamos entre experimentar y observar.

Los estudios **observacionales** permiten valorar la capacidad de **predicción** de un **indicador** —ya sea un diagnóstico (presente) o un pronóstico (futuro). Pero dejemos de lado la predicción y centrémonos en la causalidad: 1) los estudios **experimentales** estiman **efectos** de **intervenciones**; 2) algunos **observacionales** tantean las **causas** (*etiología*) entre todas las **exposiciones** —que pueden ir juntas y estar relacionadas, conformando la denominada **maraña causal.**

Para estudiar las causas, los datos *reales* observados están amenazados por grandes retos, entre los cuales destacan la confusión de efectos y el sesgo de selección.

- Por **ejemplo**, si aprueban más los alumnos de cierto grupo, podría ser debido a su profesor y también porque les pone exámenes más fáciles. Aun si el examen fuera el mismo para todos los grupos, como los alumnos no se asignan al azar, sino que eligen grupo primero los que tienen mejores notas, este hecho podría ser una explicación alternativa al profesorado. Un experimento bien diseñado y ejecutado garantiza que la **única diferencia** entre los grupos es el tratamiento en estudio.

Nota técnica. La intervención es independiente de (*ortogonal a*) cualquier explicación alternativa.

Para afrontar los grandes retos de los estudios sobre las causas, el investigador debe mostrar su ingenio: **revertir** las causas negativas (**exposiciones**) en positivas (**intervenciones**), y así poder diseñar un experimento que pueda recibir la autorización del comité correspondiente.

En el diseño de un experimento, puedes convertir un atributo de las unidades (exposición Z) en una **causa asignable** (intervención **X**), que permitirá introducir el azar (*randomization* o asignación aleatoria).

► Por **ejemplo**, razones ético-legales no permiten asignar el tabaco a personas.

Broma. No podemos decir "le ha tocado el grupo 'fumador desde los 16 a los 52 años'".

Aunque sí podríamos obtener el permiso del comité responsable para asignar una estrategia para reducir el consumo del tabaco en los fumadores.

Por **ejemplo**, para estimar la magnitud de la brecha salarial por género en un entorno concreto, mostramos historiales a posibles empleadores y les preguntamos: "¿Cuánto pagaría a éste?", aunque previamente hemos asignado al azar el sexo a cada historial.

Además, el experimento permite documentar que no hay sesgo de selección, ya que hemos seguido todos los casos **desde** el momento **inicial** —o basal. El inicio viene marcado por el momento en que decidimos intervenir o predecir.

Finalmente, el experimento facilita observar si, al aconsejar una cierta intervención, los pacientes y los clínicos actúan según lo previsto. Antes de aprobar una intervención, una agencia de regulación requiere estimar sus efectos, positivos y negativos. Y también quiere datos sobre el grado de cumplimiento, por parte de los

clínicos y pacientes, conforme a lo previsto en el protocolo de administración de la intervención.

Usamos el operador *Do/See* para recordar la distinta interpretación de una relación extraída de una observación (*See*) o de un experimento (*Do*).

> Postulamos causas en las **observaciones** (*See*) y estimamos efectos en los **experimentos** (*Do*).

1.10. Guías de publicación

Como distintos objetivos requieren diferentes metodologías, los creadores de las guías optaron por subdividirlas. A continuación, relacionamos: 1) la **pregunta** del paciente, 2) el **objetivo** clínico, 3) el **diseño** metodológico y 4) la **guía** de publicación.

Pregunta	Objetivo	Diseño	Guía		
¿Qué me pasa?	Diagnóstico	Transversal	Construcción… Validación…	…del indicador	TRIPOD STARD
¿Qué me pasará?	Pronóstico	Cohortes			
¿Por qué me pasa?	Etiología	Caso-control	Resultados		STROBE
¿Puede decirme cómo mejorar mi futuro?	Tratamiento	Ensayo Clínico	Resultados Protocolo		CONSORT SPIRIT
Acumulando toda la información, ¿qué sabemos?		Revisión sistemática	Resultados		PRISMA

Tabla1.1. Guías de publicación según el objetivo y el diseño de estudio

El **diagnóstico** clínico responde una pregunta sobre el presente: "¿Qué tengo?" Su diseño habitual recibe el nombre de **transversal** (*cross-sectional*), porque no requiere el paso del tiempo. Los demás diseños se llaman **longitudinales**, porque precisan un lapso entre las variables. La pregunta **etiológica** mira al pasado, "¿Por qué me pasa?", mientras que las otras dos, al futuro, "¿Qué me pasará?", **pronóstico**, y "¿Puede alguien mejorar mi futuro?", **intervención**.

El diseño más habitual para estudiar el pronóstico es el de **cohortes**; para la etiología, el **caso-control**, y para la intervención, el **ensayo clínico**. Mientras que el

ensayo y la cohorte siguen el orden natural de los acontecimientos, el caso-control mira hacia atrás, va al revés, quizás recurriendo a una frágil memoria.

En diagnóstico y pronóstico, la guía para validar un indicador es STARD y para construirlo, TRIPOD. Para estudiar la etiología, STROBE. Para reportar los resultados de un estudio pivote de intervención, CONSORT, y para publicar su protocolo, SPIRIT.

CONSORT y SPIRIT están en fase de unificación y simplificación. En 2017, se inició un intento de actualizar STROBE (STRATOS). SPIROS, para protocolos de observacionales, está en fase de consenso.

CONSORT, publicada en 1996 y revisada en 2001 y 2010, es la guía más antigua, quizás porque, de todos los actos clínicos, la intervención (ya sea tratamiento o prevención) es el más importante.

Para combinar la evidencia disponible y obtener una estimación más precisa de los efectos de una intervención, tenemos PRISMA, desarrollada para ensayos con asignación al azar.

Las primeras guías venían en dos artículos: una declaración (*statement*) con la lista de comprobación (*checklist*) y una "explicación y elaboración" más extensa, que incluye ejemplos de buen informe. Son textos de unas 30 páginas, que explican lo más importante de un diseño. Actualmente, las guías se publican en un único artículo.

Consulta en EQUATOR la actualización de las guías.

Diferentes objetivos clínicos tienen distintos métodos y guías de publicación.

2
Conceptos estadísticos

Veamos ahora los conceptos estadísticos básicos. Empezaremos por la gran protagonista, la *variabilidad*, la cualidad de ser *variable*. Y su adaptación para extender los resultados concretos de un estudio a toda la población objetivo: el error esperado o típico del estimador. También las propiedades que debe tener este estimador para garantizar que avanzamos en la línea correcta. Terminaremos viendo el concepto de *regresión a la media*, que ha aturdido a grandes investigadores.

2.1. Variabilidad

La variabilidad no es ni buena ni mala o, al menos, no es nuestra responsabilidad determinarlo, sino enseñar a cuantificarla e interpretarla.

Lo único constante es que nada es constante.

Definimos la **varianza** como el valor esperado del cuadrado de la distancia a la media.

Entre dos individuos escogidos al azar de una población, su discrepancia al cuadrado coincide con el doble de la varianza.

Para facilitar la interpretación de la varianza, hacemos su raíz cuadrada.

La raíz de la varianza recibe el nombre de *desviación típica* (DT).

► **Ejemplo.** Mercè disfrutó de este mundo durante 91 años, y Oriol durante 77.
La disparidad entre el tiempo de vida de Mercè y de Oriol es de 14 años.
Tomando a Mercè y Oriol como si fuera la típica disparidad entre dos individuos al azar, 196 ($=14^2$) $años^2$ representarían una varianza de 98 $años^2$.
El cuadrado detrás de los años recuerda que la varianza es una medida de los cuadrados, difícil de interpretar, lo cual aconseja trabajar con la raíz.
Así, Mercè y Oriol nos dan una primera aproximación de la dispersión de la longevidad: = 10 años.

Opinión: A algunos nos parece que el nombre *desviación* invita a interpretar que quien alcanza la *extraordinaria* longevidad de 115 años es un gran *desviado*. El término *varianza* es neutro, pero *desviación* podría tener una connotación negativa, inapropiada en ciertas circunstancias. Deberíamos acordar un nombre más neutro. Para evitar esta discusión utilizaremos solo DT, sin aclarar si hablamos de desviación, dispersión, distancia, diversión o desigualdad… típica.

La T de *típica* indica que es la distancia esperada, en promedio, con el centro.

DT es una medida excelente para variables simétricas como la presión arterial, pero no para el sueldo, que es una medida claramente asimétrica: quizás el 99% de los asalariados cobran menos de 100.000 €/año, pero el 1% restante se dispara y puede multiplicar este valor.

Cuantifica la variabilidad. Si la variable es simétrica, utiliza la DT.

2.2. Estadística descriptiva

Conviene describir de forma resumida las características de los voluntarios reclutados, ya que reportarlas todas sería tan tedioso que terminaría por no informar de nada. Además, debemos comprobar que hemos abarcado todo el espectro deseado. A continuación, explicamos cómo interpretar los *estadísticos*, o indicadores que se utilizan habitualmente para resumir los resultados.

- Por **ejemplo**, la población objetivo podría referirse a pacientes de entre 18 y 85 años, pero los participantes reclutados podrían tener entre 55 y 75 años.

El documento explicativo de CONSORT proporciona el ejemplo de la tabla. Veamos qué significan estos números, en función de si resumen variables numéricas, como el colesterol LDL, o categóricas, como la presencia previa de hipertensión o diabetes.

	Telmisartán (n=2.954)	Placebo (n=2.972)
Edad (años)	66,9 (7,3)	66,9 (7,3)
Sexo (mujer)	1.280 (43,3%)	1.267 (42,6%)
…		
Colesterol LDL	3,02 (1,01)	3,03 (1,02)
Hipertensión	2.259 (76,5%)	2.269 (76,3%)
Diabetes	1.059 (35,8%)	1.059 (35,8%)

*Los datos son promedios (DT) o números (%)

Tabla 2.1. Características iniciales de los pacientes (adaptado de CONSORT 2010)

El dato más relevante es el *denominador*. ¿Cuántos voluntarios aportan información? Por ejemplo, la cabecera de la tabla informa de que se estudiaron 2.954 y 2.972 casos en cada grupo.

En las variables cuantitativas, el resumen habitual es la *media* o *promedio*, acompañado de alguna medida de dispersión para indicar en cuánto se suelen alejar de la media.

DT informa de cuánto dista un caso, en término medio, de la media.

Una regla útil en las variables simétricas consiste en sumar y restar (±) a la media el doble de la DT para ver dónde se sitúa la gran mayoría de estos pacientes.

- Por **ejemplo**, en el caso de la variable LDL, 3 ± 2, proporciona 5 y 1: la gran mayoría de los casos incluidos tenían el valor de LDL entre 1 y 5, aproximadamente.

Si, además de simétrica, la distribución de la variable es normal, entonces podemos concretar qué significa "gran mayoría": el 95% de los casos estarán dentro de una distancia igual a 1,96, en lugar de 2 veces la desviación típica.

Si las variables numéricas no son simétricas, este cálculo falla.

Una solución consiste en recurrir al uso de *estadísticos robustos*, como la mediana y la desviación intercuartil.

La mediana es el valor del caso "central', que es el que se halla justo en el centro cuando se ordenan todos los valores. Coincide con el percentil 50, P_{50}, el que deja el 50% de las observaciones a cada lado.

La desviación intercuartil es la distancia entre P_{75} y P_{25}. En lugar de la distancia intercuartil, también pueden reportarse los valores de los percentiles P_{25} y P_{75}.

Otra solución consiste en *transformar la variable* para convertirla en simétrica, quizás utilizando logaritmos.

- Por **ejemplo**, el pH es el logaritmo decimal de la concentración de hidrogeniones. Parece muy sofisticado, pero lo utilizamos sin problemas para calcular la cantidad de ácidos o bases que debemos añadir para recuperar los valores recomendables.

Dicotomías como la "historia previa de hipertensión" se resumen reportando: 1) el numerador, 2) el denominador y 3) su porcentaje.

- Por **ejemplo**, 2.259 de los 2.954 pacientes tratados (el 76,5%) tenían antecedentes documentados de hipertensión.

Si el estudio no tiene pérdidas ni valores ausentes, el denominador será común para todas las variables y puede ponerse en la cabecera, arriba de todo, como en la tabla anterior.

2.3. Inferencia

Entendemos por *inferencia* el proceso de trasladar las estimaciones de las muestras a toda la población. Lo abordamos con mayor profundidad en el capítulo 3 (Evidencia); aquí introducimos su necesidad.

Diferentes muestras proporcionan distintos resultados: son irrepetibles.

La inferencia estadística busca la repetibilidad.

- **Ejemplo:** Listamos del 1 al 50.000 todos los pacientes atendidos en una zona de atención primaria y generamos una secuencia de 50 números al azar entre 1 y 50.000. A continuación, recogemos la presión arterial de

esos 50 pacientes, de todos ellos sin excepción. Supongamos que 12 de ellos la tienen alta: un 24% de hipertensos es la estimación puntual con esta muestra.

La inferencia estadística explica qué dice esta muestra de 50 casos sobre los 50.000.

Si la muestra es aleatoria, la inferencia estadística permite conocer el error introducido por el proceso de muestreo.

► Por **ejemplo**, las encuestas electorales de los organismos oficiales disponen de una lista de todos los potenciales electores y generan una muestra al azar.

Si todos los encuestados responden sin mentir, la teoría estadística permite conocer el error esperado y, mediante procesos que explicamos en el capítulo 3, estimar la proporción que votará a cada candidato.

La *cocina* de las encuestas busca soluciones si no se cumple la premisa "*todos responden sin mentir*".

El proceso de inferencia estadística parte de una muestra aleatoria.

2.4. Distribución de los estimadores

Las estimaciones puntuales fallan, *yerran*.

Los valores del estadístico procedente de una *muestra aleatoria* variarán en función de la muestra, y se distribuirán de alguna forma particular. Por ejemplo, no es igual la distribución de la media muestral que la de la mediana muestral. La teoría estadística define la distribución del estadístico (p. ej., la media) bajo distintas suposiciones, siendo la más habitual asumir que la variable objeto de estudio (p. ej., el peso) sigue la distribución normal definida por Gauss-Laplace, lo cual implica que la media seguirá también una distribución normal, aunque caracterizada por sus propios parámetros de posición y dispersión.

Imaginemos la distribución de todos los posibles resultados del estimador de cierto parámetro. Representamos con Θ el valor desconocido del parámetro de interés y con $\hat{\Theta}$ el valor del estimador obtenido a partir de cierta muestra. Con esta muestra, el error es la diferencia entre Θ y $\hat{\Theta}$.

Representamos por $E(\hat{\Theta})$ el promedio de los posibles valores del estimador, cuando provienen de todas las muestras posibles, dado que conocemos la distribución del estimador.

Definimos el **sesgo** (*SG*) como la diferencia entre el promedio de todos los posibles valores del estimador y el parámetro de interés Θ: $SG = E(\hat{\Theta}) - \Theta$

Calculamos la varianza (*V*) del estimador como el promedio de las distancias cuadradas entre los posibles valores del estimador $\hat{\Theta}$ y su centro:

$$V = V(\hat{\Theta}) = E\left[(\hat{\Theta} - E(\hat{\Theta}))^2\right]$$

Conocemos la raíz de *V* como *error típico* (ET) y representa la imprecisión del estimador, su oscilación aleatoria.

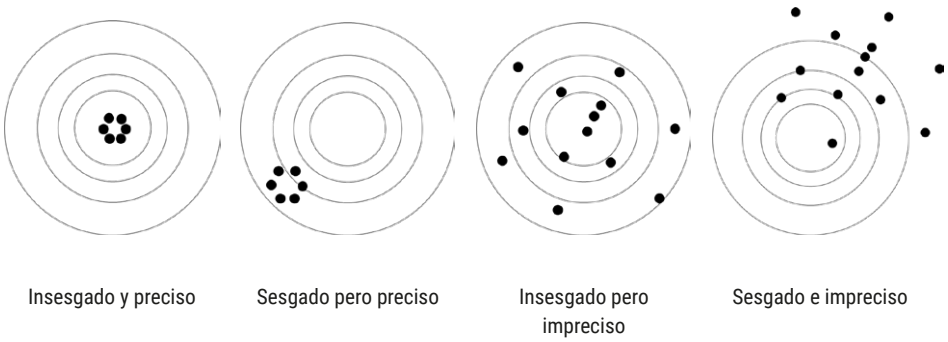


Figura 2.1. Cuatro tiradores, según su sesgo y su precisión

Entonces, la suma de cuadrados del error se descompone en el sesgo al cuadrado más la varianza del estimador: $SCE = SG^2 + V = SG^2 + ET^2$

Así, con este modelo, en muestras aleatorias obtenemos dos fuentes de error: el sesgo o *error sistemático* y la imprecisión aleatoria o *error típico*.

Un estimador obtenido en una muestra aleatoria contiene dos fuentes de error: el sistemático o sesgo y la imprecisión o error típico.

2.4.1. Imprecisión del estimador

Hemos visto que las estimaciones puntuales fallan, *yerran*, y que, si las muestras son aleatorias, podemos calcular el error en cada posible muestra, promediarlos y decir cuál será la imprecisión, el *error esperado* del conjunto de todas las muestras aleatorias de un tamaño concreto.

► **Ejemplo:** Supongamos que la población era infinita, en lugar de 50.000 pacientes. El error típico de las muestras de 50 para una proporción vale $\sqrt{P \cdot Q / n} = 0{,}060399$.

Si consideramos que la población no era infinita, sino de 50.000, el error típico vale ahora $\sqrt{(P \cdot Q / n) \cdot (N - n) / (N - 1)} = 0{,}060369$ eso es, es casi idéntico.

Debido a esta pequeña diferencia, suele considerarse la población infinita.

El *error típico* valora el error que tiene un estadístico (por ejemplo, una proporción o una media.) obtenido de una muestra aleatoria concreta.

La **DT** describe la **d**ispersión de los **d**atos —por ejemplo, de la presión arterial sistólica.

El **e**rror típico **ET** estima el **e**rror **e**sperado de un **e**stadístico —por ejemplo, de la proporción de hipertensos.

El *error típico* disminuye cuando aumenta el tamaño de la muestra al azar. En función de la magnitud del error que se puede tolerar debe fijarse el tamaño de la muestra aleatoria.

No hay fórmulas para el error en las muestras no aleatorias —por ejemplo, de conveniencia.

Puedes calcular el error esperado de una muestra aleatoria.

2.4.2. Sesgo

El sesgo es un *error sistemático*, siempre en el mismo sentido.

A diferencia de la imprecisión, que es distinta para cada muestra y hemos de recurrir al error típico para calcular su promedio o error esperado, el sesgo es una constante, un solo valor, común a todas las posibles muestras aleatorias.

- Por **ejemplo**, si estimamos la varianza de las observaciones dividiendo por n en lugar de $n-1$, tenemos un sesgo igual a $n/(n-1)$, común a todas las muestras. En este ejemplo, como el sesgo es conocido, lo podemos corregir dividiendo por '$n-1$', en lugar de n.

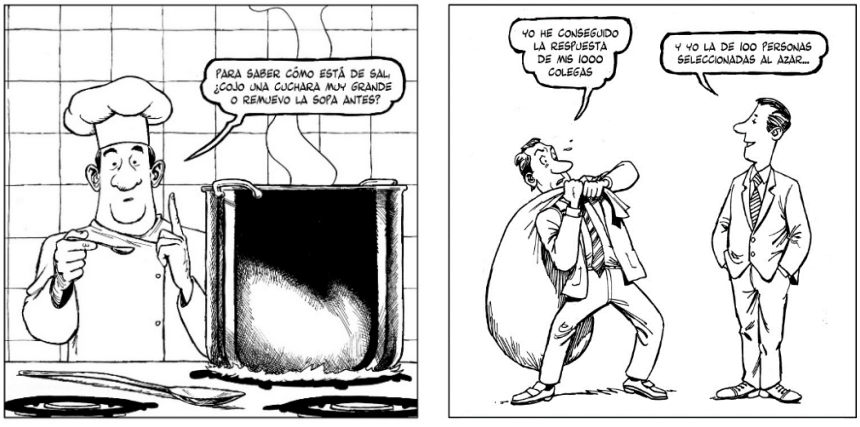
Utilizamos el sesgo y el error típico para escoger los mejores estimadores en muestras aleatorias simples.

Las muestras no aleatorias carecen de cálculo formal de su error y de su imprecisión, pero algunos estudios de simulación con datos reales muestran que la oscilación del estadístico es mayor y más errática que en las muestras aleatorias y que el ajuste estadístico no la contrarresta.

Si con datos reales (muestras no aleatorias, acaso de conveniencia) calculamos el error típico como si fueran muestras aleatorias, cometeremos un sesgo llamado *impredecible*, que indica que la oscilación del estimador está infraestimada: el error típico calculado como si las muestras fueran aleatorias es menor que el real. Por tanto, las estimaciones principales —por ejemplo, del efecto— tienen una oscilación mayor, que no se puede corregir, ya que puede ir en cualquier sentido.

El sesgo impredecible indica que los resultados oscilan más de lo que cuantifican las medidas estadísticas —que suponen muestras aleatorias.

- Por **ejemplo**, una encuesta electoral con una muestra de conveniencia puede terminar diciendo: "El candidato A obtendrá el 25% de los votos", pero *no existe ningún método* para calcular la confianza en esta *muestra de conveniencia*.



Las muestras no aleatorias no permiten cálculos de incertidumbre.

2.5. Riesgo de sesgo

La calidad de un estudio puede ser la máxima posible hoy en día y no por ello estar libre de la sospecha de riesgo de sesgo (*risk of bias,* RoB).

- Por **ejemplo**, resulta imposible enmascarar a un cirujano —el cual debe conocer la intervención que llevará a cabo. Al no estar enmascarado, este investigador podría revelar la intervención de un paciente a otros investigadores, lo cual abriría la posibilidad de influir en la evaluación y, quizás, de sesgar los resultados.

Que haya la posibilidad de sesgo no implica que exista sesgo. Las guías de publicación recomiendan hablar de *riesgo de sesgo*, en lugar de *calidad del estudio* o *sesgo* a secas. El RoB debe ponerse de manifiesto en la discusión, ya que amenaza la replicabilidad de los resultados. Si otros autores resaltan el RoB de un estudio o no logran replicar los resultados, los autores del estudio quedarán mal, salvo que lo hayan mencionado en la discusión.

Iremos explicando los riesgos de sesgo en los diferentes apartados. La buena noticia es que son bastantes menos que los sesgos contenidos en los largos listados de sesgos, como el *Catalogue of Bias*. Este catálogo incluye algunos con el nombre de quien lo describió primero en una disciplina concreta, lo cual induce a multiplicidades.

Ya hemos introducido 1) la *confusión de efectos* y 2) el *sesgo impredecible.* En la píldora 6.7, hablaremos 3) del *sesgo de selección* y ahora introducimos un cajón de sastre que incluye 4) las diferentes posibilidades de *hacer mal* un estudio.

1. El primer caso se conoce como el *sesgo por atrición* de la muestra, esto es, perder casos. Un ejemplo: por la seguridad del paciente, puede argumentarse que no es aconsejable seguir administrando un cierto tratamiento. Adelante: los pacientes pueden *abandonar el tratamiento* en estudio siempre, pero ello no significa que deban *abandonar el estudio*, ya que aportan información valiosa para futuros usuarios. Igualmente, los pacientes fallecidos generan valores ausentes (*missing*) en las variables del seguimiento posterior a su deceso. Una vez más, esta circunstancia no significa que hayan abandonado el seguimiento del estudio.

   Consideremos el fallecimiento en la definición de la variable respuesta.

- **Ejemplo:** La escala de Rankin es una valoración ordinal de la evolución tras un ictus, que oscila de 0 (sin síntomas) a 5 (estado vegetativo).

Actualmente, se utiliza una versión ampliada (*modified Rankin Score* o mRS), que incluye el valor 6 (deceso).
Evitamos el sesgo por atrición consiguiendo el valor de la respuesta para cada caso.

Nota técnica: Una sofisticada clasificación estadística categoriza los valores ausentes en 1) aleatorios, 2) completamente aleatorios y 3) no ignorables. Desgraciadamente, esta clasificación teórica no se puede aplicar en la práctica, ya que depende de variables no observables, de modo que la auténtica solución para los valores ausentes consiste en no tenerlos. (Disculpas: es más fácil decirlo que conseguirlo)
El *flujograma* introducido por la guía CONSORT, como veremos en la píldora 4.4.1, ilustra el seguimiento de los casos: ¿Cuántos se han perdido?

Broma. Un editor perezoso mira el flujograma para valorar la calidad del artículo de un ensayo clínico.

2. Un segundo caso de hacerlo mal es el *sesgo del informe selectivo* (*selective outcome reporting*), que consiste en presentar aquellos estadísticos que apoyan las conclusiones deseadas. Una forma ingenuamente habitual de dirigir un informe hacia las conclusiones deseadas consiste en escoger la variable y el análisis que más convienen.

   Broma. Busca "*spin research*" y obtendrás múltiples ejemplos.

   El sesgo del informe selectivo se previene siguiendo fielmente el análisis especificado antes de observar los resultados (esto es, de forma "independiente" a los resultados).

El protocolo del estudio debe explicar la finalidad, la metodología y las razones por las cuales el análisis elegido responderá a los objetivos. Y suele dejar los detalles técnicos, que también podrían influir en los resultados, para el plan de análisis estadístico (*statistical analysis plan* o SAP). Este SAP también debe detallarse con independencia de los resultados. Ello puede garantizarse suscribiendo el SAP en una fecha anterior a la de desvelar el tratamiento asignado a cada caso (conocido como "romper el ciego").

Tenemos 4 grandes tipos de riesgo de sesgo: la confusión, la selección, el sesgo impredecible y cualquier forma de desviación del plan previo.

2.6. Regresión a la media

Los valores extremos observados en una primera determinación no suelen repetir posteriormente.

Dicho de otra manera: en dos evaluaciones separadas un cierto tiempo, los valores que han destacado (por arriba o por abajo) en la primera no suelen presentar valores tan extremos en la segunda.

► **Ejemplo:** Francis Galton, al estudiar los rasgos hereditarios, observó que existía una relación entre la altura de los progenitores y la de sus descendientes, como se muestra en la figura siguiente.

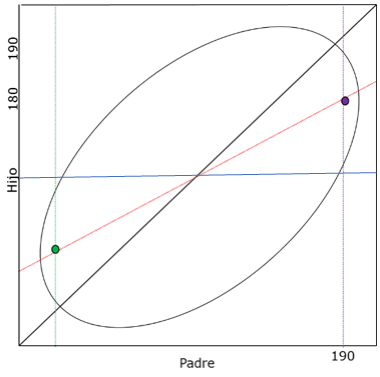


Figura 2.2. Altura de hijos (ordenadas) y padres (abscisas)

La línea negra diagonal marca la equivalencia de altura entre padres e hijos: la denominamos *recta identidad*. Por debajo, se encuentran los casos en que el padre era más alto y, por encima, aquellos en que el hijo era más alto. Cuando Galton se centraba en estudiar aquellas parejas en

que los padres eran muy altos (como los marcados con la línea vertical de la derecha), encontró que sus hijos, por término medio, eran más bajos: se alejaban de la recta *identidad* (línea negra) y se acercaban a la media de la altura de los hijos (línea horizontal). Además, en el otro extremo, los padres bajos (línea vertical de la izquierda) tuvieron hijos que también estaban más cerca de la media y ahora eran más altos. La línea roja marca la recta de regresión obtenida minimizando la distancia vertical con la recta (con el método tradicional de los mínimos cuadrados) y muestra que la magnitud de este fenómeno es mayor en los extremos.

Galton *interpretó erróneamente* que la variabilidad disminuía y habló de "regreso a la mediocridad".

En efecto, es cierto que, condicionada por la altura del padre, la altura predicha del hijo se acerca a la media de la altura de los hijos. Pero es una lectura equivocada: en sucesivas generaciones las alturas no irán confinándose en un margen cada vez más estrecho. Las medias y las dispersiones de ambas alturas coinciden en la figura, porque los descendientes de estaturas medianas suelen irse a los extremos. La variabilidad es *constante*: no ha cambiado con el tiempo.

Para que aparezca regresión a la media, es preciso que el valor observado incluya un componente aleatorio "independiente" en ambas determinaciones. Cuanto mayor sea este componente aleatorio, menor será la correlación entre ambas medidas y mayor será el fenómeno de regresión a la media.

Historieta. El curandero Asclepio conoce bien las dolencias crónicas de su conciudadano Alejandro. Sabe que son muy variables y alternan épocas buenas y malas. También sabe que Alejandro acude a él cuando ("condicionado") se halla en una de sus épocas malas —a la cual le seguirá, algún día, una buena. Por ello, le receta algo inofensivo y le pide paciencia hasta que surta efecto.

En la siguiente crisis, cuando Alejandro vuelve a acudir a Asclepio, este coge una pócima de otro color, también inofensiva, y le dice: "Bueno, la vieja pócima hizo su efecto un tiempo; ahora deberemos cambiar a otra nueva, más fuerte."

Nota técnica. Alejandro confunde la evolución natural, que es oscilante, con el efecto de la pócima porque no dispone de un control y no puede saber cuál habría sido su evolución si no hubiera tomado la pócima. Este reto es conocido en lógica como "el problema fundamental de la inferencia causal" (ver píldora 4.2). Como indica su nombre, es irresoluble a nivel individual y requiere premisas adicionales a nivel poblacional.

A menor variabilidad del componente aleatorio, menor efecto de regresión a la media: la regresión a la media es menos marcada cuanto más se aproxima a 1 la correlación entre ambas variables.

Nos hemos centrado en comentar la variabilidad y ver cómo evolucionan los casos extremos. Estos casos extremos "regresan al centro" y son compensados por otros, que se alejan de él: hijos de padres no tan extremos que tienen, ellos sí, una altura que les hace "destacar".

Si invertimos los papeles de ambas alturas, observamos el mismo fenómeno, pero al revés: los padres de hijos extremos tienen, por término medio, alturas más cercanas a la media.

Nota técnica. Supongamos que repetimos la medición Y de la presión diastólica (PAD) en dos ocasiones independientes, Y_1 e Y_2. Definimos un punto A para seleccionar a pacientes con PAD alta, es decir, por encima de la media de esta primera determinación: $A > E(Y_1)$. Entonces, asumiendo distribución normal bivariante con varianzas iguales, el valor esperado del cambio, o diferencia entre la segunda determinación y la primera, vale:

$$E[(Y_2-Y_1) \mid Y_1= A] = (A-E(Y_1))(\rho-1)$$

- **Ejemplo**: Supongamos que la PAD tiene una media de 80 y una DT de 10 mmHg; si la correlación entre dos determinaciones vale, por ejemplo, 0,75 y seleccionamos a los pacientes con PAD igual a 100 mmHg en la primera determinación, ¿cuánto cabe esperar que baje en la segunda determinación?

  $$E[(Y_2-Y_1)|Y_1 = 100] = (100-80)\ (0{,}75-1) = 20\ (-0.25) = -5$$

  En voluntarios con PAD = 100 en la primera determinación, el valor esperado en la próxima medición será 95, 5 mmHg menos.

Ejercicio

Discute si la regresión a la media puede explicar fenómenos como el "efecto placebo" o la "hipertensión de bata blanca". Vídeo explicativo de la regresión a la media.

La regresión a la media será mayor cuanto:
1. mayor sea la distancia entre el punto de corte y la media, y
2. menor sea la fiabilidad de la variable objeto de estudio.

3
Evidencia

En este capítulo, exponemos el punto de vista de las guías consensuadas de publicación sobre la inferencia estadística, proceso de inducción que aplica a toda la población los valores del estadístico obtenido en una muestra aleatoria concreta.

Opinión. Deseamos un futuro en que definamos la *inferencia* como el proceso que valora la reproducibilidad de los resultados.

Broma. Evidente era lo que no requería evidencia. Hoy ya nada es evidente.

3.1. Información aportada: la confianza

Explicamos aquí los intervalos de confianza clásicos que reflejan la **incertidumbre** remanente al terminar un estudio aleatorizado.

- **Ejemplo**. La guía CONSORT, en su punto 17, pide comunicar el tamaño del efecto (*effect size*) junto con alguna medida de incertidumbre, como el intervalo de confianza del 95% ($IC_{95\%}$). Otras guías piden intervalos similares para sus medidas más relevantes.
  En un ensayo, el intervalo de confianza refleja la incertidumbre originada por el proceso de aleatorización. Esta asignación al azar de las intervenciones en comparación genera unos resultados que habrían sido diferentes si la asignación de las intervenciones a los casos hubiera sido distinta. Los intervalos proporcionan los valores poblacionales del efecto compatibles con los resultados observados.

- **Ejemplo**. Una encuesta oficial de intención de voto, a partir del listado de los posibles electores (población objetivo), genera una muestra aleatoria. Y, respetando sus derechos, se organiza para conseguir la intención de voto de toda la muestra. A partir de sus respuestas en la muestra, obtiene

la proporción estimada de votos de cada candidato, junto con su estimación por intervalo (p. ej., $IC_{95\%}$).

- **Ejemplo**. Un ensayo valora el descenso de la mortalidad de un nuevo tratamiento. El análisis principal es el cociente, en los dos grupos de tratamiento, de las probabilidades de morir durante el seguimiento. Los resultados estiman puntualmente que este cociente vale 0,61: el nuevo tratamiento multiplica por 0,61 la mortalidad del grupo de referencia. Por tanto, la reduce un 39% (= 100%–61%). Y proporcionan un $IC_{95\%}$ que va de 0,33 a 1,11: los resultados observados son **compatibles** con que la nueva intervención multiplica la probabilidad de morir por un valor entre 0,33 y 1,11. Creemos que, si hubiéramos aplicado el tratamiento a toda la población, la mortalidad habría sido entre 0,33 y 1,11 veces la que habríamos observado si hubiéramos seguido el tratamiento de referencia. Los resultados son compatibles tanto con una reducción de la mortalidad a la tercera parte (0,33) como con un incremento del 11% (1,11).

Podríamos decir, por **analogía**, que en un ensayo el efecto poblacional es una verdad oculta y que los efectos observables en las muestras son mentiras a la vista.

El intervalo refleja los valores del parámetro probabilísticamente compatibles con el valor observado.

Debemos valorar si el intervalo aporta información, nuevo conocimiento.

- **Ejemplo.** Un estudio aleatorizado quiere valorar si una nueva intervención aumenta la probabilidad de curarse. El intervalo del cociente de probabilidades va de 0,2 a 10. En su extremo superior, diríamos que los resultados del estudio son compatibles con que la nueva intervención multiplica por 10 las probabilidades de sanar. Y, en su extremo inferior, que las divide por 5 (0,2 = 1/5). Los resultados observados son compatibles con que la nueva intervención tanto podría beneficiar (10) como perjudicar (0,2) a los pacientes. Dos extremos opuestos: podríamos interpretar que la lección aprendida es nula y etiquetar el estudio como anodino, poco informativo.

- **Ejemplo.** El $IC_{95\%}$ del cociente de las probabilidades de curar va de 20 a 30. Parece razonable decir que sí, que algo hemos aprendido, ya que el efecto observado sería relevante, tanto si fuera tan grande como 30 como si fuera tan grande como 20. Sin embargo, si un estudio previo había establecido que esa intervención podía incrementar la probabilidad de curar entre 22 y 26 veces, añadir ahora que podría subirla entre 20 y 30 veces parece poca

aportación. O quizás todo lo contrario, ya que nos está recordando lo más importante en la ciencia: que se han reproducido los resultados previos.

El IC requiere un juicio de valor final que, en un hipotético manuscrito, el propio investigador debería sostener en su discusión.

Nota técnica. El documento explicativo de STROBE, en su ítem 10, resume que el $IC_{95\%}$ aporta al lector la **precisión estadística**, que toma en consideración la incertidumbre introducida por el método aleatorio de muestreo. Y, a continuación, añade que todos los estudios pueden tener **incertidumbres adicionales** que al menos deben comentarse en la discusión. Además, en el ítem 20 de la discusión, añade que la incertidumbre global siempre será mayor.

Si no existe un proceso aleatorio en la generación de los datos, o si el estudio tiene alguna impureza, cualquier riesgo de sesgo, admite cautamente las limitaciones presentes. Y recuerda que la incertidumbre será mayor que la reflejada por el intervalo.

En **resumen**, interpreta los intervalos de confianza como informativos de 1) los valores poblacionales compatibles con los resultados observados y 2) la magnitud de la incertidumbre restante tras el estudio.

3.2. Neyman-Pearson, un pivote para decidir

Si los resultados son variables, un buen sistema de decisión debe contemplar la posibilidad de cometer errores. Jerzy Neyman y Egon Pearson propusieron un sistema (NP) que acota los dos posibles riesgos de error en una decisión binaria. Así, el sistema NP acepta 2 posibles errores, llamados de *tipo I* (permitir una intervención con eficacia cero, nula) y de *tipo II* (no permitirla, si tuviera eficacia Δ , "delta").

	Decisión D^C No autorizar	**Decisión D** Autorizar
Escenario de efecto nulo El fármaco no tiene efecto. H_0: μ_A-μ_B=0		Error de tipo I Riesgo α
Escenario de efecto alternativo Δ El fármaco tiene un efecto delta H_1: μ_A-μ_B=Δ	Error de tipo II Riesgo β	

Tabla 3.1. Riesgos y errores en la decisión de NP

La decisión de autorizar un fármaco nuevo requiere verificar que mejora la evolución de los pacientes.

Se basa en estudios aleatorizados para establecer la **superioridad** del fármaco nuevo frente al de referencia. Y se plantean dos escenarios simples: **nulo** (la diferencia "real" en sus efectos es exactamente **0**) o **alternativo** (esta diferencia vale exactamente el valor **delta** Δ).

Adjetivos usuales para describir este ensayo son *decisorio*, *final*, *confirmatorio*, *de fase III*, etc. Este ensayo pivote requiere establecer previamente:

1. la población objetivo (target) con sus criterios de elegibilidad;
2. la **respuesta** principal;
3. el **análisis** estadístico principal, en el cual se basará la decisión;
4. los riesgos aceptados, quizás α = 0,05 y β = 0,10, de cometer los errores I y II, y
5. el valor **delta** Δ, para el cual deseamos garantizar la potencia (*power*) 1-β deseada.

A partir de estos valores, calculamos el tamaño del estudio ($n_R + n_T = N$) para poder acotar los riesgos α y β.

Respuesta y análisis principal están ambos en singular, para controlar la multiplicidad o el consumo de α (ver píldora 4.3, más adelante).

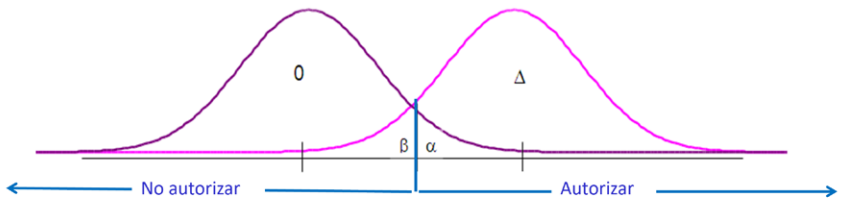


Figura 3.1. Aplicación del proceso de NP a la decisión de autorizar un fármaco

La figura 3.1 muestra la distribución del estadístico —típicamente, la diferencia de medias muestrales— en estos dos escenarios que se confrontan: 1) el efecto es nulo (curva de la izquierda) y 2) el efecto es delta (la de la derecha).

Mientras los escenarios previstos son solo dos, el resultado del estudio puede tomar muchos valores. Así, el estadístico que resume los resultados del estudio puede tomar cualquier valor. Pero, al final, la decisión es binaria (hacer D o hacer D^C), lo cual requiere un umbral, un punto de corte sobre el que pivota la decisión.

La línea vertical central marca el umbral que guiará la decisión. Si el resultado se sitúa a su derecha, actuaremos como si procedieran de la distribución con efecto delta y autorizaremos la intervención. Si el resultado se sitúa a la izquierda del umbral, actuaremos como si procediera del escenario de efecto nulo y no la autorizaremos. Cuando hay algún solape entre ambas distribuciones, la decisión no puede ser siempre correcta.

Nota técnica. Esta decisión binaria implica riesgos **unilaterales**.

Así, una pequeña proporción alfa = α de ensayos sobre intervenciones no eficaces terminará con autorización. Y, de forma similar, otra pequeña proporción beta = β de los realizados sobre intervenciones con el efecto hipotético delta, sin autorización. En la mayoría de las ocasiones, adoptaremos la decisión correcta. Hemos **controlado** (o acotado) los riesgos de tomar decisiones erróneas bajo estos dos simples escenarios. El complementario de beta es la potencia o probabilidad de autorizar una intervención con el efecto delta.

Nota técnica. Alfa y beta resumen las **características operativas** del diseño en un ensayo de muestra fija —es decir, no adaptativo (ver píldora 8.3).

Es ilustrativo, aunque simple, decir que α es el riesgo del ciudadano y β, el del promotor.

Aumentar el tamaño muestral n llevará a curvas más puntiagudas, con menor solapamiento: podemos escoger el tamaño para que los riesgos α y β sean los deseados.

Nota técnica. Esta es una de las posibles fórmulas que **NO** debes recordar:

$$n=[2\sigma^2(Z1-_{\alpha}-Z_{\beta})^2]/\Delta^2$$

Pero sí que las letras griegas representan valores previamente establecidos (disponemos de argumentos o de **evidencia** para ese valor concreto).

Nota técnica. Un detalle importante es que las letras griegas representan constantes, valores conocidos de antemano. No olvides que estás en un estudio confirmatorio, fundamentado en la evidencia de estudios previos. Hay también una letra latina, la *n*, pero esta representa el resultado del cálculo, el número de participantes por grupo.

El diseño de experimentos ayuda a controlar σ. Disminuir σ es una alternativa para conseguir la potencia necesaria sin aumentar el número de voluntarios.

Opinión. Este sistema de decisión es pobre, ya que no considera ninguna *función de pérdida* que otorgue un valor numérico (*utilidades*) a las consecuencias de tomar decisiones erróneas.

ALERTA. No debemos confundir el riesgo alfa con el valor P (ver píldora 3.5), origen de inagotables discusiones entre científicos, quizás por su frecuente mala interpretación.

> Para calcular *n* bajo NP, es preciso prestablecer el análisis y la variable principales, los valores del efecto Δ, la dispersión σ y los riesgos α y β.

3.3. Equivalencia

Veamos ahora el objetivo de establecer que dos tratamientos son similares.

La **equivalencia** incluye la igualdad absoluta y también las diferencias irrelevantes.

De forma similar, definimos los conceptos **unilaterales** de **no inferioridad** y **no superioridad**.

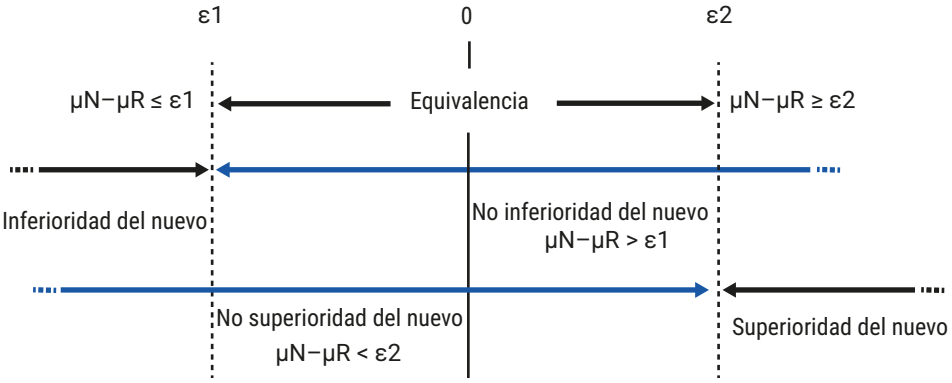


Figura 3.2. Equivalencia, no inferioridad, no superioridad y margen ε de irrelevancia

El gráfico muestra que el concepto de equivalencia es más amplio que el de estricta igualdad.

- Por **ejemplo**, consideramos que dos hipotensores son **equivalentes** si ninguno gana al otro en más 10 mmHg. Y consideramos que un nuevo hipotensor es **no inferior** si el clásico no le gana en más de 10 mmHg.

Para demostrar equivalencia, el contraste de NP sitúa en la hipótesis alternativa lo que queremos demostrar: la equivalencia. También podemos afirmar que existe equivalencia si el IC está completamente comprendido entre —ε y +ε.

- **Ejemplo**. Si definimos el intervalo, simétrico, de equivalencia con ε, entonces |μN-μR| < ε, y declaramos que existe equivalencia si los extremos del IC de la diferencia de medias no superan ε ni por un lado ni por el otro.

- **Ejemplo**. La figura 3.3 muestra equivalencia en los estudios 1 a 3, pero no en los estudios 4 a 6. Según el planteamiento clásico de diferencias, los resultados son distintos: 1, 4 y 5 muestran diferencias, pero 2, 3 y 6, no.

Nota técnica. El estudio 1 es informativo, con un IC preciso que permite establecer equivalencia y también diferencias. Mediante IC no hay contradicción: aunque hay diferencias, son irrelevantes.

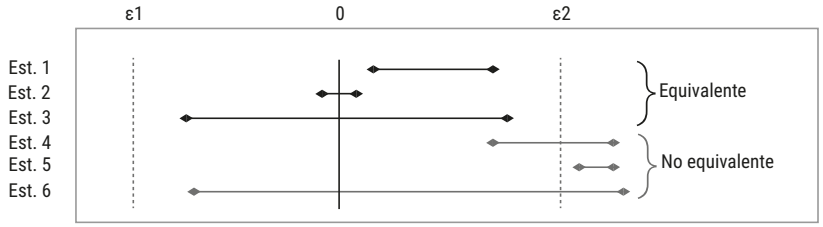


Figura 3.3. Los estudios 1 a 3, que dejan fuera ε_1 y ε_2, establecen equivalencia

- **Ejemplo**. Previamente, sabemos que un nuevo analgésico (N) tiene una tolerabilidad superior a cierto producto clásico de referencia (R). Ahora interesa demostrar que sus niveles de eficacia son parecidos. La eficacia se mide por la proporción de casos en que desaparece el dolor a los 30'. Ambos fármacos serán equivalentes en eficacia si las proporciones de desaparición del dolor no difieren en más de un 5%. El $IC_{90\%}$ de la diferencia de ambas proporciones va entre -4% y +2%. Dado que no alcanza los límites, se establece la equivalencia.

- **Ejemplo** de bioequivalencia o equivalencia en biodisponibilidad. Las agencias requieren que el cociente R/N de los niveles en sangre se encuentre entre 0,8 y 1,25. Es decir, que R no puede estar ni al 80% (4/5) ni al 125% (5/4) de N. Así, si trabajamos con la "diferencia de los logaritmos naturales", los límites de dichos cocientes son ln(0,8) = -0,223 y ln(1,25) = 0,223. En esta escala, los resultados del IC han sido -0,004 y 0,204. Ambos productos son bioequivalentes.

Así pues, la gran novedad es ε, margen, límite o diferencia clínicamente **irrelevante**. En un estudio clásico, que desea establecer diferencias, Δ es la diferencia **relevante**. En uno de equivalencia, ε representa la irrelevante, que ha de ser menor. ε se establece a priori a partir de criterios clínicos. Suele ser la mitad o la tercera parte de Δ.

Referencias: extensión de CONSORT para equivalencia, los documentos ICH-E10 de elección del grupo control e ICH-E9 de análisis estadístico, y la directriz de la EMA sobre estudios de bioequivalencia.

La **sensibilidad** es la capacidad de un ensayo concreto de distinguir entre un tratamiento eficaz y un tratamiento ineficaz o menos eficaz.

Siempre es importante: en el diseño clásico para establecer diferencias, si logra demostrarlas, queda también establecida la sensibilidad: el ensayo se auto-valida. Pero si un ensayo de equivalencia no muestra diferencias, queda la duda de si ello se explica: a) porque no existen o b) porque el estudio no ha sido capaz de establecerlas.

Varios factores pueden reducir la sensibilidad del ensayo: cambios en la población objeto de estudio (criterios de selección), cambios en las dosis y pautas de tratamiento, cambios en las variables de eficacia y su momento de evaluación, períodos de lavado preinclusión cumpliendo con la medicación, baja respuesta de los pacientes a los tratamientos, uso de tratamientos concomitantes prohibidos, pacientes que tiendan a mejorar espontáneamente, criterios diagnósticos mal aplicados (pacientes sin la patología), evaluación sesgada debida al conocimiento de que todos los pacientes reciben algún tratamiento activo, etc.

Quien proponga un diseño de equivalencia (o de no inferioridad o de no superioridad) debe aportar evidencia histórica de la sensibilidad de diseños similares a los efectos del tratamiento en estudio. Así, el diseño debe ser similar a los ensayos previos respecto a los criterios de selección, las variables, los análisis, etc. Además, su ejecución ha de ser de alta calidad: reclutamiento, seguimiento, administración de la intervención, valoración, etc.

Recuerda: La sensibilidad se puede deducir a partir de: 1) la evidencia histórica de la sensibilidad de ensayos previos a los efectos del tratamiento y 2) el diseño y la ejecución apropiados del estudio, que no limitan su capacidad para distinguir entre tratamientos.

Recuerda: El protocolo ha de: 1) definir el margen de irrelevancia y 2) aportar evidencia de la sensibilidad del diseño para establecer equivalencia.

3.4. Información acumulada. Credibilidad

Veamos escuetamente los principios de probabilidad y estadística bayesianos.

3.4.1. Teorema de Bayes

Thomas Bayes formalizó el aprendizaje, la adquisición de conocimiento, en tres partes:

1. lo que sabíamos antes, la **distribución a priori**;
2. lo que aprendemos durante el estudio, la **razón de verosimilitud**, y
3. lo que sabemos después, la **distribución a posteriori**.

El teorema de Bayes combina A y B para obtener C.

Si la pregunta es sobre dos escenarios alternativos simples y la expresamos en razones "a favor/en contra", el teorema de Bayes es:

Razón a posteriori C = razón a priori A × razón de verosimilitud B

► **Ejemplo** (adaptado de *Causal Inference in Statistics: A Primer*, de Pearl, Glymur y Jewell, Wiley, 2016).

(A) En cierto casino, hay 2 mesas de ruleta por cada mesa de dados: 2 a 1 = 2/1 = 2. La disparidad previa ruleta/dados vale 2, que indica que ruleta es 2 veces más frecuente que dados.

(B) Un visitante oye cantar un 7. La probabilidad de que salga un 7 en una ruleta es 1/36 y, en una mesa de dados, 1/6. La razón de verosimilitud de la ruleta frente a los dados, habiendo observado 7, vale 1/6 [=(1/36)/(1/6)]: escuchar un 7 es seis veces más frecuente en una mesa de dados.

(C) Multiplicar A·B proporciona C, el grado de creencia a posteriori: 2× 1/6 = 1/3. Habiendo escuchado 7, la disparidad posterior ruleta/dados vale 1/3: en este caso, la mesa de dados es tres veces más frecuente.

De esta forma, el teorema de Bayes combina la información A, disponible previamente, con la información B, aportada por el estudio, para darnos la información C, disponible al final. En la Europa continental, el teorema de Bayes suele expresarse con probabilidades y no con disparidades, lo cual hace los cálculos más engorrosos.

Las probabilidades condicionadas de las razones B y C son distintas en interpretación causal: B condiciona por la causa, mientras que C condiciona por la consecuencia.

(B) Las probabilidades de la razón de verosimilitud siguen el orden causal en que estamos entrenados a razonar: "probabilidad de la consecuencia dada la causa o P(consecuencia|causa)"; en el ejemplo, P(7|ruleta) y P(7|dados).

(C) En cambio, las probabilidades de la razón posterior invierten el orden causal: "probabilidad de la causa dada la consecuencia o P(causa|consecuencia)". En el ejemplo, P(ruleta|7) y P(dados|7). La primera vez que vemos estas probabilidades nos parecen absurdas o imposibles de obtener, pero no es así: se pueden calcular. Esta es la importancia de Bayes.

- **Ejemplo de probabilidades diagnósticas**. Simplifiquemos la pregunta clínica a una sola enfermedad: Un paciente viene a urgencias del Hospital Central a las 3 de la madrugada y refiere un dolor precordial. ¿Tiene o no tiene un infarto de miocardio (IM)?

  (A) Según los registros del Hospital Central, 1 de cada 2 pacientes que llegan al hospital con estos síntomas en esa franja horaria tiene un IM. Este ½ resume las creencias previas, la disparidad, odds, razón previa de sufrir IM: por cada 1 caso con IM hay 2 que no lo tienen.

  (B) Mediante su exploración, observamos una serie de signos y síntomas clínicos, junto con varios indicadores bioquímicos, que podemos resumir en la razón de verosimilitud: la razón de las probabilidades de observar esta clínica si el paciente tiene IM respecto a las mismas probabilidades si no lo tiene vale 10: es 10 veces más frecuente observar estos síntomas en las personas que tienen IM que en las que no lo tienen.

  (C) Multiplicando ambos valores, obtenemos la razón posterior, la firmeza con que podemos apostar a que ese paciente tiene IM: ½×10 = 5. Hemos actualizado la información previa con la aportación clínica: entre los pacientes con esos síntomas que han llegado a esas horas, hay 5 que tienen IM por cada 1 que no lo tiene.

- **Contraejemplo de probabilidades diagnósticas (cont.).** Un paciente con exactamente los mismos valores en todos los indicadores llega a una visita concertada en una consulta.

(A) Según los registros de esta consulta, 1 de cada 50 pacientes tiene IM.
(B) La razón de verosimilitud vale lo mismo, 10.
(C) Ahora, 1/50×10 = 1/5; es 5 veces más frecuente no tener IM.

Es común asumir que B, P(consecuencia|causa), sea constante en diferentes entornos o condiciones. Aun así, C, P(causa|consecuencia), no es constante; cambia en diferentes entornos en que cambian las probabilidades previas A. Al estudiar las propiedades de un indicador clínico, es más habitual obtener las probabilidades B, que siguen el orden causal natural.

► **Ejemplo.** Sensibilidad, o P(positivo|enfermo);
Especificidad, o P(negativo|sano).

En cambio, la pregunta de interés clínico, C, sigue el orden causal inverso: Dado que he observado estos signos, ¿qué enfermedad es más probable? Ahora queremos conocer la **capacidad predictiva** del resultado de un indicador.

El valor predictivo de un positivo, VP+, es P(enfermo|positivo) y
el valor predictivo de un negativo, VP-, P(sano|negativo).

La información acumulada iguala la información previa, más la aportada por el estudio actual.

3.4.2. Estadística bayesiana

El mismo razonamiento puede aplicarse al resultado de una investigación. Su aplicación a la estimación por intervalo conduce a los llamados intervalos de credibilidad.

Ejemplo. El estudio Spyral deseaba conocer la diferencia en el descenso de la presión sistólica entre la denervación renal mediante catéter y una intervención simulada, falsa, *sham*, asumida inocua. Se fijó como objetivo calcular el intervalo de credibilidad, C, de esa diferencia. A partir de un estudio idéntico **anterior**, obtuvo las probabilidades a priori **A** de esta; calculó la razón de verosimilitud **B** en los datos de **su estudio** y obtuvo **C** para resumir su conocimiento **final**, que expresó como "*posterior probability of superiority greater than 0,999 and a treatment difference of −3,9 mm Hg (95% Bayesian CI −6,2 to −1,6)*". Es decir, apostó 999 veces frente a 1 por una diferencia a favor de la intervención nueva. Y creyó que esta diferencia vale entre 1,6 y 6,2 mmHg (centro: 3,9) a favor de la denervación renal.

El resultado **C** del análisis bayesiano depende de los conocimientos previos, **A**, y de los resultados aportados por el estudio, **B**. Por tanto, es crucial especificar de dónde extraemos **A**. Y debe defenderse al diseñar el estudio. Una primera estrategia consiste en poner en **A** una distribución **a priori no informativa, vaga.** Se refiere por distribución vaga a una que presenta mucha variabilidad (las creencias previas pueden ir en cualquier sentido). Como consecuencia, influye poco en el resultado final, que se aproxima a los resultados de los métodos clásicos, "**frecuentistas**".

Una segunda estrategia presenta varios análisis de **sensibilidad** para mostrar hasta qué punto los resultados finales C dependen de las creencias previas A.

- **Ejemplo (cont.)**. El informe Spyral mostró en el apéndice varios análisis de sensibilidad, incluyendo planteamientos clásicos, frecuentistas, que arrojaron resultados similares. Además, ampliaron estos análisis secundarios para considerar diferentes escenarios alternativos, por ejemplo, para los valores ausentes — que eran muchos, de alrededor del 10%.

Opinión: Una investigación que aporta datos originales suele estar interesada en la información facilitada por el estudio y suele usar los métodos clásicos, quizás en forma de intervalos de confianza. Si al final deseas reportar cómo queda el conocimiento añadiendo la información del artículo **B** a lo que sabíamos previamente **A** ("qué sabemos al terminar el estudio"), quizás sea más apropiado añadir intervalos de credibilidad bayesianos —lo cual requiere recopilar todos los estudios previos.

Valora pausadamente qué probabilidades deseas aportar al finalizar el estudio.

3.5. Sorpresa: el valor P

El valor P indica el grado de sorpresa asumiendo cierta una hipótesis y varias premisas.

- Por **ejemplo**, en el ajuste de una muestra de datos a una distribución normal obtenemos un valor P de 0,01. Ello quiere decir que la probabilidad de obtener el resultado observado (o más extremo) es tan solo del 1%: tan poco frecuente que lo podríamos etiquetar como "raro". Podríamos decir: "¡Caramba, qué sorpresa! Estoy presenciando un evento poco probable, asumiendo como ciertas algunas premisas, como que la muestra es aleatoria."

Uno de los hándicaps de P es que se mueve en una escala antipática para las personas: no distinguimos muy bien entre 0,01 o 0,0001. No se nos da bien interpretar de forma diferencial valores distintos y cercanos a 0.

Nota técnica. Se define el valor **S** (de sorpresa) como $-\log_2 P$. El valor **S** se puede interpretar por el número de unidades de información (bits) que se requieren para expresar el desenlace de la prueba.

- Por **ejemplo**, para P = 0,05, $S = -\log_2 0{,}05 = 4{,}3$ indica una sorpresita y, para P = 0,001, $S = -\log_2 0{,}001 = 10$, una sorpresa mayor.

  Lanzar una moneda al aire y observar 4 caras seguidas, tiene una probabilidad igual a 0.0625 (0.5^4=0.0625), poco frecuente, algo raro. Pero observar 6 caras seguidas, 0.015625, muy raro, una sopresa algo mayor. Y 10 caras, 0.000977, una sorpresa aún mayor, con $S=-\log_2(0.000977)=10$.

El tamaño muestral *n* tiene mucho que ver. Si *n* es grande, el valor de P será pequeño, hacia 0, y si *n* es pequeño, P será grande, hacia 1. Así, el valor de P informa, sobre todo, del tamaño muestral.

Resaltemos que, para calcular P, necesitamos suponer como ciertas algunas premisas o asunciones. Por ejemplo, el origen aleatorio de los datos.

- **Ejemplo (cont.).** En el estudio de la normalidad con un valor P de 0,001, podríamos cuestionar si el modelo normal representa bien estos datos…, pero también su origen aleatorio.

Nota técnica. Hablamos de **modelo normal** porque se trata de eso, de un modelo que representa, de forma esquemática o simplificada, la realidad. Como *modelo*, sabemos que no es cierto, aunque sirve para esquematizar muchas variables.

Broma. *All models are wrong, but some are useful* (Box, 1976).

Los libros clásicos de estadística, en vez de hablar de modelos de probabilidad, hablaban de leyes de probabilidad, como si existiera un cierto mandato que obligara a cumplir con sus axiomas.

Broma. Las variables que no siguen la ley normal deben ir a la cárcel.

Opinión. Una pregunta más interesante podría ser: ¿Cuál es el grado de proximidad entre la distribución observada en estos datos y el modelo normal? Las medidas que indican esta proximidad (el estadístico D de Kolmogórov, el W de Shapiro-Wilk, etc.) son más informativas que un valor de P, siempre ligado al tamaño de la muestra.

Nota técnica. Si deseamos plantear pruebas clásicas, como la t de Student, en una muestra pequeña (p. ej., < 20), una premisa de normalidad falsa conducirá a inferencias erróneas, aunque en una muestra grande podría ser menos relevante.

La tradición utiliza un valor pequeño de P para cuestionar tan solo H_0, sin ni siquiera preguntarse por las premisas restantes. Este abuso ha motivado la declaración de la ASA (American Statistical Association), secundada por la RSS (Royal Statistical Society), las dos asociaciones de estadísticos más importantes y de más tradición, y que también hemos presentado en las sociedades españolas y en Medicina Clínica Práctica.

El documento que acompaña la declaración de la ASA sobre el valor P alerta sobre decenas de malinterpretaciones. Aquí las resumimos en cuatro.

Malinterpretación 1. Un valor P < 0,05 ha sido interpretado erróneamente como **garantía de ciencia**, de reproducibilidad. Sin embargo, que un suceso tenga una probabilidad ínfima, casi milagrosa, *de haber sucedido tan solo por azar* no informa de su reproducibilidad futura —tal vez al contrario.

Broma. "Disfrute de sus inesperados resultados significativos…, ¡porque no los volverá a ver!" (Frase atribuida a Stephen Senn)

Malinterpretación 2. La ausencia de pruebas no implica una prueba de ausencia: P solo aporta evidencia en contra, nunca a favor.

Analogía: Absuelto no implica inocente.

- **Ejemplo**: Supongamos que disponemos de una pequeña muestra con 15 observaciones y hacemos diversas pruebas de bondad del ajuste utilizando la distancia D de Kolmogórov: (A) En la bondad del ajuste a una normal, obtenemos un valor de P igual a 0,8. (B) La misma D con una log-normal podría retornar P = 0,9. Y (C) con una uniforme, 0,7. Si nos tomáramos el valor de P como herramienta de decisión, llegaríamos al absurdo de que la prueba demuestra simultáneamente que la muestra procede de modelos tan distintos como estos tres. En resumen, como ya hemos dicho, que no tengamos evidencia en contra de esas distribuciones no significa que la tengamos a su favor.

Malinterpretación 3. Confundir P con la probabilidad a posteriori. Ya hemos visto que no es lo mismo P(datos|hipótesis) que P(hipótesis|datos). El valor P solo habla de la primera.

Malinterpretación 4. Confundir P con el riesgo α del sistema de decisión de Neyman-Pearson (ver píldora 3.2): **es erróneo tomar decisiones** según el valor de P. Ninguna lógica ampara los umbrales, ni 0,05 ni ningún otro.

Además de acotar **α**, NP requiere: preespecificar un **criterio** de decisión y un efecto Δ en una respuesta de interés, ese efecto para el cual queremos una cierta **potencia**, y acotar β.

Broma. P es el **agujero negro del razonamiento**; absorbe las neuronas de los científicos, que parecen creer, "si algo es significativo, no es necesario pensar más". De hecho, se ha propuesto dejar de utilizar la expresión "estadísticamente significativo".

Opinión: Y quizás, en el futuro, también se abandone el término *hipótesis*.

Guías como CONSORT, STROBE o TRIPOD anteponen los intervalos de incertidumbre.

- Por **ejemplo**, CONSORT 2010 dice: "Aunque los valores de P se pueden proporcionar como complemento a los intervalos de confianza, nunca se deberían ofrecer los resultados únicamente con los valores de P."

El lector inquieto que desee utilizar el valor de P debería estudiar primero las decenas de posibles malinterpretaciones que explica el documento que acompaña la declaración de la ASA.

Interpreta el valor P **solo** como una medida de sorpresa.


EN UNA MUESTRA ALEATORIA DE CONDUCTORES, HEMOS PODIDO RECHAZAR, CON UNA P DE 0.0001 QUE EL COCHE NO FRENE.
PROBADOS CIENTÍFICAMENTE
TRAS UNA MUESTRA ALEATORIA, PODEMOS AFIRMAR, CON UNA CONFIANZA DEL 95%, QUE EL COCHE FRENA DE 60 A 0 KM ENTRE 13 Y 14 M.
PROBADOS CIENTÍFICAMENTE

4
Ensayo clínico

Un ensayo asigna al azar una intervención y su referencia para estimar, mediante la comparación de los resultados obtenidos en la variable respuesta Y, el efecto de la intervención relativo a su referencia.

En las píldoras siguientes, explicaremos el ensayo clínico con la ayuda de los puntos clave incluidos en las guías CONSORT, SPIRIT o PRISMA.

Opinión: Si quieres ahorrar costosas y laboriosas enmiendas a tu protocolo, antes de diseñar tu ensayo repasa las recomendaciones de estas guías.

Son recomendaciones, no exigencias.

4.1. Objetivos

La regla nemotécnica 2 recuerda los puntos clave del diseño:

(**P**) la **p**oblación objetivo (con sus criterios de selección), cuya evolución pretendemos mejorar,
(**I**) la **i**ntervención en el estudio,
(**C**) la intervención de referencia o el **c**omparador,
(**O**) la variable respuesta (**o***utcome*) y
(**S**) el tipo de estudio (**s***tudy*), en principio, aleatorizado.

Escribe el protocolo del estudio con **suficiente** detalle para garantizar su reproducibilidad, pero no exageres: redacta un protocolo **conciso**, en el cual sea fácil encontrar los puntos clave que aconseja SPIRIT.

Broma. El objetivo es *subjetivo*, pues depende de investigadores y patrocinadores.

Recuerda establecer los objetivos del acrónimo PICOS

4.2. Necesidad de la referencia

"*Dadme un punto de apoyo y moveré el mundo*", decía Arquímedes.

Nota histórica. Suele atribuirse a Avicena (siglo XI) el origen de la medicina experimental; a Lind (siglo XVII), el uso de un grupo de referencia, y a Hill (siglo XX), la introducción de la asignación al azar.

La referencia o comparador marca la opción "clásica", que en general deseamos mejorar y, por tanto, desplazar entre las opciones disponibles.

Nota técnica. El efecto causal en una unidad se define como la diferencia entre su respuesta cuando se le asigna la intervención nueva y cuando se le asigna la de referencia. Dado que ambas respuestas no se pueden observar a la vez y en las mismas condiciones, se habla del **problema fundamental de la inferencia causal**, irresoluble a nivel de unidad. A nivel poblacional, se afronta mediante la asignación al azar y premisas adicionales —como un efecto homogéneo en todas las unidades de la población objetivo.

La referencia debe definirse con precisión. Puede ser la pauta previamente establecida en la guía de práctica clínica. Según CONSORT, las intervenciones que se describen usualmente en los ensayos son difíciles de reproducir. Especialmente la de referencia, que suele describirse de forma vaga —por ejemplo, *standard of care*.

Ejercicio. Encuentra en EQUATOR la guía para definir de forma reproducible la referencia.

Si existen tratamientos de eficacia demostrada para una cierta patología, un objetivo clínico ético suele consistir en mejorar sus resultados, añadiendo otra intervención. Definir un grupo de control sin estos tratamientos ya establecidos debería ser de difícil aprobación para el Comité de Ética —acaso podría obtenerse para tratamientos muy rápidos y en que las consecuencias de la enfermedad fueran reversibles.

Para poder enmascarar, la intervención de referencia suele acompañarse de un simulador de la intervención objeto de estudio, acaso un placebo inocuo.

El placebo no es un tratamiento; es una mentira consentida.

Para facilitar la replicación de los resultados, debes definir con precisión tanto la intervención objeto de estudio como la de referencia, que suele ser el mejor tratamiento conocido, más un simulador del tratamiento.

4.3. Respuesta

La extensión CONSORT-outcomes especifica la definición de las respuestas en un ensayo.

Nota técnica. El punto 6a2 de CONSORT-*Outcomes* dice: "*Describe the specific measurement* ***variable*** *(e. g., systolic blood pressure), analysis* ***metric*** *(e. g., change from baseline, final value, time to event), method of* ***aggregation*** *(e. g., mean, proportion), and the* ***time point*** *for each outcome.*"

La tabla 4.1 adaptada de CONSORT-*Outcomes* pone 3 ejemplos modelo de los aspectos clave en la definición de una respuesta.

Ejemplo	Método	Variable	Escala	Estadístico	Tiempo	Principal
Presión arterial	PAD con monitor Omrom	Valor final	Continua	Media	2, 4 y 12 semanas	12 semanas
Depresión	Total puntos MADRS	Cambio desde el inicio	Binaria	Proporción	2, 4 y 8 semanas	8 semanas
Muerte	Cualquier causa	Tiempo hasta el evento		Incidencia	Continuo	Final del seguimiento

Tabla 4.1. Aspectos para definir una respuesta de forma replicable.
Adaptada de CONSORT-*Outcomes*

En un estudio confirmatorio o pivote, la respuesta principal es la variable elegida para determinar el tamaño muestral y tomar la decisión final. Es la que permite

considerar el estudio "positivo", según el argot. Diseñamos los estudios pivote bajo el paradigma de decisión de Neyman-Pearson (ver píldora 3.2).

De acuerdo con las propiedades de la medida de la píldora 1.3, la respuesta debe ser:

1. **válida** o pertinente, eso es, reflejar un objetivo de interés sanitario.
   - ► Por **ejemplo**, la cantidad o la calidad de vida.

2. **fiable**, para permitir estimaciones precisas con un menor tamaño experimental.
   - ► Por **ejemplo**: "Se determina la PAD cinco veces, separadas 5', se desechan los dos extremos y se promedian los otros tres."

En ocasiones, recurrimos a respuestas **subrogadas** o intermedias para acortar o abaratar el estudio.

- ► Por **ejemplo**, los primeros antihipertensivos, desarrollados para disminuir los eventos cardiovasculares (su auténtico objetivo), tuvieron como variable subrogada precisamente la presión arterial, que podía obtenerse mucho antes en el tiempo y requería un estudio de menor tamaño y duración que esperar a observar esos eventos.

Una variable subrogada requiere el acuerdo previo de la autoridad que aprobará la intervención.

Nota técnica. ¿Y por qué algunos fármacos fueron luego desautorizados? Pues porque estudios posteriores más amplios y extensos mostraron que tenían, además, otros efectos no deseados.

Evita definir varias respuestas principales, pues ello podría provocar multiplicidad e invalidar el estudio.

Nota técnica. La multiplicidad consiste en aumentar la probabilidad de tener un resultado "positivo" por azar. En el argot estadístico, se habla de "consumir alfa".

Broma. "Tanto va el cántaro a la fuente…"

Un ensayo puede tener otras respuestas, llamadas **secundarias**, que permiten aprender más sobre **todos** los efectos de la intervención.

La respuesta principal determina el tamaño de un ensayo **pivote**.

4.3.1. Eventos adversos. Seguridad

El estudio de la seguridad difiere esencialmente del de la eficacia. Antes de estudiar la eficacia, **confiamos** mucho en ella. Incluso en los estudios pivote sabemos lo suficiente para definir un efecto "delta" en una respuesta concreta. Pero, en el de seguridad, queremos detectar cualquier **sorpresa** y debemos estar preparados para recoger **eventos** de cualquier tipo, en cualquier momento en que se puedan presentar.

Un *evento adverso* (EA) es cualquier suceso desfavorable en un paciente. La expresión **efectos colaterales** indica que son: 1) no deseados, 2) no esperados, y 3) causados por la intervención.

Nota técnica. No abordamos aquí aquellos estudios que saben tanto de seguridad que la convierten en su objetivo principal. En este caso, es usual recurrir a la definición de una respuesta que combina la aparición de varios eventos negativos, como el *cardiovascular composite endpoint*, que engloba la defunción, el infarto de miocardio, el ictus, etc. Eventos y efectos suenan similares, pero no lo son.

Tras observar un evento y sospechar que es un efecto, el desafío consiste en saber si existe relación causa-efecto. Hay algoritmos de causalidad en farmacología (p. ej., Naranjo, Lasagna…), para atribuir la causa a una intervención cuyas categorías pueden ser segura, posible, probable o dudosa. Así, cuando se presenta un evento, suele requerirse a los clínicos que los clasifiquen según si consideran que puede atribuirse, o no, al fármaco objeto de estudio. Luego, se desvela el tratamiento asignado y se compara la frecuencia de estos eventos en ambos grupos.

Broma. Los investigadores se admiran observando los eventos en el grupo placebo que fueron etiquetados como "definitivamente causados por el fármaco".

A nivel **poblacional**, un ensayo enmascarado permite comparar los grupos para estimar cuántos eventos se pueden atribuir a la intervención.

- **Por ejemplo**, si el 10% de los pacientes refieren somnolencia en el grupo placebo y el 20%, en el grupo tratado, diremos que el efecto (negativo) del tratamiento es aumentar un 10% la somnolencia.

A nivel **individual**, es imposible inferir causalidad para convertir el *evento en efecto* del tratamiento. Requiere saber qué habría pasado en ese individuo si hubiera tenido otra exposición. Pero cambiar el pasado es imposible —es lo que

en lógica se conoce como **contrafáctico**. La extensión de CONSORT para daños (harms 2022) aconseja reportar los eventos de forma descriptiva, indicando su frecuencia, en cifras absolutas y relativas, y aclarando siempre el denominador.

Por **ejemplo**, entre los 102 casos del grupo tratado que tomaron al menos una pastilla cada día durante 3 semanas, 8 casos reportaron espontáneamente somnolencia (8%), frente a los 9 (9%) de los 103 del grupo de control.

También aconseja describir cualquier tipo de evento, ya sea **leve**, **moderado** o **grave**. Y aclarar el método de recogida, en especial si ha sido espontáneo o referido (con una pregunta concreta sobre ese evento).

Los informes estadísticos de seguridad suelen consistir en tediosos listados que atentan contra la legibilidad. Además, solo pueden incluir eventos acaecidos durante el período de seguimiento, dejando abierta y condicional la valoración positiva de la intervención.

Opinión. Como *pacientes*, preferiríamos observar los resultados en una variable que combinara toda la seguridad. Y quizás también la eficacia.

Por **ejemplo**, la escala de **Rankin** sobre la evolución del ictus combina seguridad y eficacia.

Reporta los eventos adversos de forma descriptiva.

4.4. Participantes. Población objetivo

La teoría estadística empieza con la definición de una población objetivo (*target*), a la cual queremos aplicar los resultados.

Usualmente, esta población se especifica con unos criterios de **selección o de elegibilidad** que deben cumplir, inicialmente, todos los voluntarios. Es importante resaltar que los criterios de selección se **valoran al inicio**, en el mismo momento en que decidimos intervenir, lo cual desaconseja basar estos criterios en pruebas lentas (p. ej., un cultivo microbiológico).

Las guías no aconsejan la usual distinción entre criterios de inclusión y de exclusión —quizás para evitar la frecuente malinterpretación de que, una vez incluidos, pueden ser excluidos sin inducir sesgo por atrición.

Los rangos cubiertos de características iniciales, como la edad, definen la *ventana experimental*.

Los criterios de elegibilidad son aplicables tanto a los pacientes en que se desea estudiar el efecto de la intervención, como a aquellos en que se deseará aplicar los resultados de la investigación.

- Por **ejemplo**, la *ficha técnica* de un fármaco define las características de los pacientes a los cuales va dirigido y debe reproducir los criterios de selección de los estudios que documentan sus efectos.

La **transportabilidad** (o extrapolabilidad, o validez externa, o aplicabilidad) *especula* si los rendimientos observados dentro de la ventana experimental se reproducirán también en otras variables.

Define al inicio la población objetivo con la ayuda de los criterios de elegibilidad.

4.4.1. Flujo de participantes

El primer dato que resaltar es el número de casos.

Nota técnica. En los estudios en que el número de casos se decide a priori, como los pivote, la *n* no es un resultado, sino su propia decisión. Es lógico entonces encontrarse con algunos paquetes estadísticos, como R, que no incluyen *n* en su resumen. Pero **no omitas el denominador**, el número total de casos estudiados.

Al describir el flujo de participantes, especifica los números de los que:

1. fueron **evaluados** para estudiar su elegibilidad;
2. fueron **invitados** a participar;
3. **firmaron** el consentimiento informado (CI);
4. fueron **reclutados** para el estudio;
5. fueron **asignados** al azar a cada grupo;
6. **iniciaron** la intervención asignada;
7. **completaron** la intervención asignada;
8. completaron el **seguimiento** previsto, y
9. fueron incluidos en el análisis estadístico.

Si todos los pacientes reclutados y todos los investigadores cumplen el protocolo sin desviarse, los números 4 a 9 permanecerán constantes.

La figura muestra el hipotético flujograma de un ensayo bien ejecutado, con un seguimiento impecable.

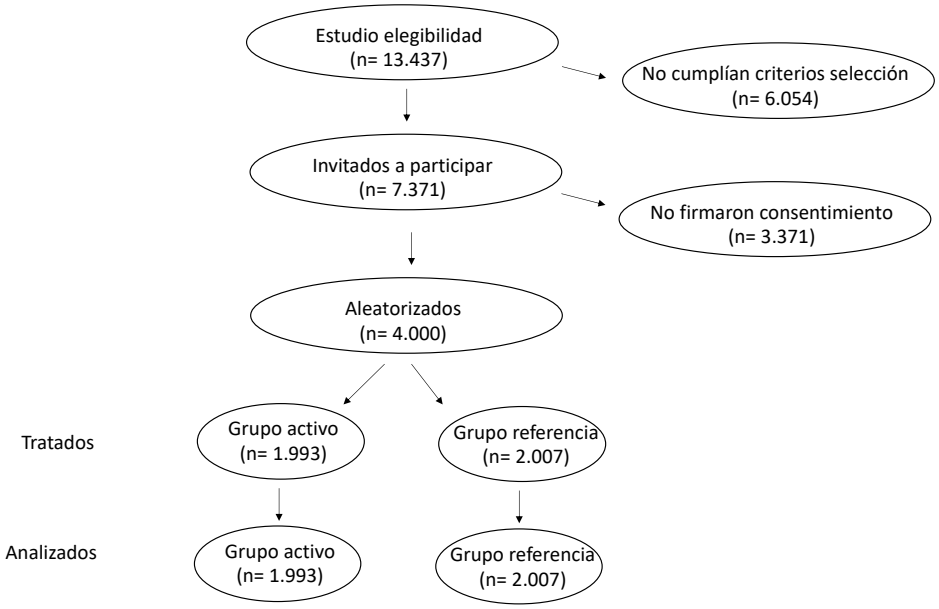


Figura 4.1. Flujograma mostrando un seguimiento modélico

Opinión. Pedir el consentimiento a un paciente grave, quizás inconsciente, en urgencias o en la ambulancia parece una broma pesada. En nuestra cultura, cuando un paciente con infarto o ictus, por ejemplo, llama a urgencias está implícitamente dando su consentimiento a que le traten de acuerdo con la guía de práctica clínica. O con las modificaciones que autorice el comité competente, quizás de ética. Y a que traten sus datos con los procesos previamente autorizados.

Aunque solo hay una forma de cumplir con el protocolo, existen infinitas formas de desviarse de él: toda pérdida a partir del punto 4 comporta un riesgo de sesgo por atrición y compromete la replicabilidad de los resultados.

Nota técnica. Seguir con el proceso de inferencia en caso de *missing* requiere la premisa no observable de "valor ausente ignorable".

Opinión. Si los pacientes o los investigadores no son capaces de seguir el protocolo, el valor de esa intervención en ese entorno disminuye.

Ya dijimos que conviene definir el final del seguimiento como la fecha de la última visita o la fecha de la defunción, para evitar convertir las defunciones en valores ausentes (*missing*).

En cualquier caso, siempre conviene definir en el protocolo los valores de la respuesta que se asignarán a todos estos "desvíos".

Nota histórica. Clásicamente, se definían las poblaciones "por intención de tratar" y "por cumplimiento del protocolo", y se analizaban ambas, lo cual contribuía a la no legibilidad del informe. CONSORT 2010 dejó claro que la población de interés es "según fueron asignados a los grupos", el punto 5 del listado anterior. Se beneficia de las propiedades aleatorias para analizar la replicabilidad de los resultados.

El diagrama de flujo permite valorar la calidad (del diseño y ejecución) del ensayo.

Minimiza los desvíos del protocolo y refléjalos en un flujograma.

4.4.2. Registros

Cada vez es más frecuente disponer de un registro de pacientes recogidos con otra finalidad.

Por **ejemplo**, el objetivo del registro puede ser administrativo, quizás para gestionar los pagos, o de investigación, quizás epidemiológica.

La extensión CONSORT-*Routine* aconseja cómo reportar los ensayos clínicos dentro de un registro.

Estos registros han recibido nombres como cohortes, registros electrónicos, bases de datos administrativas, bases rutinarias de datos, etc.

Ofrecen oportunidades adicionales de investigación.

Por **ejemplo**, podríamos definir el registro como la población objetivo y **obtener al azar** la muestra global, que será subdividida por **asignación aleatoria.**

Al recoger variables de la evolución de los pacientes o **respuestas**, se puede mejorar la calidad de su recogida mediante su comparación con las del ensayo, o incluso sustituirlas, abaratando los costes del estudio.

Algunos registros, quizás electrónicos, pueden facilitar el **reclutamiento** y la **aleatorización** de los participantes.

O la obtención del **consentimiento informado**, que podría solaparse con el necesario para el registro.

Se pueden diseñar simultáneamente estudios con diversos objetivos bajo un paraguas (*umbrella*) común.

Al diseñar un ensayo, valora las oportunidades de realizarlo dentro de un registro.

4.5. Responsabilidades

Un ensayo implica decenas de investigadores, lo cual requiere definir roles y responsabilidades.

Nota técnica. La página del ICMJE denominada *autoría* (*authorship*) establece los criterios para poder considerarse autor en una publicación —por ejemplo, de un ensayo.

El documento explicativo de SPIRIT incluye ejemplos de cómo proponen las revistas definir las responsabilidades.

Por **ejemplo**, cuáles son las del investigador principal, las del comité ejecutivo, las del comité de vigilancia no enmascarado, las del gestor de datos, las de los investigadores principales de cada centro, la del promotor y la del financiador.

También conviene especificar y puntualizar cualquier limitación de las políticas de:

- difusión de resultados a *i)* voluntarios, *ii)* profesionales y *iii)* público en general.
- acceso al *i)* protocolo completo, a la *ii)* base anonimizada de datos y al *iii)* código estadístico.

Opinión. Los comités editoriales, los de financiación, los de autorización y los de ética de la investigación deben priorizar los estudios con más posibilidades de ser replicados, con mejores estándares —quizás valorando su adherencia a guías consensuadas de investigación.

Detalla en tu protocolo las responsabilidades de todos los agentes implicados y tus políticas de difusión de resultados y de acceso a información intermedia.

4.6. Asignación aleatoria y equilibrio entre los grupos

El método experimental más importante consiste en **asignar al azar** o "aleatorizar" (*randomize*).

Nota histórica. A mediados del siglo xx, Hill convenció a los investigadores de que el azar era la forma más ética de distribuir una escasa estreptomicina entre los candidatos potenciales.

Aleatorizar garantiza el equilibrio de **todas** las variables Z previas, incluidas las aún desconocidas por la ciencia y las medidas sin suficiente precisión.

Por **ejemplo**, postulamos que el nivel de estrés influye en la aparición de accidentes cardiovasculares, aunque no existe consenso científico sobre cómo medirlo.

Conviene garantizar que el método de asignación aleatoria funciona bien. Y estar preparado para documentarlo.

Las declaraciones CONSORT y SPIRIT dedican varios puntos a la asignación aleatoria.

Broma. Asignar al azar no significa lo mismo que hacerlo al *tuntún*.

Recordemos *lo mucho que se parecen* entre sí las muestras vistas en la descriptiva de la píldora 2.2. No ha sido por buena suerte: ocurre siempre que el ensayo aleatorizado tiene un tamaño razonable y no incluye impurezas, como pérdidas de casos. Quien te diseñe un mecanismo para asignar al azar puede mostrarte mediante simulación el equilibrio esperado.

Asignar al azar garantiza que cualquier variable quede equilibrada entre los grupos a nivel poblacional. Como ambas muestras vienen de la misma población, las dos poblaciones de origen son idénticas. A nivel muestral, tenemos más garantías en las muestras grandes. Y menos a medida que reducimos el tamaño del estudio.

- **Contraejemplo**. Si asignamos 20 pacientes al azar, 10 a cada grupo, con un 50% de mujeres, la probabilidad de que este porcentaje coincida es del 17,62%; la probabilidad de diferir en un 10% es del 32,04%, y de diferir en un 20% o más, del 50,34%. Es decir, en dos muestras de diez casos, para una dicotomía equiprobable, es más fácil que la diferencia sea del 20% o superior que del 10% o inferior. Y una diferencia del 20% podría ser mayor que el efecto de la intervención, lo cual invalidaría los resultados.

Si el tamaño muestral es reducido, conviene **garantizar** el equilibrio de aquellas variables que, de quedar desequilibradas, restarían credibilidad a los resultados.

Por **ejemplo**, en un ensayo sobre una intervención con hipotético efecto hipotensor, la presión arterial inicial en el momento de reclutar al paciente suele ser un buen predictor de la presión final.

Como explicaremos en el capítulo 6 (**Control y ajuste**), para equilibrar estas variables disponemos de cuatro grandes grupos de técnicas: 1) seleccionar un **subgrupo**; 2) hacer **bloques**; 3) **modelar**, y 4) asignar de forma **dinámica**, dando una mayor probabilidad al grupo que **minimiza** los desequilibrios.

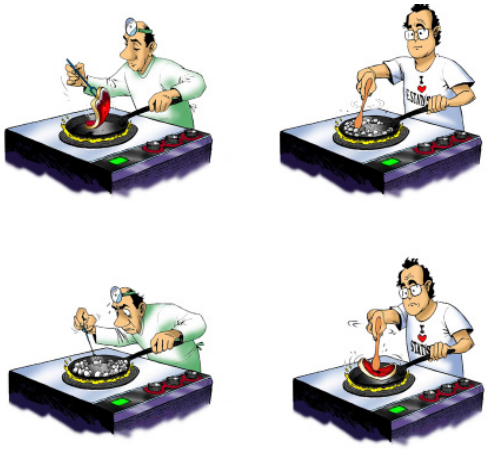


Figura 4.2. Analogía del cocinero que controla o que descansa en el azar: beneficios y riesgos

Nota técnica. Para evitar el sesgo de **selección** (ver píldora 9.2.2), las variables de control han de ser **iniciales** o basales, lo cual permite utilizarlas en el momento de la asignación.

Para evitar el riesgo de sesgo del **informe selectivo**,
especifica el método de control en el **diseño**.

Si definimos **bloques**, el período de reclutamiento termina al completar el último bloque.

- **Ejemplo**. Si ponemos el sexo como bloque en un estudio de 20 casos, tendremos la buena suerte de poder terminar el reclutamiento justo después de los 20 primeros (que sean exactamente 10 mujeres y 10 hombres), un 18% de las ocasiones, y deberemos alargar el reclutamiento con alta pro-

babilidad, del 82%. Este retardo es mayor si aumenta el número de variables que deseamos bloquear.

Los métodos computarizados de **asignación dinámica**, como la **minimización**, permiten controlar varias variables sin alargar el reclutamiento.

Ajustar por variables pronósticas tiene un beneficio colateral: mejora la precisión de la estimación del efecto de la intervención.

Si el tamaño es reducido, controla las variables pronósticas mediante bloques o minimización.

4.7. Diseño balanceado mediante bloques

Con la referencia asignada al azar, no preocupan ni la regresión a la media (ver píldora 2.8) ni la confusión de efectos (ver 9.2.1).

► **Ejemplo.** Si incluimos a pacientes en un ensayo porque presentan unas cifras (de la presión arterial, pongamos) muy altas, cabe esperar una cierta regresión a la media, pero igual en ambos grupos: tener a un grupo comparativo sometido al mismo fenómeno controla la regresión a la media.

Tampoco preocupa la confusión de efectos en un diseño balanceado. Veamos un ejemplo basado en fáciles números redondos, en el cual estamos bloqueando la tercera variable Z, "centro".

► **Ejemplo.** Supongamos (ver tabla superior izquierda) que disponemos de voluntarios, tanto en atención primaria, $\mathbf{Z}_{AP}$, como en el hospital de referencia, $\mathbf{Z}_{HR}$. En ambos centros, su evolución **Y** puede ser o positiva o negativa. Por la razón que sea, quizás porque los pacientes son más jóvenes o más incipientes, en primaria la evolución suele ir bien con más frecuencia que mal: con una disparidad de 2 a 1. En cambio, en el hospital ocurre lo contrario, de 1 a 2: es decir, la disparidad es de ½ a 1=½.

Dividiendo la disparidad de la primera fila, 2, por la de la segunda, ½, obtenemos una *odds ratio* de 4 (si damos 2 pasteles a media persona, a una entera le tocarían 4).

Vemos que el centro Z predice la respuesta Y (evolucionan mejor en AP): apostaremos 4 veces más por una mala evolución en el hospital. Por supuesto, al no ser una relación causal, no diremos: "Evita ir al hospital."

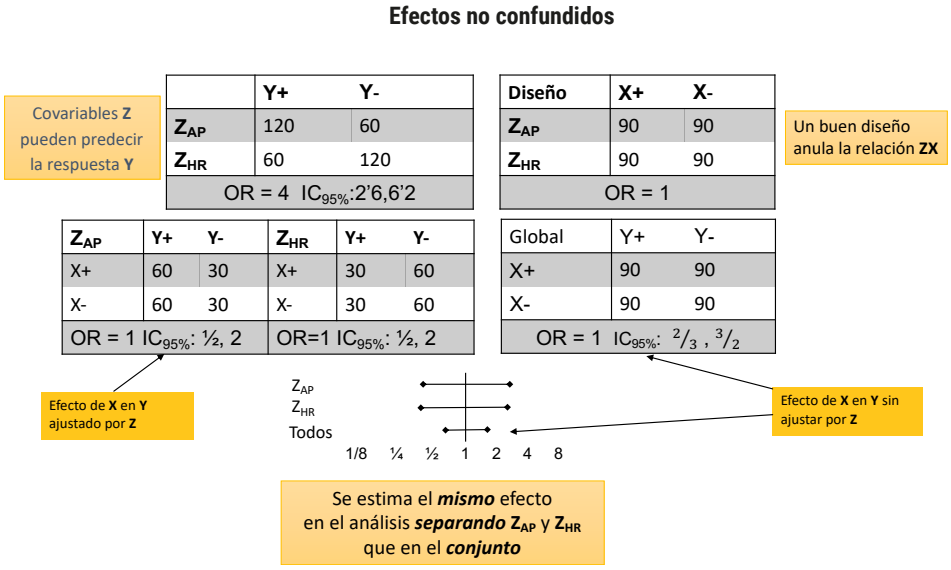


Figura 4.3. Datos inventados mostrando las ventajas de un diseño equilibrado

Eso sí, tenemos un **reto**, ya que los pacientes del hospital no son comparables con los pacientes de primaria. Por eso, **bloqueamos** (*fijamos*) el centro y **equilibramos** la intervención X dentro de los centros Z: asignamos a los voluntarios con la misma disparidad, supongamos 1 a 1, tanto en primaria como en el hospital (ver tabla superior derecha), y el resultado es un cociente de disparidades, una *odds ratio* (OR) de 1, ausencia de relación entre X y Z. No incluimos un intervalo porque no tenemos incertidumbre sobre su valor: hemos decidido una OR de 1. Así, conseguimos por diseño que X y Z sean independientes, *ortogonales*.

Para ver la ventaja de equilibrar (a menudo se denomina *diseños balanceados*), estudiemos la evolución Y de los 90 casos asignados a cada intervención en cada centro, según las tres subtablas inferiores. En primaria, teníamos 120 con Y+ y 60 con Y– (Z_{AP}, subtabla izquierda), que repartimos con esta misma disparidad (2 a 1) en ambas intervenciones X+ y X–, 60 y 30 respectivamente para Y+ e Y–, lo cual resulta en una OR = 1: sin efecto de X en Y en primaria. Si interpretamos los $IC_{95\%}$, como mucho la intervención multiplica por 2 o bien divide por 2 la disparidad a favor de una buena evolución. Si nos fijamos en el hospital de referencia (Z_{HR}, subtabla central), la *disparidad* es ½, y, de nuevo, la reproducimos en ambas intervenciones, lo cual genera también una OR = 1: relación ausente en el hospital, e idéntica incertidumbre.

Si combinamos las tablas de primaria, Z_{AP}, y hospital, Z_{HR}, obtenemos la tabla inferior derecha: sumando 60 y 30 obtenemos 90 en las 4 celdas. En esta nueva tabla global no controlamos, o ajustamos, por centro, y tampoco muestra efecto: OR = 1 ($CI_{95\%}$ de 2/3 a 3/2). En resumen, las dos ventajas del diseño equilibrado son que, al ajustar por centro Z: 1) llegamos a la misma estimación puntual que sin ajustar por Z, y 2) la incertidumbre reflejada al ajustar ($IC_{95\%}$ de ½ a 2) es más amplia que la del cálculo sin ajustar (de 2/3 a 3/2), ya que concentra todos los casos en responder la pregunta principal.

El análisis ajustado basado en subgrupos no es eficiente, ya que subdivide los casos en clases, y suele resultar en mayor incertidumbre en sus intervalos.

Bloquear en el diseño permite una misma estimación del efecto cuando ajustamos y cuando no.

En el modelo lineal clásico (p. ej., regresión lineal), condicionar por una variable Z ortogonal con la intervención X no cambia el coeficiente de X: ambas relaciones —marginal y parcial— coinciden.

Nota técnica. Las propiedades matemáticas del ajuste dependen del tipo de variable respuesta considerada. En el modelo lineal general (p. ej., regresión lineal, ANOVA, *t-test*…), el teorema de la descomposición de las sumas de cuadrados garantiza la coincidencia de la relación marginal con la parcial cuando ajustamos por variables **Z** ortogonales. Pero los modelos generalizados (regresión para respuestas binarias, tiempos de supervivencia, recuentos, etc.) no disponen de esta propiedad, de modo que surge una paradoja, conocida como la no colapsabilidad (*collapsibility*).

Recordemos que bloquear en el diseño a una variable **Z** (ortogonalidad entre **X** y **Z**) evita la posible confusión de los efectos que **X** y **Z pueden** tener en **Y**.

4.8. Tipos de ensayos

En este punto, explicamos el punto de vista clásico, muy influido por la industria farmacéutica.

Durante el último tercio del siglo xx, se impulsó la armonización de los criterios de los tres grandes mercados en aquel momento: el europeo, el americano y el japonés.

Nota histórica. El International Council for Harmonisation (ICH), patrocinado por la industria farmacéutica, consensuó las distintas normativas requeridas por las agencias de regulación y sirvió de modelo para la primera guía de publicación, la CONSORT.

Una intervención farmacológica es más fácil de estandarizar que, por ejemplo, una intervención quirúrgica o psicológica. Puede resumirse en: "Tómese 1 cada 8 horas en ayunas", lo cual permite asumir el mismo efecto en todos los pacientes, y ello facilita el análisis estadístico y la interpretación de los resultados.

Nota técnica: Un efecto constante implica las mismas distribuciones en ambas intervenciones, con variancias iguales, eso es, homocedasticidad.

CONSORT tiene adaptaciones a otro tipo de intervenciones, que ya se han introducido en la píldora 1.5.

En el capítulo 8, veremos diseños más sofisticados, que requieren un análisis estadístico diferente.

4.8.1. Ensayos clínicos según la fase de desarrollo del fármaco

La tradición clasifica los ensayos según la fase de desarrollo farmacéutico, con unas etiquetas que informan del orden: I, II, III y IV.

La fase clasifica (refiere) el desarrollo, no el estudio.

Fase I. Antes de administrar el fármaco a los pacientes, debemos estudiar su seguridad, sus efectos negativos (*Primum non nocere*: "Lo primero es no dañar"). Para ello, se suele recurrir a voluntarios sanos.

La premisa de que los efectos negativos se manifiestan por igual en sanos que en enfermos permite centrarse en los voluntarios sanos, más fáciles de reclutar. Luego, en los pacientes, estudiaremos todos los efectos positivos y negativos.

En la fase II, **exploramos** los efectos en los pacientes (p. ej., cuál debería ser la dosis).

Nota. Suele explorarse en condiciones **ideales** —es lo que se llamada "eficacia del método" o **efectividad**— a pacientes y a intervencionistas seleccionados, y se obtienen respuestas bioquímicas, intermedias o subrogadas, etc.

La fase III **confirma** estos efectos con un ensayo **pivote**, diseñado bajo el sistema NP (píldoras 3.2 y 4.8.2), con riesgos α y β controlados y evidencia previa para Δ.

Suele confirmarse en condiciones **reales**: "eficacia de **uso**" o **eficacia** a secas, con respuestas clínicas de interés para los pacientes. El reclutamiento suele ser más abierto (p. ej., multicéntricos).

Sirve para decidir si se abre el acceso general a todos los pacientes: la prescripción.

En la fase IV, una vez obtenido el permiso para su comercialización y con el producto ya en el mercado, toda esta información se complementa con un mayor seguimiento, en general observacional.

Por **ejemplo**, durante el desarrollo de las fases previas, acaso se ha visto a miles de pacientes. Pero, si un producto presenta un evento adverso grave con una frecuencia muy pequeña (quizás 1 por millón), será casi imposible observarlo durante el desarrollo, aunque sí con millones de administraciones. Y lo mismo puede decirse a propósito de los eventos tardíos, que aparezcan después del seguimiento.

Pueden incluirse cálculos de **eficiencia**, teniendo en cuenta los costes, o estudios de **calidad de vida**, para valorar mejor las implicaciones para los pacientes —en especial si el desarrollo se ha basado en variables intermedias.

La tabla 4.2 esquematiza las características de los sucesivos estudios que intervienen en el desarrollo de un fármaco.

Etiqueta	Objetivo	Fase	Diseño	Respuesta	Duración	Tamaño
Fase I	Seguridad	Investigación	Escalado	Intermedia	Días	Unidades
Fase II	Efectividad		Comparativo		Semanas	Decenas
Fase III	Eficacia	Desarrollo	Aleatorio	Finalista	Meses	Centenas
Fase IV	Eficiencia	Post-I+D			Años	Millares

Tabla 4.2. Fases de desarrollo de un fármaco

La clasificación en fases (I, II, III, IV) es aplicable al desarrollo del producto.

4.8.2. Pivote

El ensayo pivote está diseñado con el sistema NP (ver píldora 3.2) para autorizar, o no, el acceso general a una nueva intervención, quizás farmacéutica.

Por **ejemplo**, este ensayo pivote permite decidir acotando los riesgos de acciones erróneas. Como era equivalente mirar 1) si el estadístico del contraste se situaba en la región crítica y 2) si P<α; la confusión entre P y α se diseminó por todos lados. Y se creyó, erróneamente, que una P inferior al 5% sin potencia ni apoyo previo para el efecto *delta* garantizaría que ese estudio sería reproducible.

Opinión. La influencia positiva del ensayo pivote fue demasiado lejos, ya que difundió la falsa idea de que toda investigación debe basarse en los mismos métodos.

CONSORT y SPIRIT explican la metodología del ensayo pivote.

Puedes identificar si un ensayo está diseñado según los principios de NP observando la determinación del tamaño muestral en el protocolo: si acota los riesgos α y β para un determinado efecto Δ, está diseñado con el sistema NP.

El ensayo pivote requiere una investigación previa necesaria para determinar σ, sigma, y Δ, delta.

σ representa la dispersión de la respuesta numérica en las condiciones del estudio. Y Δ, el efecto para el cual deseamos tener la potencia del 80% —o quizás del 90%.

Nota técnica. σ es un parámetro secundario (*nuisance*), necesario para el diseño.

El ensayo pivote permite una decisión al final del desarrollo de una nueva intervención.

4.8.3. Piloto

Los estudios piloto o de factibilidad, en vez de aportar evidencia sobre la intervención, quieren aportar información sobre el diseño del estudio.

Por ejemplo, ¿aceptarán los destinatarios de la intervención participar en el estudio? ¿Completarán el seguimiento? Los investigadores ¿lo han entendido bien? ¿Seguirán ambos el protocolo?

En esta extensión de CONSORT, los términos *piloto* y *factibilidad* se utilizan como sinónimos. Resaltemos que esta extensión de CONSORT denomina a este estudio como **piloto** o de **factibilidad**, y el futuro **definitivo**.

El cambio más extendido se refiere a la adaptación del **resumen** a la extensión previa de CONSORT para resúmenes de artículos y presentaciones en congresos, con un total de 15 subapartados.

También hay dos grandes cambios a nivel **estadístico**, como consecuencia de su carácter exploratorio. No es preciso reportar 1) ni una justificación estadística del tamaño del estudio; 2) ni medidas de eficacia.

Se producen varios cambios sobre los **objetivos** del piloto y sobre los criterios especificados para **adaptar** el diseño del ensayo **definitivo**.

Y otros cambios son, principalmente, cómo llevarán los resultados del estudio piloto al diseño del estudio definitivo, por ejemplo, en cuanto a reclutamiento y seguimiento de los voluntarios.

Si tienes muchas dudas a la hora de diseñar tu estudio definitivo, plantéate un piloto.

4.9. Riesgos de sesgo en un ensayo

Veamos las amenazas que pueden darse en un ensayo, los riesgos de sesgo (RoB), eso es, qué características tienen los estudios que proporcionan resultados no consistentes.

El acrónimo **SABIOS** recuerda estos RoB:

S: no documentan el sistema para generar la **s**ecuencia aleatoria;
A: no ocultan el tratamiento **a**signado hasta después de completar el reparto en los grupos;
B: no enmascaran (**b**linding);
I: pierden casos (**i**ncomplete);
O: cambian de variable (**o**utcome) o de análisis tras ver los resultados, y
S: nunca hay que olvidar el cajón de **s**astre.

La Colaboración Cochrane ha impulsado una metainvestigación que ha documentado que los estudios que incurren en RoB del tipo **SABIOS**, producen esti-

maciones más optimistas del efecto. Una clara advertencia de que estos sesgos no son imparciales.

Asignación. Un buen clínico querría ofrecer la mejor intervención a los pacientes más graves, seleccionando de forma distinta a los de ambos grupos, que perderían su comparabilidad por el riesgo de **sesgo de selección**. Además, un buen investigador también debería tener argumentos clínicos para asignar la intervención de referencia. Y si no lo creyera, debería renunciar a participar en el ensayo. Se protege el estudio de este RoB **asignando al azar y ocultando** la intervención.

Solo la comparación de los grupos tal como fueron asignados está protegida por el azar.

Seguimiento. Un investigador no convencido de la ética del estudio podría facilitar intervenciones adicionales o de rescate a algún paciente. Lo prevenimos **enmascarando** el tratamiento, quizás con un placebo, y realizando un **seguimiento exhaustivo.**

- **Ejemplo**. El estudio REVASCAT logró la evolución a los tres meses de 206 pacientes con ictus. Fue mérito de los investigadores, que siguieron y persiguieron a cada paciente; de los monitores, que promovieron la calidad de los datos, y de los diseñadores, que eligieron una respuesta que incorporaba la muerte entre sus posibles valores (mRS).

Evita ser más generoso al valorar la respuesta en uno de los grupos mediante el **enmascaramiento** de los evaluadores y el uso de procesos de medición que sean independientes del investigador.

Broma. Un ciego es un invidente.

Enmascarar es un método. Que el investigador quede cegado, un resultado.

Nota técnica. Las guías aconsejan especificar a quién se enmascara: al paciente, al investigador que recluta, al que trata, al que evalúa, etc.

Nota histórica. CONSORT 2001 aconsejaba reportar si se había investigado el éxito del enmascaramiento. CONSORT 2010, no.

Informe. Finalmente, conviene no caer en la tentación de reportar el resultado estadístico que más nos conviene: este es el **sesgo del informe selectivo**. Lo protegemos publicando el plan de análisis o SAP (statistical analysis plan) antes de tener acceso a los datos.

Protege tu estudio de los **RoB** que resume el acrónimo **SABIOS**.

5
Medidas del efecto

En este capítulo, explicamos la medida del efecto (*effect size*) en un ensayo clínico.

Si hemos diseñado y ejecutado bien el ensayo, la única diferencia entre los grupos objeto de comparación será el tratamiento asignado. Por tanto, si la evolución difiere, el tratamiento es la única explicación posible.

Postulamos causas y estimamos efectos.

Según si la respuesta es numérica o binaria, utilizaremos medidas distintas del **efecto**.

5.1. Respuesta numérica

Estimamos el **efecto** por la **diferencia** de respuesta entre los grupos.

► Tomamos el **ejemplo** del documento explicativo de CONSORT 2010 sobre el dolor patelofemoral en la rodilla del corredor. Las dos intervenciones que se comparaban eran el tratamiento tradicional ("referencia") y este más **ejercicio**, ambos grupos con sus intervenciones definidas de forma reproducible. La tabla muestra los resultados de tres respuestas, en una escala de 0 a 100: 1) función articular; 2) dolor en reposo, y 3) dolor en actividad. Menos puntos indican menos dolor o menos función. La respuesta principal es **dolor en reposo**.

	Media (DT)				Diferencia en el cambio final-inicial con $IC_{95\%}$
	Ejercicio (n=65)		Referencia (n=66)		
Respuesta	Inicial	Final	Inicial	Final	
Dolor en reposo	4,14 (2,3)	1,43 (2,2)	4,03 (2,3)	2,61 (2,9)	-1,29 (-2,16 a -0,42)
Dolor en actividad	6,32 (2,2)	2,57 (2,9)	5,97 (2,3)	3,54 (3,38)	-1,19 (-2,22 a -0,16)
Función articular	64,4 (13,9)	83,2 (14,8)	65,9 (15,2)	77,8 (17,5)	4,52 (-0,73 a 9,76)

Tabla 5.1. Ejemplo adaptado de la tabla 6 del documento CONSORT 2010. Media y DT iniciales y finales (12 meses) de tres respuestas. Acompaña las estimaciones del efecto con $IC_{95\%}$

Para conocer qué ventaja tienen al final los pacientes tratados, se compara la respuesta **final**. El análisis principal ajusta por tres variables pronósticas: nivel inicial, edad y duración previa de los síntomas.

A los 12 meses (final), como respuesta principal los tratados tenían una media del dolor en reposo de 1,43 puntos, frente a los 2,61 de los controles: -1,18 puntos. Así pues, la estimación puntual de la diferencia entre tratados y controles en la respuesta final indica que **añadir** ejercicio a las intervenciones del grupo de referencia **disminuyó el dolor** medio **en 1,18 puntos.**

En vez de comparar el valor final, el objetivo podría haber sido conocer la **ventaja** del grupo con ejercicio en el **cambio** desde el valor inicial. Es decir, tener en cuenta el punto de partida.

El estimador del efecto escogido por los autores como principal (diferencia entre tratados y controles de la evolución desde el inicio hasta el final) se muestra en la última columna: los tratados evolucionaron 1,29 puntos de dolor en reposo mejor que la referencia.

Tanto estos -1,29 como los anteriores -1,18 son estimaciones puntuales que no toman en consideración la incertidumbre introducida por la asigna-

ción al azar. Nótese la amplitud del intervalo de confianza, de -2,16 a -0,42, que incluye ambas estimaciones con holgura.

Nota técnica. Si la asignación al azar ha sido correcta, ambas muestras vienen de la misma población, por lo cual ambos grupos tienen el mismo valor inicial en todas sus variables. Así, en un estudio aleatorizado, es indiferente restar por un valor idéntico inicial, y los dos estimadores tienen el mismo valor **poblacional** esperado.

El **criterio estadístico** usual propone elegir el análisis que conducirá presumiblemente al resultado **más preciso**, con un IC más estrecho. En general, este criterio lleva a proponer el análisis ajustado.

Para proteger el estudio del sesgo del informe selectivo, define en el protocolo respuesta y análisis principal.

Define en el protocolo el análisis que proporcionará la medida del efecto; para las respuestas numéricas, normalmente la diferencia de medias entre los grupos en comparación, ajustando por las variables pronósticas.

5.2. Respuesta binaria

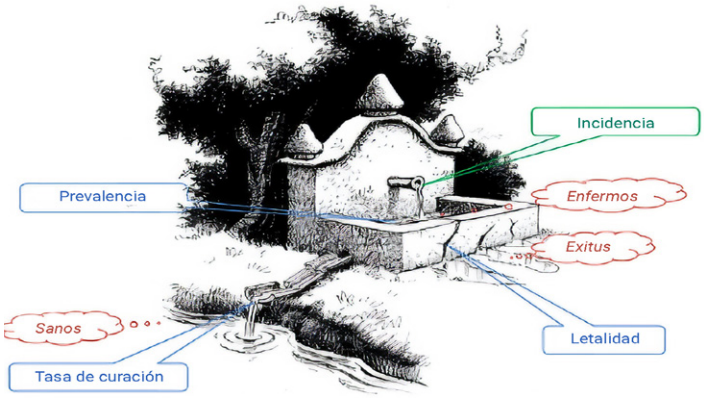
En las respuestas binarias (del tipo sí o no), comparamos las **proporciones** de éxito mediante su **diferencia** y su **cociente**.

Ratio, razón y cociente son sinónimos. Puesto que **razón** es más popular, tiene otras acepciones y podría resultar un término ambiguo, utilizamos **cociente**.

Y las disparidades (*odds*) las expresamos solo mediante su cociente: las disparidades son los casos a favor, divididos por los casos en contra.

Para tomar en consideración el paso del tiempo, hablaremos de la velocidad de aparición del evento o tasa, del mismo modo que la medida de la enfermedad distingue entre los casos actualmente presentes, o **prevalencia**, y los casos que aparecen, o **incidentes**.

Analogía. Para interpretar valores **dinámicos** que consideran el paso del tiempo, podemos imaginar una fuente de la cual brota (incidencia) agua nueva, que se acumula (prevalencia) en su recipiente. El agua que retorna al circuito (tasa de curación) más la que se pierde por filtraciones (letalidad) son los que dejan de estar enfermos.



5.2.1. Diferencia de proporciones

CONSORT proporciona el **ejemplo** de la tabla. Lograron la curación, definida por el criterio de la artritis psoriásica, 6 de 30 pacientes (86,7%) en el grupo activo y 7 de 30 (23,3%) en el de referencia: el tratamiento incrementó un 63,3% la proporción de curaciones, con un $IC_{95\%}$ de entre el 44 y el 83%: así pues, la intervención aumenta, por lo menos, un 44% la proporción de curados.

Utilizamos la **probabilidad** para referirnos al concepto (valor poblacional) y la **proporción** para lo observado (valor muestral). Las proporciones se expresan en tanto por ciento (p. ej., 23%) y las probabilidades, en tanto por uno (0,23).

Respuesta principal	Número (%)		Diferencia de proporciones ($IC_{95\%}$)
	Etanercept (n=30)	Placebo (n=30)	
PsARC alcanzado en la semana 12	26 (87%)	7(23%)	63% (44 a 83)

Tabla 5.2. Ejemplo adaptado de CONSORT 2010. Número de casos y porcentaje en cada grupo, y diferencia de proporciones como medida del efecto, con $IC_{95\%}$

CONSORT pone la etiqueta de diferencia de riesgos (*risk difference*) a lo que hemos denominado **diferencia de proporciones**. Riesgo es un término más popular, ya que la epidemiología, que suele estudiar eventos negativos, introdujo esta medida en la investigación clínica. También es común la denominación **riesgo atribuible (RA)**, que desaconsejamos, debido a las connotaciones causales del adjetivo **atribuible**.

Nota técnica. Aunque es usual identificar el riesgo con la probabilidad, muchas disciplinas, como la economía, incluyen la consecuencia en la definición de riesgo.

► Por **ejemplo**, las agencias de regulación suelen definir el riesgo mínimo como una probabilidad i) inferior a 0,1 de sufrir una complicación leve o ii) inferior a 0,001 de que sea grave.

En las respuestas binarias, una posible medida del efecto es la diferencia de proporciones.

5.2.2. Número necesario a tratar (NNT)

A partir de la diferencia de probabilidades, derivamos una medida muy intuitiva.

► En el **ejemplo** anterior, disminuir 63 eventos de los 100 tratados implica que, por cada 100 que tratemos, "evitaremos" 63 eventos. Simplificando: por cada 3 tratados, evitaremos 2 eventos.

Obsérvese, una vez más, la connotación causal de "evitaremos". Sería más correcto, aunque también más farragoso, decir: "Por cada 3 tratados, *evitaríamos* 2 eventos, si realmente esta relación fuera de causa-efecto."
El número necesario de casos a tratar (NNT) es el inverso de la diferencia de probabilidades.
En el ejemplo, 1/0,63 = 1,5873, es decir, aproximadamente 3/2.
Si la respuesta es negativa, procedemos de la misma forma.

► **Ejemplo.** La variable respuesta Y es la bronquitis crónica (o tos del fumador) y la exposición Z, el tabaco (al menos 100 cigarrillos al mes durante 20 meses, en los últimos dos años). En un determinado grupo de edad, el 60% de los fumadores tienen esta tos, frente al 10% de no fumadores, con una diferencia de proporciones del 50%; entonces, el NNT para evitar 1 evento de "tos" sería 2, es decir, por cada 2 personas que no hubieran fumado, habríamos evitado la bronquitis en 1.

Nota técnica. El NNT es una medida **sin** propiedades matemáticas elegantes. Por ejemplo, si el IC de la diferencia de probabilidades incluye el 0 (= no diferencias), es preferible no proporcionar el IC a calcular uno cuya interpretación es todo un reto.

Utiliza NNT para calcular la magnitud del beneficio de una intervención y reporta los $IC_{95\%}$ para la diferencia de probabilidades.

5.2.3. Cociente de proporciones

En las respuestas binarias, **también** utilizamos el **cociente** de proporciones.

- **Ejemplo (cont.).** Los cocientes aplicados a los dos ejemplos anteriores serían:
  - 86,7/23,3 = 3,7, el tratamiento multiplica casi por 4 la probabilidad de curarse, con un $IC_{95\%}$ de entre 1,9 y 7,2, y
  - 0,6/0,1 = 6, la exposición al tabaco multiplica por 6 la probabilidad de tener la tos del fumador (sería complicado argumentar que un efecto tan grande pueda ser explicado por otra variable confundida con tabaco).

También es usual la denominación **riesgo relativo (RR)**. Aunque no tiene connotaciones causales, induce a pensar en consecuencias negativas (**riesgo**).

Como siempre, el análisis **principal** será el que se especifique en el protocolo, aunque ahora las guías recomiendan **reportar ambos**.

A algunos investigadores les impresionará más decir que las probabilidades de curar han aumentado un determinado porcentaje, quizás un 63%. Y, a otros, que esas probabilidades se han multiplicado casi por 4. Además, en ocasiones es difícil saber a qué se refieren, razón por la cual las guías aconsejan reportar ambas.

- **Ejemplo.** Reducir un 10% los eventos adversos ¿significa que el **cociente** de probabilidades vale 0,9 o que su **diferencia** vale -10%?

No nos dejemos impresionar por expresiones del tipo: “La intervención divide por 2 la frecuencia de los eventos”. Interpreta ambas medidas del efecto junto con los datos objeto de comparación.

En las respuestas binarias:
a) proporciona las frecuencias descriptivas de cada grupo y
b) estima el efecto con **ambas** medidas, la de la diferencia y la del cociente.

5.2.4. Cociente de disparidades (*odds ratio*)

En una **proporción**, el numerador está incluido en el denominador: casos a favor/casos totales. La **disparidad** es más simple: casos a favor/casos en contra.

Nota técnica. Si el evento es raro (pocos casos a favor, como las enfermedades), la disparidad y la proporción son parecidas, por tener un denominador muy similar.

La disparidad suele tener menos incertidumbre (IC más estrechos).

A la hora de comparar disparidades, nunca utilizamos su diferencia, sino solo su **cociente**, la *odds ratio* (OR), que es la medida más popular.

La medida más popular para estimar el efecto en las respuestas binarias es la OR.

5.3. Cociente de tasas hasta el evento (*hazard ratio*)

Cuando los casos presentan el evento, sabemos el momento exacto en que apareció. Y, si no lo presentan, solo sabemos que, hasta el momento en que finalizó su seguimiento, no lo presentó y que su tiempo "libre del evento" es superior a su tiempo de seguimiento. Y entonces hablamos de **censura**. Si, como es habitual, terminamos el seguimiento antes de que todos los casos hayan presentado el evento, no podemos comparar los tiempos hasta el evento, porque en algunos casos los desconocemos: están **censurados**.

Si la información que tenemos es parcial, eso es, **censurada**, podemos calcular la velocidad de aparición de los eventos, la **tasa**, que indica la **fuerza instantánea** de aparición del evento en un lapso temporal **mínimo**, infinitesimal.

Nota histórica. Su uso más popular ha sido para el evento "muerte", por lo que se suele usar la denominación "análisis de supervivencia" para estudiar el tiempo hasta cualquier evento. Las curvas de supervivencia, como las que se ven en la figura 5.3, reflejan el declive de la proporción de "supervivientes" con el paso del tiempo. La tasa estaría relacionada con la pendiente: cuanto más abrupta, mayor es la tasa en el punto considerado.

Si disponemos de una referencia, podremos comparar los grupos de intervención mediante el cociente de tasas, más conocido por su denominación inglesa, *hazard rate ratio* o simplemente ***hazard ratio (HR)***.

Nota técnica. Podemos comparar las caídas de ambas curvas con su cociente. Y, añadiendo la premisa de proporcionalidad de riesgos (el cociente entre ambas caídas es el mismo para todos los tiempos considerados), podemos calcular una medida de resumen del efecto de la intervención, HR, común a todos los instantes de tiempo analizados.

- **Ejemplo.** La curva del grupo de control cae más rápido, antes. Tiene una mayor tasa de mortalidad: globalmente, la velocidad de aparición del evento "muerte" en el grupo tratado es 0,86 veces la del grupo de control, un 14% más lenta. Este cociente es difícil de interpretar: ¿Qué significa 0,86? ¿Es mucho o es poco?

HR informa de la reducción de la frecuencia (de la probabilidad) en el eje vertical.

- Por **ejemplo**, una HR a favor del tratamiento de 0,86 indica que, si seguimos el tratamiento, la probabilidad instantánea de fallecer baja un 14% en cualquier instante del seguimiento. Decimos "en cualquier instante del seguimiento" por la premisa de riesgos proporcionales, denominada de Cox (que es quien la propuso), para simplificar el análisis. Conviene recordar que es una premisa, asunción o presunción, no un resultado del estudio, que no demuestra que la proporcionalidad de riesgos sea cierta.

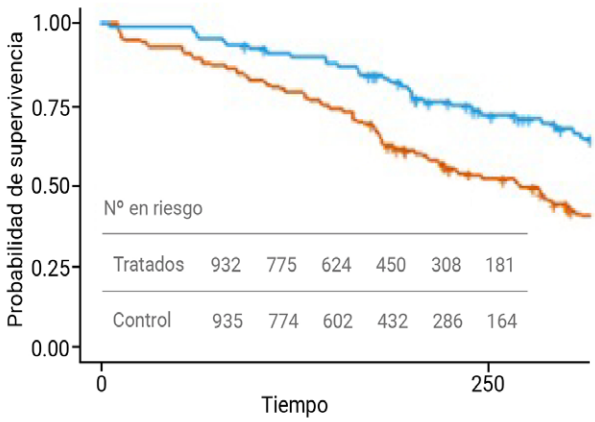


Figura 5.1. Caida de la supervivencia a lo largo del tiempo. En azul, los casos tratados. La tabla muestra la información disponible (casos) al inicio de cada período

El cociente de tasas aproxima la reducción del tiempo hasta el evento.

¿Cómo observar la reducción del tiempo? Localiza la supervivencia mediana (a la altura del 50%): para el grupo de quimioterapia, unos 4,4 años; para el grupo de control, cerca de 3,7.

Opinión. Aconsejamos interpretar la HR en la dimensión del tiempo: una HR de 0,86 (descenso del 14% en la tasa) indica que, si no seguimos el tratamiento, viviremos un 14% menos. Recuerda que esto no es una propiedad matemática.

También podemos interpretar el cociente de tasas (HR) como un cociente de proporciones, ya que, en general, el cociente de tasas es similar a los otros dos cocientes que ya hemos estudiado, el de disparidades (OR) y el de proporciones (RR).

Nota técnica. Por sus propiedades matemáticas, son tanto más parecidos cuanto 1) más corto es el seguimiento, 2) menos frecuente es el evento y 3) más próximo a 1 es el cociente. Y la divergencia va en este sentido: OR es el más extremo, luego viene HR y, después, RR.

Nota. Estos tres cocientes pueden tomar distintos nombres en distintos entornos y revistas. Además, cada uno de ellos puede ser controlado o ajustado mediante variables pronósticas, lo cual se indica añadiendo "ajustado".

Nota técnica. Debido a la no colapsabilidad de los modelos no lineales (p. ej., Cox, logística, etc.), las medidas ajustadas y las no ajustadas no son intercambiables. En general, las ajustadas tienen valores mayores.

Los tres cocientes (de probabilidades, de disparidades y de tasas) toman valores similares, en general, por lo cual algunos autores se refieren a todos ellos como *riesgo relativo*.

5.4. Homogeneidad del efecto

Por simplicidad, conviene que el efecto sea el mismo en distintas condiciones.

- Por **ejemplo**, imaginemos que una escala pronóstica recoge todos los predictores de la evolución que categorizamos en tres niveles (según si el riesgo inicial es alto, medio o bajo), que son sucesivamente más frecuentes, con 200, 400 y 800 casos. Además, presentan un cierto evento negativo, con frecuencias decrecientes del 36, el 12 y el 4%, respectivamente.

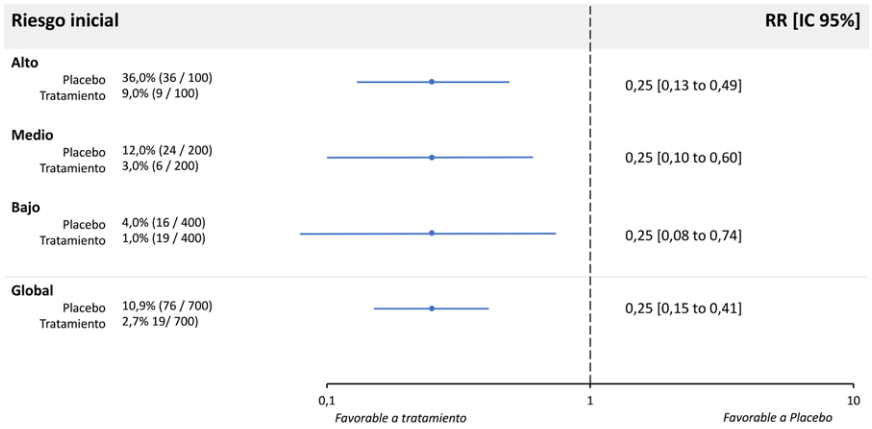


Figura 5.2. $IC_{95\%}$ del cociente de probabilidades (RR) de una buena evolución, según el nivel de riesgo

El cociente de proporciones suele mostrar la homogeneidad del efecto.

En los tres niveles, la intervención reduce la frecuencia del evento negativo a una cuarta parte, es decir, el tratamiento multiplica por ¼ la probabilidad de experimentar el evento. Esta oportuna **homogeneidad** del efecto permite combinar todas las estimaciones en una sola, más fácil de comunicar: la intervención reduce los eventos a la cuarta parte, sea cual sea la gravedad inicial. Al juntar casos, su $IC_{95\%}$ es más preciso, presenta menos incertidumbre.

Nota técnica. Entre los tres subgrupos, el IC más preciso es el del riesgo alto, a pesar de tener menos pacientes en el denominador. Esta aparente paradoja ocurre porque la incertidumbre también depende de los casos en el numerador.

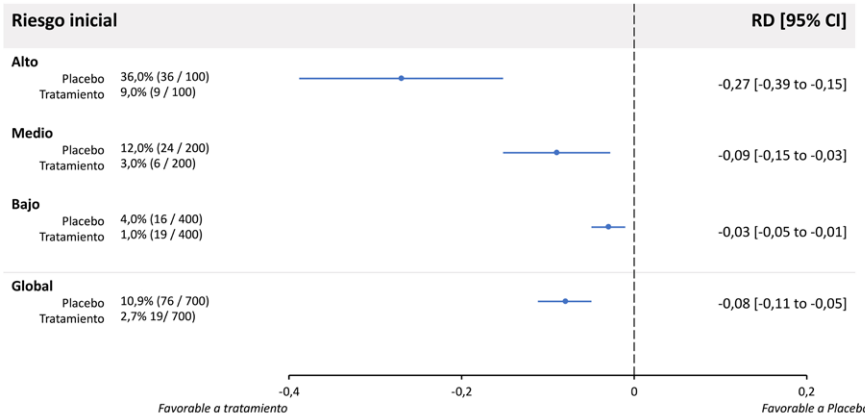


Figura 5.3. $IC_{95\%}$ de la diferencia de probabilidades (RD) de una buena evolución, según el nivel de riesgo

Y, ¿qué pasa si, en lugar del cociente, calculamos la diferencia?

Con los mismos datos, con la diferencia como medida del efecto, la frecuencia del evento baja un 27% en los casos graves, un 9% en los moderados y un 3% en los leves, lo cual muestra la **heterogeneidad** del efecto. Nótese que los NNT respectivos redondeados son 4, 10 y 33, respectivamente: mientras, en el grupo de riesgo alto, por cada 4 tratados evitaríamos 1 evento; en el de riesgo bajo, necesitaríamos tratar a 33 —y quizás no compensarían los eventos adversos.

La diferencia de proporciones suele mostrar la heterogeneidad del efecto.

Busca qué medida resume más y mejor tus datos. Y sigue los consejos de CONSORT para reportar: 1) la descriptiva de cada grupo por separado, 2) la diferencia y 3) el cociente de probabilidades.

6
Control y ajuste

Condicionar por variables pronósticas mejora la precisión de las estimaciones.

En este capítulo, veremos: 1) las herramientas que nos ofrecen *a)* el diseño de experimentos para controlar y *b)* el análisis posterior para ajustar por estas variables; 2) los peligros de la colinealidad; y 3) las diferencias entre controlar en la fase inicial de diseño y ajustar en la fase final de análisis.

6.1. Visión global

La tabla 6.1 resume de forma esquemática las propiedades de las principales técnicas para controlar en el diseño y ajustar en el análisis.

El estudio de **subgrupos** lo denominamos **bloquear** si lo planificamos en el protocolo o **estratificar** si lo decidimos durante el análisis.

Podemos recurrir a métodos de **modelado** para descontar numéricamente el efecto de variables previas medidas sin error.

- ► **Ejemplo.** Sean Y la gravedad final y Z la inicial. Podemos incluir Z en el modelo para predecir Y con el fin de responder preguntas como: "¿Cuál es el efecto de la intervención para un nivel fijo de gravedad inicial?"

Nota técnica. Con gravedad final como respuesta, un modelo de regresión que ajusta por gravedad inicial, si le exigimos que la pendiente sea 1 y la constante 0, equivale matemáticamente a considerar como respuesta el cambio inicial-final.

Ajustar en el análisis por un método de regresión abre la puerta al sesgo del informe selectivo. Para proteger el estudio, debemos preespecificar exactamente cómo vamos a ajustar. Quizás en el plan de análisis estadístico.

Opción	Fase	Nombre	Ventajas	Inconvenientes
Asignar al azar	Diseño	Aleatorización	Control completo, incluso para variables desconocidas o medidas con error	Restricciones éticas Sin garantía en muestras pequeñas
Criterios de selección	Diseño	Criterios de elegibilidad	•Control completo •Barato •Fácil de diseñar •Fácil de analizar	• Amenaza transportabilidad • Limitado número de variables • Posible confusión residual (cuando demasiadas restricciones)
	Análisis	Análisis de 1 subgrupo		
Análisis de subgrupos	Diseño	Bloques (apareamiento)	•Potencia •Eficiencia •Sin premisas •Directo •Cálculo simple	• Coste • Pérdida de flexibilidad • Diferentes estratificaciones • Dispersión de casos en estratos • Complejo para resumir
	Análisis	Estratificación (apareamiento)		
Modelado	Diseño	Covarianza	•Factible con pocos casos y "muchas" variables •Redondea los efectos menores •Permite predicciones •Permite variables continuas	• Muchas premisas • Elección del modelo • Elección de las variables • Interpretación • Parametrización del software
	Análisis	Regresión		
Global	Diseño	Minimización	•Ajusta varias var. a la vez •No reduce la generabilidad	Logística sofisticada

Tabla 6.1. Opciones para condicionar. Adaptada de Kleimbaum et al. (1991): Epidemiologic Research: Principles and Quantitative Methods

Por **ejemplo**, si deseamos ajustar por edad, conviene aclarar la relación funcional con la respuesta: ¿Una recta, una dicotomía con un umbral claramente definido, una curva cuadrática…?

Además, disponemos de métodos más sofisticados, como la aleatorización **dinámica**, que permite ajustar simultáneamente por varias variables.

Por **ejemplo**, la aleatorización con **minimización** de las diferencias entre grupos.

Para evitar el sesgo del informe selectivo, especifica previamente tus métodos o de control en el diseño, o de ajuste en el análisis.

6.2. Condicionamiento y colinealidad

Para estudiar la relación entre dos variables (sean X la intervención e Y la respuesta), fijamos ("condicionamos por") las variables Z.

Por **ejemplo**, para estudiar las relaciones complejas entre la presión, el volumen y la temperatura de un gas ideal, Boyle empezó por dejar fija la temperatura para centrarse en la presión y el volumen.

Si la relación entre X e Y fuera la misma para cualquier valor de Z, se dice que la relación XY es **homogénea** para cualquier valor de Z; en caso contrario, hay **interacción** (**heterogeneidad**) entre X, Y y Z.

Nota técnica. Para evitar el sesgo de selección, solo condicionamos por variables **previas**.

Nota técnica. La declaración STROBE señala tajantemente: "Los lectores no deben suponer que los análisis ajustados por (posibles) factores de confusión establecen la "parte causal" de una asociación."

La **colinealidad** amenaza toda investigación. Consiste en la relación, más o menos intensa, entre la intervención X y las variables Z —y de estas entre sí.

Clásicamente, todas estas variables X y Z se solían llamar erróneamente *independientes*. Si hay colinealidad, no pueden ser independientes. El término *covariables* es más acertado porque nos alerta de la posible colinealidad, aunque no distingue si son *aleatorias, estocásticas*, como las *variables* Z, o si sus valores pueden depender del investigador —es lo que hemos denominado *intervenciones X*.

La colinealidad tiene malas consecuencias.

1. Si el objetivo es predecir (p. ej., regresión múltiple), introducir un nuevo predictor muy relacionado con las variables ya incluidas:

   i) aporta poca información "nueva", con lo cual mejora poco la capacidad de predicción.

   ► Por **ejemplo**, si pretendemos añadir variables a una escala multivariante para predecir la evolución de un paciente en coma, utilizaremos variables que añadan información no aportada por otras predictoras. Así, si hemos incluido la temperatura corporal en el tiempo 0', no inclui-

remos la temperatura en el tiempo 5' para mejorar la predicción. Dos determinaciones cercanas tendrán una alta correlación o "colinealidad", quizás del 0,95, indicando que la segunda determinación aportará poca información para mejorar la predicción.

ii) y aumenta la inestabilidad de las estimaciones de los coeficientes del modelo.

► Por **ejemplo**, observemos que, en la misma situación anterior, la variación de la temperatura a los 5', con un nivel fijo de temperatura inicial, se reduce drásticamente, lo cual implica menor información para estimar el nuevo coeficiente. Se conoce como *factor de incremento de la varianza* (FIV) de la estimación del coeficiente.

2. Si el objetivo es etiológico y varias exposiciones están correlacionadas, podemos atribuir los posibles efectos a cualquiera de las variables que están relacionadas con la evolución, abriendo el abanico de interpretaciones.

► Por **ejemplo**, si buscamos las causas de la bronquitis crónica y sospechamos del tabaco, el alcohol y el sedentarismo, nos encontraremos que estas pretendidas causas están relacionadas entre ellas, son *colineales*, lo cual da pie a varias posibles interpretaciones.

En epidemiología, se habla de "deshacer la maraña o el ovillo causal".

Hemos visto que, si el objetivo es estimar efectos, el diseño de experimentos permite "aislar" las causas objeto de estudio, que no serán colineales con ninguna otra variable.

► Por **ejemplo**, asignar al azar las dos intervenciones que se comparan garantiza que ambas muestras vienen de la misma población y, por tanto, son idénticas en todas las variables previas, tanto las ya conocidas como las aún desconocidas por la ciencia.

Un estudio de observación puede controlar las variables *pronósticas* Z, *medidas sin error*. Uno experimental se sirve del azar para distribuir por igual *todas* las Z en los grupos en comparación, lo cual incluye las aún desconocidas por la ciencia.

Broma muy seria. Un ensayo aleatorizado no caduca. Las agencias reguladoras no solicitan repetirlo al descubrir una importante nueva variable pronóstica.

La colinealidad tiene dos consecuencias, según el objetivo: 1) si es predictivo, aumenta la inestabilidad de los coeficientes y, 2) si es causal, podríamos atribuir la causa a cualquier variable relacionada con la auténtica causa.

Como la garantía del azar es para muestras grandes (p. ej., de más de 500 casos), en los estudios pequeños y medianos las variables Z pronósticas importantes las "controlamos" con otras técnicas de diseño de experimentos.

Distinguimos entre relación **marginal** y relación **parcial** o condicionada por alguna variable. La relación marginal no la llamamos "no condicionada", porque en todo estudio siempre existen circunstancias condicionantes.

Al condicionar por Z, la relación parcial permite hablar de la relación entre X e Y **independientemente** de Z.

6.3. Control frente a ajuste

Utilizamos el término **ajuste** si condicionamos durante el análisis. Y **control**, si lo hacemos en el diseño.

Ajustar pretende solucionar el reto de la colinealidad. *Controlar* lo soluciona.

Ajustar requiere: 1) conocer cuáles son las variables Z por las cuales debemos ajustar; 2) medir perfectamente estas variables Z, sin ningún error de medida; y 3) conocer la relación funcional (¿línea recta?) entre la variable Z y la respuesta Y.

Controlar en el diseño requiere prever todas las sorpresas, todos los disgustos que nos podríamos encontrar.

Broma. El mejor momento para diseñar un estudio es cuando termina.

Valora si ya dispones de toda la información para diseñar el estudio **pivote** o si te conviene empezar por un estudio de factibilidad, un **piloto**.

El control **preespecificado** en el protocolo garantiza las propiedades características (p. ej., los riesgos estadísticos α y β) y la integridad del diseño. Permite, por tanto, sostener las conclusiones.

Para hacer viable tu estudio, minimiza las variables pronósticas que deseas controlar.

El ajuste, quizás improvisado en el análisis, ha cambiado las reglas. Lo utilizamos para aprender, para postular nuevas ideas.

Además de alterar las características operativas del diseño y desviarnos del plan previsto, cambiar el ajuste una vez vistos los resultados atenta contra el sesgo del informe selectivo.

Reserva tu mejor momento para diseñar el estudio.

7
Información acumulada

Una revisión sistemática (RS) recopila toda la información disponible hasta el momento sobre una pregunta concreta. Un metaanálisis (MA) combina esta información para estimar el efecto de la intervención en caso de homogeneidad. En caso de heterogeneidad cuantifica la variabilidad del efecto.

La declaración PRISMA documenta las RS con MA sobre los efectos de una intervención: la RS recopila información con un objetivo concreto; el MA la analiza.

No abordamos las revisiones de perspectiva, o *scoping reviews*, que pretenden recoger, no exhaustivamente, el "estado de la cuestión" o qué se sabe sobre cierto tema, sin centrarse en una intervención concreta. Tampoco los MA en red, más sofisticados y frágiles.

7.1. Objetivo

Cuando se inicia un ensayo no podemos saber por dónde irán los resultados. Al definir una RS, esto no es necesariamente así, ya que se basa en estudios pasados. Un experto hábil puede anticipar los resultados de una RS o de un estudio fármaco-económico, lo cual hace más difícil prevenir el sesgo del informe selectivo.

Broma. Los resultados de un estudio fármaco-económico me convencerán cuando vea un estudio "negativo".

El MA, o análisis de análisis, **combina** los efectos observados en los diferentes estudios recopilados por la RS.

Y valora si la discordancia entre las estimaciones puntuales puede explicarse por la fluctuación aleatoria de cada estimación o, en cambio, cabe sospechar una cierta **heterogeneidad** del efecto entre estudios.

En el ensayo pivote, el objetivo era tomar una decisión (sí/no) sobre el acceso al mercado y lo hacíamos prefijando el tamaño del experimento mediante NP.

Ahora el objetivo no es decidir: la intervención ya se ha administrado y disponemos de varios estudios.

Aún sigue siendo poco habitual disponer de los datos originales de los ensayos sobre una intervención. El método más popular se basa en combinar las medidas utilizadas para resumir el efecto en cada ensayo.

Acompañaremos la estimación puntual con **intervalos de incertidumbre,** para mostrar la cantidad de información que este MA proporciona sobre dicha intervención.

Un MA estima el **efecto** y su **precisión** (error típico, $IC_{95\%}$). Y cuantifica la **heterogeneidad** o variabilidad del efecto entre estudios.

7.2. Estimaciones del efecto de la intervención en cada ensayo clínico

Repasemos brevemente las medidas según el tipo de respuesta que hemos visto en el capítulo 5.

Para una respuesta numérica, la medida más habitual del efecto es la diferencia de medias.

Su error típico es $\sqrt{[(\sigma^2/n_1)+(\sigma^2/n_2)]} = \sqrt{(\sigma^2 \cdot N/n_1 n_2)}$.

Sin embargo, podría pasar que, entre diferentes estudios, la dispersión de la respuesta fuera distinta.

► **Ejemplo**. Distintos ensayos sobre la presión arterial usan distintas respuestas: unos la PAS, otros la PAD, otros el promedio de PAS y PAD, otros el promedio de..., etc.

Para combinarlas, podemos estandarizar todas las respuestas y dejarlas con una DT = 1. Ahora, la medida del efecto sería la diferencia de medias estandarizada.

Entonces su error típico es $\sqrt{((1/n_1)+(1/n_2))} = \sqrt{(N/(n_1 n_2)}$.

Como esta medida estandarizada no tiene unidad de medida, podemos combinar estudios con distintas respuestas numéricas.

En el caso de las respuestas binarias, podemos usar la diferencia (RA), el cociente de proporciones (RR) o el cociente de disparidades (OR).

$$RA = (a/n_1) - (c/n_2) \text{ y } V[DP] = p_2(1-p_2)/n_2 + p_1(1-p_1)/n_1$$

$$RR = (a/n_1)/(c/n_2) \text{ y } V[\ln(RP)] = (1-p_2)/n_2 p_2 + (1-p_1)/n_1 p_1$$

$$OR = (ad)/(bc) \text{ y } V[\ln(OR)] = (1/a) + (1/b) + (1/c) + (1/d)$$

Los valores de a, b, c y d corresponden al valor de cada celda en la tabla de efectivos de X e Y:

	Y+	Y−	
X+	a	b	n_1
X−	c	d	n_2

Nota técnica. Los logaritmos convierten en aditivas medidas proporcionales como RR y OR. Entonces, su error típico puede ser simétrico en la escala logarítmica.

Disponemos de valores 1) del efecto y 2) de su error típico en cada ensayo.

7.3. Estimación conjunta del efecto de la intervención

Distingamos dos situaciones: A) el efecto es exactamente el mismo en cada estudio y se denomina **efecto constante** o **fijo**, y (B) el efecto varía entre estudios; es un **efecto variable** o **aleatorio**.

Nota técnica. Permitimos que el efecto varíe entre estudios, pero seguimos asumiéndolo constante para los pacientes dentro de cada estudio.

Podemos asumir, de entrada, que el efecto de una intervención farmacológica será el mismo, independientemente del intervencionista, que asumimos que tiene un efecto fijo o constante.

- Por **ejemplo**, el efecto será el mismo si la indicación es "Tome 1 cada 8 horas" que si es "Tome 3 al día".

Ahora bien, si los criterios de elegibilidad varían entre estudios, este efecto farmacológico **único** podría variar.

En cambio, en una intervención **no** farmacológica, como la cirugía o la fisioterapia, tendría sentido esperar una cierta variación entre intervencionistas.

En el caso de los efectos aleatorios, además de estimar el efecto de la intervención debemos cuantificar su variabilidad.

En el caso de los efectos variables, hablamos del **efecto promedio** de la intervención.

Al variar el efecto, también debemos decir cómo varía, cuál es su **distribución** (por lo general, asumimos la **normal** de Gauss–Laplace).

Supongamos que disponemos, para cada ensayo clínico *c*, del efecto estimado Y_c, y de la varianza de esta estimación $V(Y_c)$. Si asumimos homogeneidad del efecto, definimos la ponderación (peso o *weight*) W_c como:

$$W_c = 1/V(Y_c)$$

La **estimación conjunta** ponderada **Y** de ese **efecto único** es:

$$Y = \Sigma W_c Y_c \, / \, \Sigma W_c$$

Si no asumimos homogeneidad, la ponderación W^*_c para el estudio *c* es:

$$W^*_c = 1/(\tau^2 + V(Y_c))$$

donde τ^2 es la estimación de la varianza entre estudios del efecto (existen múltiples métodos para determinar una estimación, la de DerSimonian y Laird es una de las más sencillas).

Nota técnica. Esta es la estimación de DerSimonian y Laird: $\tau^2=[\Sigma W_c(Y_c-Y)^2-(C-1)]/[(C-1)(W-S^2_W/CW)]$, siendo W el promedio de pesos; si resultara un valor negativo, $\tau^2 = 0$.

El hecho de que se tenga que estimar no un efecto concreto sino la variancia de un efecto variable implica la adopción de métodos más complejos, tales como los modelos de efectos aleatorios.

En los pesos W^*_c aparece τ^2, constante en todos los estudios, por lo cual los pesos empleados en el modelo de efectos aleatorios son más similares que en el modelo de efectos fijos: se diluye la ponderación de los ensayos más precisos.

Opinión. Al aumentar el peso de los estudios pequeños, podría aumentar también el de los mal ejecutados.

La elección del modelo, de efectos fijos o aleatorios, conviene establecerla con argumentos teóricos. Y discutir las discrepancias de resultados entre un modelo de efectos fijos y uno de aleatorios.

Si escogemos un modelo aleatorio y, contrariamente a lo esperado, la estimación de la variabilidad entre estudios τ^2 es nula, ambas fórmulas coinciden.

El error típico del estimador conjunto es similar en ambos modelos:

$$V(Y) = \Sigma W_c \quad \text{y} \quad V(Y) = \Sigma W^*_c$$

Podemos calcular los $IC_{95\%}$ del efecto combinado y estimar la variabilidad del efecto entre estudios, τ^2.

7.4. Heterogeneidad

Veamos cómo medir y cómo interpretar la heterogeneidad, y qué tipos hay.

Medimos la heterogeneidad del efecto a lo largo de distintos ensayos con el estadístico I^2, debido a Higgins, definido como la proporción que representa τ^2 respecto a su suma con la varianza aleatoria σ^2.

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Nota técnica. Si los efectos de todos los estudios tuvieran la misma precisión, $\sigma^2 = 1/W_c$.

I^2 oscila entre 0 y 1, o entre el 0 y el 100%. Cuanto mayor es I^2, mayor es la heterogeneidad. Al interpretar los resultados, PRISMA aconseja precaución si $I^2 > 0{,}25$.

Higgins puso cinco ejemplos: con un I^2 partiendo de 0, toda la oscilación del efecto es explicable por el azar, hasta 0,98, cifra absurda, explicable por un error de transcripción (de un total de tres estudios, el segundo tenía un efecto casi idéntico al resto, pero en sentido contrario).

Veamos qué fuentes clínicas y metodológicas pueden originar esta heterogeneidad.

1. Sobre las **fuentes clínicas**, distinguimos entre las **identificables** y las **aleatorias**.
   Las **identificables** o repetibles permiten estimar el efecto en ese subgrupo.
   ► Por **ejemplo**, el efecto podría ser distinto en los menores de 18 meses. Como conoceremos la edad del paciente, podemos utilizar esta fuente para anticipar el efecto.
   Y las no identificables las denominamos **aleatorias**. Para hacer la predicción, la estadística suma esta heterogeneidad τ^2 a la habitual σ^2.
   En este ejemplo, la estimación puntual de una intervención con cuidadores no profesionales de demencia es que los síntomas mejorarán y disminuirán 1/3, aunque el IC95%, que considera τ^2 además de σ^2, dice que, en un paciente tratado por cualquier cuidador, pueden disminuir entre 1/5 y 1/2 —pueden bajar entre un 20% y un 50%.

2. Otra fuente de heterogeneidad es el **método estadístico**.
   Recuerda el ejemplo de proporciones de eventos adversos, que tenía efectos idénticos si dividíamos las probabilidades (RR), pero muy diferentes si las restábamos (RA).
   Elige aquel método estadístico que presumiblemente reduce la heterogeneidad.

3. Finalmente, una mala metodología también es fuente de heterogeneidad.
   Por ejemplo, los estudios que no enmascaran suelen dar estimaciones optimistas del efecto.

No estudiamos la heterogeneidad en un solo ensayo clínico. Si así fuera, la interpretación sería tan delicada que incluso las mejores revistas discrepan en algunas de sus recomendaciones, aunque no en la fundamental: la significación de un subgrupo no salva un ensayo sin resultados globales positivos.
PRISMA, Cochrane o GRADE ayudan a perfilar la homogeneidad metodológica.

1) Al diseñar la RS, busca la homogeneidad metodológica y, 2) ante una posible heterogeneidad anónima, valora el modelo de efectos aleatorios.

7.5. Gráfico del bosque

El gráfico del bosque (*forest plot*) es un resumen magnífico de los resultados de un MA, que muestra los efectos observados en cada estudio, junto con su precisión.

- El **ejemplo** proporcionado por R, igual que los del documento explicativo PRISMA, utiliza como medida del efecto el cociente de probabilidades, con la etiqueta RR (*relative risk*). Cada fila representa un estudio. La primera columna identifica el estudio; la siguiente, los $IC_{95\%}$, y las últimas reportan los valores numéricos de estos intervalos de incertidumbre con las ponderaciones (de efectos fijos y variables) de cada estudio, que reflejan su influencia en cada estimación, junto con la medida I^2 de la variabilidad del efecto o heterogeneidad en la última fila.

Observa en el eje de abscisas la simetría en la escala logarítmica.

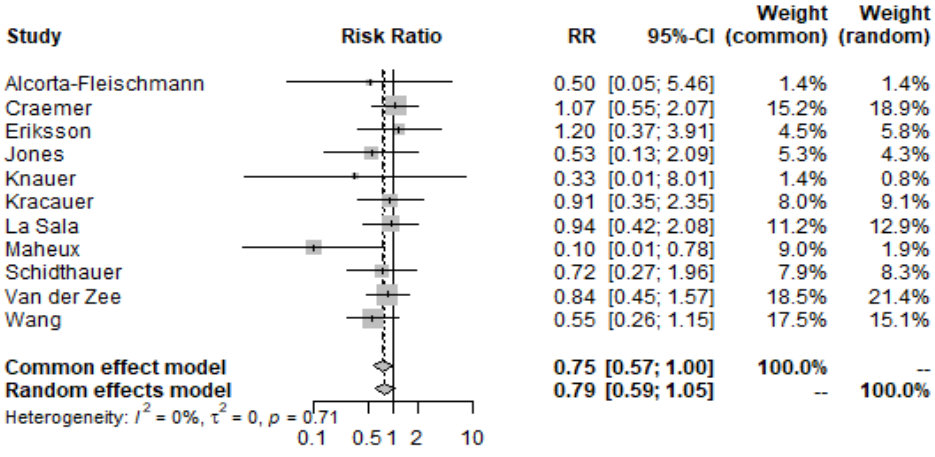


Figura 7.1. Gráfico del bosque (*forest plot*) de R, resumen de un MA

En este ejemplo, obtenido con R, la medida del efecto de la intervención es el cociente de probabilidades (RR). El efecto observado toma valores muy parecidos en ambos modelos, 0,75 y 0,79; en números redondos, la intervención reduce 1/4 o 1/5 la aparición de eventos, con intervalos de confianza amplios, de ½ a 1. Aunque las estimaciones puntuales de los diferentes estudios difieren mucho (desde 0,1 hasta 1,2, es decir, desde 1/10 hasta 6/5), su incertidumbre aleatoria es tan grande (el caso extremo va desde 0,01 o 1/100 hasta 8) que la medida de la heterogeneidad vale $I^2 = 0$: toda la variación entre estudios puede explicarse por la poca precisión de estos.

Interpreta el gráfico del bosque.

7.6. Riesgos de sesgo añadidos durante la revisión

Además de los RoB propios de cada estudio (ver píldora 4.9), existe la posibilidad de **introducir sesgos durante la RS**: es lo que PRISMA denomina RoB *entre* los estudios o a lo largo de ellos.

Una RS debe valorar los RoB de cada estudio y valorar si ha introducido RoB **entre** los estudios.

Los RoB introducidos durante la RS o el MA se refieren a información **ausente**: o bien 1) estudios completos que no se pueden localizar, quizás porque no se han publicado, o bien 2) estudios que se han omitido o han perdido parte de su información.

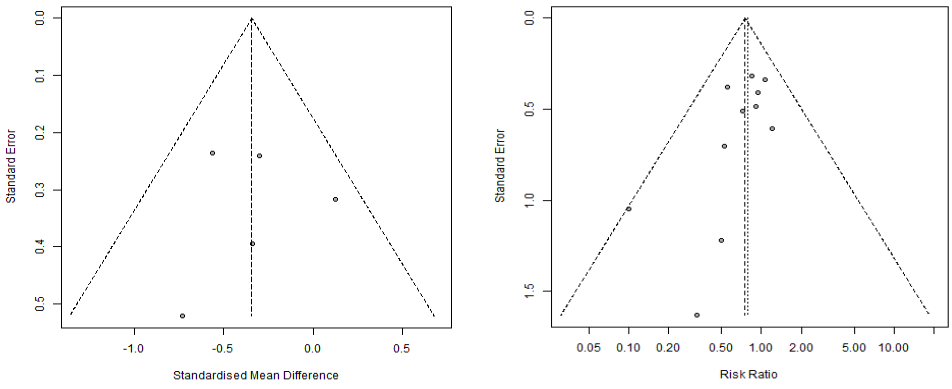


Figura 7.2. Gráficos en forma de embudo. El segundo advierte de un posible sesgo de publicación

Estudiemos gráficamente el primer grupo, el RoB por no localizar los estudios, también conocido como **sesgo de publicación**. El **gráfico en forma de embudo** (funnel plot) muestra, para cada EC, sus resultados principales, resumidos en dos ejes. La figura 7.2. muestra dos ejemplos.

En las **abscisas**, el **efecto** observado: la diferencia de medias tipificada en el ejemplo de la izquierda y el cociente de probabilidades (risk ratio) en el de la derecha.

Obsérvese la simetría de este eje. En el ejemplo de la derecha, está en escala logarítmica.

Y en las ordenadas, su precisión, recíproca del error típico de este efecto.

La línea vertical de puntos marca el efecto promedio, calculado a partir de todos los ensayos. En el ejemplo de la derecha hay dos líneas verticales, para ambos modelos, de efectos fijos y aleatorios.

En el primer gráfico, no vemos nada anómalo, pero en el segundo parece que en su base falten aquellos estudios menos precisos que arrojaron resultados en contra de la intervención, alertando de un posible sesgo de publicación. Tengamos presente que un estudio de baja precisión suele ser uno modesto, sin la cantidad mínima de participantes para conseguir una potencia suficiente. Pero también un estudio que, ante resultados inesperados, puede acabar sus días en un cajón (lo cual ocurre menos con proyectos más ambiciosos).

Los autores, los revisores y los editores tienden a no publicar aquellos pequeños estudios cuyos resultados son contrarios a lo esperado.

Una revisión debe interpretar los RoB **dentro de** cada estudio presentado. Y procurar no introducir RoB **durante dicha revisión.**

8
Ensayos con diseños especiales

A continuación, veremos los diseños apareados (8.1), que mejoran la precisión de las estimaciones; los diseños que asignan la intervención a grupos de pacientes (8.2), que pierden precisión, y otros diseños que permiten adaptaciones a resultados intermedios (8.3), acelerando el proceso de desarrollo.

8.1. Diseños apareados

CONSORT tiene extensiones para dos diseños apareados: **intrapaciente** y **con intercambio**. Valora la posibilidad de supervisar ambas extensiones a la vez.

8.1.1. Ensayos intrapaciente

En los diseños intrapaciente, los voluntarios reciben diversos tratamientos en distintas zonas. Se obtiene la respuesta en cada zona; repetida, por tanto, en el individuo.

Nota. Si son dos zonas, las respuestas vienen por parejas: **apareadas**.

Aleatorizamos los tratamientos a las zonas, en lugar de a los casos.

En el flujograma de ejemplo de CONSORT, la intervención se asigna a los ojos, no a los pacientes.

Los ensayos intrapaciente son más eficientes y conducen a estimaciones más precisas, con $IC_{95\%}$ más estrechos.

Al establecer la comparación dentro de cada paciente, eliminamos del término del error las diferencias de un caso a otro (controlamos la variabilidad entre pacientes), con lo cual tenemos menos variabilidad no explicada y el diseño es más eficiente, más preciso o más potente.

Nota técnica. Un diseño apareado permite estudiar: 1) la magnitud del efecto del caso, estudiando la correlación entre ambas medidas, y 2) el efecto de la intervención, mediante la comparación usual que puede encontrarse en los libros de estadística.

Al valorar un diseño intrapaciente, hemos de plantearnos varias preguntas, que convendrá aclarar en el informe.

- La primera ha de ser si podremos obtener **información no contaminada,** sin efectos arrastrados de una parte a otra del organismo (*carry-cross*).

- La segunda consiste en que, en ocasiones, parte de la información viene incompleta, no apareada, con solo un valor por sujeto. Si no puedes evitarlo, valora la posibilidad de renunciar a este diseño.

- La tercera es si administrar los **tratamientos de forma secuencial** o **concurrente**.

- La cuarta, si son realmente **similares** entre sí las zonas en que administraremos la intervención. Si fueran poco similares, por ejemplo, en cuanto a su gravedad o a sus condiciones iniciales, los datos estarían poco "apareados" y ganaríamos poco en eficiencia.

Y luego vienen las consideraciones técnicas.

Para que el diseño sea replicable, debemos especificar: ¿Cuándo recogeremos los valores iniciales, al aleatorizar o justo antes de administrar la intervención? ¿Cómo aleatorizaremos las intervenciones en las zonas? ¿Cómo comprobaremos que la administración se ha realizado correctamente? ¿Y cómo enmascararemos el tratamiento administrado al evaluador en cada zona?

Entre las adaptaciones de CONSORT, encontramos una gran mayoría que requieren hablar de las zonas, y no de los pacientes. Del resto, conviene resaltar la influencia que tiene la correlación entre las determinaciones del mismo paciente en el tamaño muestral, en los análisis estadísticos y en los resultados.

La extensión de CONSORT para diseños intrapaciente incluye ayudas didácticas para mejorar el informe.

El diseño intrapaciente es más eficiente a cambio de riesgos adicionales.

8.1.2. Diseños con intercambio de la intervención

En el diseño con intercambio de la intervención (*crossover*) más frecuente, AB/BA, los sujetos son asignados al azar a dos secuencias, AB o BA. Por tanto, todos los voluntarios reciben ambas intervenciones, pero en un orden diferente.

En los diseños intrapaciente, los voluntarios recibían diversos tratamientos en distintas **zonas**. En cambio, en los diseños con **intercambio**, se asignan a distintas **secuencias**.

Su gran ventaja reside en controlar las diferencias entre pacientes. Pero tienen la misma contrapartida: no podemos permitir ni una pérdida, ya que el estudio perdería el balance diseñado.

El proceso de medición ha de ser muy fiable, ya que los errores, aunque sean aleatorios, repercutirían en la eficiencia del estudio.

El flujograma de ejemplo de la extensión CONSORT al diseño crossover de dos períodos y dos tratamientos resalta que 1) los voluntarios reciben ambos tratamientos, pero en un orden distinto, y 2) los sujetos se asignan a las secuencias, en vez de a los tratamientos.

Para utilizar este diseño, tras el primer tratamiento, los pacientes deben volver a las condiciones de partida:

1. La patología ha de ser estable a lo largo del tiempo. En términos clínicos, una enfermedad crónica, como el asma o la diabetes.

2. El tratamiento ha de tener un efecto fugaz, que desaparezca en un tiempo razonable. Por ejemplo, una intervención paliativa, no curativa. Es decir, no debe haber efectos arrastrados (carry-over): el tratamiento no ha de tener efecto más allá del período en que se estudia, porque invalidaría los resultados. Conviene prevenirlo en la fase de diseño, quizás alargando el lapso intermedio, denominado período de lavado. O incluir intervenciones inocuas, que puedan desactivar las reacciones automáticas al tratamiento (p. ej., vómitos). Por supuesto, luego convendrá estudiar que no se haya presentado ningún efecto arrastrado.

Nota histórica. Una estrategia errónea consistía en poner a prueba (valor P) el efecto arrastrado y, si fuera significativo, comparar únicamente los resultados del primer período. Esta estrategia conduce a estimaciones sesgadas.

El ahorro en pacientes que supone este eficiente diseño depende de la variabilidad entre los sujetos que controlamos.

- **Ejemplo:** con diez veces menos casos: de los 61 pacientes requeridos por brazo en un diseño paralelo a los 6 voluntarios requeridos por secuencia en un diseño crossover.
  Sean delta $\Delta = 5$, la varianza entre $\sigma_\alpha^2 = 9^2$, la varianza intra $\sigma_\varepsilon^2 = 4^2$, $\alpha = 0{,}05$ bilateral y $\beta = 0{,}2$.
  $n = [2 \cdot (9^2 + 4^2)\ (1{,}96 + 0{,}84)^2] / 5^2 \approx 60{,}84 \rightarrow 61$ casos por grupo en un diseño paralelo.
  $N = [2 \cdot (4^2)\ (1{,}96 + 0{,}84)^2] / 5^2 \approx 10{,}04 \rightarrow 11$ casos totales $\rightarrow 6$ por secuencia en un diseño con intercambio.

La suposición implícita detrás de los ensayos clínicos habituales en farmacología es que el efecto es constante, es decir, exactamente el mismo en todos los pacientes pertenecientes a la población diana definida con los criterios de selección. Si no queremos descansar en esta premisa, los ensayos N=1 son una herramienta útil para determinar la efectividad de un tratamiento en un individuo determinado. Note que estos múltiples ensayos cruzados individualizados requieren determinaciones repetidas de la variable respuesta bajo los distintos tratamientos en comparación. Lo que quizás requiera desarrollar medidas subrogadas fiables de la respuesta de interés.

Puede estudiar y cuantificar la heterogeneidad del efecto siguiendo las recomendaciones de esta guía recientemente consensuada.

Entre las adaptaciones de la extensión CONSORT, resaltamos: 1) la **justificación** del **diseño** *crossover* en la introducción; 2) la **adaptación** a la comparación intra-caso de: a) el **cálculo muestral**, b) el **análisis** estadístico y c) la presentación de **resultados**, y 3) la **discusión** final de los posibles **efectos arrastrados**.

Si el diseño con intercambio es factible, valora sus riesgos en la planificación.

8.2. Aleatorizados en grupo

Ahora asignamos la intervención a una unidad que engloba al paciente.

- Por **ejemplo**, en vez de a los participantes, la asignamos a los hospitales.

Asignamos a los centros la intervención, pero valoramos la respuesta en los pacientes, por ello denominamos estas unidades de **intervención** y de **inferencia**.

Hay una relación **jerárquica** entre los elementos del estudio: los pacientes **pertenecen** a los centros. También podemos decir anidados o contenidos en los centros.

- **Ejemplo**. Para prevenir los embarazos no deseados en adolescentes, asignamos la intervención a escuelas y valoramos el éxito de los conocimientos (respuesta: un test concreto) en los adolescentes.

- **Contraejemplo**. El *second-opinion trial* quería reducir la proporción de cesáreas en distintos centros, en los cuales medía esta respuesta (% cesáreas por centro). Tanto la unidad de intervención como la de inferencia coincidían en el hospital, de modo que no estaba aleatorizado en grupo. Su única peculiaridad era que su unidad estadística no la formaban los pacientes, sino los hospitales.

8.2.1. Aleatorización simple en grupo

Su nombre clásico era **ensayo comunitario**, aunque ahora domina el término anglosajón *cluster*.

- El **ejemplo** de la tabla, tomado de la extensión CONSORT-Cluster 2004, muestra que 25 y 27 centros fueron asignados a las intervenciones en estudio y de referencia respectivamente, que resultan en 172 y 156 pacientes en cada grupo, es decir, unos 6-7 por centro.

Nivel	Información inicial por grupos de tratamiento a nivel de paciente y de centro			
			Grupo de intervención	Grupo control
Centro	n		25	27
	Tamaño	Pequeño (<1.600)	8	10
		Mediano (1.600-2.200)	8	10
		Grande (>2.200)	9	17
	Dispone de enfermera	No	8	11
		Sí	17	16
Paciente	N		172	156
	Edad media en años		66,4	64,8
	Hombres (n de n, %)		107/172, 62%	87/156, 56%
	Fumadores (n de n, %)		68/161, 42%	49/141, 35%
	Infarto de miocardio (n de n, %)		103/172, 60%	91/156, 58%

Tabla 8.1. Ejemplo de características iniciales, adaptado de CONSORT-Cluster

El flujograma también debe informar sobre ambas unidades.
El **ejemplo** de la extensión CONSORT-Cluster 2012 muestra que los centros fueron 204 y 123. En total, las mujeres fueron unas 40.000 en cada grupo de intervención.

Con el planteamiento estadístico **usual**, cada unidad aporta información **independiente**: añade información 100% nueva. En cambio, ahora los pacientes de cada **grupo** comparten alguna **información común**, lo cual debe tenerse en cuenta en el análisis, ya que el cálculo usual proporciona resultados erróneamente precisos. Por ejemplo, los intervalos de confianza son demasiado estrechos y no alcanzan la cobertura deseada.

Para considerar esta información repetida, compartida por las unidades del mismo grupo, disponemos de dos indicadores estadísticos: la **correlación intragrupo** (o intraclase) y el **efecto del diseño**, que indica por cuánto hemos de multiplicar la varianza obtenida para obtener una cobertura correcta.

Definimos el **coeficiente de correlación intragrupo (CCI)** como la proporción de la variabilidad total explicada por los grupos. Mide la información compartida por los miembros de un mismo grupo.

La extensión CONSORT-Cluster 2012 muestra un ejemplo en que la varianza compartida depende de la respuesta escogida: 1% para el colesterol, 5% para la dieta y 10% para el peso.

El **efecto del diseño (DE)** indica por cuánto hemos de multiplicar la varianza que obtenemos si efectuamos los cálculos usuales.

$$DE=1 + (n-1)\ ICC_{ngt}$$

Depende de 1) el número n de casos por grupo y 2) su correlación intragrupo. En otras palabras: 1) el número de casos con información repetida, ya aportada por otros casos, y 2) la cantidad de información compartida.

- Por **ejemplo**, el número de casos necesarios para un estudio diseñado bajo NP se multiplica por DE.

  Si la correlación es nula (CCI = 0), DE = 1: no aumenta la varianza del estimador. DE no puede ser negativo, ya que hemos definido el CCI en términos de cociente de varianzas. Los valores habituales del CCI se sitúan en el rango entre 0,01 y 0,10.

> Con un promedio de 6 pacientes por grupo, el efecto del diseño es de 1,5 para el peso corporal (la varianza aumenta un 50%) y de 1,05 para el colesterol (aumenta un 5%).

Como la estimación del CCI suele ser muy pobre, es aconsejable basarse en premisas conservadoras sobre su valor —por ejemplo, CCI = 0,10.

La información disponible para estimar el CCI depende del número de grupos G, en lugar del número total de casos (G·n).

Si el número de grupos G es pequeño (p. ej., 20), mejor **no confiar** en el equilibrio de la aleatorización basado en resultados para muestras grandes (asintóticos). Mejor estratificar o minimizar.

Nota técnica. Estratificar aumenta la eficiencia si la correlación entre la respuesta y los criterios para definir los estratos es superior a 0,5.

Los ensayos asignados en grupo presentan **riesgos adicionales de sesgo.**

Por **ejemplo**, podemos enmascarar, pero los pacientes podrían averiguar la intervención que se practica en cierto centro y no resultar cegados, lo cual obliga a ser prudentes a la hora de interpretar respuestas con componente subjetivo.

Por **ejemplo**, ocultar la intervención antes de la asignación puede resultar imposible, ya que en estos diseños es usual asignar primero la intervención y reclutar luego a los pacientes, al contrario que en los diseños tradicionales, con lo cual existe la amenaza del sesgo de selección.

Cambian los aspectos metodológicos, que han de ser reportados según la extensión CONSORT-Cluster. Y los estadísticos, ya que difieren los cálculos para obtener los errores de estimación.

La extensión CONSORT a los diseños aleatorizados en grupo requiere especificar: 1) las **razones** para hacer un diseño de clúster; 2) **ambas unidades** aleatorias en el flujograma, y 3) medidas estadísticas específicas (**CCI y DE**) y sus implicaciones en el análisis y en el cálculo del tamaño muestral.

Los diseños aleatorizados en grupo tienen sus peculiaridades metodológicas y estadísticas.

8.2.2. Asignación al azar escalonada

Podemos asignar grupos a la nueva intervención al azar, aunque de forma secuencial, progresiva y escalonada (*stepped wedge*, SW).

Este término hace referencia a los escalones de la figura siguiente, que muestra los tratamientos utilizados en cada período:

Grupos (*clusters*)	Pasos	1	2	3	4	Secuencia de tratamientos
	Período 1	Período 2	Período 3	Período 4	Período 5	
						0 0 0 0 1
						0 0 0 1 1
						0 0 1 1 1
						0 1 1 1 1

Tabla 8.2. Ejemplo de asignación secuencial adaptado de CONSORT-SW

En esta figura, hemos asumido que la transición entre períodos de tratamiento es instantánea.

Al ser un tipo de ensayo en grupo (*cluster*), hemos de considerar **la correlación intragrupo** (*intraclass correlation,* **CCI)** y **el efecto del diseño (DE)**.

- **Ejemplo.** Determinadas intervenciones, comúnmente en gestión sanitaria, no pueden incorporarse de forma simultánea a todos los centros.
- **Contraejemplo.** El Plan Bolonia fue adoptado sin ninguna evaluación pública de sus efectos. Con una implementación gradual, quizás por universidades o por países, podrían haberse estimado sus efectos.

Opinión. Al ser decisiones que se aplican en grupo, el consentimiento implica más a los grupos que a las unidades.

Como otros diseños aleatorizados en grupo, los voluntarios podrían conocer la intervención, lo cual abriría la posibilidad de riesgos de sesgo adicionales (selección, atrición, no enmascaramiento…).

Este diseño facilita una evaluación rigurosa, al tiempo que distribuye de forma equitativa (al azar) el peso de la investigación.

La contrapartida es que los efectos de la intervención **se confunden** con los del paso del tiempo. Efectivamente, si comparamos en su conjunto los períodos con la intervención nueva, más recientes en el tiempo, con los períodos con la clásica, anteriores, los efectos de la intervención están **confundidos** con todos

los efectos del paso del **tiempo**. A diferencia de los estudios antes/después, los diseños escalonados permiten estimar y descontar el efecto del tiempo (**ajustar**).

La extensión CONSORT te ayudará a diseñar, ejecutar y reportar correctamente los diseños escalonados.

Considere, en el diseño y en el análisis, los efectos del grupo y del tiempo.

8.3. Ensayos adaptativos

Resumimos nuestro vídeo previo.

Un ensayo adaptativo permite modificar el diseño según lo visto hasta el momento.

Broma. "Pero ¿cómo? Si aprendo durante el estudio, si aclaro dudas, ¿no puedo adaptar el diseño?" Bueno, si quieres explorar de cara a futuros estudios en un piloto, no hay ninguna dificultad. Pero, si haces un estudio pivote para reclamar el permiso para un nuevo fármaco, entonces has de garantizar que el diseño alcanzará un resultado positivo cuando haya un efecto real detrás. Ha de ser una garantía en términos probabilísticos, admitiendo los riesgos de error α y β. Todo lo que anticipes y pienses ahora lo ahorrarás después.

Si quedan muchas dudas, valora la posibilidad de realizar un estudio piloto. Si quedan pocas, uno adaptativo.

Hay razones **previstas** y razones **inesperadas** para terminar el estudio. En el primer caso, el estudio llega hasta el final; en el segundo, se interrumpe.

► **Ejemplo**. ¿Hemos parado el estudio antes de tiempo por alguna razón **inesperada**? Quizá por falta de pacientes o de fondos... ¿O hemos alcanzado una **regla vinculante** para terminar el estudio? Quizá por tener la información deseada de la intervención, o bien por considerar inútil seguir con el estudio.

La extensión CONSORT a diseños adaptativos (CONSORT-ACE) prevé adaptaciones para:

1. Redefinir el tamaño al mejorar la información (enmascarada) sobre parámetros secundarios (*nuisance*), como la varianza.

2. Decidir sobre los efectos de la intervención al valorar la información acumulada.

3. Reducir el número de tratamientos estudiados.

4. Cambiar la razón de asignación.

5. Acelerar el desarrollo, combinando, en un solo diseño, objetivos que suelen estudiarse en estudios sucesivos.

6. Perfilar la población objetivo, buscando la que presente homogeneidad del efecto, que permite una prescripción común ("estudios enriquecidos").

7. Cambiar los objetivos.
► Por **ejemplo**, de equivalencia a superioridad. O, dicho llanamente: desde empatar, y compartir el mercado de otra intervención ya autorizada, hasta ganar y sacarla del mercado.

8. Modificar las intervenciones.

Los diseños adaptativos requieren definir:

– los privilegios de cada investigador (p. ej., solo cierto comité puede acceder a la información sobre el tratamiento recibido);

– análisis interinos, que se basan en datos acumulados intermedios;

– adaptaciones o modificaciones del diseño, según lo previsto en el protocolo;

- reglas vinculantes, que permiten adoptar una modificación prevista;
- respuestas para las modificaciones, definidas en el protocolo, iguales o distintas a las respuestas finales;
- validez, para proporcionar estimaciones finales no sesgadas, tanto del efecto, como de su incertidumbre;
- sesgo, tanto en las estimaciones finales como en las adaptaciones intermedias;
- integridad, que se previene respetando las reglas del estudio, y
- características operativas, que incluyen, bajo distintos escenarios: los riesgos clásicos (α y β), las probabilidades de opciones intermedias y la distribución de las estimaciones finales.

Veámoslas para el ensayo secuencial REVASCAT, cuyo número final de pacientes estudiados dependía de los resultados intermedios:

			H_0: OR=1			H_1: OR=1.62		
	N total	n brazo	Inutil	Eficaz	Ambos	Inutil	Eficaz	Ambos
	174	87	29,7%	0,3%	30,0%	0,4%	22,6%	23,0%
	346	173	49,6%	0,6%	50,2%	2,1%	46,7%	48,8%
	518	259	15,2%	1,0%	16,2%	2,3%	19,6%	21,9%
	690	345	3,1%	0,5%	3,6%	1,5%	4,8%	6,3%
			97,6%	**2,4%**	100%	6,3%	**93,7%**	100%
	Prob(N>518)		3,1%	0,5%	3,6%	1,5%	4,8%	6,3%
Tamaño fijo diseño clásico						564		
Tamaño esperado secuencial				335			366	

			H_0: OR=2			H_1: OR=2.45		
	N total	n brazo	Inutil	Eficaz	Ambos	Inutil	Eficaz	Ambos
	174	87	0,0%	50,2%	50,2%	0,0%	78,5%	78,5%
	346	173	0,1%	43,1%	43,2%	0,0%	21,2%	21,2%
	518	259	0,0%	6,3%	6,3%	0,0%	0,3%	0,3%
	690	345	0,0%	0,3%	0,3%	0,0%	0,0%	0,0%
			0,1%	**99,9%**	100%	0,0%	**100%**	100%
	Prob(N>518)		0,0%	0,3%	0,3%	0,0%	0,0%	0,0%
Tamaño fijo diseño clásico				270			162	
Tamaño esperado secuencial				272			211	

Tabla 8.3. Ejemplo de características operativas del diseño secuencial REVASCAT

► **Ejemplo**. REVASCAT enfrentaba varios escenarios posibles sobre el efecto (**nulo**, OR = 1, y **delta**, con 3 opciones: efecto moderado, OR = 1,62; efecto medio, OR = 2, y efecto alto, OR = 2,45); contemplaba dos razones para terminar: por **eficacia** de la **intervención** o por **inutilidad** del **estudio**, y pla-

nificó **cuatro análisis** sucesivos, cuando 174, 346, 518 y 690 pacientes hubieran completado el seguimiento.

La suma de las probabilidades de parar en estos cuatro análisis intermedios proporciona aproximaciones de A) la **potencia 1**-β en cada escenario, 1) efecto moderado: 87,7%; 2) efecto medio: 99,9%; 3) efecto alto: 100%; y B) el **riesgo** α unilateral en el escenario de efecto nulo (OR = 1), igual a 2,4%.

Para contemplar los diseños adaptativos CONSORT-ACE, incluye muchos cambios: añade 7 puntos nuevos, renumera 4, modifica 9 y amplía, con aclaraciones, 6. Y proporciona modelos ejemplares de flujogramas para distintos diseños.

El protocolo de un diseño adaptativo especifica reglas vinculantes y características operativas.

9
Estudios observacionales

Hasta ahora, hemos estimado los efectos de las intervenciones en diseños experimentales.

Veamos ahora cómo estudiamos los otros tres grandes objetivos clínicos: etiología, diagnóstico y pronóstico, a partir de observaciones.

Los dos últimos comparten la existencia de un indicador cuya capacidad predictiva debemos valorar, aunque difieren en si el valor de la respuesta lo recogemos en el presente (diagnóstico, *gold standard*) o en el futuro (pronóstico).

Al ser observaciones (*See*), no es posible un diseño experimental (*Do*): no es posible asignar las intervenciones, ya que son exposiciones, y por tanto, tampoco al azar.

Opinión. Suponemos que el desconocimiento de estas definiciones (*Do* y *See*) está detrás del uso extendido en la práctica real de las medidas usuales de inferencia estadística (error estándar, intervalo de confianza, P valores, etc.) en situaciones que carecen de justificación formal.

Estas medidas descansan en un proceso de azar: lo requieren. Y si no existe, carecen de fundamento. Por ello, si un estudio no dispone de un proceso aleatorio y aun así presenta medidas estadísticas de incertidumbre, debería resaltar esta limitación en la discusión.

- Por **ejemplo**, "por tradición hemos calculado los intervalos de confianza que requieren una condición de azar que, en nuestro caso, no se cumple. Aun así, los hemos presentado a nivel orientativo. Al interpretarlos, conviene recordar que, por el sesgo impredecible, la incertidumbre real será mayor que la reflejada por los intervalos que nosotros aportamos".

9.1. La etiología postula causas

Hemos visto que un ensayo bien diseñado y ejecutado permite estimar los efectos de una intervención (causa) asignada al azar. Ahora explicamos por qué un estudio observacional solo permite lanzar ideas, hipótesis sobre causas.

Broma sobre la I+D. Postular causas es investigación pura; estimar efectos, un simple desarrollo.

► **Ejemplo**. A finales del siglo xx, R. Warren y B. Marshall **postularon** que la causa de las úlceras gastroduodenales era la colonización del estómago por el *Helicobacter pylori* y no por el estrés o por la comida picante, como se sostenía hasta entonces. Otros propusieron diversas intervenciones para erradicarlo, que evaluaron experimentalmente, y redujeron espectacularmente la necesidad de cirugía gástrica. En 2005, Warren y Marshall recibieron el Nobel de Medicina.

Recordemos **dos retos** de los estudios observacionales:

1. Las unidades han llegado con el valor de la causa. En el futuro, ¿podremos asignar su valor? ¿Podremos *intervenir*?

   Convertir las exposiciones en intervenciones requiere un acuerdo más político que científico.

   Historieta. Al estudiar el asma preolímpica de Barcelona, Antó, Sunyer y otros colaboradores vieron que aparecía en los días en que había descarga de soja en el puerto y el viento soplaba hacia el barrio donde ocurrían los casos. Y postularon que la causa del asma era esa combinación de viento, Z_1, con la descarga de soja, Z_2. Ninguna de las dos variables les "pertenecía": ellos no podían prohibir ni el viento ni, por razones prácticas, la descarga de soja. Al no poder asignar su valor, no podían ser intervenciones X. Así que pactaron con el responsable sanitario de Barcelona, Joan Clos, con el beneplácito del alcalde Pasqual Maragall, su "intervención": reparar y aislar los silos para la descarga de la soja.

2. La elección del valor de la causa en estas unidades ¿ha estado relacionada con el valor de la respuesta? Si estudiamos datos *reales*, podría ser que los casos con mejor pronóstico terminen en una intervención concreta.

► Por **ejemplo**, supongamos que disponemos de una gran base de datos sobre la evolución de todos los pacientes que padecen un determinado

cáncer y deseamos comparar el rendimiento de una intervención quirúrgica con el tratamiento médico habitual. Podría ser que la evolución de quienes fueron tratados con la primera resultara mucho mejor, quizás por el hecho de que solo los casos en buenas condiciones pueden ser admitidos en cirugía.

Estos ejemplos podrían etiquetarse como sesgos de confusión o de selección, según la terminología elegida. La epidemiología habla de factor de confusión o de sesgo de confusión: atribuir a la supuesta causa unos efectos que pueden ser explicados por un pronóstico inicial distinto. Como la distinta evolución es real y es replicable (en ese entorno, los operados viven y vivirán más), en estadística evitamos el término sesgo y hablamos de **confusión de efectos.** Lo estudiamos en la píldora 9.2.1.

En la terminología de los ensayos clínicos, en una comparación **no** aleatorizada, decimos que los grupos **no son comparables** ya que, desde el inicio, pueden tener distinto pronóstico.

Recuerda: Estima los efectos en los **experimentos** y postula las causas en las **observaciones**.

9.2. Dos grandes amenazas en las observaciones

Explicamos primero por separado la confusión de efectos y el sesgo de selección. Luego, introducimos los modelos causales, que permiten representarlos juntos.

9.2.1. Confusión de efectos

Veamos qué podría suceder en el entorno observacional con la presencia de colinealidad entre la causa X y alguna de las covariables Z.

- **Ejemplo**. Igual que en la píldora 4.7 sobre diseños equilibrados, partimos de pacientes diferentes en primaria y en el hospital, con un mejor pronóstico en primaria (OR = 4, subtabla superior izquierda), aunque ahora los clínicos y los pacientes pueden elegir. Supongamos (subtabla superior derecha) que los 180 voluntarios de primaria Z_{AP} tienden a recibir la intervención, con una **disparidad** de 5 a 1, mientras que en los 180 del hospital Z_{HR} ocurre justo al revés, de 1 a 5. No importan las razones de su elección: pueden ser "las cosas de la vida". El caso es que ahora hay una OR = 25 que indica fuerte colinealidad entre la covariable Z y la variable de interés X (puesto que no podemos intervenir, mejor llamarla **exposición**, en lugar de **intervención**).

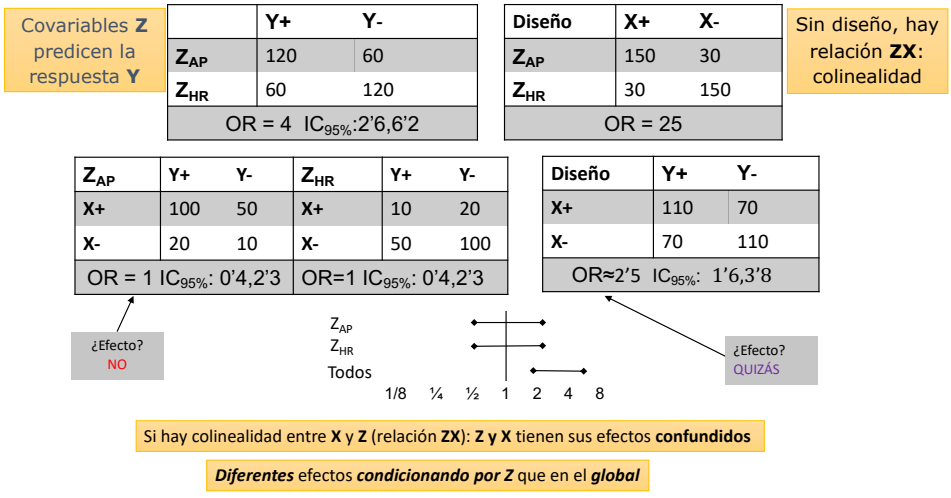


Figura 9.1. Datos inventados que muestran que la colinealidad termina en confusión de efectos

En las subtablas inferiores, repetimos el mismo análisis de la píldora 4.7. Primero, cada centro por separado: en la subtabla de la izquierda, en primaria Z_{AP}, los 150 y 30 casos se reparten también 2 a 1 para Y+ e Y- respectivamente: los 150 tratados X+ son 100 y 50; y los 30 controles X-, 20 y 10, por lo cual el efecto estimado de X en Y es otra vez nulo (OR = 1). Y lo mismo ocurre en la subtabla central del hospital Z_{HR}. Veamos ahora qué sucede en la de la derecha, cuando combinamos ambos centros: 100 + 10 dan 110, y 20 + 50, 70. Parece que la primera

fila de la subtabla superior derecha tiende a ir bien, 110 a 70, aproximadamente 3 a 2, y la segunda fila a ir mal, 70 a 110, aproximadamente 2 a 3. Con una OR aproximada de 2,5, y un intervalo que excluye el 1: con el 95% de confianza los tratados presentan una disparidad entre 1,6 y 3,8 veces mayor que la de los del grupo de control.

Es una situación **paradójica**, ya que tenemos dos estimaciones diferentes del efecto, ajustando y sin ajustar. Pero es el resultado de que X y Z vayan juntas en esta muestra: observa que la primera fila de la subtabla superior derecha incluye los tratados X+, pero también que 150 de los 180 voluntarios provienen de primaria Z_{AP}, con mejor pronóstico. Y algo parecido ocurre en la segunda fila de los controles X-: 150 de los 180 provienen del hospital Z_{HR}. Es decir, entre ambas filas, hay dos diferencias simultáneas: la exposición X y la variable Z.

La **interpretación** ahora es imprescindible: si, por determinados argumentos que no podemos imaginar, decidiéramos que el centro es una variable por la cual no hubiera que ajustar, llegaríamos a un efecto de 2.5. En cambio, si acertadamente argumentamos que los centros representan una variable previa por la cual debemos ajustar, interpretamos que la exposición X no tiene efecto en la respuesta Y.

Nota técnica. La relación entre X e Y existe y es real: conocido el valor de X, podemos anticipar, en parte, el valor de Y. Pero no es una relación causal. Hay asociación, pero no causalidad. Como la asociación es real, la estadística no dice "sesgo de confusión": habla de confusión de efectos.

Opinión. Ya hemos dicho que la tradición extendida de hacer IC sobre las estimaciones puntuales en estudios sin un proceso aleatorio detrás carece de fundamento lógico. Queremos añadir ahora que, además, tenemos muchas incertidumbres que la estadística no cuantifica: ¿Por qué variables debo ajustar? ¿Cómo debo hacerlo, de forma lineal, cuadrática o binaria? En este último caso, ¿en qué umbral se basará la dicotomía? Estas variables de ajuste ¿están determinadas sin error de medida?

Si la covariable Z está relacionada con la causa X y con la respuesta Y, entonces X tendrá sus efectos en Y confundidos con los de Z.

9.2.2. Sesgo de selección

Si estudiamos la relación entre dos variables (sean M y N) ajustando por una tercera (sea O) que es **posterior** y que está influida por M y N, aparecerá una **relación negativa**, resultado de un sesgo de selección.

Broma. Le preguntan por qué arma ese escándalo y dice que está espantando rinocerontes. Y le aclaran: "Pero si en Valladolid no hay rinocerontes", a lo que responde: "¡Por eso no hay! ¡Por eso!"

Hemos utilizado la definición de los modelos causales y de la inteligencia artificial, pero el sesgo de selección es muy frecuente y recibe muchos nombres: por ejemplo, "del superviviente". Supongamos que un libro recibe muchos "Me gusta". Antes de que la vanidad invada a los autores, que se pregunten: ¿Cuántos de los que empezaron el libro no lo terminaron y no contestaron la encuesta?

Broma. Cuentan que las amigas de Jordan Ellenberg le dijeron: "Los hombres con los cuales quedamos o son guapos o son simpáticos, es como si una cosa excluyera la otra." Y él les contestó: "Hay un sesgo de selección, ya que los feos antipáticos no han llegado a la primera cita."

Historieta. Cuentan que, como muestra del poderío de los dioses, las sacerdotisas griegas enseñaban a los visitantes los magníficos presentes de marineros que, tras rezar a los dioses, se habían salvado de terribles tormentas. Y decían: "¿Qué mayor demostración quieren del poderío de los dioses?". A lo que los visitantes escépticos respondían: "Magnífico, aunque estaría bien saber, tal vez, cuántos no se salvaron o cuántos no rezaron y se salvaron."

Si los datos que faltan difieren en distintos valores de las variables objeto de estudio, tendremos una relación aparente, *espuria*, que no se reproducirá en los estudios sin pérdidas.

Da nombre a este sesgo la **selección** de casos por una variable **posterior**.

- **Ejemplo.** Una universidad promociona a sus profesores si destacan en dos dimensiones: investigación y docencia. En su distribución (ver figura 9.2 izquierda), vemos casos por todos los lados, con una valoración entre 5 y 10. Ambos aspectos son independientes en este ejemplo hipotético: saber si alguien es bueno en investigación no informa de su capacidad docente. Los casos buenos tendrían un 10 en ambos ejes, con una suma de 20. Y los regulares, con un 5 en cada eje, sumarían 10.

Dicha universidad podría utilizar esta suma para promocionar a los profesores en cuatro categorías, como muestra la figura de la derecha. Obsérvese que, en todos ellos, hay una relación negativa: se tiende a ser bueno o en docencia o en investigación, pero no en ambas dimensiones.

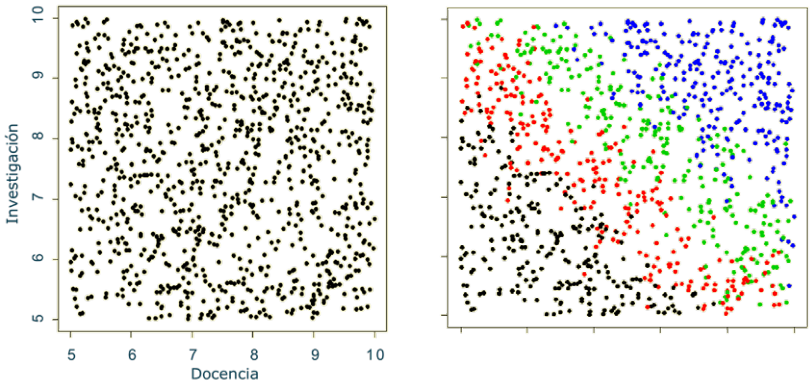


Figura 9.2. Un caso inventado para mostrar dos variables independientes (izquierda), pero con relación negativa cuando condicionamos por cierta respuesta (derecha)

Veamos un hipotético **ejemplo clínico** con dos variables binarias: los lípidos L (altos + o normales -) y un determinado gen G son independientes en toda la población (ver tabla izquierda). Ambas variables provocan eventos cardiovasculares E (tablas no mostradas). Supongamos que los pacientes que presentan dichos eventos, E+, van al hospital (ver primera subtabla de la derecha), donde recogen sus datos sobre L y G, variables previas y estables. Al analizarlos, observan disparidades aproximadas de 2 a 1 en la primera fila, y de 4 a 1 en la segunda, que dan como resultado una OR = 0,4, relación negativa: el gen previene de los lípidos altos. Y lo mismo ocurre en aquellos que no presentan eventos y no van al hospital (ver segunda subtabla de la derecha).

	L+	L-
G+	90	90
G-	90	90
OR = 1 $IC_{95\%}$:2/3 to 3/2		

En global, Gen G y Lípidos L son independientes.

E+	L+	L-	E-	L+	L-
G+	80	45	G+	10	45
G-	45	10	G-	45	80
OR≈0,4 $CI_{95\%}$: 0,18 to 0,86			OR≈0,4 $CI_{95\%}$: 0,18 to 0,86		

Dentro (E+) y fuera (E-) del hospital, gen (G) y lípidos (L) están inversamente relacionados.

Figura 9.3. Otro ejemplo fabricado para mostrar independencia (izquierda) entre dos variables previas, pero relación negativa si condicionamos por una respuesta (derecha)

Ejercicio

Como hemos dicho, no mostramos las tablas que contienen las relaciones de G y L con E. Puedes reconstruirlas. reordenando las celdas de las tablas, y calcular sus OR (en ambos casos resulta ser 5,2).

Una vez más, seleccionar por una variable posterior, relacionada con las dos variables previas, hace aparecer una relación **negativa** que no es real: hay un sesgo de selección (por negativa, queremos decir que cuando un factor presenta la modalidad - el otro tiende a presentar la +, y viceversa).

El sesgo de selección aparece al condicionar por una variable posterior relacionada con las dos en estudio. Y provoca una aparente relación negativa entre esas dos variables previas.

9.3. Diagramas causales

Los diagramas dirigidos sin ciclos (*directed acyclic graph,* DAG) permiten representar relaciones causales. Los DAG contienen flechas **dirigidas**, es decir, con inicio en una variable y final en otra. Y **no** forman **ciclos**.

► **Ejemplo.** La figura 9.4 representa tres variables sobre la presión arterial con tres relaciones causales entre los tres pares de variables. La presión inicial (basal) influye en el tratamiento y en el valor final (respuesta), que a su vez depende del tratamiento.

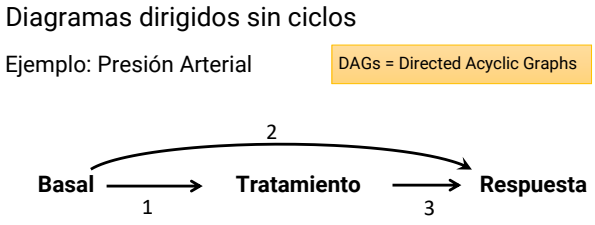


Figura 9.4. Ejemplo de diagrama causal

Las tres flechas están dirigidas (tienen principio y fin), y, si las seguimos, no permiten volver atrás: no forman ciclos, no hay bucles, ya que la presión final no influye ni en la inicial ni en el tratamiento, ambas previas.

► Ejemplo. La figura 9.5 muestra que fumar es causa común de llevar tabaco y de desarrollar cáncer. Podemos verlo, ya que salen dos flechas de la variable fumar. Pero llevar tabaco no tiene efecto causal en cáncer.

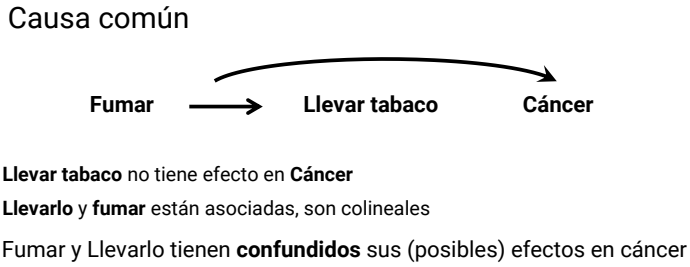


Figura 9.5. Diagrama que muestra una causa común

En cambio, por el reto de los efectos confundidos, si olvidamos el hecho de ser fumador, observaríamos una relación entre llevar tabaco y tener cáncer: quienes llevan tabaco encima tienen más *odds* de cáncer: llevar tabaco encima puede ser un indicador predictivo de la aparición de cáncer (un *chivato*).

Ahora bien, condicionar o ajustar por el hecho de fumar hace desaparecer la relación entre llevar tabaco y tener cáncer. Aquí lo llamamos **bloquear** y lo representamos por un cuadrado alrededor de la variable fijada.

Una **causa común** puede provocar la **confusión** de efectos si **no ajustamos** por esta **causa**.

Si dos flechas **salen** de una variable, **debemos condicionar** por esta variable previa, que se halla al inicio de las flechas, al estudiar la relación entre las otras dos.

Veamos ahora que, en cambio, un **efecto común** puede provocar un sesgo de **selección** si **ajustamos** por este **efecto**.

- En otro nuevo **ejemplo**, el diagrama representa dos variables con **efecto** en una misma respuesta: tanto el gen como el entorno pueden tener efecto en el cáncer. Pero el gen y el entorno son independientes entre ellos. Ahora bien, si bloqueáramos el efecto común, observaríamos una falsa relación originada por el sesgo de selección.

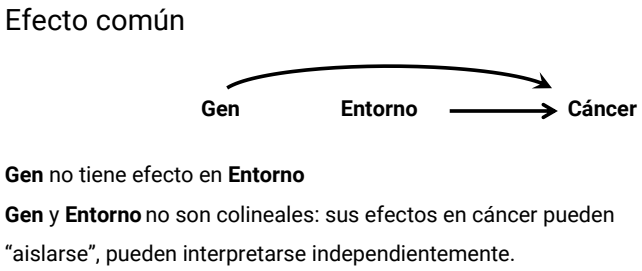


Figura 9.6. Diagrama que muestra un efecto común

Si dos flechas **llegan** a una variable posterior, **no debemos condicionar** por esta variable al final de las dos flechas cuando estudiemos la relación entre las otras dos.

Nota técnica. En terminología causal, la variable en que **colisionan** las flechas se denomina en inglés **collider**; y algunos utilizan el *collider bias* como nombre para el sesgo de selección.

Existen aplicativos gratuitos que ayudan a dibujar DAG sofisticados, con muchas variables, y orientan con las propiedades de varios análisis estadísticos, como DAGitty.

Los DAG permiten ilustrar y prevenir la confusión de efectos y el sesgo de selección.

Para interpretar causalmente una relación observacional, dibuja el DAG que explicita tus conocimientos previos.

9.4. Diagnóstico y pronóstico

El diagnóstico y el pronóstico comparten el reto estadístico de predecir la respuesta de la forma más exacta posible. La diferencia metodológica reside en el momento de recogida de esta respuesta: es simultáneo en el diagnóstico y diferido en el pronóstico.

En las píldoras siguientes, veremos las medidas más populares de la capacidad predictiva.

9.5. Medidas de la capacidad predictiva

Recuerda el ejemplo de la píldora 1.7 sobre el peso de la próxima persona que se sentará en el despacho: conocer que su altura es de 1,9 m permite perfilar mejor su peso: la predicción cambia de 70 a 90 kg y la incertidumbre, valorada con la varianza, se reduce un 51%, de 100 kg^2 a 49 kg^2.

Veamos algunas medidas para las respuestas numéricas y binarias, el reto del sobreajuste y cómo valorar el calibrado.

9.5.1. Respuesta numérica

La media minimiza la suma de los cuadrados de todos los posibles errores.

DT informa del error promedio al atribuir a cada caso el valor medio.

El coeficiente de determinación R^2 mide la reducción de los errores cuadrados de predicción y valora la reducción de la incertidumbre proporcionada por el predictor.

Ahora no preocupa si la variable predictora está relacionada causalmente con la respuesta, ni si es posible intervenir en sus valores, ni la confusión de efectos. Ni el sesgo de selección.

Ahora queremos predecir con variables fáciles de obtener y cuanto antes mejor, para poder anticipar más.

R^2 oscila entre 0 (anticipación nula) y 1 (perfecta, sin error).

Podemos utilizar R^2 para seleccionar el modelo con mejor capacidad predictiva: modelo, algoritmo, variable, escala…

► Por **ejemplo**, un cierto algoritmo informático, basado en valores conocidos una hora antes de la respuesta, tiene un R^2 del 75%. Un cierto modelo matemático, calculable un día antes, tiene un R^2 del 50%. Una cierta escala, obtenible una semana antes, un R^2 del 25%.

Valora la capacidad de anticipar respuestas numéricas con el coeficiente de determinación R^2.

9.5.2. Respuesta binaria

Si ambas variables, respuesta y predictor, son binarias (p. ej., positivo +, frente a negativo -), recurre a las probabilidades condicionadas.

Por ejemplo, las probabilidades diagnósticas. La sensibilidad y la especificidad suelen ser más fáciles de obtener en la recogida usual de datos, con dos muestras: una de enfermos, en que calculamos la sensibilidad o P(+|enfermo), y otra de sanos, en que calculamos la especificidad o P(−|sano). Pero los valores predictivos de un positivo, VP+ o P(enfermo|+), y de un negativo, VP− o P(sano|−), permiten valorar la capacidad de acierto, en términos diagnósticos.

Practica la probabilidad condicionada aquí y aquí, y VP aquí.

Si el predictor es una variable numérica (p. ej., la altura) o una escala ordinal (p. ej., escala de Rankin del ictus), puedes resumir tu capacidad de discriminación

con el área bajo la curva (*area under the curve*, AUC), que mide las características operativas del receptor (*receiver operating characteristics*, ROC).

Como el R^2, ROC oscila entre 0 y 1, aunque ahora un valor de 0,5 corresponde a la predicción de una moneda al azar.

Broma. Valores de ROC inferiores a 0,5 indican que la predicción va al revés. Como esos críticos de cine cuyas recomendaciones nos ayudan a decidir qué películas no ir a ver.

ROC valora la separación entre las distribuciones que se desea distinguir o *discriminar*.

Visualiza nuestro vídeo sobre ROC y practica sobre ROC aquí.

TRIPOD utiliza dos conceptos: discriminación y calibrado.

En resumen, valora la capacidad de anticipar respuestas binarias: *a*) para predictores también binarios, con la probabilidad condicionada, y *b*) para otros tipos, con la curva ROC.

9.6. Sobreajuste

El modelo podría adaptarse excesivamente a los datos utilizados para construirlo.

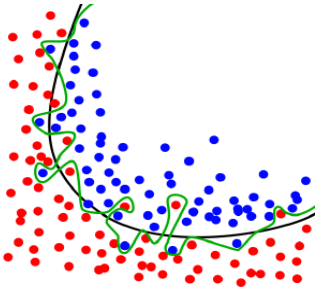


Figura 9.7. Gráfico de Thomas A. Gerd (Universidad de Copenhague), Summer School del MESIO UPC-UB

La imagen muestra una línea verde que hace una clasificación perfecta, aunque la línea negra, más simple, quizás tenga más posibilidades de replicar su capacidad predictiva en el futuro.

El riesgo de sobreajuste es mayor cuanto menor es la información disponible (p. ej., el número de casos) y cuanto más automática es la selección de variables y de sus umbrales. Por tanto, siempre conviene reestimar la capacidad predictiva en nuevos datos. Para ello, existen sofisticados métodos estadísticos, que tienen en cuenta el azar, como el remuestreo o la validación cruzada. Sin embargo, para valorar la capacidad predictiva futura en casos que no se van a seleccionar al azar, TRIPOD propone utilizar datos futuros sin selección aleatoria.

Distingue entre la muestra de **aprendizaje**, para construir el modelo, y la de **validación**, para valorar su rendimiento futuro.

9.7. Calibrado

Calibrar estudia la correspondencia entre los valores teóricos predichos y los valores medios observados. En el caso de predecir una respuesta binaria, hay que procurar que las probabilidades predichas por el modelo cuadren con las proporciones observadas.

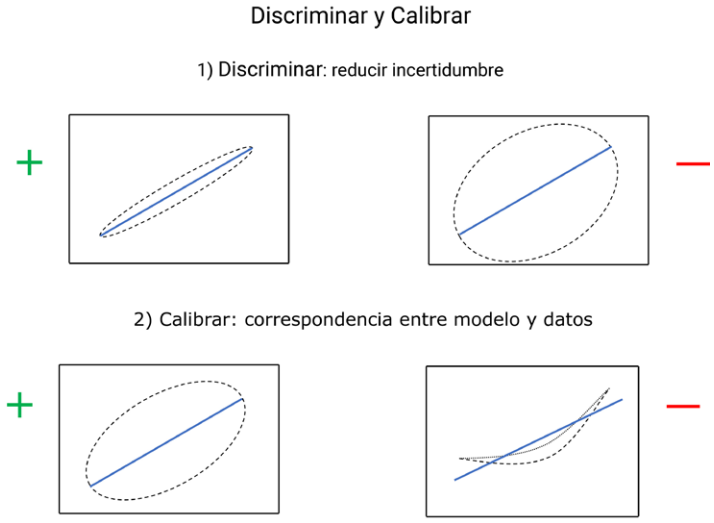


Figura 9.8. Diferencia entre discriminación y calibrado tal como se utilizan en TRIPOD: la línea azul indica el modelo y la línea puntuada negra, los datos

Los gráficos representan diversas posibilidades de modelado con una respuesta numérica: colocamos el valor predicho en las abscisas, y el valor observado en

las ordenadas. Los dos superiores muestran buena (izquierda) y mala (derecha) discriminación, y los dos inferiores, buen y mal calibrado respectivamente. El gráfico inferior derecho muestra un mal calibrado que, en cambio, permitiría una buena discriminación, ya que acertaría bastante en sus predicciones.

► **Ejemplo.** La tabla 9.1 muestra el calibrado de una escala para la aparición de neumonía posoperatoria. Proporciona la correspondencia entre el resultado predicho por el modelo ("promedio de probabilidades **predichas**", quinta fila) y las proporciones estimadas en las muestras utilizadas para generar el modelo ("**aprendizaje**", sexta fila) y para confirmar su rendimiento ("**validación**", séptima fila). Las cinco columnas muestran el riesgo predicho para distintos valores de su escala: a la izquierda, con bajo riesgo, entre 0 y 15 puntos, y, a la derecha, con alto riesgo, más de 55 puntos. Obsérvese la mayor frecuencia de casos con menor riesgo a la izquierda (casi 70.000 casos) y la muy menor a la derecha (menos de 100). Y la perfecta concordancia entre las probabilidades predichas y las proporciones observadas en la primera columna de bajo riesgo: 0,0024, así como la buena concordancia en la columna de la derecha: una probabilidad de 0,153 **predicha** por el modelo, y proporciones observadas de 0,158 y 0,159 en las muestras de **aprendizaje** y de **validación**.

Grupo de riesgo	1	2	3	4	5
Puntos en la escala original	0-15	16-25	26-40	41-55	>55
Número de pacientes en muestra aprendizaje %	69.333 43%	44.757 35%	32.103 20%	3.517 2%	95 0,1%
Promedio de las probabilidades predichas por el modelo	0,0024	0,0120	0,040	0,094	0,153
Proporción observada en la muestra de aprendizaje	0,0024	0,0119	0,040	0,094	0,158
Proporción observada en la muestra de validación	0,0024	0,0118	0,046	0,108	0,159

Tabla 9.1. La tabla adaptada de Arozullah et al. muestra un calibrado excelente

Nota histórica. Fisher observó una correspondencia similar, también excelente, entre los datos y el modelo de Mendel. Y sospechó que alguien había *maquillado* los datos.

La prueba científica del algodón es que otros autores en otros datos repliquen nuestros resultados previos.

En la guía TRIPOD se encuentran ejemplos de calibrado y su interpretación.

Opinión. Existen demasiados estudios para proponer nuevos indicadores y muy pocos para estudiar el rendimiento y la reproducibilidad de indicadores ya definidos.

En un estudio de predicción, estudia el calibrado mirando si las predicciones del modelo coinciden con los resultados observados.

Despedida

Deseamos que este texto facilite la comunicación entre investigadores teóricos y aplicados. Y que todos avancemos hacia el consenso científico.

Opinión. Sí, es muy bueno discutir. Pero no sobre los temas en los que ya hemos llegado a un acuerdo, reflejado en las guías de publicación promovidas por la red EQUATOR y por el consejo editor ICJME. Puedes mandar alternativas y mejoras a las guías a su página de contacto.

Recuerda mantenerte al día de las guías de publicación.
Ojea las novedades en EQUATOR.