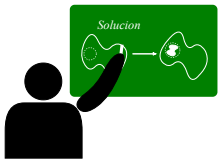
Matemáticas e Estatística II

José Carlos Díaz Ramos

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Matemáticas e Estatística II

Matemáticas e Estatística II



Contidos

0. Preliminares
1. Introdución á inferencia estadística. Estimación
   - Poboación e mostra.
   - Parámetro. Estatístico.
   - Distribución de diferentes estatísticos. Teorema central do límite.
   - Estimación puntual. Propiedades dos estimadores.
   - Estimación por intervalos de confianza: conceptos básicos. Nivel de confianza.
   - Intervalos de confianza para a media, varianza e proporción.
   - Determinación do tamaño da mostra.
2. Contrastes de hipóteses
   - Hipótese estatística. Formulación e método.
   - Tipos de erro. Criterios de decisión. Nivel crítico ou P-valor. Potencia dun contraste.
   - Interpretación dun contraste de hipóteses. Relación entre intervalos de confianza e contrastes de hipóteses.
   - Contrastes cunha mostra: para unha media, para unha proporción e para unha varianza.
3. Comparación de dúas poboacións
   - Intervalos de confianza para o cociente de varianzas, diferencia de medias e diferencia de proporcións.
   - Contrastes con dúas mostras: comparación de dúas varianzas; comparación de dúas medias (mostras independentes, mostras emparelladas); comparación de dúas proporcións.
4. A proba chi-cadrado
   - Contrastes para datos categóricos: táboas de continxencia. Test χ^2. Tablas 2×2. Deseño de estudos. Contrastes de homoxeneidade. Contrastes de independencia.
   - Contrastes de bondade de axuste: o contraste χ^2 de Pearson; o contraste de Kolmogorov-Smirnov; contrastes de normalidade.
5. Regresión e correlación

- Regresión: método de mínimos cadrados, rectas de regresión.
- Varianza total. Varianza residual e varianza explicada.
- Correlación: coeficiente de correlación linear.
- Outros modelos de regresión: o modelo exponencial e o modelo potencial.
- Contraste de hipóteses para os parámetros da regresión.

Material para o curso

Problemas para as clases interactivas.

Exames resoltos de anos anteriores.

Táboas de valores das distribucións estatísticas empregadas neste curso.

Bibliografía básica

- J. S. Milton, *Estadística para Biología y Ciencias de la Salud* 3ª Edición. McGraw-Hill Interamericana, Madrid, 2007.
- M. Samuels, J. A. Witmer, A. Schaffner, *Fundamentos de Estadística para las Ciencias de la Vida* 4ª Edición. Pearson Educación, S.A., Madrid, 2012.

Preliminares

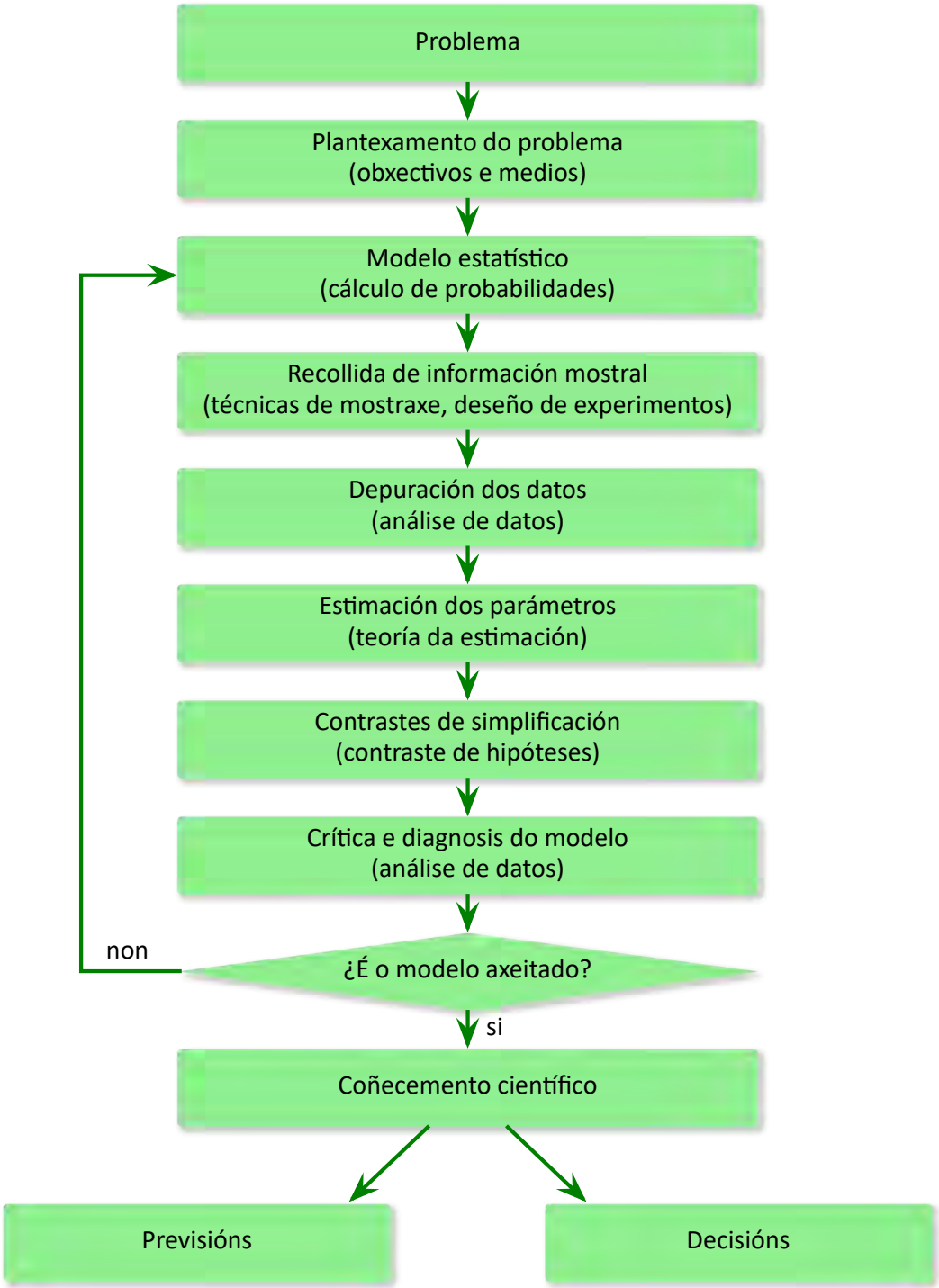


Diagrama de fluxo do método estatístico

A Estatística é a rama das Matemáticas que estuda e interpreta os procesos aleatorios, para permitir deducir propiedades dunha poboación a partir dun subconxunto pequeno da mesma. Ademais de ter un corpo

formal como parte das Matemáticas, a estatística é a miúdo empregada noutras ciencias co obxectivo de permitir establecer correlacións e dependencias entre diversos fenómenos físicos ou naturais.

Estatística descriptiva

A estatística descriptiva é a técnica matemática que organiza e describe un conxunto de datos co propósito de poder entendelos con máis facilidade. A continuación presentamos algúns conceptos relavantes na estatística descriptiva.

Tipos de datos

Os datos poden ter diversa natureza:

- **Datos nominais**, que son etiquetas para distinguir a uns de outros, como as provincias de nacemento.
- **Escalas ordinais**, nas que se asigna unha orde, pero na que o número en si non ten relevancia, como a posición dun competidor nunha liga.
- **Escalas de intervalo**, que son medicións cuantitativas nas que se mide a diferencia entre dúas variables, como a temperatura en graos Celsius.
- **Escalas de razón**, que son escalas de intervalo cun cero absoluto, como a temperatura en graos Kelvin.

Precisión

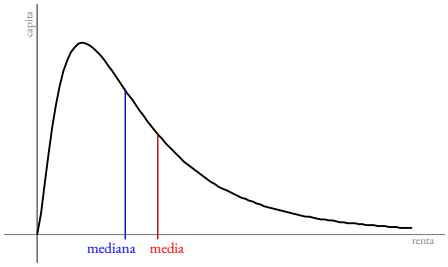
A precisión entenderémola como o número de cifras representativas empregadas para expresar unha medida. Aínda que neste curso non faremos especial fincapé neste tema, convén ter en conta que os erros de precisión nos números se van propagando a medida que imos facendo operacións aritméticas. Para certos cálculos (p. ex. o coeficiente de correlación) é necesario empregar unha cantidade suficiente de decimais para non chegar a resultados absurdos.

Medidas de tendencia central

- **Moda**: valor máis frecuente.
- **Mediana**: valor Md tal que, unha vez ordeados os datos, divide a estes pola metade.
- **Media**: é unha medida para datos obtidos como escalas de intervalo ou de razón, e que vén definida do seguinte xeito:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Observación. A media e a mediana dan lugar a medidas similares en variables que se distribúen de xeito aproximadamente normal, como as alturas e os pesos dos seres vivos dunha determinada especie, ou os erros de medición. Para outro tipo de variables poden dar resultados moi distintos. Aínda que a media goza de máis popularidade e é sinxela de entender, hai ocasións en que a mediana resulta moito máis informativa e veraz.



Por exemplo, en términos económicos, a media soe dar información moi distinta á mediana. Unha medida habitual da economía é o producto interior bruto, ou a renta per capita. Esta última, que vén a ser unha media das rentas das persoas dun país, está sesgada cara ás élites dos ricos. Se por exemplo o 90% da xente perde poder adquisitivo, pero o 10% dos ricos se convirten en moito máis ricos, é perfectamente posible que a renta per capita aumente, dando impresión de que a economía mellora, a pesar de que ó 90% da poboación lle vai peor. Non obstante, a mediana reflicte moito mellor a economía da maioría da xente, xa que nos dá a renta que divide á poboación en dúas metades do mesmo tamaño: a metade da poboación ten unha renta inferior a esa cifra, e a outra metade, superior. No caso anterior, a mediana da renta disminuiría, xa que a maior parte da xente perde poder adquisitivo. Con esta medida quedaría máis claro que é o que lle pasa á meirande parte da poboación.

Medidas de posición

- **Cuartís**: análogo á mediana, pero dividindo a distribución en cuartos Q_1, Q_2 e Q_3.
- **Percentís**: análogo á mediana e ós cuartís, pero dividindo a distribución en cen partes.

Medidas de dispersión

- **Rango**: diferencia entre o máximo e o mínimo.
- **Amplitude intercuartil**: diferencia entre Q_3 e Q_1.
- **Desviación mediana**: mediana de $|X - Md|$.
- **Varianza**:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2.$$

- **Desviación típica**: raíz cadrada da varianza.
- **Cuasi-varianza**:

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{n}{n-1}s_n^2.$$

- **Cuasi-desviación típica**: raíz cadrada da cuasi-varianza.

Salvo que se especifique o contrario, neste curso asumiremos que s denota a cuasi-desviación típica, e s^2 a cuasi-varianza.

Proposición. Séguese das fórmulas anteriores:

- A varianza, a cuasi-varianza, a desviación típica e a cuasi-desviación típica non poden ser negativas.
- Son cero se e só se tódolos datos son iguais á media.

Transformación de datos

Se $Y = aX + b$ entón,

$$\begin{aligned} \overline{Y} &= a\overline{X} + b, \\ s_{n,Y}^2 &= a^2 s_{n,X}^2, \\ s_{n,Y} &= |a| s_{n,X}. \end{aligned}$$

A miúdo se empregan cambios para modifica-la media e a varianza:

- **Puntuacións desviadas**: $x = X - \overline{X}$. Así, $\overline{x} = 0$ e $s_{n,x} = s_{n,X}$.
- **Puntuacións tipificadas**: $z = \frac{1}{s_{n,X}}(X - \overline{X})$. Así $\overline{z} = 0$ e $s_{n,z} = 1$.

Variables aleatorias

Unha variable aletoria pode describirse informalmente como unha variable que mide unha determinada característica numérica dunha poboación, de xeito que os seus valores dependen do resultado dun experimento aleatorio. Ó longo desta sección suporemos que X é unha *variable aleatoria absolutamente continua*, o que vén a querer dicir que dita variable toma os seus valores nun intervalo.

Toda variable aleatoria ten asociada unha **función de distribución** que vén dada por $F(x) = P(X \leq x)$, é dicir, $F(x)$ é a probabilidade de que a variable aleatoria X tome un valor menor ou igual ca x.

A **función de densidade** dunha variable aleatoria absolutamente continua é a derivada da súa función de distribución, $f(x) = F'(x)$.

A área baixo a gráfica da función da densidade nun determinado intervalo $[a, b]$ expresa a seguinte probabilidade:

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

En particular, $P(X \leq b) = F(b) = \int_{-\infty}^b f(x)dx$.

Neste curso calcularanse moitas veces probabilidades do estilo

$$P(X \geq x) = \int_x^\infty f(x)dx.$$

Á función $P(X \geq x) = 1 - F(x)$ tamén se lle chama *función de supervivencia* da variable aleatoria X.

A media ou **esperanza** de X defínese como

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

A media ou esperanza da distribución denótase por μ.

A **varianza** de X defínese como

$$V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx = E(X^2) - E(X)^2.$$

A varianza dunha distribución denótase por σ^2, e σ denotará a súa desviación típica.

O seguinte teorema dá unha idea de como se concentra a probabilidade dunha variable aleatoria arredor da media, sexa cal sexa a súa distribución.

Teorema. (Desigualdade de Chebyshev) Para unha variable aleatoria X satisfaise

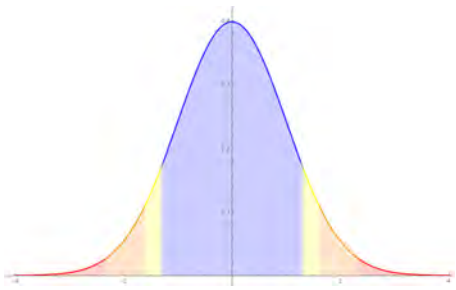
$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Por exemplo, poñendo $k = 2$ na desigualdade de Chebyshev, obtemos $P(|X-\mu| \geq 2\sigma) \leq 1/4$, é dicir, que polo menos tres cuartas partes da probabilidade dunha variable aleatoria arredor da media están contidas entre $(\mu - 2\sigma, \mu + 2\sigma)$.

Distribución normal

O exemplo máis coñecido e máis útil de variable aleatoria continua vén dado pola *distribución normal* ou campá de Gauss de media μ e desviación típica σ. Denótase por $N(\mu, \sigma)$ e está definida mediante a función de densidade

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$



Distribución normal estándar

A función de densidade da distribución normal está definida e é positiva en toda a recta real. Ademais, é simétrica respecto da súa media.

A distribución normal apareceu como un xeito de estima-las desviacións debidas a erros de medida. Tal propiedade está xustificada matematicamente polo teorema central do límite:

Teorema. (Teorema central do límite) O promedio de moitas variables aleatorias arbitrarias independentes e coa mesma distribución ten, aproximadamente, unha distribución normal.

Introdución á inferencia estadística. Estimación

Poboación e mostra

A **poboación** é o conxunto de individuos ou obxectos que queremos estudar.

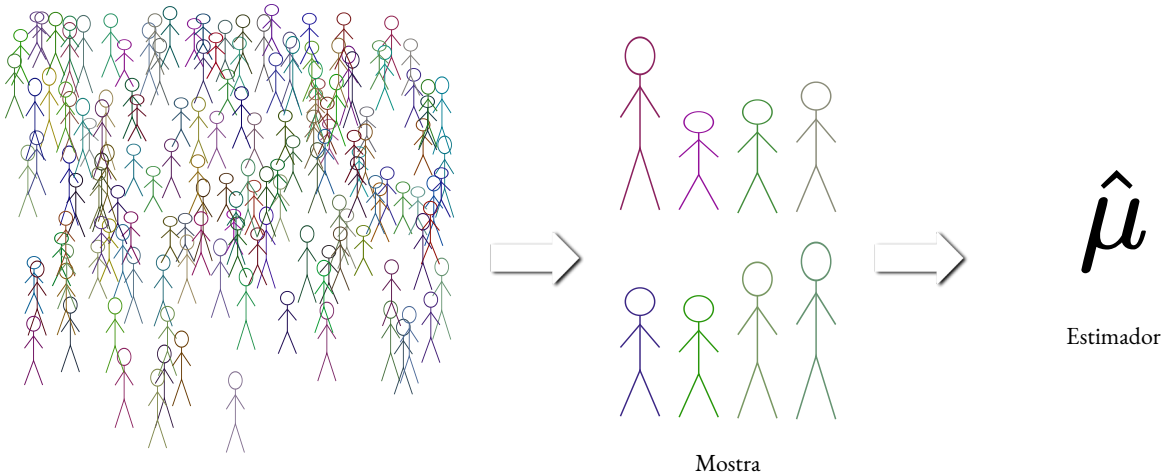
A nosa *hipótese* de partida é que a nosa poboación ten unha característica que pretendemos estudar (por exemplo, estatura, peso, etc.) que segue unha distribución da que coñecemos a súa forma xeral (modelo) pero da que descoñecemos os seus parámetros. Por exemplo, sábese que a estatura segue (aproximadamente) unha distribución normal, pero non coñecemos nin a media nin a desviación típica dunha poboación dada.

Unha mostra aleatoria é un experimento consistente en tomar n individuos da poboación. Suporemos que a mostra aleatoria se consigue extraendo individuos de xeito *independente*, de modo que tódolos individuos teñan a *mesma probabilidade de ser elexidos en cada momento*. Por tanto, construímos así n variables aleatorias $X_1, \ldots, X_n$ independentes e coa mesma distribución de probabilidade cá da poboación. Nótese que despois de face-lo experimento teremos uns valores concretos $x_1, \ldots, x_n$, pero mentres deseñámo-lo experimento eses resultados son descoñecidos e por iso son tratados como *variables aleatorias* en vez de como números; en efecto, antes de realiza-lo experimento estamos extraendo un individuo descoñecido da poboación, e por tanto, a característica que lle estudamos ten a mesma distribución cá da poboación. Dise que n é o **tamaño mostral**, e que $X_1, \ldots, X_n$ é unha **mostra aleatoria simple**.

É imposible, sen empregar teoría da probabilidade, decidir de xeito científico o tamaño mostral. Por iso diremos que este é n, e máis adiante intentaremos decidir como se calcula de xeito concreto este valor.

Un **estatístico** é unha función dunha mostra aleatoria simple que expresa unha determinada característica da mostra. Son exemplos de estatísticos a media, a varianza, a cuasivarianza e outras medidas que definimos con anterioridade.

Un **estimador puntual** é un estatístico que toma valores no espazo de parámetros. A súa misión será a de aproximar un parámetro. Un estatístico que ten como misión estimar un parámetro θ denótase $\hat{\theta}$. Por exemplo, se a poboación segue unha distribución normal $N(\mu, \sigma)$, $\hat{\mu}$ será un estimador puntual da media, e $\hat{\sigma}$ un estimador puntual da desviación típica.



Existen varios xeitos de escoller estimadores puntuais. Neste curso non enfatizarémo-la súa construcción, pero si que prestaremos atención a estimadores *insesgados* (aqueles para os que a súa media coincide co valor do parámetro que se pretende estimar) e *consistentes* (aqueles para os que o erro de medida se aproxima a cero cando o tamaño da mostra tende a infinito).

Cando temos uns datos para unha mostra concreta, un estimador puntual dános unha aproximación do parámetro que pretendemos estimar. O problema dun estimador puntual é que non temos idea de se o valor obtido está preto ou lonxe do valor real. Sería interesante ter unha idea do erro cometido coa estimación e acotar probabilisticamente ese erro. Para iso empréganse os chamados intervalos de confianza.

Chámase **intervalo de confianza** a un par de estatíticos T_1 e T_2, entre os cales se estima que estará certo parámetro descoñecido θ dunha distribución, cunha certa probabilidade de acerto determinada pola condición

$$P\big(T_1(X_1,\ldots,X_n)\leq\theta\leq T_2(X_1,\ldots,X_n)\big)\geq 1-\alpha,$$

ou ben,

$$P\Big(\theta\in\big[T_1(X_1,\ldots,X_n),\,T_2(X_1,\ldots,X_n)\big]\Big)\geq 1-\alpha,$$

onde $X_1,\ldots,X_n$ é unha mostra aleatoria simple. A probabilidade de éxito na estimación $1-\alpha$ denomínase **nivel de confianza**. Nestas circunstancias, α é o erro aleatorio ou **nivel de significación**.

Na descripción dun intervalo de confianza fálase de que a probabilidade de que un parámetro estea entre dous estatísticos sexa $1-\alpha$. Esta é a formulación correcta do problema e o xeito de construí-lo intervalo a nivel teórico. Para datos concretos dunha mostra, os estatísticos transfórmanse en dous valores entre os que se cre que o parámetro buscado está con *confianza* $1-\alpha$. Insistimos en que para valores concretos se fala de confianza, non de probabilidade. Se por exemplo $\alpha=0.1$, temos unha confianza do 90% de que o valor real se atope no intervalo calculado, é dicir, que en 90 de cada 100 mostras o intervalo conterá o valor real. Non se pode falar de probabilidade con datos concretos, xa que non hai variables aleatorias e tódolos valores son xa coñecidos.

Estimación da media poboacional

O problema que tratamos de resolver nesta sección é o de estima-la media dunha poboación que sabemos que segue unha distribución normal de media μ e desviación típica σ (que en principio son o que queremos estimar). Para iso extraemos unha mostra aleatoria simple $X_1, \ldots, X_n$.

Estimación puntual

Un xeito obvio de estima-la media da poboación é empregra-la **media da mostraxe**, é dicir, tomaremos $\hat{\mu} = \overline{X}$ onde

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Como $X_1, \ldots, X_n$ teñen a mesma distribución $N(\mu, \sigma)$ e son independentes, temos

$$E(\overline{X}) = \mu, \qquad V(\overline{X}) = \frac{\sigma^2}{n}.$$

A media mostral é un estimador *insesgado* e *consistente*.

Estimación por intervalos

Coñecida a varianza poboacional

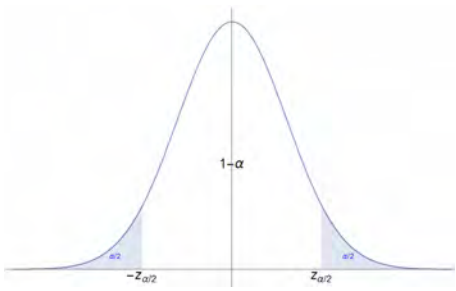
Supoñamos que a distribución poboacional segue unha distribución normal $N(\mu, \sigma)$ onde a varianza σ^2 é *coñecida*. Se $X_1, \ldots, X_n$ é unha mostra aleatoria simple, entón tomámo-lo estatístico

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim Z$$

que segue unha distribución normal estándar $Z = N(0, 1)$.

Fixemos agora un nivel de significación α (ou un nivel de confianza $1 - \alpha$).

Como a distribución normal é simétrica respecto da media, o noso intervalo de confianza tomarémolo da forma $\left[\overline{X} - \epsilon, \overline{X} + \epsilon\right]$, onde ϵ é o *erro* arredor da media que permitimos cometer. Así pois necesitamos



Valor para determina-lo intervalo de confianza

$$P(\mu \in [\overline{X} - \epsilon, \overline{X} + \epsilon]) = 1 - \alpha.$$

Tomámo-lo valor $Z_{\alpha/2}$ para o que $P(Z \geq Z_{\alpha/2}) = \alpha/2$.

Así pois témo-la cadea de igualdades

$$\begin{aligned} 1-\alpha &= P\big(|\overline{X}-\mu|\leq \epsilon\big) \\ &= P\Big(-\frac{\epsilon}{\sigma/\sqrt{n}} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{\epsilon}{\sigma/\sqrt{n}}\Big) \\ &= 1-2P\Big(Z>\frac{\epsilon}{\sigma/\sqrt{n}}\Big), \end{aligned}$$

de onde se deduce $Z_{\alpha/2}=\frac{\epsilon}{\sigma/\sqrt{n}}$. Despexando ϵ, témo-lo intervalo de confianza

$$\Big[\overline{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big].$$

Equivalentemente, resulta máis sinxelo recordar que a partir do estatístico o intervalo de confianza se obtén despexando μ da inecuación

$$\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| \leq Z_{\alpha/2},$$

ou ben,

$$-Z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}.$$

Regras para manipular inecuacións

Sexan x, y números. Supoñamos $x \leq y$. Entón;

Para calquera a,
$x+a \leq y+a$.

Se $a>0$, entón $ax \leq ay$.

Se $a<0$, entón $ax \geq ay$.

Outro xeito de escribi-lo intervalo de confianza anterior (aproveitando a simetría do mesmo) é mediante a expresión

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Problema. Desexamos estima-lo número medio de latexos por minuto para unha certa poboación. Para iso elíxense aleatoriamente 15 individuos e obtéñense os seguintes resultados:

78 95 70 97 81 85 102 75 78 85 115 80 98 101 92

Supoñendo que a distribución da poboación é normal con desviación típica de 10 latexos por minuto, calcula-lo intervalo de confianza do 99% para a media poboacional de número de latexos por minuto.

Solución. Considerámo-la variable aleatoria X="número de latexos por minuto". Temos que X ten distribución $N(\mu, 10)$, con μ descoñecido.

En primeiro lugar organizámo-los cálculos para calcula-la media mostral.

	X
	78
	95
	70
	97
	81
	85
	102
	75
	78
	85
	115
	80
	98
	101
	92
Σ	1332

Tamaño mostral $n = 15$. Estimación puntual da media $\overline{X} = 1332/15 = 88.8$ latexos.

Nivel de significación: $\alpha = 0.01$. Buscámo-lo valor $Z_{0.005}$ tal que $P(z \geq Z_{0.005}) = 0.005$. Aproximadamente, $Z_{0.005} = 2.576$.

O intervalo de confianza buscado é entón

$$88.8 \pm 2.576 \cdot \frac{10}{\sqrt{15}} = 88.8 \pm 6.65,$$

que resulta ser $[82.1, 95.5]$.

Conclusión: cunha confianza do 99%, o número medio de latexos por minuto da poboación estudada atópase entre 82.1 e 95.5. ■

Observación. En ocasións queremos limita-lo erro de estimación para que non sobrepase certo límite. En tal caso hai que tomar unha mostra suficientemente grande. Como o erro vén dado por $Z_{\alpha/2}\,\sigma/\sqrt{n}$, se queremos que sexa menor ca ϵ, entón, despexando, obtemos

$$n \geq \left(\frac{Z_{\alpha/2}\,\sigma}{\epsilon}\right)^2.$$

Observación. En caso de que a distribución da poboación non se poida garantir que sexa normal, se o tamaño da mostra é grande, o teorema central do límite dinos que podemos supoñe-la normalidade de $\overline{X}$, e por tanto, os métodos desta sección seguen sendo aproximadamente válidos. Nos apartados seguintes, se

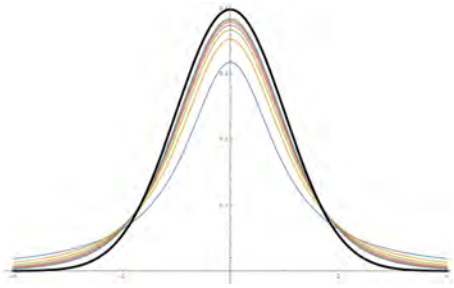
a distribución poboacional non é normal, non se aplica o teorema central do límite aínda que o tamaño da mostra sexa grande, así que neses casos habería que empregar outras técnicas que están máis aló dos obxectivos deste curso.

Descoñecida a varianza poboacional

Supoñamos agora que a distribución poboacional segue unha distribución normal $N(\mu, \sigma)$ onde a varianza σ^2 é descoñecida (o cal é o habitual). Sexa $X_1, \ldots, X_n$ é unha mostra aleatoria simple.

Recordemos que a *cuasi-varianza* ou *varianza mostral* (en contraposición a "varianza poboacional") vén definida mediante

$$\begin{aligned} s_{n-1}^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\overline{X}^2. \end{aligned}$$

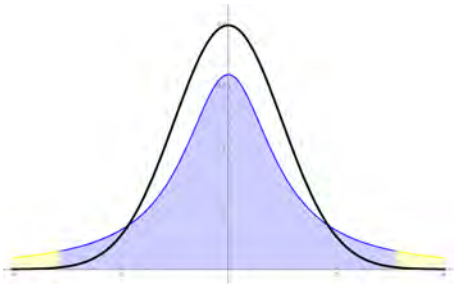


Funcións de densidade da t-Student comparadas coa normal estándar

Así, a *cuasi-desviación típica* ou *desviación típica mostral*, s_{n-1}, é a raíz cadrada da cuasi-varianza. Neste curso s denotará, salvo que se diga o contrario, a cuasi-desviación típica s_{n-1}.

Para estima-la media cando a varianza poboacional non é coñecida tómase o estatístico

$$\frac{\overline{X} - \mu}{s_{n-1}/\sqrt{n}} \sim t_{n-1}.$$



Animación das funcións de densidade de varias t-Student e as súas colas

Este estatístico resulta seguir unha distribución **t-Student de $n-1$ graos de liberdade**.

A distribución t-Student é nova e resulta ter como función de densidade

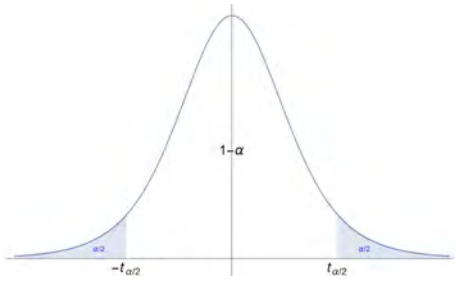
$$f(x) = c_{n-1}\left(1 + \frac{t^2}{n-1}\right)^{\frac{n}{2}},$$

sendo c_{n-1} unha constante que non especificaremos.

Proposición. Algunhas propiedades da t-Student:

- $E(t_n) = 0$ e $V(t_n) = n/(n-2)$.
- É simétrica respecto da media.
- Ten unha forma parecida á da normal, pero ten cuantiles máis grandes (por tanto produce intervalos de confianza máis grandes).
- Se $n \geq 100$, t_n pode aproximarse por unha $N(0,1)$.

Para o cálculo dun intervalo de confianza, o razoamento sería similar ó do anterior apartado. Para un nivel de significación α, o intervalo de confianza para a media vén determinado pola ecuación



Valor para determina-lo intervalo de confianza

$$-t_{n-1,\,\alpha/2} \leq \frac{\overline{X}-\mu}{s_{n-1}/\sqrt{n}} \leq t_{n-1,\,\alpha/2},$$

sendo $t_{n-1,\,\alpha/2}$ o valor tal que $P(t_{n-1} \geq t_{n-1,\,\alpha/2}) = \alpha/2$. Despexando o valor de μ obtemos

$$\left[\overline{X} - t_{n-1,\,\alpha/2}\frac{s_{n-1}}{\sqrt{n}},\, \overline{X} + t_{n-1,\,\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right],$$

ou ben,

$$\overline{X} \pm t_{n-1,\,\alpha/2}\frac{s_{n-1}}{\sqrt{n}}.$$

Problema. Os pesos ó nacer (en gramos) de 10 nenos, elexidos aleatoriamente nun hospital, son:

2750 3316 3969 2211 2806 4195 3061 3827 3572 3430

Supoñendo que a poboación segue unha distribución normal, calcular un intervalo de confianza do 95% para a media do peso ó nacer dos nenos dese hospital.

Solución. Considerámo-la variable aleatoria X="peso ó nacer". Temos que X ten distribución $N(\mu, \sigma)$, con μ e σ descoñecidos.

En primeiro lugar, organizámo-los cálculos para a media e cuasi-varianza mostrais.

	X	X^2
	2750	7562500
	3316	10995856
	3969	15752961
	2211	4888521
	2806	7873636
	4195	17598025
	3061	9369721
	3827	14645929
	3572	12759184
	3430	11764900
Σ	33137	113211233

Tamaño mostral $n = 10$. Estimación puntual da media $\overline{X} = 33137/10 = 3313.7$. A cuasi-varianza calcúlase como

$$s_n^2 = \frac{113211233}{10} - 3313.7^2 = 340516,$$
$$s_{n-1}^2 = \frac{10}{9}340516 = 378351.$$

Extraendo a raíz cadrada obtemos $s_{n-1} = 615.10$.

Nivel de significación: $\alpha = 0.05$. Buscámo-lo valor $t_{9,0.025}$ tal que $P(t_9 > t_{9,0.025}) = 0.025$. Aproximadamente, $t_{9,0.025} = 2.262$.

O intervalo de confianza buscado é entón

$$3313.7 \pm 2.262 \cdot \frac{615.10}{\sqrt{10}} = 3313.7 \pm 440.02,$$

ou explicitamente, $[2873.68, 3753.72]$.

Conclusión: cunha confianza do 95%, o peso medio ó nacer dos nenos do hospital estudado atópase entre 2873.68 e 3753.72 gramos. ■

Estimación da varianza poboacional

Nesta sección o problema será o de estima-la varianza dunha poboación que segue unha distribución normal. Tomamos unha mostra aleatoria simple $X_1, \ldots, X_n$.

Estimación puntual

Se a media da poboación é coñecida, un estimador para a varianza é

$$s_\mu^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2.$$

Entón,

$$E(s_\mu^2) = \sigma^2,$$

é dicir, que s_μ^2 é *insesgado*.

Se a media da poboación é descoñecida, o cal é o que sucede habitualmente, cabería pensar que un estimador para a varianza podería ser $s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. Isto resulta non se-la mellor idea pois

$$E(s_n^2) = \frac{n-1}{n}\sigma^2,$$

é dicir, que este estimador non é insesgado (ten tendencia a infraestima-la varianza.)

Un xeito máis correcto de estima-la varianza da poboación é emprega-la *cuasi-varianza* da mostraxe ou *varianza mostral*.

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Neste caso,

$$E(s_{n-1}^2) = \sigma^2.$$

A cuasi-varianza mostral é un estimador *insesgado*.

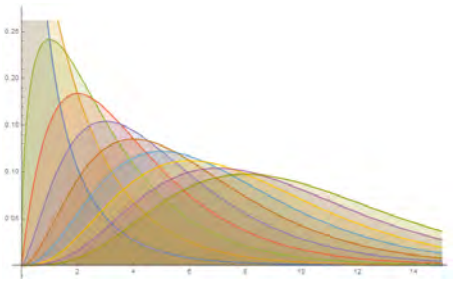
Estimación por intervalos

Coñecida a media poboacional

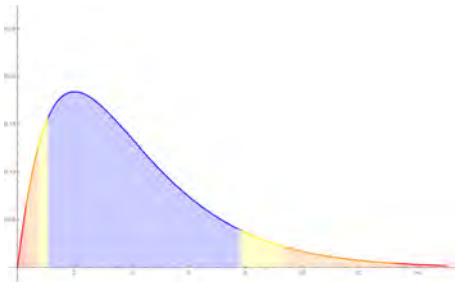
Supoñemos, aínda que normalmente non sucede, que a media poboacional μ é coñecida. É preferible, por tanto, empregar-lo estimador s_μ en lugar da cuasi-varianza mostral. En consecuencia, neste caso podemos toma-lo estatístico

$$\frac{ns_\mu^2}{\sigma^2} \sim \chi_n^2,$$

que segue unha distribución **χ-cadrado de Pearson con n graos de liberdade**.



Funcións de densidade da χ_n^2 de Pearson



Animación das funcións de densidade de varias χ^2 de Pearson e as súas colas

A distribución χ^2 de Pearson ten como función de densidade de probabilidade

$$f(x) = c_n\, x^{n/2-1}e^{-x/2},\ x > 0,$$

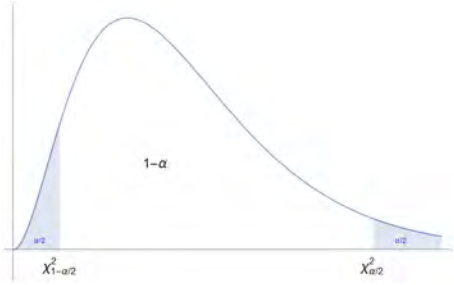
onde c_n é unha constante.

Proposición. Algunhas propiedades da χ^2 de Pearson:

- $E(\chi_n^2) = n$ e $V(\chi_n^2) = 2n$.
- Só está definida para valores positivos e non é simétrica.
- Se $n > 30$, χ_n^2 pode aproximarse por unha normal $N(n, \sqrt{2n})$; unha aproximación aínda mellor é $\sqrt{2\chi_n^2} - \sqrt{2n-1} \cong N(0,1)$.

Dado que a distribución χ^2 de Pearson non é simétrica, o intervalo de confianza que construímos tampouco o será. Fixado un nivel de significación α, buscamos dous extremos de intervalo a e b de xeito que á esquerda de a e á dereita de b quede probabilidade $\alpha/2$. É dicir, buscámo-los valores

$a = \chi^2_{n,\,1-\alpha/2}$ e $b = \chi^2_{n,\,\alpha/2}$ tales que $P(\chi^2_n \geq \chi^2_{n,\,1-\alpha/2}) = 1-\alpha/2$ e $P(\chi^2_n \geq \chi^2_{n,\,\alpha/2}) = \alpha/2$.



Valores para determina-lo intervalo de confianza

Así, o intervalo de confianza vén dado pola inecuación

$$\chi^2_{n,\,1-\alpha/2} \leq \frac{ns^2_\mu}{\sigma^2} \leq \chi^2_{n,\,\alpha/2},$$

de onde, despexando σ^2, obtemos

$$\left[\frac{ns^2_\mu}{\chi^2_{n,\,\alpha/2}},\ \frac{ns^2_\mu}{\chi^2_{n,\,1-\alpha/2}}\right].$$

Descoñecida a media poboacional

O procedemento é similar ó caso anterior, pero agora temos que emprega-la cuasi-varianza mostral.

Resulta que o estatístico

$$\frac{(n-1)s^2_{n-1}}{\sigma^2} \sim \chi^2_{n-1},$$

segue unha distribución χ^2 de Pearson con $n-1$ graos de liberdade.

Por tanto, o intervalo de confianza buscado, para unha nivel de significación α, é determinado por

$$\chi^2_{n-1,\,1-\alpha/2} \leq \frac{(n-1)s^2_{n-1}}{\sigma^2} \leq \chi^2_{n-1,\,\alpha/2}$$

Despexando σ^2 obtemos:

$$\left[\frac{(n-1)s^2_{n-1}}{\chi^2_{n-1,\,\alpha/2}},\ \frac{(n-1)s^2_{n-1}}{\chi^2_{n-1,\,1-\alpha/2}}\right].$$

Problema. Obtense unha mostra aleatoria de 100 adultos aparentemente sans co fin de establecer un patrón con respecto ó que se considerará unha lectura normal de calcio. Extráese unha mostra de sangue de cada adulto. A variable estudada é X="contido de calcio en mg/dl de sangue", que se supón que presenta unha distribución aproximadamente normal. Obtívose unha media mostral de 9.5mg/dl e unha varianza $s^2_n = 0.2475$. Calcular intervalos de confianza do 99% para a media e a desviación típica da poboación.

Solución. Considerámo-la variable aleatoria X="contido de calcio en mg/dl de sangue".

Os datos que temos no enunciado son o tamaño da mostra $n = 100$, a media mostral $\overline{X} = 9.5$ e a varianza $s^2_n = 0.2475$. A cuasi-varianza é $s^2_{n-1} = \frac{100}{99} \cdot 0.2475 = 0.25$; logo $s_{n-1} = 0.5$. O nivel

de significación é $\alpha = 0.01$.

Para o cálculo dun intervalo de confianza para a media buscámo-lo valor $t_{99,0.005} = 2.63$. Así un intervalo para a media é

$$9.5 \pm 2.63 \cdot \frac{0.5}{\sqrt{100}} = 9.5 \pm 0.13,$$

ou ben, $[9.37, 9.63]$.

A continuación pasamos á varianza. Temos que buscar *dous* valores da χ^2: $\chi^2_{99,0.005} = 138.99$ e $\chi^2_{99,0.995} = 66.51$. O intervalo de confianza para a varianza é

$$\left[\frac{99 \cdot 0.25}{138.99}, \frac{99 \cdot 0.25}{66.51}\right] = [0.18, 0.37].$$

Extraendo raíces cadradas temos un intervalo de confianza para a desviación típica $[0.42, 0.61]$.

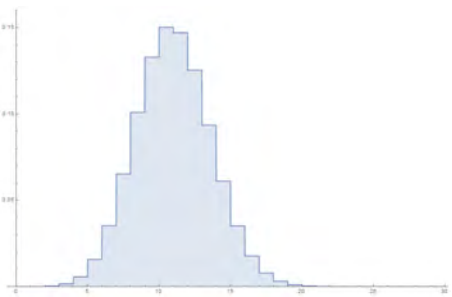
Conclusión: cunha confianza do 99%, o contido en calcio en sangue medido en mg/dl na poboación estudada ten unha media que está comprendida entre 9.37 e 9.63, e unha desviación típica entre 0.42 e 0.61. ■

Estimación dunha proporción

Supoñamos que temos unha variable con dous posibles valores. Temos unha poboación na que queremos estima-la proporción p de individuos que teñen un deses valores. Unha mostra individual desa poboación seguirá pois unha distribución de Bernoulli de parámetro p, mentres que a poboación segue unha distribución *binomial* de parámetros N (número de elementos) e p.

Recordemos que a distribución binomial de parámetros N e p é unha distribución discreta con función de masa

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}.$$



Función de masas dunha binomial (30, 0.35)

A súa media e a súa varianza son

$$E(X) = Np, \qquad V(X) = Np(1-p).$$

Estimación puntual

Queremos construír un estimador $\hat{p}$ de p. Para iso definímo-la variable aleatoria X que lle asigna 1 ó valor que queremos medir, e 0 ó outro. Escollemos unha mostra aleatoria simple $X_1, \ldots, X_n$.

É razoable toma-lo estimador puntual

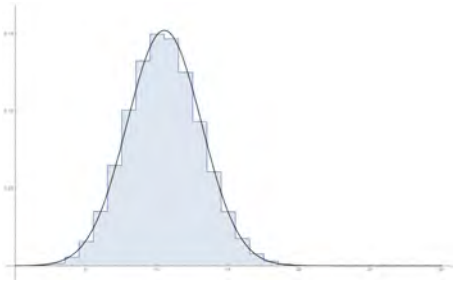
$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

que aproxima a proporción da característica que queremos medir na mostra escollida.

Temos que $n\hat{p} = \sum_{i=1}^{n} X_i$ segue unha distribución binomial de parámetros n e p. No caso de que a mostra sexa grande (con $np, n(1-p) \geq 5$ acostuma ser suficiente), podemos aproxima-la binomial por unha normal. Por tanto, habitualmente consideraremos que a distribución na mostraxe

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim Z$$

é (aproximadamente) unha $N(0,1)$.



A anterior distribución binomial comparada cunha normal da mesma media e varianza

Satisfaise que

$$E(\hat{p}) = p, \qquad V(\hat{p}) = \frac{p(1-p)}{n},$$

e por tanto, dise que $\hat{p}$ é un estimador *insesgado* e *consistente* de p.

Estimación por intervalos

O procedemento para atopar un intervalo de confianza é similar ó explicado para a media, aínda que hai algunha dificultade que presentamos a continuación. Sexa α o nivel de significación. Tomamos $Z_{\alpha/2}$ tal que $P(z \geq Z_{\alpha/2}) = \alpha/2$. En principio o cálculo dun intervalo de confianza viría expresado despexando p na fórmula

$$\left| \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| \leq Z_{\alpha/2}.$$

O problema é que o denominador $\sqrt{p(1-p)/n}$ depende de p, que é xusto o que queremos estimar. En consecuencia, aproximaremos $\sqrt{p(1-p)/n}$ por $\sqrt{\hat{p}(1-\hat{p})/n}$.

Así un intervalo de confianza para a proporción vén dado pola expresión

$$-Z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq Z_{\alpha/2}.$$

Despexando p obtemos.

$$\left[\hat{p}-Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

ou ben,

$$\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Problema. Un laboratorio desexa averigua-la proporción de cápsulas defectuosas que produce dun determinado medicamento. Para iso selecciona e proba 2000 unidades e descubre un total de 200 unidades defectuosas. Estima-la proporción de cápsulas defectuosas na produción. Calcular un intervalo de confianza ó 95% para a proporción.

Solución. Considerámo-la variable aleatoria X que asigna o valor 1 ás cápsulas defectuosas e 0 ás correctas.

Tamaño mostral $n=2000$. Estimación puntual da proporción $\hat{p}=200/2000=0.1=10\%$.

Nivel de significación: $\alpha=0.05$. Buscámo-lo valor $Z_{0.025}$ tal que $P(z\geq Z_{0.025})=0.025$. Aproximadamente, $Z_{0.025}=1.96$.

O intervalo de confianza buscado é entón

$$0.1\pm 1.96\sqrt{\frac{0.1(1-0.1)}{2000}}=0.1\pm 0.0131,$$

que explicitamente, en termos de porcentaxes, é $[8.69\%,11.31\%]$.

Conclusión: cunha confianza do 95%, a porcentaxe de cápsulas defectuosas na producción do laboratorio sitúase entre o 8.96% e o 11.31%. ■

Determinación do tamaño da mostra

En vista do intervalo de confianza construído para a proporción, o erro cometido ó tomar $\hat{p}$ en lugar do valor verdadeiro p estímase que é

$$Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

que depende do tamaño mostral n, do nivel de confianza α, e de $\sqrt{\hat{p}(1-\hat{p})}$. Se coñecemos (ou podemos estimar con precisión) o valor de $\hat{p}$, bastaría impoñer que a anterior fórmula é $<\epsilon$ e despexar n.

Cando o valor de $\hat{p}$ non é coñecido pode estimarse o tamaño da mostra necesario para limita-lo erro, se ben o valor obtido será máis grande que cando $\hat{p}$ é coñecido. No intervalo $[0,1]$ pode verse, empregando as técnicas do cálculo, que o máximo de $\sqrt{x(1-x)}$ está en $x=1/2$, de xeito que teremos sempre

$$Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq Z_{\alpha/2}\sqrt{\frac{0.5(1-0.5)}{n}}.$$

Se queremos que o erro sexa menor ca ϵ, basta entón impoñe-la condición

$$Z_{\alpha/2}\sqrt{\frac{0.5(1-0.5)}{n}} < \epsilon,$$

de onde resulta

$$n > \frac{Z_{\alpha/2}^2}{4\epsilon^2}.$$

Problema. Para toma-la decisión de someter ou non a referendo unha lei, o goberno dun certo país necesita encargar un estudo sobre a porcentaxe de votantes que a apoiaría. Dada a importancia política da mesma e a polémica xurdida, necesita unha estimación do voto cun erro menor do 1%. ¿Cal sería o tamaño mostral mínimo requerido para un nivel de confianza do 99%?

Solución. Considerámo-la variable aleatoria X="intención de voto".

Para estima-lo tamaño da mostra para unha proporción, empregámo-lo estatístico

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}},$$

que segue unha distribución normal estándar. Despexando p da desigualdade

$$\left|\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right| \leq Z_{\alpha/2},$$

obtense a fórmula

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

A estimación do erro é

$$Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Neste caso non temos unha estimación da proporción $\hat{p}$. É sinxelo ver que a función $x \mapsto \sqrt{x(1-x)}$ alcanza o seu máximo no intervalo $[0,1]$ no punto $x = 1/2$. Por tanto, necesitamos despexar n da desigualdade $Z_{\alpha/2}\sqrt{\frac{0.5(1-0.5)}{n}} \leq \epsilon$, onde ϵ é o valor fixado polo problema. Así, obtense $n \geq \left(\frac{Z_{\alpha/2}}{2\epsilon}\right)^2$.

O nivel de significación é $\alpha = 0.01$. Calculamos $Z_{0.005} = 2.5758$. Neste caso $\epsilon = 0.01$. Substituíndo na fórmula, $n \geq \left(\frac{2.5758}{2 \cdot 0.01}\right)^2 = 16587.2415$.

Conclusión: para que a diferencia entre a proporción mostral e a proporción poboacional de intención de voto sexa como moito de $\pm 0.01\%$ cun nivel de confianza do 99.0%, teriamos que tomar unha mostra de polo menos 16588 persoas. ■

Resumo de estimadores

Táboa resumo cos resultados explicados neste capítulo.

Estimación da media poboacional

Varianza poboacional coñecida

Estimador puntual

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Distribución na mostraxe

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim Z = N(0,1)$$

Inecuación

$$-Z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}$$

Intervalo de confianza

$$\left[\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Valores en táboa

$$P\big(N(0,1) \geq Z_{\alpha/2}\big) = \frac{\alpha}{2}$$

Varianza poboacional descoñecida

Estimador puntual

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Distribución na mostraxe

$$\frac{\overline{X} - \mu}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$$

Inecuación

$$-t_{n-1,\,\alpha/2} \leq \frac{\overline{X} - \mu}{s_{n-1}/\sqrt{n}} \leq t_{n-1,\,\alpha/2}$$

Intervalo de confianza

$$\left[\overline{X} - t_{n-1,\,\alpha/2}\frac{s_{n-1}}{\sqrt{n}}, \overline{X} + t_{n-1,\,\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right]$$

Valores en táboa $P\left(t_{n-1} \geq t_{n-1,\,\alpha/2}\right) = \dfrac{\alpha}{2}$

Estimación da varianza poboacional

Media poboacional coñecida

Estimador puntual

$$s_{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

Distribución na mostraxe

$$\frac{ns_{\mu}^2}{\sigma^2} \sim \chi_n^2$$

Inecuación

$$\chi_{n,\,1-\alpha/2}^2 \leq \frac{ns_{\mu}^2}{\sigma^2} \leq \chi_{n,\,\alpha/2}^2$$

Intervalo de confianza

$$\left[\frac{ns_{\mu}^2}{\chi_{n,\,\alpha/2}^2}, \frac{ns_{\mu}^2}{\chi_{n,\,1-\alpha/2}^2}\right]$$

Valores en táboa

$$P\left(\chi_n^2 \geq \chi_{n,\,\alpha/2}\right) = \frac{\alpha}{2}$$
$$P\left(\chi_n^2 \geq \chi_{n,\,1-\alpha/2}\right) = 1-\frac{\alpha}{2}$$

Media poboacional descoñecida

Estimador puntual

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

Distribución na mostraxe

$$\frac{(n-1)s_{n-1}^2}{\sigma^2} \sim \chi_{n-1}^2$$

Inecuación

$$\chi_{n-1,\,1-\alpha/2}^2 \leq \frac{(n-1)s_{n-1}^2}{\sigma^2} \leq \chi_{n-1,\,\alpha/2}^2$$

Intervalo de confianza

$$\left[\frac{(n-1)s_{n-1}^2}{\chi_{n-1,\,\alpha/2}^2}, \frac{(n-1)s_{n-1}^2}{\chi_{n-1,\,1-\alpha/2}^2}\right]$$

Valores en táboa

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

Estimación dunha proporción

Estimador puntual

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Distribución na mostraxe

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim Z = N(0,1)$$

Inecuación

$$-Z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq Z_{\alpha/2}$$

Intervalo de confianza

$$\left[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Valores en táboa

$$P\left(N(0,1) \geq Z_{\alpha/2}\right) = \frac{\alpha}{2}$$

Contraste de hipóteses

A finalidade do contraste de hipóteses é decidir se unha determinada hipótese ou afirmación sobre a distribución da poboación pode ser invalidada estatisticamente a partir das observacións contidas nunha mostra.

A hipótese sobre a distribución da poboación denomínase xenericamente **hipótese nula** e desígnase por H_0. Esta pretende contrastarse fronte a unha segunda hipótese chamada **hipótese alternativa** H_1, que agrupa a tódalas posibles poboacións nas que H_0 non é certa.

O contraste de hipóteses non ten normalmente un comportamento imparcial fronte a H_0 e H_1, xa que o problema consiste, non en decidir cal das dúas suposicións é máis verosímil en vista dos datos, senón en decidir se a mostra proporciona ou non evidencia suficiente para descartar H_0 en favor de H_1.

Nun problema de contraste de hipóteses os *dous únicos resultados posibles* consisten en *rexeitar* H_0 ou non rexeitar (ou aceptar) H_0. En xeral, o obxectivo cando se fai un contraste de hipóteses é tratar de rexeitar H_0, é dicir, de intentar dar evidencia estatística suficiente para concluír que a hipótese alternativa H_1 é certa. Por exemplo, se queremos probar estatisticamente que un determinado medicamento é útil para curar unha enfermidade, a nosa hipótese nula H_0 será formular matematicamente que o medicamento non é útil, e a nosa hipótese alternativa, que si que o é. Se conseguimos rexeitar H_0 teremos probado estatisticamente que o medicamento é útil. En caso contrario, aceptaremos H_0 e concluiremos que non hai evidencia de que o medicamento en cuestión sirva para cura-la enfermidade.

A decisión de rexeitar ou non H_0 deberá facerse en vista dos valores obtidos nunha mostra dalgún estatístico que ten unha distribución de probabilidade que, baixo a presunción de que H_0 é certa, é coñecido. Este estatístico denomínase **estatístico de contraste**.

Por tanto, un contraste de hipóteses consiste en dividi-lo espazo mostral en dúas rexións disxuntas. Unha dela chámase **rexión crítica** ou de rexeitamento, e se a mostra pertence a ela, rexéitase H_0 para inclinarse por H_1. A outra chámase **rexión de aceptación**, na que H_0 é aceptada en caso de que a mostra pertenza a ela.

Tal e como está presentado o problema existen dúas disxuntivas: a veracidade ou falsedade da hipótese nula, e aceptar ou rexeitar esta. Así, temos a seguinte casuística:

	H_0 é certa	H_0 é falsa
rexeitar H_0	**erro de tipo I**	decisión correcta
aceptar H_0	decisión correcta	**erro de tipo II**

Observación. Para o contraste de hipóteses resulta ás veces interesante facer un símil co sistema xurídico americano. O veredicto dun xurado con respecto a un crime ten dúas posibles decisións: "culpable" ou "non culpable". Nunca se dictamina que alguén é "inocente": a inocencia presuponse (hipótese nula H_0) e non é necesario probala. O que si é necesario probar é a culpabilidade (hipótese alternativa H_1). Neste sistema intenta minimizase que os inocentes sexan condenados (erro de tipo I), aínda a costa de que haxa culpables que queden impunes (erro de tipo II).

Como en xeral é imposible minimizar simultaneamente os tipos de erro I e II, o criterio tradicional na teoría de contrastes consiste en:

1. Fixar un límite para a probabilidade de cometer un erro de tipo I, chamado **nivel de significación**

$$\alpha = P(\text{rexeitar } H_0 \mid H_0 \text{ é certa}).$$

   A $1-\alpha$ chámaselle nivel de confianza.
2. Rexeitar todos aqueles tests que impoñen que a probabilidade de rexeitar H_0 cando sexa certa non supere o valor α do nivel de significación.
3. Entre tódolos test non excluídos anteriormente, tratar de minimiza-la probabilidade de erro de tipo II. Chámase **potencia** á probabilidade de detectar que unha hipótese é falsa,

$$\begin{aligned}\beta &= P(\text{rexeitar } H_0 \mid H_0 \text{ é falsa}) \\ &= 1 - P(\text{erro de tipo II}),\end{aligned}$$

   e por tanto preténdese maximiza-la potencia do método.

Tal procedemento outorga, en principio, prioridade a rebaixa-lo risco de erro de tipo I por debaixo do nivel de significación. De aí que o tratamento que reciben ambas hipóteses sexa asimétrico e estas non sexan intercambiables. De feito, no contraste de hipóteses considérase que H_0 é a hipótese establecida, que ten presunción de veracidade, e contra a cal é necesario esgrimir unha grande evidencia para poder invalidala. Así, emprégase un carácter conservador a favor da hipótese H_0: o nivel de significación que se fixa intenta garantir que sexa moi infrecuente rexeita-la hipótese correcta. A preocupación por deixar vixente unha hipótese nula falsa (erro de tipo II) é menor, polo que pode aceptarse nese caso un risco máis alto. En consecuencia, se o resultado dun contraste de hipóteses é acepta-la hipótese nula, debe interpretarse que as observacións non aportaron suficiente evidencia para descartala. Pola contra, se se rexeita é porque se está razoablemente seguro de que H_0 é falsa e H_1 é verdadeira.

O rango de valores α debe estar adaptado á importancia ou trascendencia do problema. A elección do nivel de significación é unha cuestión delicada e importante á que se lle debe prestar atención. Fixémonos cal é a razón de chamarlle a α "nivel de significación". Cando rexeitamos a hipótese nula, é porque obtivemos unha mostra que dá evidencia clara de que esta é falsa. Aínda cabería a posibilidade de que a mostra elixida fose "mala", no sentido de que non representa realmente a poboación. Non obstante, a probabilidade de que iso sucedese é menor ca α, e por tanto considérase moi improbable: é dificil que tal mostra aporte eses datos como consecuencia razoable das fluctuacións aleatorias debidas á súa elección. En consecuencia, decídese que a mostra é *significativa*, e rexéitase a hipótese nula.

Nos problemas estatísticos *paramétricos* nos que a distribución da poboación pertence a unha familia con parámetros nun conxunto Θ, tanto a hipótese nula como a alternativa serán especificadas mediante subconxuntos disxuntos Θ_0 e Θ_1 tales que $\Theta_0 \cup \Theta_1 = \Theta$. Deste xeito o contraste de hipóteses escríbese como

$$H_0 \colon \theta \in \Theta_0,$$
$$H_1 \colon \theta \in \Theta_1.$$

Habitualmente os contrastes de hipóteses estudados correspóndense con dúas posibles situacións: os **contrastes bilaterais** que nós tomaremos da forma $H_0 \colon \theta = \theta_0, H_1 \colon \theta \neq \theta_0$, e os **contrastes unilaterais** $H_0 \colon \theta \leq \theta_0, H_1 \colon \theta > \theta_0$ (ou coas desigualdades invertidas).

Os métodos estatísticos para o deseño de tests de hipóteses son complicados e están fóra dos obxectivos deste curso. Non obstante presentaremos os procedementos para realizar contrastes de hipóteses para poboacións normais nos que se contrastan as características máis habituais.

Contrate de hipóteses para a media da poboación

Supoñamos que temos unha determinada poboación que se rixe por unha distribución de probabilidade normal. Temos unha certa suposición sobre a media e queremos contrasta-la súa veracidade. Para iso tomamos unha mostra aleatoria simple $X_1, \ldots, X_n$.

Contrastes bilaterais

Empezamos co caso en que contrastamos un determinado valor da media. Así,

$$H_0 \colon \mu = \mu_0, \qquad H_1 \colon \mu \neq \mu_0.$$

Supoñendo que H_0 fose certa, tomámo-lo estatístico de contraste

$$\frac{\overline{X} - \mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$$

que ten, como vimos na sección dedicada ó cálculo de intervalos de confianza para a media con varianza descoñecida, unha distribución t-Student con $n-1$ graos de liberdade. (En caso de que a varianza poboacional σ fose coñecida tomariámo-lo estatístico $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim Z$, que ten distribución $Z = N(0,1)$, como consta na sección dedicada ó cálculo de intervalos de confianza para a media con varianza coñecida.)

Tomamos un nivel de significación α.

- A rexión crítica é $(-\infty, \ -t_{n-1,\,\alpha/2}) \cup (t_{n-1,\,\alpha/2}, \ +\infty)$, é dicir, cando

$$\left| \frac{\overline{X} - \mu_0}{s_{n-1}/\sqrt{n}} \right| > t_{n-1,\,\alpha/2}.$$

- A rexión de aceptación é por tanto $[-t_{n-1,\,\alpha/2},\ t_{n-1,\,\alpha/2}]$.

(En caso de que a varianza sexa coñecida, a rexión crítica é $(-\infty,\ -Z_{\alpha/2}) \cup (Z_{\alpha/2},\ +\infty)$, e a rexión de aceptación é $[-Z_{\alpha/2},\ Z_{\alpha/2}]$.)

Rexión de aceptación para un contraste bilateral

Cando o valor obtido na mostra está dentro do intervalo de aceptación, aceptamos H_0. Cando está na rexión crítica, é dicir, fóra do intervalo de aceptación, rexeitamos H_0. Neste caso sempre existe a pequena posibilidade α de que a mostra tomada non sexa representativa da poboación e cometamos un erro de tipo I (rexeitar un modelo correcto); non obstante, a probabilidade disto é pequena, e en vista dos datos deberemos de rexeita-la hipótese nula.

Problema. Estudámo-lo crecemento anual dos abetos. Cremos que o valor medio desta variable é $\mu_0 = 7.25$ Non obstante, nunha mostra de 50 árbores obtívose o valor $\overline{X} = 7.27$ e $s_{n-1} = 0.03$. ¿É este resultado compatible coa nosa suposición cun nivel de confianza do 95%?

Solución. Estudámo-la variable aleatoria X="crecemento anual dos abetos".

Neste caso témo-lo contraste de hipóteses

$$H_0\colon \mu = 7.25, \qquad H_1\colon \mu \neq 7.25.$$

O nivel de significación é $\alpha = 0.05$. Dannos como datos $n = 50$, $\overline{X} = 7.27$, $s_{n-1} = 0.03$.

Xa que a varianza da poboación non é coñecida, empregamos un estatístico $\frac{\overline{X}-\mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$ que ten distribución t-Student, e obtemos $t_{49,\,0.025} = 2.01$. Como

$$\frac{7.27 - 7.25}{0.03/\sqrt{50}} = 4.71 \notin [-2.01,\ 2.01],$$

o valor obtido está fóra do intervalo de aceptación.

Conclusión: *rexeitamos* H_0 e deducimos que hai evidencia significativa, polo menos do 95%, de que o valor medio de crecemento anual dos abetos non é $\mu_0 = 7.25$ metros. ■

Contrastes unilaterais

Neste caso a hipótese nula establece un límite superior ou inferior para a media. Así escribiremos

$$H_0\colon \mu \leq \mu_0, \qquad H_1\colon \mu > \mu_0,$$

que é un *contraste unilateral dereito*, ou ben,

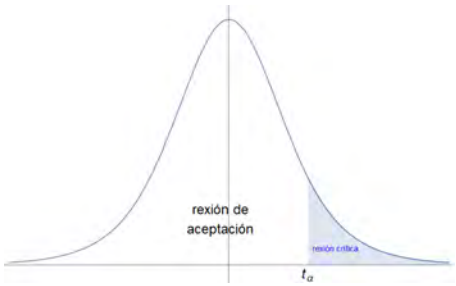
$$H_0\colon \mu \geq \mu_0, \qquad H_1\colon \mu < \mu_0,$$

para un *contraste unilateral esquerdo*.

De novo, tomámo-lo estatístico

$$\frac{\overline{X}-\mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1},$$

que ten, unha distribución t-Student con $n-1$ graos de liberdade.



Rexión de aceptación para un contraste unilateral dereito

Tomamos un nivel de significación α.

- A rexión crítica é $(t_{n-1,\,\alpha},\ +\infty)$ para un contraste unilateral dereito, e $(-\infty, -t_{n-1,\,\alpha})$ para un contraste unilateral esquerdo.
- A rexión de aceptación é por tanto $(-\infty,\ t_{n-1,\,\alpha}]$ para un contraste unilateral dereito, e $[-t_{n-1,\,\alpha},\ +\infty)$ para un contraste unilateral esquerdo.

(En caso de que a varianza poboacional σ fose coñecida tomaríamo-lo estatístico $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim Z$, que ten distribución normal $N(0,1)$. A rexión crítica é $(Z_\alpha,\ +\infty)$ para o contraste unilateral dereito, e $(-\infty,\ -Z_\alpha)$ para o contraste unilateral esquerdo. As rexións de aceptación son, respectivamente, $(-\infty,\ Z_\alpha]$ e $[-Z_\alpha,\ +\infty)$.)

Igual ca no caso anterior, cando o valor obtido na mostra está dentro do intervalo de aceptación, aceptamos H_0. Cando está na rexión crítica, é dicir, fóra do intervalo de aceptación, rexeitamos H_0.

Problema. A consellería de pesca considera que non se deben extraer ameixas se o número medio de bacterias por centímetro cúbico na auga sobrepasa 70. Como norma xeral, as rías galegas están por debaixo dese nivel de concentración. Fíxose unha mostraxe en 9 lugares da ría e obtívose un reconto $\overline{X}=71.7$, con $s_{n-1}=2.3$. ¿Que decisión deben toma-los inspectores con nivel de confianza 99%?

Solución. Considérase a variable aleatoria X="número medio de bacterias por centímetro cúbico na auga".

O contraste de hipóteses a considerar é

$$H_0\colon \mu \leq 70, \qquad H_1\colon \mu > 70.$$

O nivel de significación é $\alpha=0.01$, e temos $\mu_0=70, n=9, \overline{X}=71.7, s_{n-1}=2.3$.

Empregámo-lo estatístico $\frac{\overline{X}-\mu_0}{s_{n-1}/\sqrt{n}}$, que ten distribución t-Student, e obtemos $t_{8,0.01}=2.896$. Como

$$\frac{71.7-70}{2.3/\sqrt{9}} = 2.22 \in (-\infty,\ 2.896],$$

o valor obtido está dentro do intervalo de aceptación. Así, este número anormalmente alto non é significativo e probablemente se deba á elección da mostra.

Conclusión: *aceptamos* H_0, o cal quere dicir que non hai evidencia significativa, polo menos do 99%, de que o número medio de bacterias por centímetro cúbico de auga é menor ou igual ca 70. En consecuencia

as ameixas son aptas para o consumo. ■

Para a elección dunha hipótese nula nun contraste unilateral debe considerarse aquela desigualdade para a que se desexe minimiza-la probabilidade de erro de tipo I (rexeitar H_0 sendo certa). É dicir, que H_0 é a hipótese contra a que hai que esgrimir unha evidencia contundente para rexeitala. Recordemos que nun contraste de hipóteses aquilo que queremos probar debe estar contido na hipótese nula.

Problema. A normativa cambia e a consellería de pesca require evidencia significativa de que o número medio de bacterias por centímetro cúbico na auga sexa menor ca 70 para permiti-la extración de ameixas; é fundamental asegurarse de que tal número non é sobrepasado. Coa mostra de 9 lugares da ría obtida de $\overline{X}=71.7$, e $s_{n-1}=2.3$, ¿que decisión deben toma-los inspectores con nivel de confianza 99%? ¿E se fose $\overline{X}=68.7$?

Solución. A variable aleatoria considerada segue sendo X="número medio de bacterias por centímetro cúbico na auga".

Como agora é importante non sobrepasa-lo valor 70, e cómpre dar evidencia concluínte diso, o contraste de hipóteses a considerar é

$$H_0\colon \mu \geq 70, \qquad H_1\colon \mu < 70.$$

Igual ca antes, o nivel de significación é $\alpha=0.01$, e temos $\mu_0=70, n=9, \overline{X}=71.7, s_{n-1}=2.3$. Empregámo-lo estatístico $\frac{\overline{X}-\mu_0}{s_{n-1}/\sqrt{n}}$, que ten distribución t-Student, e obtemos $t_{8,-0.01}=2.896$. Como

$$\frac{71.7-70}{2.3/\sqrt{9}}=2.22\in[-2.896,\ +\infty),$$

o valor obtido está dentro do intervalo de aceptación.

Conclusión: *aceptamos* H_0, co que non hai evidencia significativa cunha confianza do 99% de que o número medio de bacterias por centímetro cúbico da auga sexa menor ou igual ca 70. Por tanto, hai que *prohibi-la extracción de ameixa*.

Nota: en realidade resulta superfluo facer un contraste de hipóteses para este caso, xa que a mostra non dá evidencia en contra da hipótese nula (satisfai $\overline{X}\leq\mu_0=70$). Non obstante, vemos que os cálculos claramente confirman esta afirmación.

Para $\overline{X}=68.7$ teriamos

$$\frac{68.7-70}{2.3/\sqrt{9}}=-1.696\in[-2.896,\ +\infty),$$

co que aínda neste caso *aceptamos* H_0.

Conclusión: *aceptamos* H_0, co que non hai evidencia significativa cunha confianza do 99% de que o número medio de bacterias por centímetro cúbico da auga sexa menor ou igual ca 70. Por tanto, tamén neste caso habería que *prohibi-la extracción de ameixa*.

Nótese que neste caso é importante que o nivel medio de bacterias sexa menor ca 70, e por tanto é necesario asegurarse que un valor medio pequeno na mostra non é froito do azar ó escollela. ■

O valor P ou valor crítico

Intuitivamente o **valor** P é un número que dá o grao de sorpresa que un experimento causaría nun partidario da hipótese nula. Para un contraste unilateral dereito correspóndese coa área baixo a curva da función de densidade dunha variable aleatoria X cara á dereita do valor observado polo estatístico de contraste, é dicir,

$$P = P(X \geq \text{valor no estatístico}).$$

Para un contraste unilateral esquerdo é a área baixo a curva da función de densidade cara á esquerda do valor observado polo estatístico de contraste. Por tanto, rexeitamos H_0 cando cremos que o valor P é demasiado pequeno para terse producido razoablemente polo azar.

Problema. Un estudo dun ecosistema dun bosque de folla caduca indica que o promedio neto de transformacións de nitróxeno en nitrato presenta un incremento de 2Kg por hectárea e ano. Os enxeñeiros de montes cren que unha desfoliación da maleza do bosque conduciría a un descenso dese valor. Arráncase a maleza nun área de 15 hectáreas dun bosque experimental. Límpase a área para impedi-lo crecemento. Despois dun ano determinouse o cambio de nitróxeno a nitrato, por hectárea, analizando a auga da chuvia en 15 puntos dentro do bosque. Obtivéronse os seguintes resultados: $\overline{X} = -3$, $s_{n-1} = 7.5$. ¿Proba isto que arranca-la maleza do bosque provoca un descenso no cambio medio neto de nitróxeno a nitrato por hectárea e ano?

Solución. A variable aleatoria a considerar é X="cambio neto de nitróxeno a nitrato por hectárea e ano".

O contraste a considerar é

$$H_0\colon \mu \geq 2, \qquad H_1\colon \mu < 2.$$

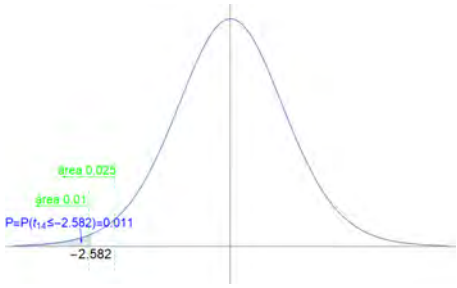
Temos como datos $\mu_0 = 2, n = 15, \overline{X} = -3, s_{n-1} = 7.5$.

Empregámo-lo estatístico de contraste $\frac{\overline{X}-\mu_0}{s_{n-1}/\sqrt{n}}$, que ten distribución t-Student.

Substituíndo,

$$\frac{-3-2}{7.5/\sqrt{15}} = -2.582.$$

Resulta entón que o P-valor é $P = P(t_{14} \leq -2.582)$. Buscando nas táboas (hai que empregar a simetría da t-Student e mira-lo valor á dereita de 2.582) obsérvase que $0.01 < P < 0.025$. Empregando software informático tense, de feito, $P = 0.011$.



valor P

Conclusión: como o valor P obtido é pequeno, *rexeitamos* H_0 e concluímos que hai evidencia significativa, polo menos do 97.5%, de que a retirada de maleza do bosque deu como resultado un incremento inferior a 2Kg por hectárea e ano da concentración media de nitróxeno en forma de nitratos.

(Nótese que, se no enunciado do problema nos tivesen pedido un nivel de confianza do 99%, teriamos que ter aceptado a hipótese nula, mentres que se o nivel de confianza fose do 97.5% teriamos que tela rexeitado. Como o valor P se aproxima bastante ó 1%, e cometer un erro de tipo I non parece que vaia ocasionar problemas graves, decidimos que o valor obtido é suficiente para rexeita-la hipótese nula.)

Cómpre enfatizar que como resultado da resolución deste problema, acabamos de probar que o *incremento de transformación de nitróxeno en nitrato é menor ca* $2Kg$. *Non* estamos probando que diminúe a transformación de nitróxeno en nitrato (a pesar de que iso é o que pasa na mostra). Se quixeramos probar isto último, teriamos que face-lo contraste de hipóteses

$$H_0\colon \mu \geq 0, \qquad H_1\colon \mu < 0.$$

Neste caso, o valor no estatístico resulta

$$\frac{-3-0}{7.5/\sqrt{15}} = -1.549,$$

e o valor P é $P = P(t_{14} < -1.549) = 0.0718$, que é un valor relativamente grande. Por tanto, para este problema teriamos que aceptar H_0, e concluiriamos que non habería evidencia significativa de que a as transformacións medias de nitróxeno en nitrato por hectárea e ano diminuísen. ■

Cando facemos contrastes bilaterais o procedemento máis común para un estatístico simétrico é considerar un valor P de dúas colas como dúas veces o valor P dunha cola. Non obstante, non existe consenso para calcula-lo valor nestes casos, especialmente se o estatístico non é simétrico.

Contraste de hipóteses para a varianza

Neste caso trátase de facer un contraste de hipótese sobre a varianza dunha poboación normal despois de ter elixido unha mostra aleatoria simple $X_1, \ldots, X_n$.

Contrastes bilaterais

O contraste de hipóteses é neste caso

$$H_0\colon \sigma^2 = \sigma_0^2, \qquad H_1\colon \sigma^2 \neq \sigma_0^2.$$

Suposto que H_0 fose certa, tomámo-lo estatístico de contraste

$$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} \sim \chi_{n-1}^2,$$

que, como vimos na sección dedicada ó cálculo de intervalos de confianza para a varianza, ten unha distribución χ^2 con $n-1$ graos de liberdade.

Para un nivel de significación α temos:

- Rexión crítica: $[0,\ \chi_{1-\alpha/2}^2) \cup (\chi_{\alpha/2}^2,\ +\infty)$.
- Rexión de aceptación: $[\chi_{1-\alpha/2}^2,\ \chi_{\alpha/2}^2]$.

Se a media é coñecida empregáse o estatítico ns_μ^2/σ_0^2 e procédese de xeito análogo.

Contrastes unilaterais

Supoñemos agora que facemos un contraste unilateral dereito. Tamén suporemos que a media é descoñecida. Se non fose así tomariámo-lo estatístico ns_μ^2/σ_0^2 e procederiamos similarmente. O contraste é entón

$$H_0\colon \sigma^2 \leq \sigma_0^2, \qquad H_1\colon \sigma^2 > \sigma_0^2.$$

Tomámo-lo mesmo estatístico de contraste ca no caso anterior, co que para un nivel de significación α temos:

- Rexión crítica: $(\chi_{n-1,\,\alpha}^2,\ +\infty)$.
- Rexión de aceptación é $[0,\ \chi_{n-1,\,\alpha}^2]$.

Nótese que como esta distribución non é simétrica, para o contraste unilateral esquerdo haberá que tomar:

- Rexión crítica: $[0,\ \chi_{n-1,\,1-\alpha}^2)$.
- Rexión de aceptación: $[\chi_{n-1,\,1-\alpha}^2,\ +\infty)$.

Ó igual que sucedía co contraste de hipóteses unilateral para a media, unha alternativa para aceptar ou rexeita-la hipótese nula é calcula-lo valor P e comprobar se este número é pequeno ou non.

Contraste de hipóteses para unha proporción

Temos unha poboación na que unha determinada propiedade se dá con probabilidade p, tomamos unha mostra aleatoria simple $X_1, \ldots, X_n$, e denotamos por $\hat{p}$ á proporción desa propiedade que se dá na mostra.

Contrastes bilaterais

O contraste de hipóteses é neste caso

$$H_0\colon p = p_0, \qquad H_1\colon p \neq p_0.$$

Suposto que H_0 fose certa, tomámo-lo estatístico de contraste

$$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim Z,$$

que, como vimos na sección dedicada ó cálculo de intervalos de confianza para a proporción, pode supoñerse que ten unha distribución normal $N(0,1)$ se o tamaño da mostra n é suficientemente grande.

Para un nivel de significación α temos:

- Rexión crítica: $(-\infty,\ -Z_{\alpha/2}) \cup (Z_{\alpha/2},\ +\infty)$, é dicir, cando $\left|\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right| > Z_{\alpha/2}$.
- Rexión de aceptación: $[-Z_{\alpha/2},\ Z_{\alpha/2}]$.

Contrastes unilaterais

Supoñemos agora que facemos un contraste unilateral dereito (para un contraste unilateral esquerdo procederíase de xeito análogo) da forma

$$H_0 \colon p \leq p_0, \qquad H_1 \colon p > p_0.$$

Tomámo-lo mesmo estatístico de contraste ca no caso anterior, co que para un nivel de significación α temos:

- Rexión crítica: $(Z_\alpha,\ +\infty)$.
- Rexión de aceptación: $(-\infty,\ Z_\alpha]$.

Ó igual ca noutros casos, unha alternativa para aceptar ou rexeita-la hipótese nula é calcula-lo valor P e comprobar se este número é pequeno ou non.

Problema. Unha empresa farmacéutica quere comercializar un medicamento que cura certa doenza. Sábese que o 40% dos doentes se curan sen toma-lo medicamento. A empresa debe probar que o seu medicamento é eficaz, e para iso adminístrao a 100 doentes, dos cales se curan 50. ¿É realmente eficaz o medicamento?

Solución. A variable aletoria a estudar é X, doentes que se curan despois de tomar certo medicamento.

A cuestión é se o medicamento cura máis ca non tomar nada. Para iso necesitarase evidencia concluínte de que na mostra se obtiveron resultados positivos. Por tanto, o contraste sobre proporcións é

$$H_0 \colon p \leq 0.4, \qquad H_1 \colon p > 0.4.$$

Temos $p_0 = 0.4, n = 100$, e tomámo-lo estatístico de contraste $\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Substituíndo os datos:

$$\frac{0.5-0.4}{\sqrt{\frac{0.4(1-0.4)}{100}}} = 2.04.$$

O valor P é por tanto $P = P(z \geq 2.04) = 0.0206$. En consecuencia, o resultado é significativo ó 5%, pero non ó 1%.

Conclusión: é dubidoso, pero como o nivel crítico é bastante pequeno, poderiamos *rexeita-la hipótese nula* e aceptar que existe evidencia significativa, polo menos do 2.5%, de que a proporción de curacións entre as persoas que toman o medicamento é maior có 40%. ■

Observación. Unha cuestión que pode ser interesante é ternos preguntado, con anterioridade a ve-los resultados da mostraxe, polo número de casos que satisfán a propiedade buscada para que haxa que rexeita-la hipótese nula.

Poñamos, por exemplo, que témo-lo contraste de hipóteses unilateral esquerdo $H_0 \colon p \geq p_0$, $H_1 \colon p < p_0$. Supoñamos que o tamaño mostral é n, e que o nivel de significación é α. Temos que calcula-lo número máximo k de individuos para os que poderiamos rexeita-la hipótese nula H_0.

Calculamos en primeiro lugar o valor Z_α. Así, para rexeitar H_0, necesitamos que o valor no estatístico estea na rexión crítica $(-\infty, -Z_\alpha)$, é dicir,

$$\frac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -Z_\alpha.$$

Despexando k na anterior inecuación obtemos

$$k < n\left(p_0 - Z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right).$$

Problema. Para analiza-lo risco de sufrir un aborto espontáneo nos embarazos de mulleres hipertensas tratadas con inhibidores de enzima convertidora de angiotensina (IECA) durante o primeiro trimestre do embarazo, estudáronse 329 casos nos que se observaron 47 abortos espontáneos. Se a taxa de abortos espontáneos na poboación fose do 10%,

1. ¿Poderíase afirmar que o tratamento con IECA no primeiro trimestre de embarazo incrementa a porcentaxe de abortos espontáneos?
2. ¿Cantos casos de abortos espontáneos terían que terse observado na mostra anterior para poder afirmar, cun nivel de significación do 0.05, que a taxa de abortos espontáneos en mulleres hipertensas sometidas a tratamento con IECA no primeiro trimestre de embarazo supera o 20%?

Solución. Sexa X a variable aleatoria "abortos espontáneos en mulleres hipertensas sometidas a tratamento con IECA no primeiro trimestre de embarazo".

Para a primeira parte do problema debemos face-lo contraste de proporcións

$$H_0 \colon p \leq 0.1, \qquad H_1 \colon p > 0.1.$$

Os datos do problema dinnos que $p_0 = 0.1$, $n = 329$ e $\hat{p} = 47/329 = 0.143$. Substituíndo no estatístico de contraste obtemos

$$\frac{0.143 - 0.1}{\sqrt{\frac{0.1(1-0.1)}{329}}} = 2.59.$$

O valor P é $P = P(Z \geq 2.59) = 0.0048$, que é menor có 0.5%.

Conclusión: *rexeitámo-la hipótese nula* e concluímos que hai evidencia significativa, polo menos do 99.5%, de que o tratamento con IECA no primeiro trimestre de embarazo provoca que a porcentaxe de abortos espontáneos sexa maior có 10%.

Para a segunda parte do problema témo-lo novo contraste de hipóteses

$$H_0: p \leq 0.2, \qquad H_1: p > 0.2.$$

O nivel de significación é $\alpha = 0.05$. Así, $Z_\alpha = 1.6449$. Por tanto necesitamos atopar k na inecuación

$$\frac{\frac{k}{329} - 0.2}{\sqrt{\frac{0.2(1-0.2)}{329}}} > 1.6449.$$

Despexando obtemos $k > 77.73$.

Conclusión: necesitariamos ter rexistrado polo menos 78 casos de abortos espontáneos nunha mostra de 329 mulleres para ter evidencia significativa, polo menos do 95%, de que a taxa de abortos espontáneos en mulleres hipertensas sometidas a tratamento on IECA no primeiro trimestre de embarazo supera o 20%. ■

Resumo de contrastes de hipóteses para unha poboación

A continuación preséntase unha táboa resumo cos resultados deste capítulo.

Contrastes para a media

Distribución na mostraxe $$\frac{\overline{X}-\mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$$

$$H_0\colon\ \mu=\mu_0, \qquad H_1\colon\ \mu\neq\mu_0.$$

Rexión crítica $$\left(-\infty,\ -t_{n-1,\,\alpha/2}\right)\cup\left(t_{n-1,\,\alpha/2},\ +\infty\right)$$

Rexión de aceptación $$\left[-t_{n-1,\,\alpha/2},\ t_{n-1,\,\alpha/2}\right]$$

$$H_0\colon\ \mu\leq\mu_0, \qquad H_1\colon\ \mu>\mu_0.$$

Rexión crítica $$\left(t_{n-1,\,\alpha},\ +\infty\right)$$

Rexión de aceptación $$\left(-\infty,\ t_{n-1,\,\alpha}\right]$$

Contrastes para a varianza

Distribución na mostraxe $$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

$$H_0\colon\ \sigma^2 = \sigma_0^2, \qquad H_1\colon\ \sigma^2 \neq \sigma_0^2.$$

Rexión crítica $$\left[0,\ \chi^2_{n-1,\,1-\alpha/2}\right) \cup \left(\chi^2_{n-1,\,\alpha/2},\ +\infty\right)$$

Rexión de aceptación $$\left[\chi^2_{n-1,\,1-\alpha/2},\ \chi^2_{n-1,\,\alpha/2}\right]$$

$$H_0\colon\ \sigma^2 \leq \sigma_0^2, \qquad H_1\colon\ \sigma^2 > \sigma_0^2.$$

Rexión crítica $$\left(\chi^2_{n-1,\,\alpha},\ +\infty\right)$$

Rexión de aceptación $$\left[0,\ \chi^2_{n-1,\,\alpha}\right]$$

$$H_0\colon\ \sigma^2 \geq \sigma_0^2, \qquad H_1\colon\ \sigma^2 < \sigma_0^2.$$

Rexión crítica $$\left[0,\ \chi^2_{n-1,\,1-\alpha}\right)$$

Rexión de aceptación $$\left(\chi^2_{n-1,\,1-\alpha},\ +\infty\right)$$

Contrastes para a proporción

Distribución na mostraxe $$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

$$H_0\colon\ p = p_0, \qquad H_1\colon\ p \neq p_0.$$

Rexión crítica $$\left(-\infty,\ -Z_{\alpha/2}\right) \cup \left(Z_{\alpha/2},\ +\infty\right)$$

Rexión de aceptación $$\left[-Z_{\alpha/2},\ Z_{\alpha/2}\right]$$

$$H_0\colon\ p \leq p_0, \qquad H_1\colon\ p > p_0.$$

Rexión crítica $$\left(Z_{\alpha},\ +\infty\right)$$

Rexión de aceptación $$\left(-\infty,\ Z_{\alpha}\right]$$

Comparación de dúas poboacións

Neste capítulo centrarémosnos no problema de comparar dúas poboacións. A situación xeral é que temos dúas poboacións de interese, e en ambas se trata de estuda-la mesma característica. A cuestión é que as dúas poboacións se atopan, por así dicilo, en circunstancias distintas, e interesa comparalas para saber como afectan esas circunstancias particulares á medida da característica que se estuda. Colleremos unha mostra aleatoria simple en cada unha desas poboacións, e a partir delas, trataremos de tomar unha decisión sobre a característica que estamos estudando.

En principio preséntanse dúas posibilidades para as mostras: que sexan independentes, ou que estean emparelladas.

Se as mostras son *independentes*, entón temos dúas poboacións para as cales estudamos respectivamente dúas variables aleatorias X e Y *independentes* cunhas distribucións de probabilidade que pertencen á mesma familia.

Extraemos unha mostra aleatorias simple $X_1, \ldots, X_{n_1}$ da primeira poboación, e $Y_1, \ldots, Y_{n_2}$ da segunda. Supoñemos que as dúas mostras son *independentes*, é dicir, que os obxectos ou individuos da mostra da primeira poboación non teñen relación algunha cos da segunda. Nótese que os tamaños mostrais n_1 e n_2 non teñen por que ser iguais.

Cando as mostras están emparelladas, o procedemento é distinto e tratarase máis adiante neste capítulo.

Comparación das medias de dúas poboacións con mostras independentes

Supoñamos que X e Y seguen as dúas unha distribución normal, $N(\mu_1, \sigma_1)$ e $N(\mu_2, \sigma_2)$, respectivamente. Un xeito de compara-las medias poboacionais é restalas e compara-la súa diferencia. (Aínda que as distribucións non sexan normais, se a mostra é suficientemente grande recordemos que podemos asumi-los resultados que seguen en virtude do teorema central do límite.)

En primeiro lugar centrámonos na estimación puntual. Se as medias mostrais son $\overline{X}$ e $\overline{Y}$ respectivamente, entón está claro que un estimador para a diferencia das medias é $\widehat{\mu_1 - \mu_2} = \overline{X} - \overline{Y}$.

Para o resto de consideracións desta sección temos varios casos.

Coñecidas as varianzas poboacionais

Se σ_1 e σ_2 son coñecidas, cousa que habitualmente non sucede, entón o estatístico

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim Z,$$

ten unha distribución normal estándar.

Coñecido tal estatístico podemos tanto calcular intervalos de confianza (seguindo o mesmo procedemento estudado no capítulo dedicado a intervalos de confianza), como facer contrastes de hipótese (seguindo o procedemento do capítulo dedicado a contrastes de hipóteses).

Intervalos de confianza

Por exemplo, un intervalo de confianza de nivel de significación α para a diferencia de medias vén determinado pola expresión

$$\left|\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right|\leq Z_{\alpha/2}.$$

Por tanto, tal intervalo é da forma

$$\left[(\overline{X}-\overline{Y})-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\,(\overline{X}-\overline{Y})+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right].$$

ou ben,

$$(\overline{X}-\overline{Y})\pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

Contraste de hipóteses

Para un contraste bilateral

$$H_0\colon \mu_1-\mu_2=(\mu_1-\mu_2)_0,\qquad H_1\colon \mu_1-\mu_2\neq(\mu_1-\mu_2)_0,$$

con nivel de significación α, temos:

- Rexión crítica: $(-\infty,\ -Z_{\alpha/2})\cup(Z_{\alpha/2},\ +\infty)$.
- Rexión de aceptación: $[-Z_{\alpha/2},\,Z_{\alpha/2}]$.

Para un contraste unilateral

$$H_0\colon \mu_1-\mu_2\leq(\mu_1-\mu_2)_0,\qquad H_1\colon \mu_1-\mu_2>(\mu_1-\mu_2)_0,$$

con nivel de significación α, temos:

- Rexión crítica: $(Z_\alpha,\ +\infty)$.

- Rexión de aceptación: $(-\infty,\ Z_\alpha]$.

Para un contraste unilateral esquerdo procederíase de xeito análogo.

Descoñecidas as varianzas poboacionais, pero supostas iguais

Supoñamos agora que $\sigma^2 = \sigma_1^2 = \sigma_2^2$ é a varianza (coincidente) das dúas poboacións. Non obstante, σ é descoñecida.

En primeiro lugar temos que estima-la varianza. Para iso temos dous estimadores da mesma cantidade, $s_1^2 := s_{X,\,n_1-1}^2$ e $s_2^2 := s_{Y,\,n_2-1}^2$, obtidos a partir das dúas mostras. Para uni-la información obtida por ambos, calculámo-la *cuasi-varianza mostral conxunta*

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2},$$

que é a media ponderada das cuasi-varianzas das mostras.

Entón, o estatístico

$$\frac{(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

ten unha distribución t-Student con n_1+n_2-2 graos de liberdade.

Observación. Para determinar se podemos considera-la varianzas de dúas poboacións iguais ou non, unha posibilidade é empregar un test de hipóteses sobre a varianza, tal e como se describe na sección dedicada á comparación das varianzas de dúas poboacións con mostras independentes.

Outra posibilidade empregada habitualmente consiste en aceptar que $\sigma_1 = \sigma_2$ cando se ten

$$\frac{1}{2} \leq \frac{s_1}{s_2} \leq 2,$$

é dicir, cando ningunha das cuasi-desviacións típicas é máis do doble da outra.

Intervalos de confianza

Agora un intervalo de confianza de nivel de significación α para a diferencia de medias vén determinado pola expresión

$$\left|\frac{(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right| \leq t_{n_1+n_2-2,\,\alpha/2}.$$

Por tanto, tal intervalo é da forma (omitímo-los graos de liberdade)

$$\left[(\overline{X}-\overline{Y})-t_{\alpha/2}\,s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\,(\overline{X}-\overline{Y})+t_{\alpha/2}\,s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right],$$

ou ben,

$$(\overline{X}-\overline{Y})\pm t_{\alpha/2}\,s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Contraste de hipóteses

Para un contraste bilateral

$$H_0\colon \mu_1-\mu_2=(\mu_1-\mu_2)_0,\qquad H_1\colon \mu_1-\mu_2\neq(\mu_1-\mu_2)_0,$$

con nivel de significación α, temos:

- Rexión crítica: $(-\infty,\ -t_{\alpha/2})\cup(t_{\alpha/2},\ +\infty)$.
- Rexión de aceptación: $[-t_{\alpha/2},\,t_{\alpha/2}]$.

Para un contraste unilateral

$$H_0\colon \mu_1-\mu_2\leq(\mu_1-\mu_2)_0,\qquad H_1\colon \mu_1-\mu_2>(\mu_1-\mu_2)_0,$$

con nivel de significación α, temos:

- Rexión crítica: $(t_\alpha,\ +\infty)$.
- Rexión de aceptación: $(-\infty,\,t_\alpha]$.

En lugar de fixar un nivel de significación poderiamos ter calculado o valor P.

Para un contraste unilateral esquerdo procederíase de xeito análogo.

Descoñecidas as varianzas poboacionais

Se σ_1^2 e σ_2^2 son descoñecidas e non poden ser supostas iguais, entón non hai solución exacta para o problema de determina-la distribución na mostraxe de $\overline{X}-\overline{Y}$, o cal obriga a adoptar solucións aproximadas.

Cando os tamaños das mostraxes son grandes, cabe argumentar que s_1^2 e s_2^2 son boas aproximacións de σ_1^2 e σ_2^2. Así, o estatístico que empregaremos

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}},$$

terá unha distribución aproximadamente normal.

Non obstante, neste curso empregarémo-lo feito de que o estatístico anterior está mellor aproximado por unha t-Student con γ graos de liberdade, onde

$$\gamma \sim \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}.$$

Como γ ten que ser enteiro, tomarase como valor a parte enteira do valor obtido no cálculo. Esta aproximación é debida a Welch-Smith-Satterthwaite.

Intervalos de confianza

No caso máis xeral, un intervalo de confianza de nivel de significación α para a diferencia de medias vén determinado pola expresión

$$\left|\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right| \leq t_{\gamma,\,\alpha/2}.$$

Por tanto, tal intervalo é da forma

$$\left[(\overline{X}-\overline{Y})-t_{\gamma,\,\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\,(\overline{X}-\overline{Y})+t_{\gamma,\,\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right],$$

ou ben,

$$(\overline{X}-\overline{Y})\pm t_{\gamma,\,\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Contraste de hipóteses

Para un contraste bilateral

$$H_0\colon \mu_1-\mu_2=(\mu_1-\mu_2)_0, \qquad H_1\colon \mu_1-\mu_2\neq(\mu_1-\mu_2)_0,$$

con nivel de significación α, temos:

- Rexión crítica: $(-\infty,\ -t_{\gamma,\,\alpha/2})\cup(t_{\gamma,\,\alpha/2},\ +\infty)$.
- Rexión de aceptación: $\left[-t_{\gamma,\,\alpha/2},\,t_{\gamma,\,\alpha/2}\right]$.

Para un contraste unilateral

$$H_0\colon \mu_1-\mu_2\leq(\mu_1-\mu_2)_0, \qquad H_1\colon \mu_1-\mu_2>(\mu_1-\mu_2)_0,$$

con nivel de significación α, temos:

- Rexión crítica: $(t_{\gamma,\,\alpha},\ +\infty)$.
- Rexión de aceptación: $(-\infty,\ t_{\gamma,\,\alpha}]$.

En troques de fixar un nivel de significación poderiamos ter calculado o valor P como

$$P = P(t_\gamma \geq \text{valor no estatístico}),$$

e decidir, se tal valor é moi pequeno, que rexeitámo-la hipótese nula; de non ser así, aceptámola.

Para un contraste unilateral esquerdo procederíase de xeito análogo.

Problema. Un isótopo radioactivo (Sr-90) acumúlase nos ósos por medio do leite de vaca consumido. Quérese coñecer se o nivel de isótopo nos nenos é distinto ca nos adultos. Para iso tómase:

- unha mostra aleatoria de 121 nenos; obtense unha concentración media de 2.6 picocurios/g, cunha cuasi-desviación típica de 1.2.
- outra mostra de 61 adultos; para estes tense como media 0.4 e cuasi-desviación típica 0.11.

Solución. Denotemos por X á variable aleatoria que mide o nivel de isótopo nos nenos, e por Y á dos adultos. Asumímo-las notacións que vimos empregando nesta sección. Trátase pois dun problema de contraste de hipóteses bilateral, é dicir,

$$H_0\colon \mu_1 = \mu_2, \qquad H_1\colon \mu_1 \neq \mu_2.$$

Equivalentemente, miramos se a diferencia $\mu_1 - \mu_2$ é nula.

O problema dános como datos: $n_1 = 121, \overline{X} = 2.6, s_1 = 1.2, n_2 = 61, \overline{Y} = 0.4, s_2 = 0.11$.

Xa que non hai coñecemento das varianzas, e $s_1/s_2 = 1.2/0.11 = 10.9 > 2$, empregámo-la fórmula de Welch, co que substituíndo:

$$\gamma \sim \frac{\left(\frac{1.2^2}{121} + \frac{0.11^2}{61}\right)^2}{\frac{\left(1.2^2/121\right)^2}{120} + \frac{\left(0.11^2/61\right)^2}{60}} = 123.965,$$

Así que tomamos $\gamma = 124$. Realmente a distribución t_{124} é moi parecida á normal estándar.

Substituíndo no estatístico obtemos:

$$\frac{(2.6 - 0.4) - 0}{\sqrt{\frac{1.2^2}{121} + \frac{0.4^2}{61}}} = 18.255.$$

Este valor está fóra de intervalos da forma $[-t_{124,\,\alpha/2},\, t_{124,\,\alpha/2}]$ para valores moito menores a $\alpha = 0.1\%$. (Por exemplo, $t_{124,\,0.0005} = 3.37072$.) O cálculo do valor P con software informático daría (nótese que é un contraste bilateral cun estatístico simétrico) $P = 2P(t_{124} > 18.255) = 6.09 \cdot 10^{-37}$, que é un valor moi pequeno.

Conclusión: *rexeitámo-la hipótese nula* e concluímos que hai evidencia significativa, cunha confianza moi próxima ó 100%, de que o nivel de isótopo en nenos é distinto ó dos adultos. ■

Problema. En vista dos resultados obtidos no problema anterior, preguntámonos agora se hai evidencia significativa de que a media obtida para os nenos sexa maior cá dos adultos. ¿É maior ca 2 picocurios/g?

Solución. As variables aleatorias a considerar son as mesmas cá no problema anterior.

Como necesitamos evidencia concluínte de que a dos nenos é *significativamente* maior, temos que estuda-lo contraste:

$$H_0\colon \mu_1 \leq \mu_2, \qquad H_1\colon \mu_1 > \mu_2.$$

Neste caso calculámo-lo valor P substituíndo no estatístico:

$$\begin{aligned} P &= P\left(t_{124} \geq \frac{(2.6-0.4)-0}{\sqrt{\frac{1.2^2}{121}+\frac{0.4^2}{61}}}\right) \\ &= P(t_{124} \geq 18.255) = 3.0 \cdot 10^{-37}. \end{aligned}$$

Conclusión: *rexeitámo-la hipótese nula* e concluímos que hai evidencia significativa, cun nivel de confianza moi alto, de que o nivel de isótopo en nenos é superior ó dos adultos.

A última pregunta correspóndese a un contraste de hipóteses

$$H_0\colon \mu_1 - \mu_2 \leq 2, \qquad H_1\colon \mu_1 - \mu_2 > 2.$$

Procédese igual, pero agora o valor no estatístico é

$$\frac{(2.6-0.4)-2}{\sqrt{\frac{1.2^2}{121}+\frac{0.4^2}{61}}} = 1.66,$$

co cal, o valor P é $P = P(t_{124} \geq 1.66) = 0.0498$.

Agora $0.025 < P < 0.05$, así que parece que podemos rexeita-la hipótese nula con nivel de confianza do 95%, pero non do 97.5%.

Conlusión: *rexeitámo-la hipótese nula* e por tanto concluímos que hai evidencia significativa, polo menos ó 95%, de que a diferencia de concentración de isótopo Sr-90 en nenos é superior a 2 picocurios/g con respecto á dos adultos. ■

Comparación das varianzas de dúas poboacións con mostras independentes

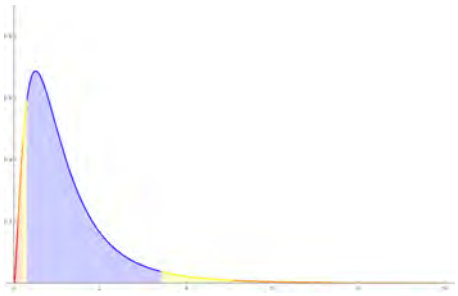
Na sección anterior vimos que é máis sinxelo, e que non hai que facer aproximacións, estima-la diferencia de medias se supoñemos que as varianzas poboacionais son iguais. A utilidade desta sección é precisamente

dar un test de hipóteses para contrastar se as varianzas poboacionais son iguais.

De novo suporemos que X e Y seguen distribucións normais, $N(\mu_1, \sigma_1)$ e $N(\mu_2, \sigma_2)$, respectivamente. Extraemos dúas mostras independentes de cada poboación, $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$.

Como anteriormente, denotamos por s_1^2 e s_2^2 as cuasi-varianzas das mostras anteriores. Para comparar σ_1^2 e σ_2^2, non é conveniente considera-lo estatístico $s_1^2 - s_2^2$. Por varias razóns é preferible empregar

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,\, n_2-1},$$



Varias densidades da F de Fisher-Snedecor

que segue unha distribución F **de Fisher-Snedecor con** $(n_1 - 1,\ n_2 - 1)$ **graos de liberdade**.

A distribución F de Snedecor depende de dous parámetros. A distribución de probabilidade da $F_{m,n}$ ten como función de densidade unha función da forma

$$f(x) = c_{m,n}\, x^{\frac{m}{2}-1}(n + mx)^{-\frac{m+n}{2}}, \quad x > 0,$$

sendo $c_{m,n}$ unha determinada constante.

Proposición. Algunhas propiedades da F de Snedecor:

- $E(F_{m,n}) = \frac{n}{n-2}$ se $n > 2$.
- Só está definida para valores positivos e non é simétrica.
- Se $F_{m,\,n,\,\alpha}$ é o valor para o que $P(F_{m,\,n} > F_{m,\,n,\,\alpha}) = \alpha$, entón

$$F_{n,\,m,\,\alpha} = \frac{1}{F_{m,\,n,\,1-\alpha}}.$$

Observación. Nótese que nas táboas da F de Snedecor empregadas neste curso, o primeiro índice é o das columnas.

Contraste de hipóteses

Estamos interesados no contraste de hipóteses

$$H_0\colon \sigma_1^2 = \sigma_2^2, \qquad H_1\colon \sigma_1^2 \neq \sigma_2^2.$$

Como por hipótese, $\sigma_1^2 = \sigma_2^2$, o estatístico anterior simplifícase e temos que empregar s_1^2/s_2^2 que, como vimos, ten distribución $F_{n_1-1,\,n_2-1}$. En consecuencia, para un nivel de significación α temos:

- Rexión crítica: $\left(0,\ \frac{1}{F_{n_2-1,\,n_1-1,\,\alpha/2}}\right) \cup \left(F_{n_1-1,\,n_2-1,\,\alpha/2},\ +\infty\right)$.

- Rexión de aceptación: $\left[\frac{1}{F_{n_2-1,\,n_1-1,\,\alpha/2}},\, F_{n_1-1,\,n_2-1,\,\alpha/2}\right]$.

Problema. Realizouse un estudo sobre as necesidades enerxéticas para o crecemento e mantemento dun niño de avións en Perthshire, Escocia. Obtivéronse os seguintes resultados para as observacións de dúas mostras independentes da variable normal "número de kilocalorías por gramo e hora que se requiren por paxaro".

Adultos incubando	Adultos precriando
$n_1 = 57$	$n_2 = 12$
$\overline{X} = 0.0167$	$\overline{Y} = 0.0144$
$s_1 = 0.0042$	$s_2 = 0.0024$

¿Indican estes datos que o número de kilocalorías requerido por adultos que están incubando é maior có requerido polos adultos que están precriando? Razoalo empregando o valor P. (Facer un contraste de hipóteses para determina-la igualdade das varianzas empregando para iso $\alpha = 0.1$.)

Solución. Sexa X a variable aleatoria que mide o número de kilocalorías por gramo e hora que se requiren por paxaro en adultos incubando, e Y a que mide o mesmo valor en adultos precriando.

Trátase dun problema de contraste de hipóteses da media de dúas poboacións, que coa notación que vimos empregando, se escribe como

$$H_0\colon \mu_1 \leq \mu_2, \qquad H_1\colon \mu_1 > \mu_2.$$

En primeiro lugar teremos que decidir se podemos supoñer que as varianzas poboacionais son ou non iguais. Isto require un contraste de hipóteses previo:

$$H_0\colon \sigma_1^2 = \sigma_2^2, \qquad H_1\colon \sigma_2^2 \neq \sigma_2^2.$$

Substituíndo no estatístico s_1^2/s_2^2 obtemos $0.0042^2/0.0024^2 = 3.0625$. Por outro lado, para $\alpha = 0.1$ temos $F_{56,11,0.05} = 2.4960$ e $F_{56,11,0.95} = 0.5091$. Claramente $3.0625 \notin [0.5091, 2.4960]$, co que debemos rexeita-la última hipótese, e por tanto non podemos supoñer que σ_1 e σ_2 sexan iguais.

Temos que empregar entón a aproximación de Welch. Primeiro calculámo-los graos de liberdade:

$$\frac{\left(\frac{0.0042^2}{57} + \frac{0.0024^2}{12}\right)^2}{\frac{(0.0042^2/57)^2}{57-1} + \frac{(0.0024^2/12)^2}{12-1}} = 27.51,$$

así que tomamos $\gamma = 27$.

Substituímos agora no estatístico:

$$\frac{(0.0167-0.0144)-0}{\sqrt{\frac{0.0042^2}{57}+\frac{0.0024^2}{12}}}=2.589.$$

Entón o valor P é $P=P(t_{27}\geq 2.589)=0.0077$, é dicir, $0.005<P<0.01$.

Conclusión: *rexeitámo-la hipótese nula* e concluímos que existe evidencia significativa, polo menos do 99%, de que o número de kilocalorías requerido por adultos incubando é maior có requerido polos adultos que están precriando. ■

Problema. Nun estudo sobre hábitos de alimentación de morcegos, márcanse 25 femias e 11 machos e rastréanse por radio. A variable de interese é "distancia que percorren voando en busca de alimento". O experimento proporciona a seguinte información:

Femias	Machos
$n_1=25$	$n_2=11$
$\overline{X}=205$	$\overline{Y}=135$
$s_1=100$	$s_2=95$

Calcular un intervalo de confianza para a diferencia de medias cun nivel de confianza do 90%.

Solución. Sexa X a variable aleatoria que mide a distancia que percorre un morcego femia voando en busca de alimento, e Y a que mide a mesma distancia para morcegos macho.

Para calcula-lo intervalo de confianza pedido, primeiro temos que decidir que estatístico empregamos. Por tanto, hai que determinar se podemos supoñer que as varianzas poboacionais poden ser consideradas iguais ou non.

Así, investigámo-lo contraste de hipóteses

$$H_0\colon \sigma_1^2=\sigma_2^2, \qquad H_1\colon \sigma_2^2\neq\sigma_2^2.$$

Subsituíndo no estatístico s_1^2/s_2^2 obtemos $100^2/95^2=1.108$. Por outro lado, para $\alpha=0.1$, $F_{24,10,0.05}=2.737$ e $F_{24,10,0.95}=0.4435$. Dado que o valor no estatístico cae entre os dous, aceptamos que as dúas varianzas poboacionais son iguais.

Podemos, por tanto, toma-la varianza conxunta

$$s_p^2=\frac{(15-1)\cdot 100^2+(11-1)\cdot 95^2}{25+11-2}=9713.24.$$

Dado que $t_{34,0.05}=1.691$ o intervalo de confianza buscado é

$$(205-135)\pm 1.691\sqrt{9713.24}\sqrt{\frac{1}{25}+\frac{1}{11}}=70\pm 60.30,$$

que despois de face-los cálculos resulta $[9.70, 130.30]$.

Conclusión: cun nivel de confianza do 90%, a diferencia entre a distancia media percorrida por un morcego femia e un morcego macho en busca de comida sitúase entre 9.70 e 130.30 metros.

Fixémonos que os resultados obtidos supoñendo que as varianzas poboacionais son iguais non é moi distinto do que se obtería se empregarámo-lo método xeral. Para o método xeral o número de graos de liberdade estímase por

$$\frac{\left(\frac{100^2}{25}+\frac{95^2}{11}\right)^2}{\frac{(100^2/25)^2}{25-1}+\frac{(95^2/12)^2}{11-1}}=20.13,$$

de xeito que tomamos $\gamma=20$.

A continuación calculamos $t_{20,0.05}=1.72472$. Por tanto, o intervalo calcúlase como

$$(205-135)\pm 1.72472\sqrt{\frac{100^2}{25}+\frac{95^2}{11}},$$

e facendo os cálculos resulta $[9.75, 130.25]$, que é moi similar ó obtido anteriormente. ■

Comparación de proporcións de dúas poboacións con mostras independentes

Supoñamos que hai dúas poboacións, nas que unha determinada propiedade se dá con probabilidades p_1 e p_2, respectivamente. Tomamos dúas mostras aleatorias simples $X_1,\ldots,X_{n_1}$ e $Y_1,\ldots,Y_{n_2}$ de cada poboación.

De novo, o estimador puntual é o esperado: $\widehat{p_1-p_2}=\hat{p}_1-\hat{p}_2$.

Tómase o estatístico

$$\frac{(\widehat{p_1-p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\sim Z,$$

que ten distribución normal $N(0,1)$.

Intervalos de confianza

Para un nivel de significación α, un intervalo de confianza para a diferencia de proporcións vén determinado pola expresión

$$\left|\frac{(\widehat{p_1-p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\right|\leq Z_{\alpha/2}.$$

Por tanto, un intervalo de confianza será da forma:

$$(\widehat{p_1 - p_2}) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Contraste de hipóteses

Chamamos *valor nulo* ó valor frente ó que contrastámo-la hipótese (o que en analoxía con outros casos chamariamos $(p_1 - p_2)_0$). Dependendo de se este valor é cero ou non, podemos facer unha pequena simplificación. En particular, se o valor nulo é cero, non tomarémo-lo mesmo estatístico que para o cálculo dun intervalo de confianza.

Valor nulo distinto de cero

Para un contraste bilateral

$$H_0\colon p_1 - p_2 = (p_1 - p_2)_0, \qquad H_1\colon p_1 - p_2 \neq (p_1 - p_2)_0,$$

con $(p_1 - p_2)_0 \neq 0$, e nivel de significación α, temos:

- Rexión crítica: $(-\infty,\ -Z_{\alpha/2}) \cup (Z_{\alpha/2},\ +\infty)$.
- Rexión de aceptación: $[-Z_{\alpha/2},\ Z_{\alpha/2}]$.

Para un contraste unilateral

$$H_0\colon p_1 - p_2 \leq (p_1 - p_2)_0, \qquad H_1\colon p_1 - p_2 > (p_1 - p_2)_0,$$

con $(\hat{p}_1 - \hat{p}_2)_0 \neq 0$ e nivel de significación α, temos:

- Rexión crítica: $(Z_{\alpha},\ +\infty)$.
- Rexión de aceptación: $(-\infty,\ Z_{\alpha}]$.

En troques de fixar un nivel de significación poderiamos calcula-lo valor P como temos feito en exemplos anteriores.

Para un contraste unilateral esquerdo procederíase de xeito análogo.

Valor nulo cero

Nun contraste bilateral escribimos

$$H_0\colon p_1 - p_2 = 0, \qquad H_1\colon p_1 - p_2 \neq 0,$$

Como por hipótese as proporcións reais son as mesmas, $\hat{p}_1$ e $\hat{p}_2$ estiman a mesma cantidade. Así, facemos como que as dúas mostras proveñen dunha mesma poboación con proporción descoñecida p, e tomámo-la media ponderada das proporcións estimadas en cada mostra, é dicir,

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

O estatístico de contraste neste caso simplifícase un pouco e tomamos

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim Z.$$

O resto das cuestións son análogas ó caso anterior.

Problema. Entre marzo e agosto de 1998 fíxose en Baltimore un ensaio clínico aleatorizado e doble cego para comproba-la eficacia do paracetamol como analxésico para trata-la migraña. Un grupo de voluntarios da zona dividiuse aleatoriamente en dous grupos. A un deles subministróuselle paracetamol e ó outro un placebo. Entre os 147 pacientes que recibiron o paracetamol, 85 notaron disminución de dor ás dúas horas de tomalo, frente a 56 de 142 no caso do grupo de control (o que tomou o placebo). A partir dos resultados deste estudo clínico, ¿hay evidencia suficiente, con nivel de confianza 99%, para afirmar que o paracetamol é un analxésico útil á hora de trata-los síntomas da migraña? ¿É a diferencia de efectividade maior ó 10%?

Solución. Sexa X a variable aleatoria que mide a eficacia do paracetamol como analxésico para trata-la migraña, Y a do placebo.

Trátese dun problema de contraste de hipóteses:

$$H_0\colon p_1-p_2\leq 0,\qquad H_1\colon p_1-p_2>0,$$

que por tanto ten valor nulo cero.

Témo-los datos: $n_1=147$, $\hat{p}_1=85/147=57.8\%$, $n_2=142$, $\hat{p}_2=56/142=39.4\%$. Así, a media ponderada das proporcións é:

$$\hat{p}=\frac{147\cdot 0.578+142\cdot 0.394}{147+142}=0.488.$$

Substituíndo no estatístico de contraste:

$$\frac{0.578-0.394}{\sqrt{0.49481\,(1-0.49481)\left(\frac{1}{147}+\frac{1}{142}\right)}}=3.126.$$

Por outra banda, $P=P(z\geq 3.126)=0.000886<0.01$ é un valor máis pequeno ca α.

Conclusión: *rexéitase a hipótese nula*, e concluímos que existe evidencia significativa, polo menos do 99%, de que o paracetamol é útil no tratamento da migraña.

Con respecto á segunda pregunta, agora témo-lo contraste de hipóteses

$$H_0\colon p_1-p_2\leq 0.1,\qquad H_1\colon p_1-p_2>0.1.$$

Como o valor nulo non é cero, substituímos no estatístico xeral para obter

$$\frac{(0.578-0.394)-0.1}{\sqrt{\frac{0.578(1-0.578)}{147}+\frac{0.394(1-0.394)}{142}}}=1.4509$$

Agora $P=P(z\geq 1.4509)=0.0734>0.01$, é dicir, $P>0.01$. Non podemos rexeita-la hipótese nula para o nivel de significación dado.

Conclusión: *aceptámo-la hipótese nula* e concluímos que non hai evidencia significativa de que a proporción de enfermos de migraña que melloran ás 2 horas de tomar paracetamol supere no 10% ós que tomaron o placebo.

En consecuencia, hai evidencias de que, ante un ataque de migraña, é mellor tomar paracetamol que non tomar nada. Non obstante, a diferencia de proporcións entre enfermos que toman paracetamol e que non toman nada, e melloran ás 2 horas, non supera o 10%. ■

Tamaño da mostra

Podemos facer un argumento similar a cando estimámo-lo tamaño da mostra para unha proporción, e así, se facemos $n\leq n_1,n_2$, para un erro máximo ϵ obtemos (vendo que o máximo de $x(1-x)$ está en $x=1/2$)

$$Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\leq Z_{\alpha/2}\sqrt{\frac{1}{4n}+\frac{1}{4n}}\leq\epsilon.$$

Despexando, $n\geq\frac{Z^2_{\alpha/2}}{2\epsilon^2}$. En consecuencia, teremos que tomar

$$n_1,n_2\geq\frac{Z^2_{\alpha/2}}{2\epsilon^2}.$$

Comparación da media con mostras emparelladas

Nesta sección estudamos un caso que se presenta con bastante frecuencia. É aquel no que as dúas mostras están emparelladas, é dicir, que para cada individuo dunha lle corresponde un da outra, asociado de xeito natural ou a propósito. Isto dáse por exemplo cando se estuda o comportamento de dous xemelgos, nais e fillas, ou o comportamento dunha persoa antes e despois de tomar un medicamento. Nestes casos faise un único sorteo e a segunda mostra dedúcese da primeira.

Os métodos das seccións anteriores *non* son aplicables xa que, para calcula-los estatísticos correspondentes, facíase uso da independencia entre mostras.

Temos, por tanto, dúas poboacións nas que medimos unha mesma característica, e denotamos por X e Y as variables aleatorias de cada unha. Tomamos $X_1,\ldots,X_n$ unha mostra aletoria simple, que está emparellada con $Y_1,\ldots,Y_n$, non independente da anterior, e do mesmo tamaño. Considerámo-la variable aleatoria $D=X-Y$, que supoñemos que está normalmente distribuída. Así, temos unha mostra das diferencias $D_1,\ldots,D_n$, onde $D_i=X_i-Y_i$.

En consecuencia, acabamos de reduci-lo problema de dúas mostras a facer inferencia sobre a súa diferencia. No caso da diferencia de medias temos que estudar entón $\mu_D = \mu_1 - \mu_2$. Como vimos, o estatístico necesario para estuda-la media é

$$\frac{\overline{D} - \mu_D}{s_D/\sqrt{n}} \sim t_{n-1},$$

que segue unha distribución t-Student con $n-1$ graos de liberdade.

O procedemento para obter intervalos de confianza e facer contrastes de hipóteses é, por tanto, o mesmo có discutido para o cálculo de intervalos de confianza para a media, e contraste de hipóteses para a media, respectivamente.

Observación. Cabe resaltar que, se ben $\mu_D = \mu_1 - \mu_2$ e $\overline{D} = \overline{X} - \overline{Y}$, pola contra $s_D^2 \neq s_1^2 \pm s_2^2$. En consecuencia, *a varianza da diferencia non pode ser deducida das varianzas das dúas variables.*

Problema. Realízase un estudo para investiga-lo efecto do exercicio no nivel de colesterol no sangue (mg/dl). Tomáronse mostras de sangue nos participantes. Despois someteuse ós individuos a un programa de exercicios, e volvéronse tomar mostras de sangue. Obtivéronse os seguintes datos:

Persoa	Nivel previo	Nivel posterior
1	182	198
2	232	210
3	191	194
4	200	220
5	148	138
6	249	220
7	276	219
8	213	161
9	241	210
10	480	313
11	262	226

Construír un intervalo de confianza para a diferencia de medias con nivel de confianza do 90%.

Solución. En primeiro lugar organizámo-los cálculos para a diferencia, o cadrado das diferencias, e sumámo-los resultados. Denotamos por X á variable aleatoria que mide o nivel de colesterol en sangue antes do exercicio, e Y o nivel despois do exercicio. Tomámo-la diferencia $D = X - Y$.

Persoa	X	Y	D	D^2
1	182	198	-16	256
2	232	210	22	484
3	191	194	-3	9
4	200	220	-20	400
5	148	138	10	100
6	249	220	29	841
7	276	219	57	3249
8	213	161	52	2704
9	241	210	31	961
10	480	313	167	27889
11	262	226	36	1296
Σ	2674	2309	365	38189

En vista da táboa temos

$$\begin{aligned}\overline{D} &= \frac{1}{n}\sum_i D_i = \frac{365}{11} = 33.182,\\ s_D^2 &= \frac{1}{n-1}\sum_i D_i^2 - \frac{n}{n-1}\overline{D}^2 = \frac{38189}{10} - \frac{11}{10}33.182^2 = 2607.76.\end{aligned}$$

Ademais, $t_{10,\,0.05} = 1.812$. Así, o intervalo de confianza vén dado por

$$33.18 \pm 1.81\frac{\sqrt{2607.76}}{\sqrt{11}},$$

o que, facendo os cálculos resulta $[5.28,\ 61.09]$.

Conclusión: cun nivel de confianza do 90%, a diferencia media de nivel de colesterol entre a xente que fai exercicio e a que non sitúase entre 5.28 e 61.09. ■

Resumo de contrastes de hipóteses para dúas poboacións

A continuación preséntase unha táboa resumo cos resultados deste capítulo ata agora.

Comparación da media de dúas poboacións

Varianzas poboacionais coñecidas

Estatístico

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim z$$

Intervalos de confianza

Inecuación

$$\left|\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right|\leq Z_{\alpha/2}$$

Fórmula

$$(\overline{X}-\overline{Y})\pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$H_0\colon\ \mu_1-\mu_2=(\mu_1-\mu_2)_0,\qquad H_1\colon\ \mu_1-\mu_2\neq(\mu_1-\mu_2)_0.$$

Rexión crítica

$$\left(-\infty,\ -Z_{\alpha/2}\right)\cup\left(Z_{\alpha/2},\ +\infty\right)$$

$$H_0\colon\ \mu_1-\mu_2\leq(\mu_1-\mu_2)_0,\qquad H_1\colon\ \mu_1-\mu_2>(\mu_1-\mu_2)_0.$$

Rexión crítica

$$\left(Z_{\alpha},\ +\infty\right)$$

Varianzas poboacionais descoñecidas pero iguais

Estatístico

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n_1+n_2-2}$$

Cuasi-varianza conxunta

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

Intervalos de confianza

Inecuación $$\left| \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right| \leq t_{n_1+n_2-2\,\alpha/2}$$

Fórmula $$(\overline{X}-\overline{Y}) \pm t_{n_1+n_2-2,\,\alpha/2}\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$H_0\colon\ \mu_1-\mu_2=(\mu_1-\mu_2)_0, \qquad H_1\colon\ \mu_1-\mu_2\neq(\mu_1-\mu_2)_0.$$

Rexión crítica $$\left(-\infty,\ -t_{n_1+n_2-2,\,\alpha/2}\right)\cup\left(t_{n_1+n_2-2,\,\alpha/2},\ +\infty\right)$$

$$H_0\colon\ \mu_1-\mu_2\leq(\mu_1-\mu_2)_0, \qquad H_1\colon\ \mu_1-\mu_2>(\mu_1-\mu_2)_0.$$

Rexión crítica $$\left(t_{n_1+n_2-2,\,\alpha},\ +\infty\right)$$

Varianzas poboacionais descoñecidas

Estatístico $$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\sim t_\gamma$$

Graos de liberdade $$\gamma\sim\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}.$$

Intervalos de confianza

Inecuación $$\left| \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \right| \leq t_{\gamma,\,\alpha/2}$$

Fórmula $$(\overline{X}-\overline{Y}) \pm t_{\gamma,\,\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$H_0\colon\ \mu_1-\mu_2=(\mu_1-\mu_2)_0, \qquad H_1\colon\ \mu_1-\mu_2\neq(\mu_1-\mu_2)_0.$$

Rexión crítica $\left(-\infty,\ -t_{\gamma,\,\alpha/2}\right) \cup \left(t_{\gamma,\,\alpha/2},\ +\infty\right)$

$$H_0\colon\ \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0, \qquad H_1\colon\ \mu_1 - \mu_2 > (\mu_1 - \mu_2)_0.$$

Rexión crítica $\left(t_{\gamma,\,\alpha},\ +\infty\right)$

Comparación da varianza de dúas poboacións

Estatístico $\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \in F_{n_1-1,\,n_2-1}$

$$H_0\colon\ \sigma_1^2 = \sigma_2^2, \qquad H_1\colon\ \sigma_1^2 \neq \sigma_2^2.$$

Rexión crítica $\left(0,\ \dfrac{1}{F_{n_2-1,\,n_1-1,\,\alpha/2}}\right) \cup \left(F_{n_1-1,\,n_2-1,\,\alpha/2},\ +\infty\right)$

Comparación de proporcións de dúas poboacións

Estatístico

$$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\sim z$$

Intervalos de confianza

Inecuación

$$\left|\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\right|\leq Z_{\alpha/2}$$

Fórmula

$$(\hat{p}_1-\hat{p}_2)\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$H_0\colon\ p_1-p_2=(p_1-p_2)_0\neq 0,\qquad H_1\colon\ p_1-p_2\neq (p_1-p_2)_0.$$

Rexión crítica

$$\left(-\infty,\ -Z_{\alpha/2}\right)\cup\left(Z_{\alpha/2},\ +\infty\right)$$

$$H_0\colon\ p_1-p_2\leq(p_1-p_2)_0\neq 0,\qquad H_1\colon\ p_1-p_2>(p_1-p_2)_0.$$

Rexión crítica

$$\left(Z_{\alpha},\ +\infty\right)$$

Valor nulo cero

Estatístico

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim z$$

Proporción ponderada

$$\hat{p}=\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$$

$$H_0\colon\ p_1=p_2,\qquad H_1\colon\ p_1\neq p_2.$$

Rexión crítica

$$\left(-\infty,\ -Z_{\alpha/2}\right)\cup\left(Z_{\alpha/2},\ +\infty\right)$$

$$H_0\colon\ p_1 \leq p_2, \qquad H_1\colon\ p_1 > p_2.$$

Rexión crítica $\left(Z_\alpha,\ +\infty\right)$

A proba chi-cadrado

A proba chi-cadrado, ou proba χ^2, é un contraste de hipóteses introducido por Pearson para determinar se a discrepancia entre as frecuencias esperadas e as frecuencias observadas nunha táboa de continxencia é estatisticamente significativa.

Neste capítulo, as variables estatísticas son discretas: só toman un número finito de valores, divididos en categorías. Distinguiremos dous tipos de tests, que computacionalmente son practicamente iguais, pero que conceptualmente son un pouco distintos.

Contrastes de independencia para datos categóricos

Supoñamos que nunha poboación estamos interesados en observar dúas características X e Y que se corresponden con datos categóricos, é dicir, que son datos nominais que soamente poden tomar valores concretos, chamados categorías. Cada valor está nunha, e só nunha, categoría (é dicir, as categorías son disxuntas). Poñamos que X pode tomar f valores distintos $A_1, \ldots, A_f$, e que Y pode tomar c valores $B_1, \ldots, B_c$. O problema ó que nos enfrentamos agora é o de determinar se as dúas características X e Y son ou non independentes. De feito, o que queremos é ver se hai (ou non) evidencia significativa de que as dúas características non son independentes.

Tomamos unha mostra aleatoria simple bidimensional (é dicir, medindo as dúas características) na poboación, $(X_1, Y_1), \ldots, (X_n, Y_n)$. Denotamos por n_{ij} ó número de observacións na mostra de tal xeito que o valor de X se atopa en A_i e o valor de Y en B_j. Os valores poden por tanto dispoñerse nunha *táboa de continxencia*, que consiste en organiza-los datos do seguinte xeito:

$X \setminus Y$	B_1	B_2	$\ldots$	B_c	Σ
A_1	n_{11}	n_{12}	$\ldots$	n_{1c}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\ldots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_f	n_{f1}	n_{f2}	$\ldots$	n_{fc}	$n_{f\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot c}$	n

Nesta táboa empregouse a notación:

$$n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ic},$$
$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{fj},$$

que son os números totais de observacións que se atopan nos conxuntos A_i e B_j respectivamente. Obviamente, $n_{1\cdot} + \cdots + n_{f\cdot} = n_{\cdot 1} + \cdots + n_{\cdot c} = n$.

As probabilidades reais da poboación son denotadas como

$$p_{ij} = P\big((X \in A_i) \cap (Y \in B_j)\big),$$

e así, por se-las categorías disxuntas,

$$\begin{aligned} p_{i\cdot} &= P(X \in A_i) = p_{i1} + p_{i2} + \cdots + p_{ic}, \\ p_{\cdot j} &= P(Y \in B_j) = p_{1j} + p_{2j} + \cdots + p_{fj}. \end{aligned}$$

Isto podería organizarse tamén nunha táboa de probabilidades:

$X \setminus Y$	B_1	B_2	$\cdots$	B_c	Σ
A_1	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_f	p_{f1}	p_{f2}	$\cdots$	p_{fc}	$p_{f\cdot}$
Σ	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot c}$	1

En caso de que as dúas características fosen realmente independentes, teriamos

$$p_{ij} = P\big((X \in A_i) \cap (Y \in B_j)\big) = P(X \in A_i) P(Y \in B_j) = p_{i\cdot}\, p_{\cdot j}$$

para calquera para i, j. En consecuencia, o contraste de hipóteses que pretendemos estudar é:

$$H_0\colon\ p_{ij} = p_{i\cdot}\, p_{\cdot j}, \quad \forall i \in \{1, \ldots, f\},\ \forall j \in \{1, \ldots, c\}.$$

Por tanto, o que resta por facer é estima-las probabilidades p_{ij} e atopar un estatístico convinte que nos permita decidir se cos valores da mostra podemos ou non descartar H_0.

A partir da mostra, os estimadores obvios das probabilidades son

$$\widehat{p_{ij}} = \frac{n_{ij}}{n}, \qquad \widehat{p_{i\cdot}} = \frac{n_{i\cdot}}{n}, \qquad \widehat{p_{\cdot j}} = \frac{n_{\cdot j}}{n}.$$

Por outra banda, baixo a hipótese de independencia, o valor da celda (i, j) da táboa de continxencia debería ser

$$E_{ij} = n p_{ij} = n\, p_{i\cdot}\, p_{\cdot j},$$

que por tanto se estima por

$$\widehat{E_{ij}} = n\, \widehat{p_{i\cdot}}\, \widehat{p_{\cdot j}} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}.$$

Para determinar se os n_{ij} están suficientemente próximos a $\widehat{E_{ij}}$, empregámo-lo estatístico

$$\sum_{i,j} \frac{(n_{ij} - \widehat{E_{ij}})^2}{\widehat{E_{ij}}} \sim \chi^2_{(f-1)(c-1)},$$

que segue aproximadamente unha distribución χ^2 de Pearson con $(f-1)(c-1)$ graos de liberdade cando a mostra é suficientemente grande.

Este contraste é unilateral dereito, así que para un nivel de significación α, temos

- Rexión crítica: $(\chi^2_{(f-1)(c-1),\,\alpha},\, \infty)$.
- Rexión de aceptación: $[0, \chi^2_{(f-1)(c-1),\,\alpha}]$.

Obviamente, tamén se podería calcula-lo valor P e rexeita-la hipótese nula cando este valor sexa moi pequeno.

Problema. Realízase un estudo para investiga-la asociación entre a cor e a fragancia das azaleas silvestres. Obsérvanse 200 prantas floridas seleccionadas aleatoriamente, e clasifícase cada unha delas segundo a cor e a presencia de fragancia.

fragancia \ cor	branca	rosa	naranxa
si	12	60	58
non	50	10	10

¿Hai probas significativas de asociación entre a cor das flores e a súa fragancia?

Solución. Denotemos por X a fragancia dunha azalea, e por Y a súa cor. En primeiro lugar construímo-la táboa de continxencia:

fragancia \ cor	branca	rosa	naranxa	Σ
si	12	60	58	130
non	50	10	10	70
Σ	62	70	68	200

O problema consiste en facer un contraste de hipóteses de independencia para datos categóricos, é dicir,

$$H_0\colon p_{ij} = p_{i\cdot}p_{\cdot j}, \quad \forall i \in \{1,2\},\ \forall j \in \{1,2,3\}.$$

Veremos máis adiante que a razón de que este sexa un contraste de independencia é que o investigador simplemente clasifica os datos do total da mostra en dúas categorías (neste caso, fragancia e cor das azaleas).

Para resolve-lo problema, calculámo-los valores esperados, no suposto de que houbese independencia das variables, mediante a fórmula $\widehat{E_{ij}} = \frac{n_{i\cdot}n_{\cdot j}}{n}$ (en verde), e tamén os valores $(n_{ij} - \widehat{E_{ij}})^2/\widehat{E_{ij}}$ (en vermello), obtendo:

fragancia \ cor	branca	rosa	naranxa	Σ
si	12 40.3 19.87	60 45.5 4.62	58 44.2 4.31	130
non	50 21.7 36.91	10 24.5 8.58	10 23.8 8.00	70
Σ	62	70	68	200

Finalmente aprovéitanse todos estes cálculos para determina-lo valor no estatístico, (que consiste en suma-los valores vermellos), para obter 82.29.

O estatístico segue unha distribución χ^2 con $(2-1)(3-1)=2$ graos de liberdade. Xa que non nos dan un nivel de significación, calculámo-lo valor P como $P=P(\chi^2_2 \geq 82.29) < 0.001$. (Utilizando software informático obtense $P=1.35\cdot 10^{-18}$.)

Conclusión: *rexeitámo-la hipótese nula*, e concluímos que si hai evidencia significativa, cun nivel de confianza moi alto (maior có 99.9%), de que existe relación entre a cor da flor e a súa fragancia. ■

Contrastes de homoxeneidade para datos categóricos

Este contraste é bastante parecido ó da sección anterior, polo menos no que a cálculos se refire, anque o obxectivo é bastante distinto.

Supoñamos que temos f poboacións nas que se observa unha determinada característica que pode tomar un valor de entre c valores distintos $A_1,\ldots,A_c$. O problema ó que nos enfrentamos é o de determinar se a distribución de probabilidade desa característica é a mesma en todas esas poboacións, ou se polo contrario, ditas poboacións son heteroxéneas con distintas distribucións de probabilidade.

Tomamos unha mostra aleatoria simple en cada unha das poboacións, con tamaños $n_1,\ldots,n_f$, respectivamente. Denotamos por n_{ij} o número de observacións na mostra i que se atopa en A_j. Os datos poden dispoñerse nunha *táboa de continxencia*, organizada do seguinte xeito:

Mostra	A_1	A_2	$\ldots$	A_c	tamaño
1	n_{11}	n_{12}	$\ldots$	n_{1c}	n_1
2	n_{21}	n_{22}	$\ldots$	n_{2c}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
f	n_{f1}	n_{f2}	$\ldots$	n_{fc}	n_f
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot c}$	n

De novo empregouse a notación:

$$n_{\cdot j}=n_{1j}+n_{2j}+\cdots+n_{fj},$$

que son os números totais de observacións que se atopan nos conxuntos A_i. Ademais, $n=n_1+\cdots+n_f$ é o tamaño que se obtén ó xuntar tódalas mostras.

A hipótese de homoxeneidade significa que cada conxunto A_j ten unha probabilidade p_j *independente* da poboación i. Por tanto, se p_{ij} é a probabilidade de A_j na poboación i, a hipótese nula é

$$H_0\colon p_{1j} = p_{2j} = \cdots = p_{fj}\,(= p_j), \quad \forall j \in \{1, \ldots, c\}.$$

As probabilidades p_j poden estimarse mediante

$$\widehat{p_j} = \frac{n_{\cdot j}}{n}.$$

Baixo a hipótese de homoxeneidade, a frecuencia teórica de A_j na poboación i é

$$E_{ij} = n_i p_j,$$

que por tanto se estima mediante

$$\widehat{E_{ij}} = \frac{n_i n_{\cdot j}}{n}.$$

Para determinar se os n_{ij} están suficientemente próximos a $\widehat{E_{ij}}$ empregámo-lo estatístico

$$\sum_{i,j} \frac{(n_{ij} - \widehat{E_{ij}})^2}{\widehat{E_{ij}}} \sim \chi^2_{(f-1)(c-1)},$$

que segue aproximadamente unha distribución χ^2 de Pearson con $(f-1)(c-1)$ graos de liberdade.

Este contraste é unilateral dereito, así que para un nivel de significación α, temos

- Rexión crítica: $(\chi^2_{(f-1)(c-1),\,\alpha},\, \infty)$.
- Rexión de aceptación: $[0, \chi^2_{(f-1)(c-1),\,\alpha}]$.

O contraste de independencia e o contraste de homoxeneidade son moi similares en canto a cálculos e interpretación. A diferencia fundamental está no xeito de selecciona-las mostras, xa que no contraste de homoxeneidade, o tamaño das mostras (é dicir, o total das filas) está fixado polo experimentador, mentres que no contraste de independencia é arbitrario.

Problema. Para probar unha nova vacina contra a hepatite, tómanse 549 voluntarios ós que se lles administra a vacina, e 534 ós que non. Ó cabo dun tempo obsérvanse os seguinte casos de enfermidade:

mostra	ten hepatite	tamaño
vacinado	11	549
non vacinado	70	534

¿É a vacina eficaz?

Solución. Para ver se a vacina é eficaz temos que dar evidencia significativa de que a proporción de enfermos de hepatite é menor na poboación dos vacinados. Por tanto, é un contraste de hipóteses sobre

homoxeneidade no que pretendemos refuta-la hipótese nula de que a distribución de probabilidade do número de enfermos de hepatite é a mesma para as dúas poboacións.

En primeiro lugar completámo-la táboa de continxencia:

	hepatite si	hepatite non	tamaño
vacinado	11	538	549
non vacinado	70	464	534
Σ	81	1002	1083

Temos que facer un contraste de hipóteses de homoxeneidade para datos categóricos:

$$H_0: p_{11} = p_{21},\ p_{12} = p_{22}.$$

A continuación calculámo-los valores esperados no suposto de que houbese homoxeneidade nas poboacións mediante a fórmula $\widehat{E_{ij}} = \frac{n_i n_{\cdot j}}{n}$ (en verde), e tamén os valores $(n_{ij} - \widehat{E_{ij}})^2 / \widehat{E_{ij}}$ (en vermello), obtendo:

vacinado \ hepatite	si	non	tamaño
si	11 41.06 22.01	538 507.94 1.78	549
non	70 39.94 22.63	464 494.06 1.83	534
Σ	81	1002	1083

Finalmente aprovéitanse todas estes contas para calcula-lo valor no estatístico, (que consiste en suma-los valores vermellos), para obter 48.24.

O estatístico segue unha distribución χ^2 con $(2-1)(2-1) = 1$ grao de liberdade. Xa que non nos dan un nivel de significación, calculámo-lo valor P como $P = P(\chi^2_1 \geq 48.24) < 0.001$. (Empregando software informático obtense $P = 3.7 \cdot 10^{-12}$.)

Conclusión: *rexeitámo-la hipótese nula*, e concluímos que hai evidencia significativa, cun nivel de confianza superior ó 99.9%, de que a proporción de enfermos de hepatite é distinta dependendo de se estamos na poboación de individuos vacinados ou non vacinados. Por tanto, tendo en conta os datos da táboa, onde se observa que a proporción de enfermos de hepatite na poboación dos individuos vacinados é menor cá esperada, conluímos que a vacina é eficaz. ■

Os contrastes de independencia e homoxeneidade son bastante populares e empréganse a miúdo como estudos preliminares para ver se hai relación entre dúas ou máis variables. Nótese non obstante, que estes contrastes non nos din *cal* é a relación entre as variables, aínda que mirando os valores da táboa podemos sacar algunha conclusión. Para atopar unha relación que explique como se relaciona unha variable con outra necesítanse outras técnicas estatísticas como a regresión.

É moi típico que, por erro, descoñecemento, ou por tratar de influencia-la opinión da xente, se extraian dun test deste estilo conclusións distintas (aínda que aparentemente relacionadas) ás que en realidade se fan

no estudo. Tamén é típico extrapolar causalidade (un suceso implica outro), cando só hai correlación (dous sucesos pasan ó mesmo tempo).

Problema. Realízase un estudo de mercado consistente en clasifica-la poboación de acordo co seu poder adquisitivo en nivel alto, medio e baixo. Tómase unha mostra de 50 persoas de cada clase social e mírase se posúen un reloxio de marca Rolex. Constátase que da clase alta teñen un 30 persoas, 14 de clase media, e 5 de clase baixa.

- Realiza-lo correspondente contraste de hipóteses para ver se existe relación entre a clase social e ser posuidor dun Rolex.
- ¿Podemos afirmar que hai evidencia estatística de que mercar un Rolex aumenta o poder adquisitivo?

Solución. Temos 3 poboacións, dependendo do "poder adquisitivo", e a variable aleatoria Y="ter un Rolex".

En primeiro lugar construímo-la táboa de continxencia:

renta \ Rolex	si	non	tamaño
alta	30	20	50
media	14	36	50
baixa	5	45	50
Σ	49	101	150

Temos que face-lo contraste de hipóteses:

$$H_0: p_{11} = p_{21} = p_{31}, p_{12} = p_{22} = p_{32}.$$

Este é un contraste de hipóteses para homoxeneidade de datos categóricos, xa que o tamaño da mostra en cada poboación é fixado polo investigador. Para iso empregámo-lo estatístico

$$\sum_{i,j} \frac{\left(n_{ij} - \widehat{E_{ij}}\right)^2}{\widehat{E_{ij}}},$$

que segue unha distribución χ^2 de Pearson con $(f-1)(c-1)$ graos de liberdade.

O número de graos de liberdade da distribución é $(3-1)(2-1) = 2$.

A continuación calculámo-las frecuencias esperadas, no suposto de que a hipótese nula sexa certa, mediante a fórmula $\widehat{E_{ij}} = \frac{n_{i\cdot} n_{\cdot j}}{n}$:

renta \ Rolex	si	non	tamaño
alta	16.33	33.67	50
media	16.33	33.67	50
baixa	16.33	33.67	50
Σ	49	101	150

Agora calculámo-los valores intermedios do estatístico $(n_{ij} - \widehat{E_{ij}})^2 / \widehat{E_{ij}}$:

renta \ Rolex	si	non	Σ
alta	11.435	5.548	
media	0.333	0.162	
baixa	7.864	3.815	
Σ			29.157

A suma dos valores intermedios, que coincide co valor no estatístico, é 29.157.

Calculámo-lo valor P como $P = P(\chi^2_2 > 29.157) = 0.5 \cdot 10^{-6}$, que é un valor pequeno.

Conclusión: Rexeitamos H_0, e concluímos que hai evidencia significativa, cun nivel de confianza moi elevado, de que ter un Rolex depende do poder adquisitivo da persoa.

Non obstante, os datos que se dan neste exercicio non están encamiñados a responde-la segunda pregunta. Podemos deducir que hai relación entre "ter poder adquisitivo alto" e "ter un Rolex", pero non podemos dicir nada con respecto á relación entre "*mercar* un Rolex" e "*aumenta*-lo poder adquisitivo". Aínda que a lóxica indica a pensar que a resposta á segunda pregunta é negativa, os datos do problema non o confirman nin o refutan. ■

Bondade do axuste

Ata agora os contrastes de hipóteses foron empregados para decidi-la veracidade dunha hipótese sobre os parámetros dunha distribución. En ocasións, non obstante, é necesario emitir un xuízo sobre a distribución poboacional no seu conxunto. O problema da bondade do axuste consiste en decidir, á vista dos datos dunha mostra aleatoria simple dunha poboación, se pode admitirse que a distribución poboacional coincide cunha certa distribución dada (no noso caso unha $N(0,1)$). Nótese que este é un problema *non paramétrico*.

Supoñamos que queremos averiguar se a distribución F dunha poboación se axusta a unha distribución normal $N(\mu, \sigma)$. Supoñemos que temos unha mostra aleatoria simple $X_1, \ldots, X_n$. O noso contraste é por tanto

$$H_0 \colon F = N(\mu, \sigma), \qquad H_1 \colon F \neq N(\mu, \sigma).$$

En primeiro lugar teremos que estima-los valores dos parámetros. Empregaremos para iso os estimadores insesgados $\hat{\mu} = \overline{X}$ e $\hat{\sigma} = s_{n-1}$.

O segundo paso deste contraste consiste en agrupa-los datos en intervalos. Para iso realízase o seguinte procedemento:

- Busca-los valores máis pequeno e máis grande. Tomar valores "redondos" un pouco máis pequenos có máis pequeno, e un pouco máis grande có máis grande tendo en conta a precisión dos datos.
- Decidir cantos intervalos se van empregar. Dividiranse os datos en intervalos do mesmo tamaño (preferentemente os extremos deberían ser enteiros, ou números "redondos"). Hai varias regras para decidir este número. Unha posibilidade é tomar aproximadamente $\sqrt{n}$ intervalos. É conveniente que cada intervalo resultante teña polo menos 5 valores. O número de tales intervalos denotámolo por r.

- Calcula-los límites dos intervalos tendo en conta os datos anteriores.
- Face-lo reconto de valores en cada intervalo.

Unha vez que témo-los datos divididos en intervalos, podemos calcula-la frecuencia observada o_i destes en cada intervalo. Este datos compáranse coa probabilidade de que a distribución $N(\mu,\sigma)$ estea entre cada un dos valores dos intervalos, multiplicada por n. Isto é o que se chama a frecuencia teórica e_i. Utilízase para iso a estimación de μ e σ.

Para decidir se as discrepancias entre as frecuencias mostrais e as teóricas son significativas, emprégase a proba χ^2 de Pearson. Tomamos por tanto o estatístico

$$\chi^2_{r-k-1}=\sum_{i=1}^r \frac{(o_i-e_i)^2}{e_i},$$

que segue unha distribución χ^2 de Pearson con $r-k-1$ graos de liberdade, onde k é o número de parámetros que tivemos que estimar para precisa-la distribución teórica (se son μ e σ, entón $k=2$).

O estatístico anterior úsase para facer un *contraste unilateral dereito*.

Problema. Unha máquina produce pezas cunha determinada lonxitude, a cal se quere saber se segue unha distribución normal. Obtense a seguinte mostra:

10.39	10.66	10.12	10.32	10.25	10.91	10.52	10.83	10.72	10.28
10.35	10.46	10.54	10.72	10.23	10.18	10.62	10.49	10.32	10.61
10.64	10.23	10.29	10.78	10.81	10.39	10.34	10.62	10.75	10.34
10.41	10.81	10.64	10.53	10.31	10.46	10.47	10.43	10.57	10.74

Deséxase saber se esta mostra avala a hipótese de que a máquina produce pezas cunha lonxitude que efectivamente é normal.

Solución. Sexa X a variable aleatoria "lonxitude das pezas que produce a máquina".

Vemos que temos $n=40$ datos. En primeiro lugar estimamos puntualmente a media e a desviación típica. Para iso obtemos $\overline{X}=10.502$ e $s_{n-1}=0.205$.

Trátase por tanto do contraste de hipóteses

$$H_0\colon F=N(\overline{X},s_{n-1}),\qquad H_1\colon F\neq N(\overline{X},s_{n-1}),$$

pero agora non contrastamos ou estimámo-lo valor dos parámetros, se non o feito de que a distribución sexa ou non normal.

Para face-lo contraste de χ^2 de bondade de axuste, primeiramente temos que dividi-lo percorrido dos valores en intervalos. Como hai 40 datos, dividimos en 7 intervalos, o que é aproximadamente $\sqrt{40}$. O mínimo é 10.12 e o máximo 10.91. Podemos tomar como rango $[10,11]$ e dividilo en 7. Isto dá $(11-10)/7=0.1428$, así que redondeamos a 0.15 e repartímo-lo exceso $0.15\cdot 7-1=0.05$ a cada lado. Así, os intervalos serían:

(9.975, 10.125] (10.125, 10.275] (10.275, 10.425] (10.425, 10.575] (10.575, 10.725] (10.725, 10.875] (10.875, 11.025]

A continuación contámo-lo número de elementos en cada subintervalo:

Intervalo	o_i
(9.975, 10.125]	1
(10.125, 10.275]	4
(10.275, 10.425]	11
(10.425, 10.575]	9
(10.575, 10.725]	8
(10.725, 10.875]	6
(10.875, 11.025]	1

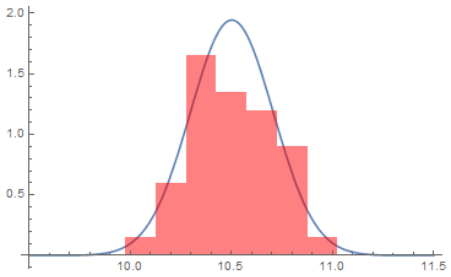


Gráfico de barras de frecuencias

Temos $r = 7$. Ademais, co propósito de comparar coa distribución teórica, que é $N(10.502,\ 0.205)$, temos que toma-las colas ata $-\infty$ e $+\infty$. Completámo-la última columna cos valores teóricos e_i, que corresponden coa probabilidade de que a distribución estea no intervalo, multiplicada por n.

Intervalo	o_i	e_i
$(-\infty, 10.125]$	1	1.3223
$(10.125, 10.275]$	4	4.04803
$(10.275, 10.425]$	11	8.77801
$(10.425, 10.575]$	9	11.4123
$(10.575, 10.725]$	8	8.89851
$(10.725, 10.875]$	6	4.16001
$(10.875, \infty]$	1	1.38084
Σ	40	40.

Unha vez temos tódolos datos, só queda substituír no estatístico de contraste con $k = 2$, $r = 7$. Este estatístico segue unha distribución χ^2 con $r - k - 1 = 4$ graos de liberdade. Así

$$\sum_{i=1}^{7} \frac{(o_i - e_i)^2}{e_i} = 2.16109,$$

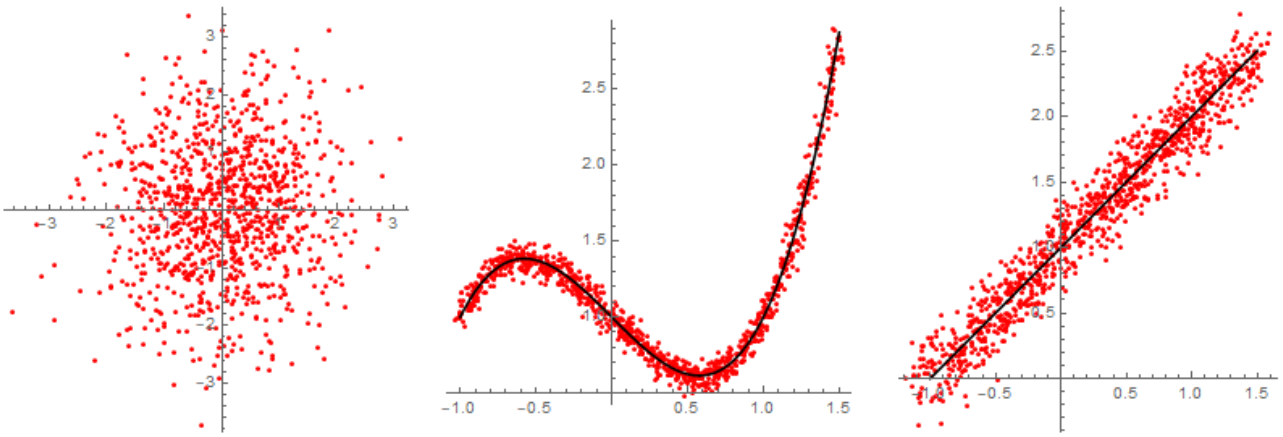
e obtemos $P(\chi^2_4 \geq 2.16109) = 0.706158$, que é un valor moi alto.

Conclusión: aceptámo-la hipótese de que a lonxitude das pezas da máquina segue unha distribución normal. ■

Regresión e correlación

O obxectivo deste capítulo é tratar de establece-la dependencia dunhas variables aleatorias con outras. En principio asumiremos que un determinado efecto se pode explicar mediante unhas causas e un erro. Asumiremos que temos dúas variables aleatorias X e Y. O obxectivo é atopar unha función f tal que $Y = f(X) + \epsilon$. Así, a Y chámaselle *resposta*, a f a *explicación*, e ϵ é o *erro*.

O seguinte gráfico amosa tres nubes de puntos distintas obtidas despois de tomar unha mostra aleatoria de dúas variables X e Y. No primeiro caso é evidente que non existe moita relación entre as dúas variables. No segundo caso parece que as variables están bastante relacionadas, e salvo un pequeno erro, dá a impresión de que Y se explica como dependente de X a través dunha ecuación polinómica. Finalmente, a terceira nube de puntos semella que se axusta a unha recta, aínda que o erro cometido é considerablemente máis grande ca no segundo exemplo.

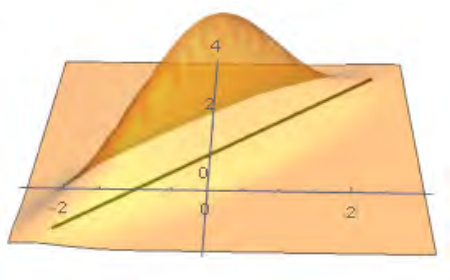


Nos dous últimos debuxos anteriores, é claro que existe unha dependencia (máis ou menos forte) entre X e Y. O obxectivo dun modelo de regresión é:

- Coñecer de que xeito a variable Y depende de X. Isto é o que se chama construír un *modelo de regresión*.
- Unha vez construído o modelo de regresión, empregar este para determina-lo valor de Y cando o valor de X é coñecido.

Neste capítulo consideraremos soamente o caso do *modelo de regresión linear simple*, que é aquel no que as variables X e Y son unidimensionais (como habitualmente), e que Y se explica a partir de X mediante a ecuación dunha recta (coma no terceiro debuxo). Tamén se tratarán outros modelos que se reducen facilmente do de regresión linear.

Regresión linear



Sexan X e Y dúas variables aleatorias. O modelo de regresión linear consiste en atopa-la recta $y = \alpha + \beta x$ que minimiza

$$E\big[(Y - (\alpha + \beta X))^2\big],$$

onde o que se trata é de atopar α e β. Non é difícil ver que estes dous valores se poden calcular simplemente derivando a anterior expresión con respecto de α e de β e igualando a cero. (Analogamente a como se fai para calcula-lo mínimo dunha función, pero con dúas variables.) Esta recta chámase a recta de regresión mínimo-cuadrática, porque na práctica se obtén despois de minimiza-la distancia cuadrática media dos puntos dunha mostra a dita recta.

Despois de face-los cálculos resulta que a ecuación da recta buscada é

$$Y - \mu_Y = \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X) + \epsilon,$$

onde

- $\mu_X = E[X]$ é a media de X,
- $\mu_Y = E[Y]$ é a media de Y,
- $\sigma_X^2 = E[(X - \mu_X)^2]$ é a varianza de X,
- $\sigma_Y^2 = E[(Y - \mu_Y)^2]$ é a varianza de Y,
- $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$ é a **covarianza** entre X e Y,
- ϵ é unha variable aleatoria que representa o erro cometido.

É consecuencia da construción do modelo que o erro ten media cero $\mu_\epsilon = E[\epsilon] = 0$, e que a súa varianza $\sigma_\epsilon^2 = V[\epsilon]$ é mínima.

Defínese o **coeficiente de correlación** de Pearson coma o cociente

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

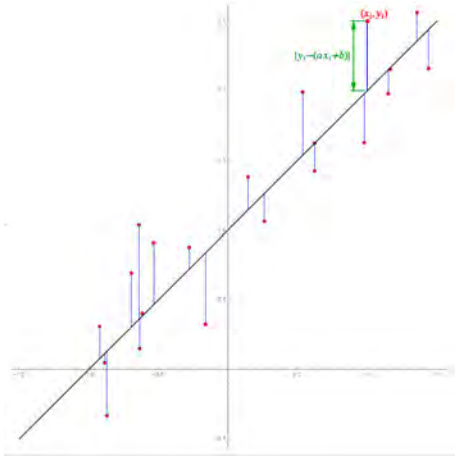
que satisfai $-1 \leq \rho \leq 1$, e dá unha idea do bo que é o axuste.

Estimación dos valores

Na práctica as variables aleatorias X e Y non son coñecidas e son estimadas por valores concretos $(x_1, y_1), \ldots, (x_n, y_n)$ dunha mostra. Por iso, o modelo de regresión linear estímase como

$$Y - \overline{y} = \frac{s_{XY}}{s_X^2}(X - \overline{x}) + \epsilon,$$

onde agora



Distancia vertical entre a recta de regresión e un punto da mostra

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\
\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\
s_X^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2, \\
s_Y^2 &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2, \\
s_{XY} &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}.
\end{aligned}$$

Unha estimación equivalente para a recta de regresión é:

$$Y = a + bX + \epsilon,$$

onde

$$\begin{aligned}
b = \hat{\beta} &= \frac{s_{XY}}{s_X^2}, \\
a = \hat{\alpha} &= \overline{y} - b\,\overline{x}.
\end{aligned}$$

Nótese que esta recta de regresión sempre pasa por $(\overline{x}, \overline{y})$.

Covarianza e correlación

A covarianza é a forma máis común de medi-la relación linear entre dúas variables. Para datos concretos recordemos que se estima por

$$\begin{aligned} s_{XY} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{x}\,\overline{y}. \end{aligned}$$

A covarianza non se ve afectada por cambios de posición, pero si de escala. De feito,

$$s_{aX+b,\,cY+d} = ac\,s_{XY}.$$

Para obter unha medida da relación linear entre dúas variables que non dependa da escala introduciuse o coeficiente de correlación, que para datos concretos se estima mediante

$$r = \frac{s_{XY}}{s_X s_Y}.$$

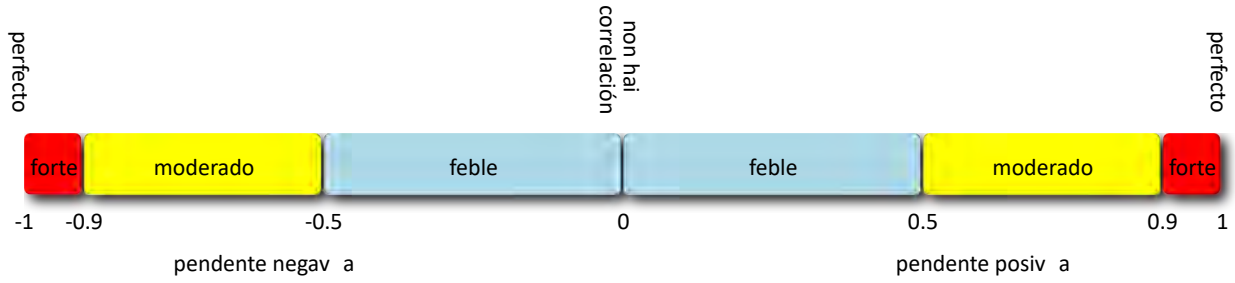
Unha versión equivalente, pero máis estable numericamente, é

$$r = \frac{n\displaystyle\sum_{i=1}^{n}x_iy_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\displaystyle\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\displaystyle\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2}}.$$

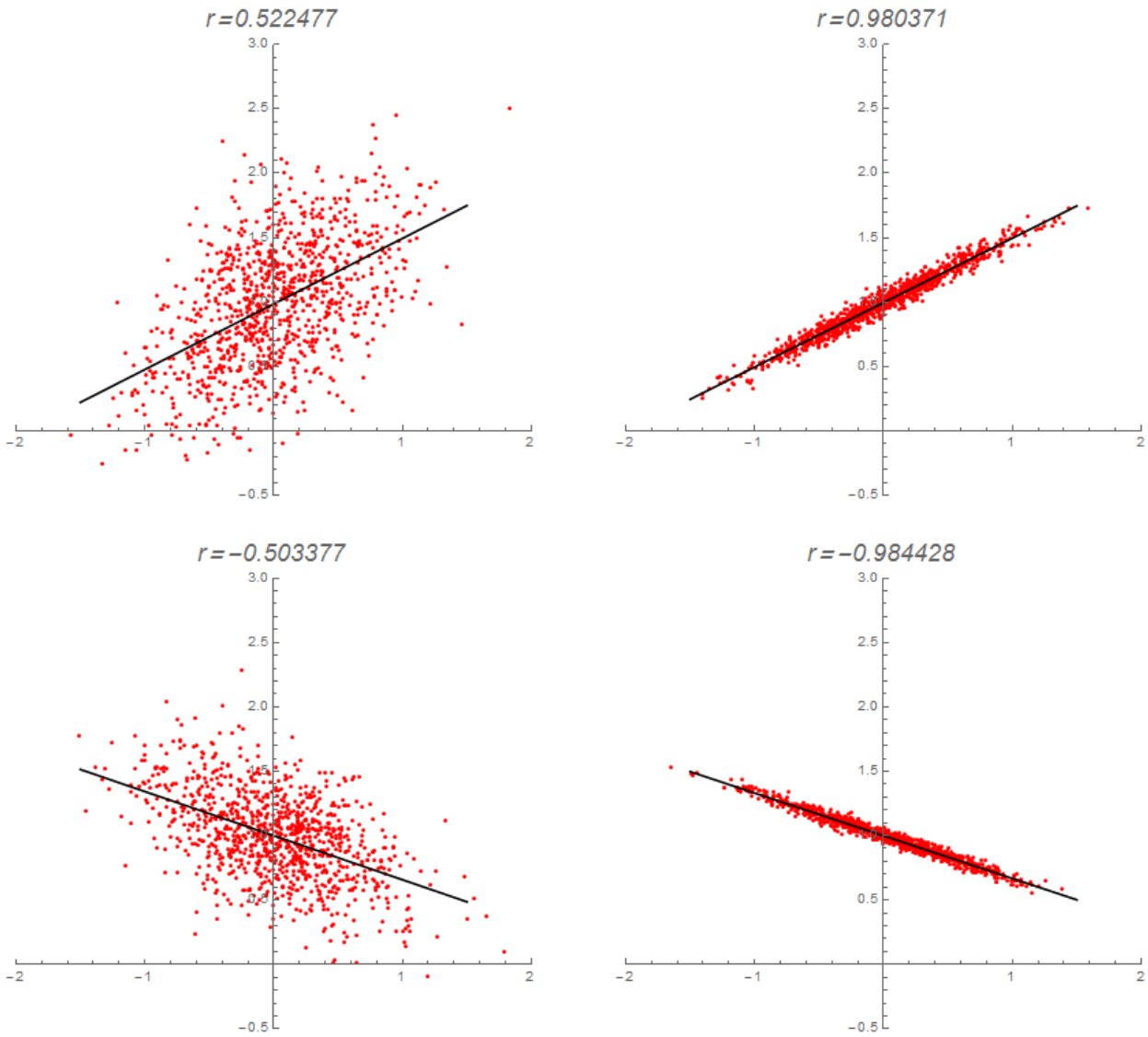
Proposición. O coeficiente de correlación satisfai as seguintes propiedades:

- O coeficiente de correlación ten o mesmo signo cá pendente da recta de regresión.
- $-1 \leq r \leq 1$; valores próximos a 0 indican que o axuste é malo, valores próximos a 1 indican que o axuste é bo e que a relación é crecente, mentres que valores próximos a -1 indican que o axuste é bo e que a relación é decrecente.

Un rango de valores para a bondade do axuste en función de r pode se-lo seguinte:



Algúns exemplos de coeficientes de correlación:



Coeficiente de correlación e calidade do axuste

Problema. Os seguintes datos correspóndense co tempo transcorrido e a velocidade de caída dun obxecto:

tempo	velocidade
1	20.52
2	29.14
3	36.76
4	47.80
5	58.72

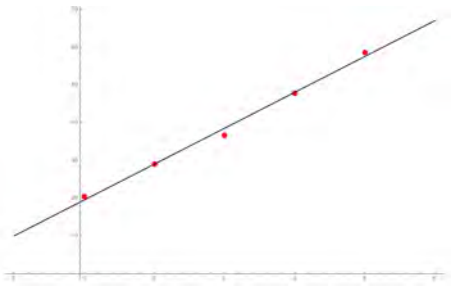
Calcula-la recta de regresión e dar unha aproximación da aceleración da gravidade. ¿Como de bo é o axuste?

Solución. Temos dúas variables que chamaremos t (tempo) e v (velocidade). En primeiro lugar dispoñémo-los cálculos:

	t	v	t^2	tv	v^2
	1.	20.52	1.	20.52	421.07
	2.	29.14	4.	58.28	849.14
	3.	36.76	9.	110.28	1351.30
	4.	47.80	16.	191.20	2284.84
	5.	58.72	25.	293.60	3448.04
Σ	15.	192.94	55.	673.88	8354.39

Entón

$$\begin{aligned}
\bar{t} &= \frac{15.}{5} = 3.0, \\
\bar{v} &= \frac{192.94}{5} = 38.59, \\
s_t^2 &= \frac{55.}{5} - 3^2 = 2.0, \\
s_v^2 &= \frac{8354.39}{5} - 38.59^2 = 181.84, \\
s_{tv} &= \frac{673.88}{5} - 3 \cdot 38.59 = 19.01,
\end{aligned}$$



Os puntos e a súa recta de regresión

co que, substituíndo na fórmula, obtémo-la recta de regresión $v - 38.59 = \frac{19.01}{2}(t-3)$, ou ben, como $b = 19.01/2.0 = 9.51$ e $a = 38.59 - 9.51 \cdot 3.0 = 10.07$, que

$$v = 10.07 + 9.51t,$$

de onde ademais se deduce que, en vista do resultado coñecido de física $v = v_0 + gt$, que $g = 9.51$ é unha aproximación da aceleración da gravidade.

Finalmente calculámo-lo coeficiente de correlación:

$$r = \frac{19.01}{\sqrt{2}\sqrt{181.84}} = 0.997,$$

o cal quere dicir que o axuste é bo. ■

Regresión exponencial

O procedemento para calcular unha regresión linear pode ser empregado tamén noutros contextos simplemente facendo un pequeno cambio de variable. Por exemplo, supoñamos que temos dúas variables aleatorias Z e T, e cremos que Z se explica a partir de T a través dunha fórmula exponencial:

$$Z = z_0\, e^{-kT},$$

onde z_0 e k son os parámetros que queremos determinar. Entón, tomando logaritmos (neperianos)

$$\log Z = \log\left(z_0 e^{-kT}\right) = \log z_0 - kT.$$

Chamando $Y = \log Z$, $X = T$, $b = -k$, $a = \log z_0$, estamos exactamente na situación $Y = a + bX$ do principio. Por tanto, este tipo de axuste exponencial redúcese a un axuste linear, que xa sabemos resolver.

Problema. Inxectamos por vía intravenosa $125mg$ dun medicamento. Témo-las seguintes concentracións plasmáticas a medida que pasa o tempo:

tempo	concentración
1	5.0
2	3.0
3	2.0
4	1.5

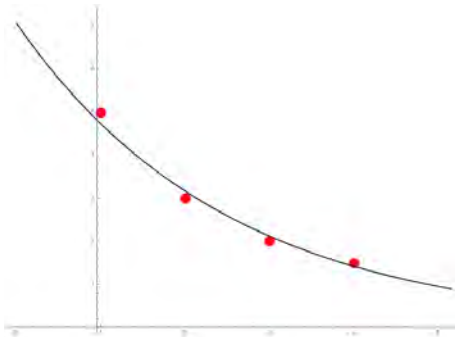
Queremos estima-la curva exponencial da concentración de medicamento en sangue.

Solución. É sabido que a evolución da concentración teórica C dun medicamento en sangue ó longo do tempo t segue unha curva exponencial $C = c_0\, e^{-kt}$. Despois de tomar logaritmos neperianos temos $\log C = \log c_0 - kt$, así que para calcula-la recta de regresión destes datos organizámo-los cálculos do seguinte xeito:

	$X = t$	C	$Y = \log C$	X^2	XY	Y^2
	1.	5.0	1.61	1.	1.61	1.59
	2.	3.0	1.10	4.0	2.20	1.21
	3.	2.0	0.69	9.0	2.08	0.48
	4.	1.5	0.41	16.0	1.62	0.16
Σ	10.	11.5	3.81	30.0	7.51	4.44

Entón

$$\begin{aligned}
\overline{X} &= \frac{10.}{4} = 2.5,\\
\overline{Y} &= \frac{3.81}{4} = 0.95,\\
s_X^2 &= \frac{30.0}{4} - 2.5^2 = 1.25,\\
s_Y^2 &= \frac{4.44}{4} - 0.95^2 = 0.20,\\
s_{XY} &= \frac{7.51}{4} - 2.5 \cdot 0.95 = -0.50,
\end{aligned}$$



Os puntos e a súa regresión exponencial

co que, substituíndo na fórmula, obtémo-la recta de regresión:

$$Y - 0.95 = -\frac{0.50}{1.25}(X - 2.5).$$

Equivalentemente, obtense $b = -0.50/1.25 = -0.40$, $a = 0.95 + 0.40 \cdot 2.5 = 1.96$, de onde se deduce $Y = 1.96 - 0.40X$, ou $\log C = 1.96 - 0.40t$. Desfacendo o cambio de variable obtemos

$$C = 7.07e^{-0.40t}.$$

Pódese ver ademais que o coeficiente de correlación é

$$r = \frac{-0.50}{\sqrt{1.25}\sqrt{0.20}} = -0.992,$$

o que, ademais dun bo axuste, indica que a variable Y (ou a concentración C) decrece en función do tempo. ■

Regresión potencial

A regresión potencial é un caso bastante parecido ó da regresión exponencial. Neste caso hai dúas variables P e A que están relacionadas mediante a fórmula

$$P = \alpha A^{\beta}.$$

Para resolver isto, tomamos coma na sección anterior logaritmos e obtemos

$$\log P = \log\left(\alpha A^{\beta}\right) = \log \alpha + \beta \log A.$$

Así, chamando $Y = \log P$ e $X = \log A$ volvemos estar nun caso de axuste linear, que xa vimos como se resolve.

Análise da varianza

O obxectivo desta sección é estudar con máis profundidade se o modelo de regresión construído é correcto e útil. Para iso imos empregar un método coñecido como ANOVA (analysis of variance).

En primeiro lugar recordamos que $Y = \alpha + \beta X + \epsilon$, onde $\hat{Y} = \alpha + \beta X$ será a estimación dada pola recta de regresión, e $\epsilon = Y - \hat{Y}$ é o erro. Un cálculo non trivial amosa que as varianzas están relacionadas mediante

$$\sigma_Y^2 = \sigma_{\hat{Y}}^2 + \sigma_{\epsilon}^2.$$

Isto significa que a *variabilidade da variable dependente* Y, σ_Y^2, se descompón como

- A *variabilidade explicada*, $\sigma_{\hat{Y}}^2$, que é aquela que se pode explicar en base ó modelo de regresión. De feito, como $\hat{Y} = \alpha + \beta X$, entón $\sigma_{\hat{Y}}^2 = \beta^2 \sigma_X^2$.
- A *variabilidade residual*, σ_{ϵ}^2, que é a que non explica o modelo de regresión.

Chámase **coeficiente de determinación** á proporción entre a variabilidade explicada e a variabilidade da variable dependente. Por tanto,

$$\frac{\sigma^2_{\hat{Y}}}{\sigma^2_Y} = \frac{\beta^2\sigma^2_X}{\sigma^2_Y} = \frac{\sigma^2_{XY}}{\sigma^2_X\sigma^2_Y} = \rho^2,$$

que é o cadrado do coeficiente de correlación. Por tanto, $0 \leq \rho^2 \leq 1$.

Como xa sucedía co coeficiente de correlación, se $\rho^2 = 1$ (é dicir, se $\rho = \pm 1$) entón o axuste é perfecto. Valores de ρ^2 próximos a 1 significan que o axuste é bo, mentres que valores próximos a 0 indican un axuste malo.

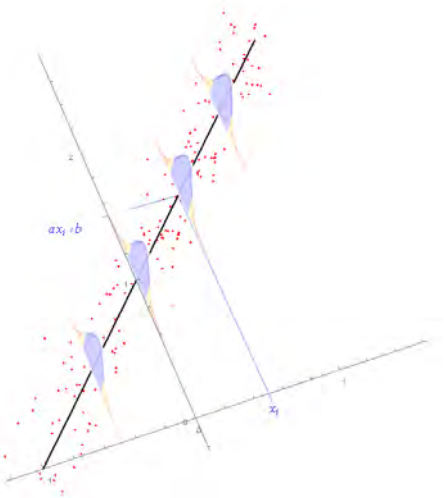
Ademais, das fórmulas anteriores temos que

$$\sigma^2_{\hat{Y}} = \rho^2\sigma^2_Y,$$
$$\sigma^2_\epsilon = \sigma^2_Y - \sigma^2_{\hat{Y}} = \sigma^2_Y - \rho^2\sigma^2_Y = (1-\rho^2)\sigma^2_Y.$$

ANOVA

En realidade, os cálculos da sección anterior son teóricos, porque en xeral as distribucións X e Y non son coñecidas. Na práctica tómase unha mostra e utilízanse as estimacións escritas con anterioridade.

Para continuar supoñamos que estamos traballando con n valores específicos $x_1, \ldots, x_n$. Por tanto, os valores da variable explicativa están fixados polo experimentador e non son aleatorios. Só é aleatorio o erro, e en consecuencia a variable resposta. Unha mostra resultante deste tipo de experimento (chamado de deseño fixo), é do tipo $(x_1, Y_1), \ldots, (x_n, Y_n)$. Asumimos que as variables aleatorias $Y \mid X = x_1, \ldots, Y \mid X = x_n$ seguen distribucións normais independentes coa mesma varianza σ^2. Se a regresión linear é válida, as medias destas variables están xustamente en $a + bx_i$, é dicir, $(Y \mid X = x_i) \in N(a + bx_i, \sigma)$.



$Y \mid X = x_i$ está normalmente distribuída

O valor $\hat{Y}_i$ será o valor predecido pola estimación do modelo, é dicir, $\hat{Y}_i = a + bx_i$. Nótese en particular que $\overline{\hat{Y}} = \overline{Y} = a + b\overline{x}$.

Así, despois de multiplicar por n, a variabilidade é estimada mediante

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Se agora denotamos

$$SS_Y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$
$$SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = b^2\sum_{i=1}^{n}(x_i - \overline{x})^2,$$
$$SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

entón, a expresión anterior pode escribirse como

SS_Y	=	SS_R	+	SS_E
(variabilidade total)		(variabilidade debida á regresión)		(variabilidade non explicada)

As fórmulas anteriores refírense ás "sumas de cadrados". Se en lugar diso querémo-las varianzas, simplemente hai que dividir polo tamaño mostral n:

$$s_Y^2 = \frac{1}{n}SS_Y, \qquad s_R^2 = \frac{1}{n}SS_R, \qquad s_E^2 = \frac{1}{n}SS_E.$$

Estas cantidades son unha estimación das varianzas teóricas obtidas no apartado anterior. Por outra banda, o coeficiente de determinación estímase mediante r^2, de xeito que temos

$$r^2 = \frac{s_R^2}{s_Y^2} = \frac{SS_R}{SS_Y}.$$

Así, a estimación do coeficiente de determinación r^2 interprétase como a proporción da variabilidade da variable aleatoria Y que é explicada por X mediante o modelo de regresión.

Esta técnica de análise da varianza utilízase para comprobar se unha liña recta mostra unha cantidade significativa de variabilidade observada de Y. Se o suposto é que a regresión é válida, entón o que terá que suceder é que a maior parte da variabilidade terá que ser explicada por SS_R, sendo a parte non explicada pequena.

Obsérvese agora a equivalencia das seguintes condicións

$$\beta = 0 \Leftrightarrow \rho = 0,$$

é dicir, toda a variabilidade é aleatoria (non explicada), e por tanto non hai regresión linear. Así pois, o test que temos que facer para comproba-la validez do modelo é

$$H_0\colon\ \rho = 0, \qquad H_1\colon\ \rho \neq 0.$$

Baixo as hipóteses anteriores, este contraste emprega dous estatísticos que pasamos a describir a continuación. En primeiro lugar, para SS_Y hai n datos e un valor estimado, $\overline{y}$, o que deixa $n-1$ graos

de liberdade.

- Para SS_E hai n datos, pero dous valores estimados, a e b, o que nos deixa $n-2$ graos de liberdade. Así, empregamos como estatístico o *cadrado medio do erro*: $MS_E = \dfrac{SS_E}{n-2}$.
- Iso significa que para SS_R queda un só grao de liberdade. O estatístico empregado é pois o *cadrado medio da regresión*: $MS_R = \dfrac{SS_R}{1}$.

No suposto de que a hipótese nula sexa certa, o estatístico

$$\frac{MS_R}{MS_E} = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2/(n-2)} \sim F_{1,\, n-2}$$

segue unha distribución F de Snedecor con $(1, n-2)$ graos de liberdade.

Se a hipótese nula é certa, o valor observado no estatístico estará próximo a 1. Noutro caso será moito maior e rexeitarase a hipótese nula se o valor é demasiado grande. Trátase por tanto de facer un *contraste unilateral dereito*.

Os cálculos necesarios para empregar ANOVA á hora de contrastar $H_0\colon \rho = 0$ (non hai regresión linear), dispóñense nunha táboa como a seguinte:

variabilidade	g.l.	SS	MS	cociente
regresión	1	$SS_R = \displaystyle\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 = nr^2 s_Y^2$	$MS_R = \dfrac{SS_R}{1}$	$F_{1,\, n-2} = \dfrac{MS_R}{MS_E}$
erro	$n-2$	$SS_E = \displaystyle\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = n(1-r^2)s_Y^2$	$MS_E = \dfrac{SS_E}{n-2}$	
total	$n-1$	$SS_Y = \displaystyle\sum_{i=1}^n (Y_i - \overline{Y})^2 = ns_Y^2$		

Cando, despois de face-lo anterior contraste de hipóteses, cheguemos á conclusión de que se rexeita a hipótese nula H_0, iso quererá dicir que unha parte significativa da variabilidade de Y se pode explicar mediante o modelo de regresión linear. Iso non quere dicir que o modelo linear sexa o mellor para explicar dita variabilidade, senón que é razoable emprega-lo modelo para explicala.

Problema. Realízase un experimento para estuda-la relación entre a altura e a lonxitude da concha de Patelloida pygmaea (en mm). Téñense os seguinte datos:

altura	lonxitude
0.9	3.1
1.5	3.6
1.6	4.3
1.7	4.7
1.7	5.5
1.8	5.7
1.8	5.2
1.9	5.0
1.9	5.3
1.9	5.7
2.0	4.4
2.0	5.2
2.0	5.3
2.1	5.4
2.1	5.6
2.1	5.7
2.1	5.8
2.2	5.2
2.2	5.3
2.2	5.6
2.2	5.8
2.3	5.8
2.3	6.2
2.3	6.3
2.3	6.4
2.4	6.4
2.4	6.3
2.7	6.3

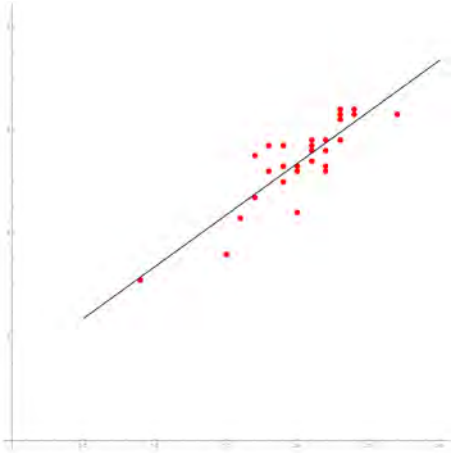
Estima-la recta de regresión da lonxitude como función da altura. Calcula-lo coeficiente de determinación e interpreta-lo seu valor. ¿Hay evidencia estatística de que o modelo de regresión linear é válido?

Solución. Chamemos X á altura e Y á lonxitude. Organizámo-los cálculos nunha táboa.

	X	Y	X^2	XY	Y^2
	0.90	3.10	0.81	2.79	9.61
	1.50	3.60	2.25	5.40	12.96
	1.60	4.30	2.56	6.88	18.49
	1.70	4.70	2.89	7.99	22.09
	1.70	5.50	2.89	9.35	30.25
	1.80	5.70	3.24	10.26	32.49
	1.80	5.20	3.24	9.36	27.04
	1.90	5.00	3.61	9.50	25.00
	1.90	5.30	3.61	10.07	28.09
	1.90	5.70	3.61	10.83	32.49
	2.00	4.40	4.00	8.80	19.36
	2.00	5.20	4.00	10.40	27.04
	2.00	5.30	4.00	10.60	28.09
	2.10	5.40	4.41	11.34	29.16
	2.10	5.60	4.41	11.76	31.36
	2.10	5.70	4.41	11.97	32.49
	2.10	5.80	4.41	12.18	33.64
	2.20	5.20	4.84	11.44	27.04
	2.20	5.30	4.84	11.66	28.09
	2.20	5.60	4.84	12.32	31.36
	2.20	5.80	4.84	12.76	33.64
	2.30	5.80	5.29	13.34	33.64
	2.30	6.20	5.29	14.26	38.44
	2.30	6.30	5.29	14.49	39.69
	2.30	6.40	5.29	14.72	40.96
	2.40	6.40	5.76	15.36	40.96
	2.40	6.30	5.76	15.12	39.69
	2.70	6.30	7.29	17.01	39.69
Σ	56.60	151.10	117.68	311.96	832.85

Entón temos $n = 28$ datos e

$$\begin{aligned}
\overline{X} &= \frac{56.6}{28} = 2.021,\\
\overline{Y} &= \frac{151.10}{28} = 5.396,\\
s_X^2 &= \frac{117.68}{28} - 2.021^2 = 0.117,\\
s_Y^2 &= \frac{832.85}{28} - 5.396^2 = 0.623,\\
s_{XY} &= \frac{311.96}{28} - 2.021 \cdot 5.396 = 0.233.
\end{aligned}$$



Os puntos e a súa recta de regresión

Obtemos $b = 0.233/0.117 = 1.996$ e $a = 5.396 - 1.996 \cdot 2.020 = 1.361$ co que a ecuación da recta de regresión é

$$y - 5.396 = 1.996(x - 2.0214),$$

ou ben,

$$y = 1.361 + 1.996x.$$

A estimación do coeficiente de correlación é

$$r = \frac{0.233}{\sqrt{0.117 \cdot 0.623}} = 0.8638,$$

de xeito que a calidade da aproximación parece moderada.

A estimación do coeficiente de determinación é $r^2 = 0.746$. Isto interprétase do seguinte xeito: o 74.6% da variabilidade da variable Y está explicada polo modelo de regresión.

Para asegurarnos, intentaremos dar evidencia significativa de que o modelo de regresión é válido. Isto significa face-lo contraste de hipóteses

$$H_0\colon \rho = 0, \qquad H_1\colon \rho \neq 0.$$

Empregamos pois a técnica de análise da varianza, ANOVA. Os datos necesarios están recollidos na seguinte táboa:

variabilidade	g.l.	SS	MS	cociente
regresión	1	$SS_R = 28 \cdot 0.864^2 \cdot 0.623 = 13.02$	$MS_R = 13.02$	76.42
erro	26	$SS_E = 28(1 - 0.864^2)0.623 = 4.43$	$MS_E = \frac{4.43}{26} = 0.17$	
total	27	$SS_Y = 28 \cdot 0.623 = 17.45$		

Como $P = P(F_{1,26} \geq 76.42) < 0.01$ é un número moi pequeno (de feito, empregando software estatístico temos $P = 3.2 \cdot 10^{-9}$), rexeitámo-la hipótese nula. Concluímos que hai evidencia

significativa de que o modelo de regresión linear é válido. ■

Intervalos de estimación

Por completitude incluímos nesta sección a estimación por intervalos de diversos valores que apareceron no noso modelo de regresión linear.

En primeiro lugar pódese ver que unha estimación puntual da desviación típica do erro σ^2 vén dada por

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2} = \frac{1}{n-2}\sum_{i=1}^n (Y_i - a - bx_i)^2,$$

tendo o estatístico

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$$

unha distribución χ^2 con $n-2$ graos de liberdade.

Tomemos un nivel de significación α.

- Ordenada na orixe:

$$\alpha = a \pm t_{n-2,\,\alpha/2}\,\frac{\sqrt{MS_E}}{S_X}\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}.$$

- Pendente da recta de regresión:

$$\beta = b \pm t_{n-2,\,\alpha/2}\,\frac{\sqrt{MS_E}}{S_X}.$$

- Resposta media para un valor de X dado:

$$\mu_{Y|X=x} = \hat{Y} \pm t_{n-2,\,\alpha/2}\,\sqrt{MS_E}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{s_X^2}}.$$

- Intervalo de predicción da resposta individual para un valor de X dado:

$$\hat{Y} \pm t_{n-2,\,\alpha/2}\sqrt{MS_E}\,\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{s_X^2}}.$$

Problemas para as clases interactivas

▼ Intervalos de confianza

1. Estas son as alturas (en metros) de vinte piñeiros da especie "Pinus strobus". Estima-la media desa especie de piñeiros cun nivel de confianza do 95%.

17.16	22.00	10.08	15.00
7.02	10.67	11.16	10.92
11.10	4.05	15.93	7.22
8.19	16.45	7.38	10.00
14.10	10.26	11.96	10.00

   [Milton 6.3.1]

2. Nunha mostra de tamaño 30 mediuse a porcentaxe de aumento de alcohol en sangue tras beber catro cervexas. Obtívose $\overline{X} = 41.2$ (media) e $s = 2.1$ (cuasi-desviación típica).
   1. Calcular un intervalo de confianza do 90% para a porcentaxe media de aumento en tódalas persoas que beben catro cervexas;
   2. Se se calcula un intervalo de confianza do 95% para μ, ¿será máis ou menos amplo có anterior?

   [Milton 6.3.7]

3. As granxas de patos contaminan a agua debido ó nitróxeno en forma de "ácido úrico". A seguinte é unha mostra aleatoria de nove observacións da variable X, número de kilos de nitróxeno producidos por granxa e día.

   4.9 5.8 5.9 6.5 5.5 5.0 5.6 6.0 5.7

   Supoñendo que X é normal, construír un intervalo de confianza do 99% para a media poboacional μ.

   [Milton 6.3.9]

4. A calor parece afecta-la mobilidade dos caracois. En 20 caracois sometidos a unha temperatura de 29°C observamos unha distancia media percorrida de $\overline{X} = 4.855$cm, con $s_{n-1} = 0.7178$. Dar un intervalo de confianza ($\alpha = 5\%$) para a distancia media percorrida por un caracol.

   [Milton 6.3.6 p. 222]

5. Queremos estima-lo peso medio ó nacer (en Kg) de fillos de mulleres adictas á heroína. Nun estudio previo obtívose que $\sigma = 2.5$. Queremos deseña-lo experimento de modo que o nivel de confianza sexa do 95%, e que o erro de estimación non supere 1Kg. ¿Que tamaño de mostra necesitamos?

   [Milton 6.6.1 p. 236]

6. No río Mississippi estudouse en 61 lugares a variable X, anchura de terreno inundable, obténdose $\overline{X} = 3400$ metros e $s_{n-1} = 100$ metros. Dar un intervalo de estimación para a desviación típica de X cun nivel de confianza do 90%.
   [Milton 7.1.7 p. 253]

7. Nun reconto no microscopio contabilizáronse 200 leucocitos, dos cales 125 eran neutrófilos. Dar un intervalo de confianza do 90% para a proporción de neutrófilos en sangue.
   [Milton 8.2.3 p. 266]

8. Nun estudo sobre obesidade infantil averíguase que a idade media de inicio da enfermidade dunha mostra de 26 nenos é de 4 anos, cunha desviación típica mostral de 1.5 anos. Determinar un intervalo de confianza do 95% para a desviación típica da poboación.
   [Milton Exemplo 7.1.6]

9. ¿Que tamaño de mostra faría falla para estima-la proporción de mortes debidas a un problema cardíaco, se traballamos cun nivel de significación do 5%, e non queremos que o erro de estimación supere o 2%?
   [Milton 8.3.2 p. 270]

10. Un investigador médico quere estima-lo nivel medio de colesterol en homes de idade avanzada. A estimación debe ter unha precisión de 6mg/dl ou menos, cun 95% de confianza. Ademais, o investigador cre, por estudos previos, que a desviación típica do colesterol na poboación ronda os 40mg/dl. ¿Que tamaño de mostra debe tomar?
    [Samuels 6.4.2]

11. Nun estudo atopouse que 40 de 400 estudantes eran zurdos. Construír un intervalo de confianza do 90% para a proporción de estudantes zurdos na poboación.
    [Samuels 9.3.1]

12. Unha bodega produce 720000 botellas de viño cada ano e desexa estima-la proporción de botellas que teñen o corcho defectuoso (o viño estropéase se hai un fallo no corcho). Nun estudo previo calcúlase que esta proporción ronda o 4%, pero agora queremos, cun nivel de confianza do 90%, que o erro de estimación non supere o 1%. ¿Cantas botellas de viño debemos comprobar?
    [Samuels 9.S.6]

▼ Contrastes de hipóteses

1. Sospéitase que o insecticida DDT provoca diminución no grosor das cáscaras dos ovos dos paxaros. Para combrobar isto, alimentouse a 16 gabiáns cunha mistura que contiña 15ppm de DDT, e atopouse unha diminución do grosor do 8%. A desviación típica mostral foi de $s = 0.05$. Contrasta-la hipótese de que houbo unha diminución no grosor en toda a poboación (nivel de confianza do 95%).

   [Milton 6.5.4 p. 233]

2. Realizouse un experimento para estuda-lo efecto do exercicio físico no nivel de colesterol de pacientes obesos. En 80 pacientes sometidos a un réxime específico de actividade, observouse unha diminución media do nivel de colesterol de $\overline{X} = 27$ puntos. A desviación estándar foi de $s = 18$. ¿Pode afirmarse, cun nivel de confianza do 90%, que ese réxime provoca, en media, unha diminución superior a 25 puntos?

   [Milton 6.5.7 p. 234]

3. A concentración media de dióxido de carbono no aire é do 0.035%. Preténdese demostrar que inmediatamente por riba da superficie do chan dita concentración é maior. Analizáronse 144 mostras de aire seleccionado aleatoriamente e tomadas á distancia de 30cm do chan. Resultou unha media mostral do 0.09% e unha cuasi-desviación típica mostral do 0.25%. ¿Cal é o valor P do contraste? ¿Comprobouse estatisticamente o argumento establecido?

   [Milton 6.5.5 p. 233]

4. En certa especie de vagalumes, a luz que producen consta dun escintileo curto seguido dun período de repouso. Quérese probar que o período de repouso ten unha duración media de menos de catro segundos. Nunha mostra de 16 insectos obtivemos unha media de 3.77 segundos, con $s = 0.30$ segundos. Por outro lado, dámonos conta de que un erro de tipo I non ten consecuencias fatais, así que fixamos un $\alpha = 10\%$ bastante alto. ¿Apoian os datos experimentais a nosa suposición sobre o escintileo?

   [Milton 6.5.7 p. 232]

5. Ó estuda-lo crecemento de abetos, sábese que a varianza poboacional acostuma ser $1.56cm^2$. Non obstante, en 50 árbores crecidos en condicións de seca observamos unha cuasi-desviación típica de 0.375cm. ¿Afectou a seca ó parámetro σ? Dar un intervalo de confianza do 95% para a desviación típica da poboación.

   [Milton 7.2.4 p. 256]

6. A concentración sanguínea de calcio nos mamíferos acostuma ser de 6mg/100ml. A desviación típica debe ser de 1mg/100ml, xa que unha variabilidade maior ocasiona trastornos de coagulación. Nunha serie de nove probas realizadas a un paciente, atopouse unha concentración media de 6.2 e unha cuasi-desviación típica de 2. Tomando un nivel de significación $\alpha = 0.05$, ¿hai evidencia de que a desviación típica sexa maior da normal?
   [Milton 7.2.2 p. 256]

7. Estímase xeralmente que o 90% dos enfermos de cancro de pulmón morren no prazo de 3 anos. Nun estudo recente no que se proban uns novos tratamentos, atopouse que 128 pacientes morreron dun total de 150 enfermos. ¿Pode dicirse que hai probas suficientes de que o emprego dos novos métodos de tratamento reduciron a taxa de falecementos?
   [Milton 8.4.4 p. 273]

8. Un 20% dos enfermos de corazón tratados cronicamente con digoxina sofre unha reacción adversa. Para evitalo, a 30 pacientes asocióuselles outro medicamento, e conseguiuse que só tres tivesen a reacción. ¿Pode afirmarse que o tratamento é eficaz cun nivel de confianza do 99%?
   [Milton 9.7]

9. O método usual para trata-la leucemia mieloblástica aguda consiste en somete-lo paciente a quimioterapia intensiva no momento do diagnóstico. Historicamente, isto produciu unha taxa de remisión do 70%. Estudando un novo método de tratamento utilizáronse 50 voluntarios. ¿Cantos dos pacientes deberían ter remitido para que os investigadores puidesen afirmar, con nivel de significación $\alpha = 0.025$, que o novo método produce remisións máis altas có antigo?

10. Os votos en contra da construción dunha presa nunha mostra de 500 persoas foi de 270. Estima-la proporción de persoas que están en contra en toda a poboación, cun nivel de confianza do 95%.
    [Milton Exemplo 8.4.1]

11. Estase probando a eficacia dun tipo de exercicio para mellora-los síntomas da artrite reumatoide. O grupo no que se proba dito tratamento é de 160 pacientes. Para un nivel de significación do 2,5%, ¿cantos pacientes terían que mellorar para que se poida afirmar que a porcentaxe de pacientes que melloran é superior ó 50%?

▼ Contrastes de hipóteses para dúas poboacións

1. Comprobouse o peso de ovos de tartaruga en dúas illas diferentes. Suponse que a variación é normal. Á vista dos datos obtidos en dúas mostras aleatorias, ¿hai evidencia de que os ovos na illa "Malabar" son máis pesados cós da illa "Grande-Terre" cun nivel de significación do 1%?

   Datos da illa "Grande-Terre": Tamaño da mostra $n_1 = 31$; peso medio $\overline{X}_1 = 64.0\text{g}$; cuasi-desviación típica $s_1 = 6.5\text{g}$.

   Datos da illa "Malabar": Tamaño da mostra $n_2 = 148$; peso medio $\overline{X}_2 = 82.7\text{g}$; cuasi-desviación típica $s_2 = 3.6\text{g}$.

   (Facer un contraste de hipóteses para a igualdade das varianzas para poder determinar se podemos asumir que ambas sexan iguais.)

   [Milton 9.4.3 p. 311]

2. Ó estuda-la velocidade de voo de dúas especies de paxaros, obtivémo-los seguintes datos:
   - (*Haematopus palliatus*): $n_1 = 9$, $\overline{X}_1 = 26.05$, $s_1 = 6.34$;
   - (*Pelecanus occidentalis*): $n_2 = 12$, $\overline{X}_2 = 30.19$, $s_2 = 3.20$;

   Face-lo contraste necesario para saber se as varianzas poboacionais se poden supoñer iguais. ¿Hai evidencia de que a velocidade de voo das dúas especies de paxaros sexa diferente? (Para todo o problema, tomar un nivel de confianza do 95%.)

   [Milton 9.2.1 p. 298]

3. Estudouse nunha mostra de $n_1 = 33$ homes novos fumadores a idade media á que empezan a fumar, obténdose $\overline{X}_1 = 11.3$ anos. A cuasi-varianza mostral foi de 4 anos. O mesmo estudo en mozas deu lugar ós seguintes datos: $n_2 = 14$, $\overline{X}_2 = 12.6$, $s_2^2 = 3.5$. Pídese, cun nivel de significación $\alpha = 5\%$:
   1. Facer unha proba F para concluír que podemos supoñer $\sigma_1^2 = \sigma_2^2$;
   2. Dar un intervalo de estimación para a diferencia de medias poboacionais entre mozos e mozas.

   [Milton 9.3.11 p. 309]

4. Un laboratorio quere compara-los efectos secundarios dun medicamento novo cos do producto da competencia. Usaremos un nivel de significación do 1%. Obtivéronse os seguintes datos sobre a porcentaxe de persoas que presentaban diarrea:

	Laboratorio	Competencia
Número de suxeitos	465	195
Número de casos de diarrea	9	1

   1. ¿Podemos afirmar que as porcentaxes son significativamente diferentes?
   2. Dar un intervalo de confianza para a diferencia de porcentaxes.

   [Milton 8.6.6 p. 285]

5. En 1970 fixéronse 759 análises de sangue e atopáronse 46 casos de infección. En 1975 outro estudo semellante descubriu 109 infeccións en 838 análises. Baseándose nestas dúas mostras, ¿podemos estar seguros de que a proporción de casos de infección aumentou en máis de 6 puntos porcentuais neses cinco anos? (Usar nivel de confianza do 90%.)

   [Milton 8.6.4]

6. A partir dos corenta anos, o cancro de mama pode detectarse a través dunha mamografía. Comprobamos que en 31 mulleres novas afectadas (idade 40-49 anos) houbo 6 casos descubertos a través de mamografía. Por outra parte, nun grupo de 101 mulleres de máis idade, a mamografía foi eficaz en 38 casos. Cun nivel de confianza do 95%, ¿podemos afirmar que a mamografía é menos eficaz nas mulleres novas?

   [Milton 8.6.3 p. 285]

7. Para ver se un medidor portátil de glucosa é útil para os diabéticos, mediuse para cada paciente o nivel de glucosa en sangue antes de aprender a usalo, e unhas semanas despois. Nunha mostra aleatoria de 36 individuos atopouse unha diferencia de 2.78mmol/l entre "antes" e "despois", con cuasi-desviación típica das diferencias igual a 6.05. ¿Quere dicir isto que o medidor é efectivo para axudar a reduci-los niveis de glucosa?

   [Milton 9.5.3 p. 319]

8. Os datos de temperatura en 1000 estacións meteorolóxicas en todo o mundo deron unha temperatura media de 57 graos Farenheit en 1950, e de 57.6 en 1988, con $s_D = 4.1$. ¿Quere isto dicir que a temperatura media do globo aumentou? (Emprega-lo valor P.) Dar un intervalo para o aumento global medio (para un nivel de confianza do 90%).

   [Milton 9.5.5 p. 320]

▼ Problemas de repaso de estimación e contraste de hipóteses

1. Para que un peixe sobreviva, a cantidade de osíxeno disolto na auga non debe ter unha desviación típica maior cá 1.2 partes por millón. Tomamos mostras de auga en 25 lugares aleatoriamente escollidos dun lago e obtemos $s = 1.7$ppm. ¿Evidencia isto que a variabilidade do osíxeno aumentou por riba do parámetro aceptable $\sigma = 1.2$?

   [Milton 7.2.3 p. 256]

2. Nun estudo sobre rexeneración de células nerviosas en monos *rhesus* mediuse o contido en creatinina fosfato na parte esquerda e na parte dereita da espiña dorsal (medido en mg de CF por cada 100g de tecido). Para un nivel de significación do 10%, ¿existe unha evidencia significativa na cantidade de CF entre os dous datos? Os datos son os seguintes:

Animal	1	2	3	4	5	6	7	8
Lado dereito	16.3	4.8	10.9	14.2	16.3	9.9	29.2	22.4
Lado esquerdo	11.5	3.6	12.5	6.3	15.2	8.1	16.6	13.1

   [Samuels p. 333]

3. Crese que a maioría dos fumadores empezan a fumar despois dos 18 anos. Nunha mostraxe con 60 individuos, atopouse que o 49% empezou a fumar despois desa idade.
   1. Decidir se hai evidencia de que na poboación a proporción de fumadores que empeza despois dos 18 é menor có 50% (cun nivel de significación do 1%).
   2. Explica-las consecuencias económicas e sanitarias de cometer un erro de tipo I ou un erro de tipo II.

   [Milton 6.4.6 p. 226]

4. Existe a teoría de que a vitamina C é beneficiosa no tratamento do cancro. Os que a defenden din que hai unha melloría superior ó 4% de casos. Fixemos dous grupos independentes de 75 individuos cada un. Ós primeiros démoslle 10g diarios de vitamina C; ós outros, nada. Ó cabo de catro semanas, no primeiro grupo 47 pacientes presentaron algunha melloría, mentres que este número foi soamente de 43 no segundo grupo. Pídese face-lo contraste $H_0 \colon p_1 - p_2 \leq 0.04$ e interpreta-lo resultado (emprega-lo valor P).

   [Milton 8.6.1 p. 280 e 8.6.2 p. 281]

5. Estase probando a eficacia de dous tipos de exercicio para mellora-los síntomas da artrite reumatoide. O primeiro tratamento (T1) foi probado en 150 pacientes con esta enfermidade obtenendo que 87 deles melloran tras un mes de práctica. O segundo tratamento (T2) foi probado en 160 pacientes dos que 72 melloraron tras un mes de práctica. ¿Podemos asegurar que hai evidencia significativa de que a proporción de pacientes que melloran co tratamiento T1 é superior á do T2? Realiza-lo correspondente contraste de hipóteses.

▼ Probas de homoxeneidade e independencia

1. Investígase a eficacia dunha nova vacina contra a gripe. Elíxese unha mostra de 900 persoas, e clasifícanse segundo que foran ou non vacinadas, e segundo contraeran a gripe durante o último ano ou non. Pídese, cun nivel de confianza do 95%, decidir se hai asociación ou non entre as dúas variables.

Vacinado \ gripe	si	non
si	150	200
non	300	250

[Milton 12.1.2 p. 449]

2. Cremos que existe relación entre o número de cloroplastos das follas das árbores e o nivel de SO_2 no aire. Selecciónanse 60 árbores, e clasifícanse en función do nivel de dióxido de azufre da súa zona e o nivel de cloroplastos das súas follas. Obtéñense os seguintes datos:

SO_2 \ Cloroplastos	alto	normal	baixo
alto	5	4	13
normal	5	10	5
baixo	7	9	2

   1. ¿Trátase dunha proba de independencia ou de homoxeneidade?
   2. ¿Que conclusións poden sacarse dos datos? Enuncia a hipótese nula apropiada e razoa en función do valor P obtido.

[Milton 12.2.6 p. 457]

3. Co obxectivo de provoca-la unión dos ósos en fracturas, aplícanse campos electromagnéticos pulsantes. Nunha mostra de 62 fracturas de tibia, 26 de húmero, e 18 de fémur, observouse que o tratamento só tivo éxito en 34, 16, e 10 delas, respectivamente.
   - Construí-la táboa de continxencia axeitada.
   - Á vista dos resultados obtidos na mostra, ¿pódese concluír que o éxito do tratamento depende do tipo de óso que se está tratando?

4. Realízase un pequeno estudo piloto para determinar se hai asociación entre a aparición de leucemia e os antecedentes de alerxia. Selecciónase unha mostra de 19 pacientes con leucemia e outro grupo de control de 17 persoas, e determínase se hai antecedentes de alerxia ou non.

grupo \ antecedentes	si	non
paciente	17	2
control	5	12

   Calcula-la frecuencia esperada para cada celda e contrastar se a distribución de casos de alerxia é homoxénea nos dous grupos. Explica-la resposta baseándose no valor P do contraste.

   [Milton 12.1.5 p. 449]

5. Nun estudo sobre quimioterapia no cancro de pulmón administráronse simultaneamente catro medicamentos a 16 pacientes, mentres que a outro grupo de 11 pacientes déronselle os medicamentos de xeito secuencial. Observouse unha resposta positiva ó tratamento en 11 pacientes do primeiro grupo, e en 3 dos tratados secuencialmente. ¿Proporcionan estes datos evidencia de que unha forma de tratamento é superior á outra?

   [Samuels 10.2.10]

▼ Regresión linear e ANOVA

1. Realízase un estudo para estima-la relación entre o índice de obesidade X e a taxa metabólica en repouso Y. A partir dos datos de 43 individuos obtemos

$$\sum X = 1482.5; \qquad \sum Y = 10719;$$
$$\sum X^2 = 53515.25; \qquad \sum Y^2 = 2736063; \qquad \sum XY = 379207.5.$$

   1. ¿Que taxa metabólica correspondería a un índice de obesidade $X = 40$?
   2. Calcular e interpreta-lo coeficiente de determinación.
   3. Contrasta-lo modelo de regresión linear.

   [Milton 11.3.4 p. 414]

2. A seguinte táboa recolle os datos de presións sistólicas (P) de cinco individuos en función da súa idade (t):

t idade (anos)	20	30	40	50	60
P presión (mm Hg)	125	128	131	133	138

   1. ¿Que ecuación linear nos permite estimar P para un individuo de 25 anos?
   2. Calcula-lo coeficiente de determinación e interpreta-lo resultado.
   3. Contrasta-lo modelo de regresión linear.

3. Realizouse un experimento para estima-la concentración plasmática Y dunha substancia a partir da súa concentración X na saliva. Os datos experimentais foron:

X	7.4	7.5	8.5	9.0	11.0	13.0	14.0	14.5	16.0	17.0
Y	30.0	25.0	31.5	27.5	40.2	48.0	52.0	54.0	56.5	58.0

   Calcula-la recta de regresión e contrasta-lo modelo de regresión linear (ANOVA).

4. A cantidade de arsénico no arroz (variable Y, en µg/kg) parece estar relacionada coa de silicio na palla de arroz (variable X, en g/kg). Ó estudar 32 plantas obtémo-los seguintes datos:

$$\overline{X} = 29.85, \quad s_X = 10.04, \quad \overline{Y} = 122.25 \quad s_Y = 44.50, \quad r = -0.556.$$

   1. ¿Que cantidade de arsénico estimamos cando $X = 12$?
   2. Calcula-la varianza residual dos erros de estimación.
   3. ¿Que proporción de varianza da concentración de arsénico está explicada pola relación linear co contido de silicio?

   [Samuels p. 505]

5. Aplicáronse dous cuestionarios a 670 persoas: un medía o nivel de estrés ó que estiveran sometidas X, e o outro detectaba posibles trastornos de saúde Y. Ó calcula-lo coeficiente de correlación de Pearson obtívose $r = 0.24$. ¿É compatible este resultado coa hipótese $\rho = 0$? (tomar $\alpha = 5\%$)

6. Déronse distintas doses dunha substancia velenosa a sete grupos de 26 ratos, e observáronse os seguintes resultados:

X doses (mg)	4	6	8	10	12	14	16
Y número de mortes	1	3	6	8	14	16	20

   1. Calcula-la ecuación da recta de mínimos cadrados axustada a estes datos.
   2. Estima-lo número de mortes nun grupo de 26 ratos que recibiron unha dose de 7mg deste veleno.
   3. Contrasta-lo modelo de regresión linear.

7. Lévase a cabo un estudo sobre as características corporais e o modo de actuar de levantadores de peso olímpicos. Estúdanse as variables X, peso corporal, e Y, mellor levantamento, obtendo:

X	134	138	154	178	176	190	190	205	205	206
Y	185	238	260	290	312	336	339	341	358	359

   1. Debuxa-la nube de puntos. Baseándose nela, ¿pódese esperar que b sexa positivo ou negativo?
   2. Calcular e interpreta-lo coeficiente de determinación.
   3. Comproba-la idoneidade do modelo de regresión linear. Se é axeitado calcula-la liña de regresión de X sobre Y, estima-lo mellor levantamento dun atleta que pesa 200 libras.

   [Milton 11.4.1]

Exames resoltos

Exame 1

Problema. Ó estuda-la coagulación do sangue utilízase a variable normal X, tempo parcial activado en segundos da tromboplastina. Os valores seguintes representan unha mostra aleatoria de 10 observacións sobre X para un determinado paciente:

45 40 47 46 42 50 47 48 49 49.

1. Construír un intervalo para o tempo parcial medio da tromboplastina para ese paciente, cun nivel de confianza do 99%.
2. Se a varianza poboacional é 9, ¿cal ten que se-lo tamaño da mostra para que a diferenza entre a media mostral e a media poboacional sexa como moito de ±1 segundo, cun nivel de confianza do 99%?

Solución. Sexa pois X="tempo parcial activado en segundos da tromboplastina".

Para o primeiro apartado temos que calcular un intervalo de confianza para a media empregando o estatístico

$$\frac{\overline{X}-\mu}{s_{n-1}/\sqrt{n}},$$

que segue unha distribución t_{n-1}. Despexando μ da inecuación

$$\left|\frac{\overline{X}-\mu}{s_{n-1}/\sqrt{n}}\right| \leq t_{n-1,\,\alpha/2}$$

obtense a fórmula

$$\overline{X} \pm t_{n-1,\,\alpha/2}\frac{s_{n-1}}{\sqrt{n}}.$$

Temos $n=10$. Organizámo-los cálculos para calcula-la media e cuasi-varianza mostral:

	X	X^2
	45	2025
	40	1600
	47	2209
	46	2116
	42	1764
	50	2500
	47	2209
	48	2304
	49	2401
	49	2401
Σ	463	21529

De aquí obtemos $\overline{X} = 463/10 = 46.3, \quad s_n^2 = 21529/10 - 46.3^2 = 9.21,$ e así, $s_{n-1} = \sqrt{\frac{10}{9}\, 9.21} = 3.20.$

Nivel de significación $\alpha = 0.01$. Buscámo-lo valor $t_{9,\, 0.005} = 3.25$ nas táboas. Aplicando a fórmula

$$46.3 \pm 3.25 \frac{3.20}{\sqrt{10}} = 46.3 \pm 3.29,$$

de onde se deduce o intervalo $[43.01,\ 49.59]$.

Conclusión: cun nivel de confianza do 99%, o tempo parcial activado medio da tromboplastina atópase entre 43.01 e 49.59 segundos.

Para o segundo apartado témo-lo dato $\sigma^2 = 9$. Como a varianza poboacional é coñecida, empregámo-lo estatístico

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

que ten distribución normal estándar. Despexando μ da inecuación

$$\left| \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \right| \leq Z_{\alpha/2}$$

obtémo-la fórmula

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

A estimación do erro é $Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, e queremos $Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \epsilon$, onde ϵ é o valor fixado polo problema. Despexando n obtense $n \geq (Z_{\alpha/2}\, \sigma/\epsilon)^2$.

O nivel de confianza é $\alpha = 0.01$. Mirando as táboas obtemos $Z_{0.005} = 2.58$. Neste caso $\epsilon = 1$. Substituíndo na fórmula $n \geq (2.58 \cdot 3/1)^2 = 59.7$.

Conclusión: para que a diferenza entre a media mostral e a media poboacional no tempo parcial activado en segundos da tromboplastina sexa como moito de ±1 segundo cun nivel de confianza do 99%, teriamos que tomar unha mostra de polo menos 60 elementos. ■

Problema. Estase a probar un antibiótico chamado DOXICICLINA para previr a "diarrea do viaxeiro". O fármaco foi probado sobre 64 voluntarios que foron a Kenya. A unha metade déuselle doxiciclina e á outra un placebo. Dos que recibiron doxiciclina, 24 libráronse do trastorno, mentres que só 16 dos do outro grupo se libraron.

1. Construír un intervalo de confianza do 95% para a diferenza entre as porcentaxes de protección entre aqueles que utilizaron doxiciclina e os que non a utilizaron. Interpreta-lo intervalo.
2. ¿Pódese asegurar que a doxiciclina contribúe a proporcionar protección contra a diarrea do viaxeiro? Explicalo sobre a base do valor P.

Solución. As variables aleatorias a considerar son X, non ter diarrea do viaxeiro entre voluntarios que tomaron doxiciclina, e Y, non ter diarrea do viaxeiro entre voluntarios que tomaron placebo.

Para o primeiro apartado temos que calcular un intervalo de confianza para a diferencia de porcentaxes empregando o estatístico

$$\frac{(\widehat{p_1 - p_2}) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

que segue unha distribución normal estándar. Nótese que as poboacións non están emparelladas. O intervalo de confianza pedido obtense despexando $p_1 - p_2$ da desigualdade

$$\left| \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \right| \leq Z_{\alpha/2},$$

de onde se obtén a fórmula

$$(\widehat{p_1} - \widehat{p_2}) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Temos como datos $n_1 = 32, \widehat{p_1} = 24/32 = 0.75, n_2 = 32, \widehat{p_2} = 16/32 = 0.5$.

Nivel de significación $\alpha = 0.05$. Buscamos na táboa $Z_{0.025} = 1.96$. Substituíndo na fórmula:

$$(0.75 - 0.5) \pm 1.96 \sqrt{\frac{0.75(1-0.75)}{32} + \frac{0.5(1-0.5)}{32}} = 0.25 \pm 0.229,$$

de onde se obtén o intervalo $[0.021,\ 0.479]$.

Conclusión: cun nivel de confianza do 95%, a diferencia de proporción de viaxantes a Kenya que non tiveron a diarrea do viaxeiro entre os que tomaron doxiciclina e os que tomaron placebo sitúase entre o 2.1% e o 47.9%.

Para a segunda cuestión temos que face-lo contraste de hipóteses

$$H_0: p_1 \leq p_2, \qquad H_1: p_1 > p_2.$$

Este contraste ten cero como valor nulo. En consecuencia, agora temos que empregar-lo estatístico

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

que tamén segue unha distribución normal estándar, e onde

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

Substituíndo temos, en primeiro lugar

$$\hat{p} = \frac{32 \cdot 0.75 + 32 \cdot 0.5}{32 + 32} = 0.625,$$

o cal nos dá o valor no estatístico

$$\frac{0.75 - 0.5}{\sqrt{0.625(1-0.625)\left(\frac{1}{32} + \frac{1}{32}\right)}} = 2.07.$$

Calculamos agora o valor P mirando a táboa da distribución normal: $P = P(z > 2.07) = 0.01923$. Temos que $1\% < P < 2.5\%$.

Conclusión: rexeitámo-la hipótese nula e concluímos que existe evidencia significativa, polo menos do 97.5%, de que a doxiciclina aumenta a proporción de viaxantes a Kenya que non teñen diarrea do viaxeiro fronte a aqueles que tomaron placebo. Por tanto, a doxiciclina contribúe a proporcionar protección contra a diarrea do viaxeiro. ■

Problema. A seguinte táboa representa as presións sanguíneas sistólicas (mm Hg) de 10 individuos alcohólicos rehabilitados, antes e despois de deixa-la bebida

Individuo	1	2	3	4	5	6	7	8	9	10
Antes	140	165	160	160	175	190	170	175	155	160
Despois	145	150	150	155	170	175	160	165	145	170

Supoñendo que as poboacións están distribuídas normalmente,

1. Estimar mediante un intervalo de confianza do 95% o cambio da presión sistólica que produce o abandono do alcohol. Interpretar o devandito intervalo.
2. ¿Hai evidencias suficientes, cun nivel de significación do 5%, para dicir que a presión sanguínea sistólica diminúe despois de deixa-la bebida?

Solución. As variables aleatorias a considerar son X, presión sanguínea sistólica dun alcohólico antes de deixa-la bebida, e Y, presión sanguínea sistólica dun alcohólico despois de deixa-la bebida. Obviamente trátase dun problema de comparación de dúas poboacións con mostras emparelladas, así que debemos toma-la variable diferencia $D = X - Y$.

O estatístico que temos que tomar é $\frac{\overline{D}-\mu_D}{s_D/\sqrt{n}}$, que segue unha distribución t_{n-1}. O primeiro que facemos é dispoñe-los datos para calcula-los elementos da fórmula:

	X	Y	D	D^2
	140	145	-5	25
	165	150	15	225
	160	150	10	100
	160	155	5	25
	175	170	5	25
	190	175	15	225
	170	160	10	100
	175	165	10	100
	155	145	10	100
	160	170	-10	100
Σ	1650	1585	65	1025

Temos $n = 10$. Por tanto, $\overline{D} = 65/10 = 6.5$, $s^2_{n,\,D} = 1025/10 - 6.5^2 = 60.25$, e $s_{n-1,\,D} = \sqrt{\frac{10}{9}\,60.25} = 8.18$.

Como no primeiro apartado temos que calcular un intervalo de confianza, despexamos μ_D da desigualdade

$$\left| \frac{\overline{D} - \mu_D}{s_D/\sqrt{n}} \right| \leq t_{n-1,\,\alpha/2},$$

de onde obtemos $\overline{D} \pm t_{n-1,\,\alpha/2}\frac{s_D}{\sqrt{n}}$.

Nivel de significación $\alpha = 0.05$. Mirámo-lo valor $t_{9,\,0.025} = 2.2622$ nas táboas. Substituíndo na fórmula anterior obtemos

$$6.5 \pm 2.26\frac{8.18}{\sqrt{10}} = 6.5 \pm 5.85,$$

o que nos dá un intervalo $[0.64,\ 12.35]$.

Conclusión: cun nivel de confianza do 95%, a diferencia media das presións sanguíneas sitólicas dun alcohólico rehabilitado entre antes e despois de deixa-la bebida sitúase entre 0.6 e 12.3mm Hg.

Para a segunda parte do exercicio, temos que face-lo seguinte contraste de hipóteses:

$$H_0{:}\mu_D \leq 0, \qquad H_1{:}\mu_D > 0.$$

Como xa calculámo-los datos, substituímos no estatístico

$$\frac{6.5-0}{8.18/\sqrt{10}} = 2.51.$$

Pero agora necesitamos mirar na táboa $t_{9,\,0.05} = 1.83$, que é menor ca 2.51.

Conclusión: rexeitamos H_0 e concluímos que hai evidencia significativa, ó 95% de confianza, de que a presión sanguínea sistólica dun alcohólico rehabilitado dimínúe despois de deixa-la bebida. ■

Problema. Deseñouse un estudo para analiza-la posible relación entre o medio no que viven e a incidencia de trastorno depresivo das persoas no paro. Seleccionáronse suxeitos pertencentes a medios rurais, semiurbanos e urbanos. De cada medio seleccionouse unha mostra aleatoria de 100 suxeitos no paro, obtendo que 12 do rural, 16 do semiurbano e 32 do urbano presentaban trastorno depresivo.

1. Construí-la táboa de continxencia axeitada. ¿Trátase dunha proba de independencia ou de homoxeneidade?
2. ¿Pode afirmarse, cun 1% de nivel de significación, que na poboación de desempregados existe relación entre o tipo de medio no que se vive e padecer ou non trastorno depresivo?

Solución. Temos tres poboacións dependendo do medio no que viven, e a variable aleatoria Y="incidencia de trastorno depresivo". En primeiro lugar construímo-la táboa de continxencia:

medio \ trastorno	si	non	tamaño
rural	12	88	100
semiurbano	16	84	100
urbano	32	68	100
Σ	60	240	300

O tamaño da mostra en cada medio está fixado polo investigador, trátase dunha proba de homoxeneidade para datos categóricos. Por tanto, temos que face-lo contraste de hipóteses:

$$H_0{:}\, p_{11} = p_{21} = p_{31},\ p_{12} = p_{22} = p_{32}.$$

A continuación calculámo-los valores esperados no suposto de que houbese homoxeneidade nas poboacións mediante a fórmula $\widehat{E_{ij}} = \frac{n_i n_{\cdot j}}{n}$ (en verde), e tamén os valores $(n_{ij} - \widehat{E_{ij}})^2/\widehat{E_{ij}}$ (en vermello), obtendo:

medio \ trastorno	si	non	Σ
rural	12 20 3.2	88 80 0.8	100
semiurbano	16 20 0.8	84 80 0.2	100
urbano	32 20 7.2	68 80 1.8	100
Σ	60	240	300

Finalmente aprovéitanse todas estes contas para calcula-lo valor no estatístico, (que consiste en suma-los valores vermellos), para obter 14.

O estatístico segue unha distribución χ^2 con $(3-1)(2-1)=2$ graos de liberdade. Dannos un nivel de significación $\alpha=0.01$, así que índonos ás táboas obtemos $\chi^2_{2,\,0.01}=9.21$, que é menor ca 14.

Conclusión: rexeitámo-la hipótese nula, e concluímos que hai evidencia significativa, cun nivel de confianza do 99%, de que a incidencia de trastorno depresivo nas persoas en paro é distinto dependendo de se o medio no que viven é rural, semiurbano ou urbano. ■

Problema. Os seguintes datos corresponden a idade (X en anos) e a conduta agresiva (Y medida nunha escala de 0 a 10) dun grupo de 10 nenos, de entre 6 e 9 anos, elexidos ó azar

$$\sum X=75,\ \sum Y=49,\ \sum X^2=570.72,\ \sum Y^2=313,\ \sum XY=345.2.$$

1. Estima-la recta de regresion que permita predicir o valor da conduta agresiva en funcion da idade do neno.
2. Calcula-lo coeficiente de determinacion r^2 e interpreta-lo seu resultado.
3. Contrasta-lo modelo de regresion lineal.

Solución. Estamos chamando X á idade en anos, e Y á conducta agresiva dos nenos. Temos que calcula-la recta de regresión de Y sobre X.

Entón temos $n=10$ datos e

$$\begin{aligned}
\overline{X}&=\frac{75}{10}=7.5,\\
\overline{Y}&=\frac{49}{10}=4.9,\\
s_X^2&=\frac{570.72}{10}-7.5^2=0.82,\\
s_Y^2&=\frac{313}{10}-4.9^2=7.29,\\
s_{XY}&=\frac{345.2}{10}-7.5\cdot 4.9=-2.23.
\end{aligned}$$

Temos $b = -2.23/0.82 = -2.71$ e $a = 4.9 + 2.71 \cdot 7.5 = 25.25$ co que a ecuación da recta de regresión é

$$y = 25.25 - 2.71x.$$

A estimación do coeficiente de correlación é

$$r = \frac{-2.23}{\sqrt{0.82 \cdot 7.29}} = -0.91,$$

de xeito que a calidade da aproximación parece bastante boa.

A estimación do coeficiente de determinación é $r^2 = 0.830$. Isto interprétase do seguinte xeito: o 83% da variabilidade da variable Y está explicada polo modelo de regresión.

Para contrasta-lo modelo de regresión linear temos que facer

$$H_0\colon \rho = 0, \qquad H_1\colon \rho \neq 0.$$

Empregamos pois a técnica de análise da varianza, ANOVA. Os datos necesarios están recollidos na seguinte táboa:

variabilidade	g.l.	SS	MS	cociente
regresión	1	$SS_R = 10 \cdot 0.83 \cdot 7.29 = 60.50$	$MS_R = 60.50$	39.02
erro	8	$SS_E = 10(1 - 0.83)7.29 = 12.40$	$MS_E = \frac{12.40}{8} = 1.55$	
total	9	$SS_Y = 10 \cdot 7.29 = 72.9$		

Como $P = P(F_{1,8} \geq 39.02) < 0.01$ é un número moi pequeno (de feito, empregando software estatístico temos $P = 0.00025$), rexeitámo-la hipótese nula. Concluímos que hai evidencia significativa de que o modelo de regresión linear é válido. ■

Exame de xuño de 2019

Problema. Moi recentemente, o xornal THE SUN publicou os resultados dun estudo sobre o peso dos paquetes de patacas fritas que as distintas cadeas de comida rápida serven en Inglaterra. O estudo consistiu en comprar tres paquetes de patacas de cada cadea en diferentes establecementos da mesma. En particular, para unha das cadeas, os resultados obtidos foron: 106g, 102g e 108g.

1. A partir da mostra, calcula un intervalo de confianza, cun nivel de confianza do 95%, para o peso medio dos paquetes de patacas na devandita cadea.
2. Pódese afirmar, desde o punto de vista estatístico, que o peso medio real dos paquetes de patacas fritas nesa cadea é inferior a 108g?

Solución. Considerámo-la variable aleatoria X="peso dun paquete de patacas fritas".

Organizámo-los cálculos para obte-la media e cuasi-varianza mostral:

	X	X^2
	106	11236
	102	10404
	108	11664
Σ	316	33304

De aquí obtemos $n=3$, $\overline{X}=\frac{316}{3}=105.333$, $s_n^2=\frac{33304}{3}-105.333^2=6.222$, e así, $s_{n-1}=\sqrt{\frac{3}{2}\cdot 6.222}=3.055$.

Calculamos un intervalo de confianza para unha media empregando o estatístico

$$\frac{\overline{X}-\mu}{s_{n-1}/\sqrt{n}},$$

que segue unha distribución t_{n-1}. Despexando μ da desigualdade

$$\left|\frac{\overline{X}-\mu}{s_{n-1}/\sqrt{n}}\right|\leq t_{n-1,\,\alpha/2},$$

obtense a fórmula

$$\overline{X}\pm t_{n-1,\,\alpha/2}\frac{s_{n-1}}{\sqrt{n}}.$$

O nivel de significación é $\alpha=0.05$. Calculamos $t_{2,\,0.025}=4.303$. Substituíndo na fórmula

$$105.333\pm 4.303\cdot\frac{3.055}{\sqrt{3}}=105.333\pm 7.589,$$

de onde se obtén o intervalo $[97.744,\ 112.922]$.

Conclusión: cun nivel de confianza do 95.0%, a media do peso dun paquete de patacas fritas atópase entre 97.744 e 112.922.

Agora facémo-lo contraste de hipóteses

$$H_0\colon \mu \geq 108, \qquad H_1\colon \mu < 108.$$

Este é un contraste de hipóteses para unha media. Para iso empregámo-lo estatístico

$$\frac{\overline{X} - \mu}{s_{n-1}/\sqrt{n}},$$

que segue unha distribución t_{n-1}.

O valor no estatístico é

$$\frac{105.333 - 108}{3.055/\sqrt{3}} = -1.512.$$

Calculámo-lo valor P como $P = P(t_2 < -1.512) = 0.1349$, que é un valor relativamente grande.

Conclusión: Aceptamos H_0, e concluímos que non hai evidencia significativa, ata un nivel de confianza do 86.5%, de que a media de peso dun paquete de patacas fritas sexa menor ca 108. ■

Problema. Para saber se o olor a lavanda na sala de espera dos dentistas diminúe a ansiedade dos pacientes, un equipo de investigadores seleccionou a 597 pacientes que dividiu aleatoriamente en dous grupos. Os do primeiro grupo (310 pacientes), que chamaremos "grupo de control", esperaron en salas sen aroma especial, mentres que os do segundo grupo (287 pacientes), que chamaremos "grupo de tratamento", esperaron en salas con aroma a lavanda. Para determina-lo nivel de ansiedade, tódolos pacientes se someteron a diferentes test psicolóxicos que permiten medilo. Se nos test de ansiedade a media do grupo de control foi de 15.40 cunha cuasi-desviación típica de 4.18, e no grupo de tratamento a media mostral foi 11.74 cunha cuasi-desviación típica de 4.10, ¿podemos afirmar que o aroma de lavanda nas salas de espera dos dentistas axuda a reduci-lo nivel de ansiedade nos pacientes? NOTA: supoñede que as varianzas poboacionais son iguais.

Solución. Considerámo-las variables aleatorias X="nivel de ansiedade no grupo de control" e Y="nivel de ansiedade no grupo de tratamento".

Temos $n_1 = 310, \overline{X} = 15.4, s_1 = 4.18$ e $n_2 = 287, \overline{Y} = 11.74, s_2 = 4.1$.

Asumimos que as varianzas das dúas poboacións son iguais.

Facémo-lo contraste de hipóteses

$$H_0\colon \mu_1 - \mu_2 \leq 0, \qquad H_1\colon \mu_1 - \mu_2 > 0.$$

Este é un contraste de hipóteses para unha diferencia de medias. Para iso empregámo-lo estatístico

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

que segue unha distribución $t_{n_1+n_2-2}$.

Aquí considerámo-la cuasi-varianza ponderada, que se define como

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}.$$

Substituíndo na fórmula da cuasi-varianza ponderada obtemos

$$s_p=\sqrt{\frac{(310-1)\cdot 4.18^2+(287-1)\cdot 4.1^2}{310+287-2}}=4.142.$$

O valor no estatístico é

$$\frac{(15.4-11.74)-0}{4.142\sqrt{\frac{1}{310}+\frac{1}{287}}}=10.788.$$

Calculámo-lo valor P como $P=P(t_{595}>10.788)=0.3\cdot 10^{-24}$, que é un valor pequeno.

Conclusión: Rexeitamos H_0, e concluímos que hai evidencia significativa, cun nivel de confianza do 99.9%, de que, en media, o nivel de ansiedade no grupo de control é maior có nivel de ansiedade no grupo de tratamento.

Por tanto, o aroma a lavanda na sala de espera dos dentiastas axuda a reduci-lo nivel de ansiedade nos pacientes. ■

Problema. Para analiza-lo risco de sufrir un aborto espontáneo nos embarazos de mulleres hipertensas tratadas con inhibidores da encima convertidora de anxiotensina (IECA) durante o primeiro trimestre do embarazo, estudáronse 329 casos nos que se observaron 47 abortos espontáneos.

1. Se a taxa de abortos espontáneos na poboación fose do 10%, poderíase afirmar que o tratamento con IECA no primeiro trimestre de embarazo incrementa a porcentaxe de abortos espontáneos?
2. Cal tería que se-lo tamaño mostral mínimo para poder estimar, a un nivel de confianza do 95.5%, a proporción de abortos espontáneos na poboación cun erro inferior ó 2%?

Solución. Considerámo-la variable aleatoria X="abortos espontáneos de mulleres hipertensas tratadas con IECA durante o primeiro trimestre do embarazo".

Temos $n=329$, e $\hat{p}=0.143$.

Facémo-lo contraste de hipóteses

$$H_0: p\leq 0.1, \qquad H_1: p>0.1.$$

Este é un contraste de hipóteses para unha proporción. Para iso empregámo-lo estatístico

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}},$$

que segue unha distribución normal estándar.

O valor no estatístico é

$$\frac{0.143-0.1}{\sqrt{\frac{0.1(1-0.1)}{329}}}=2.591.$$

Calculámo-lo valor P como $P=P(Z>2.591)=0.0048$, que é un valor pequeno.

Conclusión: Rexeitamos H_0, e concluímos que hai evidencia significativa, cun nivel de confianza do 99.5%, de que a proporción de abortos espontáneos de mulleres hipertensas tratadas con IECA durante o primeiro trimestre do embarazo é maior ca 10.0%.

Para estima-lo tamaño da mostra para unha proporción, empregámo-lo estatístico

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}},$$

que ten distribución normal estándar. Despexando p da desigualdade

$$\left|\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right|\leq Z_{\alpha/2},$$

obtémo-la fórmula

$$\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

A estimación do erro é $Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Neste case non temos unha estimación da proporción $\hat{p}$. É sinxelo ver que a función $x\mapsto\sqrt{x(1-x)}$ alzanza o seu máximo no intervalo $[0,1]$ no punto $x=1/2$. Por tanto, necesitamos despexar n da desigualdade $Z_{\alpha/2}\sqrt{\frac{0.5(1-0.5)}{n}}<\epsilon$, onde ϵ é o valor fixado polo problema. Así, obtense $n>\left(\frac{Z_{\alpha/2}}{2\epsilon}\right)^2$.

O nivel de significación é $\alpha=0.045$. Mirando as táboas obtemos $Z_{0.0225}=2.005$. Neste caso $\epsilon=0.02$. Substituíndo na fórmula $n>\left(\frac{2.005}{2\cdot 0.02}\right)^2=2511.65$.

Conclusión: para que a diferenza entre a proporción mostral e a proporción poboacional de abortos espontáneos de mulleres hipertensas tratadas con IECA durante o primeiro trimestre do embarazo sexa como moito de ±0.02 cun nivel de confianza do 95.5%, teriamos que tomar unha mostra de polo menos 2512 elementos. ■

Problema. Co obxectivo de estuda-la relación entre a aparición de depresión post-parto e o nivel de seguridade alimentaria, observáronse 325 casos de mulleres seleccionadas aleatoriamente en centros de saúde no oeste da cidade de Teherán (Irán). Clasificouse, de acordo coa seguridade alimentaria, ós fogares das devanditas mulleres en tres niveles: A1: Alimentación asegurada, A2: Alimentación non asegurada pero sen fame, A3: Alimentación non asegurada e con fame moderada ou severa. Dos 325 casos, 214 eran de fogares do tipo A1, 56 do tipo A2, e 55 do tipo A3. Dos 115 casos de depresión post-parto, 51 eran en mulleres con fogares de nivel A1, e 24 en mulleres con fogares de nivel A2.

1. Constrúe a táboa de continxencia e realiza o test estatístico adecuado para comprobar se hai relación entre a seguridade alimentaria no fogar e o feito de sufrir de depresión post-parto entre as mulleres da cidade de Teherán.
2. O test anterior, ¿é unha proba de independencia ou é unha proba de homoxeneidade? Razoa a resposta.

Solución. Temos tres poboacións dependendendo da seguridade alimentaria e a variable aleatoria Y ="depresión postparto". En primeiro lugar construímo-la táboa de continxencia:

Alimentación \ depresión	si	non	Σ
A1	51	163	214
A2	24	32	56
A3	40	15	55
Σ	115	210	325

Como o tamaño da mostra está determinado en toda a poboación, e o investigador simplemente clasifica os datos en dúas categorías, trátase dun contraste de independencia para datos categóricos. Por tanto, temos que face-lo contraste de hipóteses:

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j},\ i \in \{1,2,3\},\ j \in \{1,2\}.$$

A continuación calculámo-las frecuencias esperadas, no suposto de que a hipótese nula sexa certa, mediante a fórmula $\widehat{E_{ij}} = \frac{n_{i\cdot}n_{\cdot j}}{n}$ (en verde), e tamén os valores intermedios do estatístico $(n_{ij} - \widehat{E_{ij}})^2/\widehat{E_{ij}}$ (en vermello), obtendo:

Alimentación \ depresión	si	non	Σ
A1	51 75.72 8.07	163 138.28 4.42	214
A2	24 19.82 0.88	32 36.18 0.48	56
A3	40 19.46 21.67	15 35.54 11.87	55
Σ	115	210	325

Calcúlase o valor no estatístico, que consiste en suma-los valores vermellos. O resultado é 47.4.

O estatístico $\sum_{i,j} \frac{(n_{ij}-\widehat{E_{ij}})^2}{\widehat{E_{ij}}}$ segue unha distribución χ^2 con $(3-1)(2-1)=2$ graos de liberdade. Calculando o valor P temos $P=P(\chi^2_2 \geq 47.4)=0.5\cdot 10^{-10}$.

Conlusión: rexeitámo-la hipótese nula, e por tanto, temos evidencia significativa, de que hai relación entre as dúas variables. ■

Problema. Co obxectivo de facer un modelo linear para predici-la altura dunha persoa a partir da lonxitude da súa tibia, nunha mostra aleatoria de 20 persoas medíronse en centímetros tanto a súa tibia dereita (variable X), como a súa altura (variable Y) obténdose os seguintes valores:

$$\sum X=72.27; \qquad \sum Y=322.48;$$
$$\sum X^2=262.29; \qquad \sum XY=1168.05; \qquad \sum Y^2=5206.53.$$

1. Calcula a recta de regresión.
2. Calcula o coeficiente de determinación r^2 e interpreta o seu resultado.
3. Contrasta o modelo de regresión.

Nota: tomar 4 díxitos de precisión nos cálculos.

Solución. Considerámo-las variables aleatorias X="lonxitude da tibia dereita" e Y="altura".

Organizámo-los cálculos nunha táboa.

	X	Y	X^2	XY	Y^2
Σ	72.27	322.48	262.29	1168.05	5206.53

Temos $n=20$ datos e

$$\begin{aligned}
\overline{X} &= \frac{72.27}{20} = 3.613,\\
\overline{Y} &= \frac{322.48}{20} = 16.124,\\
s_X^2 &= \frac{262.29}{20} - 3.613^2 = 0.057,\\
s_Y^2 &= \frac{5206.53}{20} - 16.124^2 = 0.343,\\
s_{XY} &= \frac{1168.05}{20} - 3.613 \cdot 16.124 = 0.138.
\end{aligned}$$

De aquí obtemos $b = 0.138/0.057 = 2.424$ e $a = 16.124 - 2.424 \cdot 3.613 = 7.367$, co que a ecuación da recta de regresión é

$$y = 7.367 + 2.424x.$$

A estimación do coeficiente de correlación é

$$r = \frac{0.138}{\sqrt{0.057 \cdot 0.343}} = 0.989.$$

A calidade da aproximación é forte.

O coeficiente de determinación vén dado por $r^2 = 0.978$. Isto interprétase do seguinte xeito: o 97.8% da variabilidade da variable Y está explicada polo modelo de regresión.

Contrastámo-la validez do modelo de regresión linear. Para iso facémo-lo contraste de hipóteses

$$H_0\colon \rho = 0, \qquad H_1\colon \rho \neq 0.$$

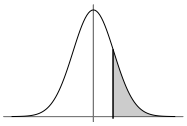
Empregamos pois a técnica de análise da varianza, ANOVA. Os datos necesarios están recollidos na seguinte táboa:

variabilidade	g.l.	SS	MS	cociente
regresión	1	$SS_R = 20 \cdot 0.978 \cdot 0.343 = 6.71$	$MS_R = 6.71$	789.789
erro	18	$SS_E = 20 \cdot (1 - 0.978) \cdot 0.343 = 0.153$	$MS_E = \frac{0.153}{18} = 0.008$	
total	19	$SS_Y = 20 \cdot 0.343 = 6.862$		

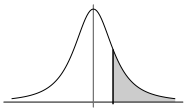
Como $P = P(F_{1,18} \geq 789.789) = 0.3 \cdot 10^{-15}$ é un valor pequeno, rexeitámo-la hipótese nula. Concluímos que hai evidencia significativa de que o modelo de regresión linear é válido. ■

Táboas estatísticas

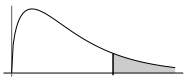
Táboas estatísticas que se empregan neste curso. Tamén se poden baixar en formato pdf.



Táboa da distribución normal

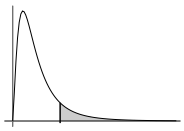


Táboa da distribución t de Student



Táboas da distribución χ^2 de Pearson

$\alpha \leq 0.4$ $\alpha \geq 0.45$



Táboas da distribución F de Fisher-Snedecor

$\alpha = 0.1$ $\alpha = 0.05$ $\alpha = 0.025$ $\alpha = 0.01$