Estadística fundamental Intervalos de confianza, inferencia y contrastes

Máster en Data Science y Big Data (Finanzas)

Pilar Barrios

Diciembre 2021

Índice

1 Inferencia estadística

1.a Introducción

Estamos interesados en responder preguntas del tipo siguiente:

Ejemplo 1: La media de ventas de un producto de una empresa es de 250 unidades diarias. Se realiza una campaña de promoción del producto y se anotan las ventas en los 8 días siguientes. Los resultados de las ventas son los siguientes: 270, 320, 180, 200, 270, 240, 280 y 200.

¿Se puede afirmar que han aumentado las ventas del producto?

Ejemplo 2 : Se tiene la serie de los rendimientos diarios de un activo. ¿Podemos suponer que siguen una distribución normal? Hemos ajustado un modelo a unos datos y disponemos de los residuos del modelo. ¿Son normales? ¿Presentan algún tipo de estructura?

1.a Introducción

Uno de los objetivos fundamentales de la Estadística es **inferir** las características de una población que no es completamente observable analizando una parte de ella, llamada **muestra**.

La selección de la muestra debe ser adecuada para garantizar su **representatividad**.

En estas sesiones se presentan los conceptos básicos de inferencia para la estimación de medias, varianzas y proporciones y se plantea el problema del contraste de hipótesis.

Una **población** es un conjunto de elementos homogéneo respecto a una variable que se desea estudiar. Suponemos que en cada elemento de la población se ha definido una variable y que se desea conocer su distribución entre los elementos de la población. Generalmente estamos interesados en conocer los **parámetros** de la población, aquellas características que sirven para determinarla (**contrastes paramétricos**). Si la muestra está bien escogida, podemos obtener una información similar a la del censo (estudio exhaustivo de todos los elementos de la población) con mayor rapidez y menor coste.

1.a Introducción

El **estimador** es un valor que puede calcularse a partir de los datos muestrales y que proporciona información sobre el valor del parámetro. Por ejemplo, la media muestral es un estimador de la media poblacional.

Otro aspecto clave en el estudio de estimadores es conocer su **precisión**, es decir, su capacidad de informarnos con exactitud del valor del parámetro.

En otras ocasiones estaremos interesados en contrastar la hipótesis de que una cierta muestra está extraída de una distribución de probabilidad dada. En este caso, estamos hablando de **contrastes no paramétricos**.

1.b Inferencia estadística. Estimadores

- La **Inferencia Estadística** es el proceso de calcular estimaciones y hacer las pruebas pertinentes (contrastes) sobre esas estimaciones con el objetivo de inferir características de la población.

Inferencia estadística

Estimación

Contraste de hipótesis

- Mediante este procedimiento, los analistas toman decisiones sobre la base de un análisis estadístico de la muestra de las diferentes variables aleatorias.

1.b Inferencia estadística. Estimadores

Una **muestra** es un subconjunto de la población.

Basándonos en la muestra estimamos los **parámetros muestrales** y hacemos inferencia sobre los parámetros poblacionales.

Estimadores puntuales: Valores que permiten inferior los parámetros poblacionales

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \longrightarrow$$

La media muestral es una estimación de la media poblacional μ. Es el estimador puntual de la media.

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \longrightarrow$$

En muchos casos se utiliza "error estándar“ en lugar de “desviación típica”.
La varianza muestral es un estimador de la varianza poblacional. Es el estimador puntual de la varianza.

Intervalo de confianza: se utiliza para estimar un resultado dentro de un rango de valores en el que el valor del parámetro estará con una probabilidad 1-α. α se llama el **nivel de significación** y 1-α es el nivel de confianza.

Ejemplo: La media poblacional estará entre los valores 10 y 15 con un nivel de confianza del 95%.

1.b Inferencia estadística. Estimadores

¿Cuándo se usan valores de la normal, de la t o de otras distribuciones?

Los intervalos de confianza se construyen así:

Estimador puntual +/- (factor de confianza* error estándar)

Valor del estimador puntual

Depende de la distribución de la estimación puntual y del nivel de confianza 1-α

Error estándar del estimador puntual

1.b Inferencia estadística. Estimadores

Propiedades de un estimador (o **estimador muestral**):

Estimador **insesgado**: El valor esperado del estimador es igual al verdadero valor del parámetro que desea estimar. Es decir, el valor esperado de la media muestral es igual a la media poblacional.

Estimador insesgado que también es **eficiente**, si la varianza de la distribución muestral es menor que la varianza de cualquier otro estimador insesgado del parámetro que se está intentando estimar.

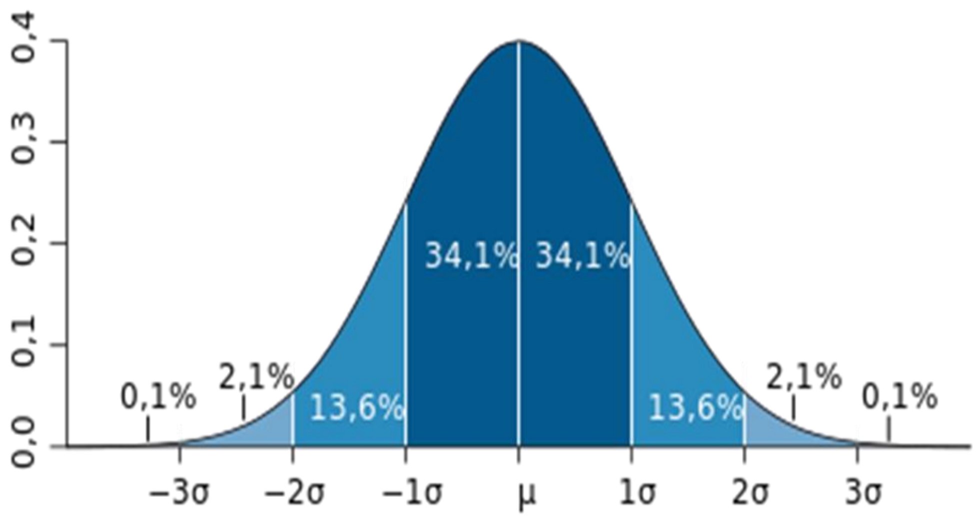
Estimador **consistente**, la precisión de la estimación aumenta a medida que el tamaño muestral aumenta.

En muchas ocasiones el estimador puntual es un estimador lineal. Si además es el mejor posible, es lineal y es insesgado se dice que es el mejor estimador lineal insesgado (BLUE, Best Linear Unbiased Estimator).

Los cálculos serían una función lineal de las observaciones (que es siempre el caso en problemas de Gauss) o no lineal.

1.b Inferencia estadística. Estimadores

La distribución normal, Z, se utiliza en la construcción de intervalos de confianza para la media de poblaciones con **varianza conocida**.



$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \equiv N(0,1)$$

$$Si \;\; X \sim N(\mu,\sigma)$$

$$\boldsymbol{P}(\boldsymbol{X} \le x) = \mathrm{P}(\mathrm{Z} \le \frac{x-\mu}{\sigma})$$

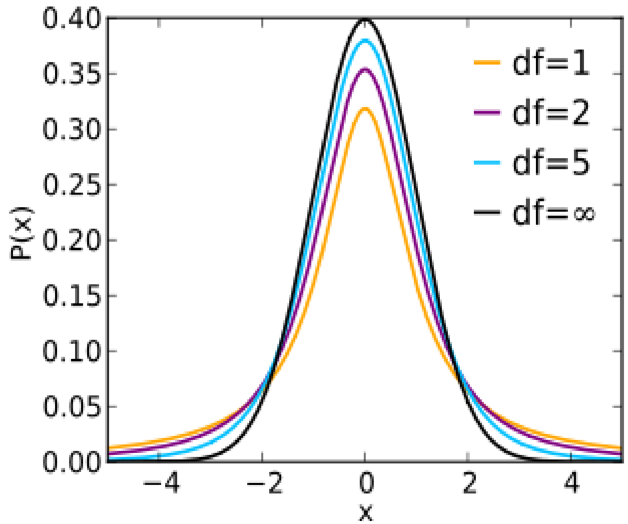
Valores críticos

El IC al 90% es $\bar{X} \pm 1.65\sigma_{\bar{X}}$
El IC al 95% es $\bar{X} \pm 1.96\sigma_{\bar{X}}$
El IC al 99% es $\bar{X} \pm 2.58\sigma_{\bar{X}}$

1.b Inferencia estadística. Estimadores

La distribución t, t_n, se utiliza en la construcción de intervalos de confianza para la media de poblaciones basados en muestras pequeñas con **varianza desconocida y distribución normal o aproximadamente normal**. También se usa cuando el tamaño de la muestra es grande por lo que hemos utilizado el TCL y digamos que la distribución muestral tiende a la normalidad.

$$f(x) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+x^2/k)^{(k+1)/2}}$$



Valores críticos con un nivel de significación del 2.5% y 5%

	One-Tailed Probabilities, p	
df	p = 0.05	p = 0.025
5	2.015	2.571
10	1.812	2.228
15	1.753	2.131
20	1.725	2.086
25	1.708	2.060
30	1.697	2.042
40	1.684	2.021
50	1.676	2.009
∞	1.645	1.960

1.b Inferencia estadística. Estimadores

Media muestral, varianza muestral, desviación típica muestral y error estándar

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$ → **Media poblacional** (N número de elementos en la población total)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$ → **Media muestral** (n tamaño muestral)

Sigue una distribución normal o t_{n-1}

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$ → **Varianza poblacional**

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$ → **Varianza muestral** (Utiliza n-1 en lugar de n para evitar subestimar el estimador, especialmente para tamaños muestrales pequeños)

Sigue una distribución χ^2_{n-1}

1.b Inferencia estadística. Estimadores

Media muestral, varianza muestral, desviación típica muestral y error estándar

La desviación estándar de la distribución de la media = Error estándar de la media muestral

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Desviación estándar de la población

Tamaño muestral

Conviene tener en cuenta que la desviación típica de la población habitualmente no se conoce.

Para calcular el error estándar de la media muestral se usa la desviación típica de la muestra.

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{n}}$$

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

1.b Inferencia estadística. Estimadores

Media muestral, varianza muestral, desviación típica muestral y error estándar

El teorema central del límite establece que para muestras aleatorias simples de tamaño n de una población con media μ y varianza σ^2, la distribución de la media muestral se aproxima a una distribución normal con media μ y varianza igual a σ^2 /n, cuando el tamaño muestral es suficientemente grande

$$\bar{X} \overset{aprox.}{\sim} N(\mu, \sigma / \sqrt{n})$$

En otras palabras…

Teorema central del límite: la media de n variables i.i.d. converge a una distribución normal a medida que el número de observaciones, n, aumenta.

Nota:

Para estimar la media poblacional se usa la media muestral (tamaño muestral, n).

Para estimar la proporción poblacional se usa la proporción muestral (tamaño muestral, n), cuyo error estándar es $\sqrt{\frac{p(1-p)}{n}}$

Cuanto mayor sea la muestra menor será la variabilidad de los estimadores y mejor serán las estimaciones.

1.c Inferencia estadística. Intervalos

Intervalos de confianza para la media poblacional

Consideremos una población con distribución normal y varianza conocida:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$ → Intervalo de confianza de nivel 1- α para la **media poblacional**

$donde:$

$\bar{x}$ → Media muestral (estimador puntual de la media muestral)

$z_{\alpha/2}$ → Factor de confianza de la normal. Valor de la distribución normal estándar para el que la probabilidad a la derecha es α/2

$\frac{\sigma}{\sqrt{n}}$ → Desviación típica de la media muestral

$$\mathbb{P}_\mu \left\{ -z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right\} = 1 - \alpha \qquad \mathbb{P}_\mu \left\{ \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

1.c Inferencia estadística. Intervalos

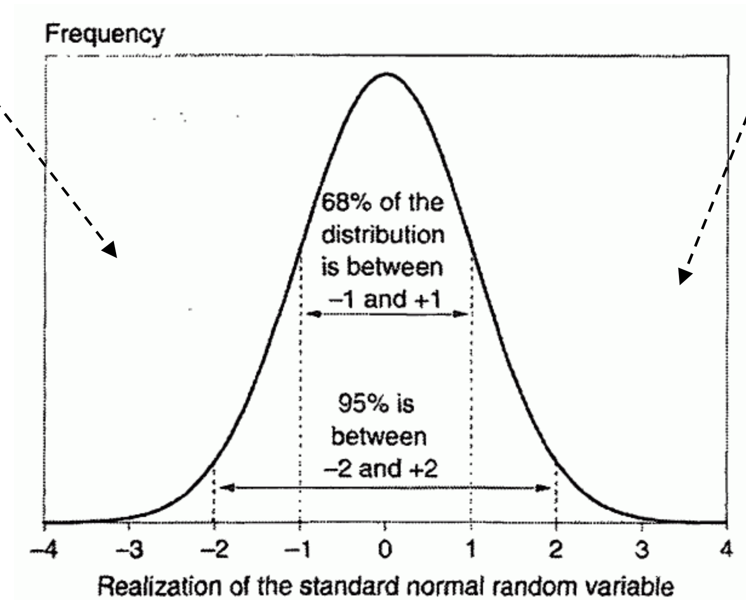
Intervalos de confianza para la media poblacional

Valores críticos de la normal más relevantes:

$z_{\alpha/2}$ = 1.645 para el intervalo de confianza al 90% (10% nivel de significación)

$z_{\alpha/2}$ = 1.960 para el intervalo de confianza al 95% (5% nivel de significación)

$z_{\alpha/2}$ = 2.575 para el intervalo de confianza al 99% (1% nivel de significación)



1.c Inferencia estadística. Intervalos

Intervalos de confianza para la media poblacional

Ejemplo

Nivel de scoring de 100 préstamos
Tamaño muestral: 36 préstamos
Media muestral de los 36 préstamos: 80
Desviación típica poblacional: 15

Intervalo de confianza al 99% para la media:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 80 \pm 2.575\frac{15}{\sqrt{36}} = 80 \pm 6.4$$

Interpretación práctica: Tenemos una confianza del 99% de que la media poblacional esté en el rango 80+/-6.4
Interpretación estadística: Después de tomar muestras de los préstamos en varias ocasiones y construir intervalos de confianza del 99% para la media muestral, el 99% de los intervalos de confianza contendrá al valor de la media poblacional.

1.c Inferencia estadística. Intervalos

Intervalos de confianza para la media poblacional

Ejemplo

Cuando aceptamos que el modelo que generó los datos de una muestra es normal, lo habitual es suponer que la media y la desviación típica son desconocidas y hay que estimarlas a partir de los datos. Por ello, R no tiene una orden para calcular intervalos de confianza para la media de una normal con varianza σ^2 conocida, pero se puede hacer fácilmente.

```
norm.interval = function(datos, varianza = var(datos),
     nivel.conf = 0.95)
{
z = qnorm((1 - nivel.conf)/2, lower.tail = FALSE)
m = mean(datos)
dt = sqrt(varianza/length(datos))
c(m - z * dt, m + z * dt)
}
```

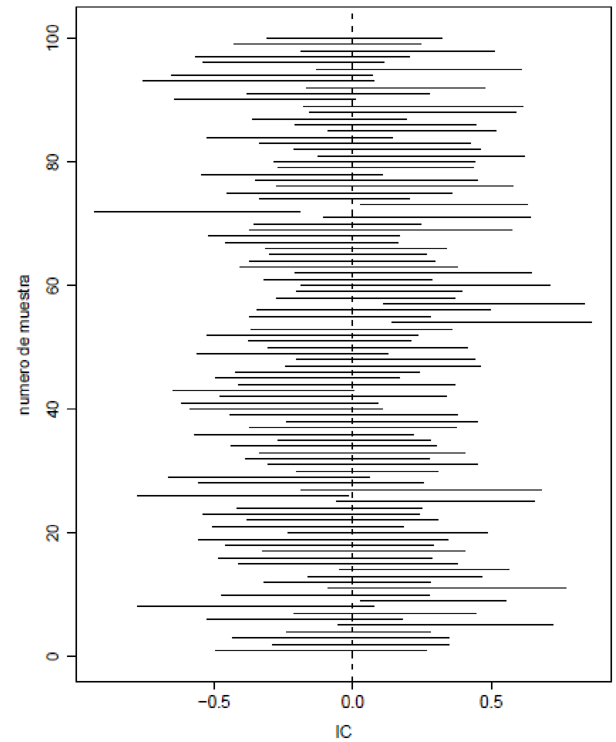

1.c Inferencia estadística. Intervalos

Intervalos de confianza para la media poblacional

Ejemplo: Muestrear 1000 intervalos de confianza y hacer un gráfico

```
nMC=1000;n=30
mu=0;sigma=1
muestras=matrix(rnorm(nMC*n,mu,sigma),n)
int. conf = apply(muestras,2,norm.interval)
sum(int.conf[1,]<=mu&int.conf[2,]>=mu)

plot(range(int.conf), c(0, 1+nMC), type = "n",
xlab = "IC", ylab = "numero de muestra")
for (i in 1:nMC) {
lines(int.conf[, i], rep(i,2),
lwd=2)
}
abline(v=0,lwd=2,lty=2)
```



1.c Inferencia estadística. Intervalos

Intervalos de confianza para la media poblacional: Población con varianza desconocida

Si la distribución de la población es normal con varianza desconocida → Usamos la distribución t para calcular los intervalos de confianza

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$ → Intervalo de confianza

$donde:$

$\bar{x}$ → Media muestral (Estimador puntual de la media poblacional)

$t_{\alpha/2}$ → Valor crítico de la t con n-1 grados de libertad.
La probabilidad que deja a la derecha $t_{\alpha/2}$ es $\alpha/2$

$\frac{s}{\sqrt{n}}$ → Error estándar de la media muestral

s → Desviación estándar muestral

NOTA: Los valores de la t son mayores que los de la normal, puesto que la t tiene colas más gruesas

1.c Inferencia estadística. Intervalos

Intervalos de confianza. El método de la cantidad pivotal

Una metodología general para obtener un intervalo de confianza para θ consiste en encontrar una función $Q(\theta; X_1,\ldots,X_n)$ (llamada "cantidad pivotal") cuya distribución no dependa de θ y sea conocida (al menos de modo aproximado). A partir de esta distribución, fijado un valor $\alpha \in (0; 1)$ se obtienen dos valores $q_1(\alpha)$ y $q_2(\alpha)$ tales que:

$$\mathbb{P}_\theta\{q_1(\alpha) < Q(\theta; X_1, \ldots, X_n) < q_2(\alpha)\} = 1 - \alpha$$

Despejando θ se obtiene una expresión del tipo:

$$\mathbb{P}_\theta\{T_n^{(1)}(X_1, \ldots, X_n) < \theta < T_n^{(2)}(X_1, \ldots, X_n)\} = 1 - \alpha$$

que ya proporciona directamente el **intervalo de confianza.**

1.c Inferencia estadística. Intervalos

Intervalos de confianza para la varianza poblacional en una normal

Se puede demostrar que si $X_1,\ldots,X_n$ son v.a. i.i.d. $N(\mu,\sigma)$ y

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

entonces

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

Este resultado proporciona directamente una cantidad pivotal y, en consecuencia, un intervalo de confianza de nivel 1 - α para σ^2

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

donde $\chi^2_{k;\beta}$ denota el valor que deja a la derecha una probabilidad β en la distribución χ^2_k

1.c Inferencia estadística. Intervalos

Intervalos de confianza para la varianza poblacional en una normal

Ejemplo

Se consideran las cotizaciones de una muestra aleatoria de bonos senior corporativos españoles obteniéndose los siguientes datos:

100 100.5 101 100.7 100.8 102 101.5 99.1 101.3 99.9 .

Suponiendo una distribución normal de las cotizaciones en la población de bonos, hallar un intervalo de confianza al nivel del 90% para la varianza σ^2 de esta población.

```
var.interval = function(datos, nivel.conf = 0.95) {
gl = length(datos) - 1
chiinf = qchisq((1 - nivel.conf)/2, gl)
chisup = qchisq((1 - nivel.conf)/2, gl, lower.tail=FALSE)
v = var(datos)
c(gl * v/chisup, gl * v/chiinf)
}
```

1.c Inferencia estadística. Intervalos

Intervalos de confianza para la media poblacional: Aplicación del TCL

Sean $X_1, \ldots, X_n$ i.i.d Bernoulli(p). Por el TCL

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{aprox.}{\sim} N(0,1)$$

y reemplazando p por su estimador natural $\hat{p} = \bar{X}$, obtenemos que el **intervalo de confianza** aproximado para p es,

$$\left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right)$$

1.c Inferencia estadística. Intervalos

Una sola muestra

Parámetro	Población	Estadístico	Distribución	Intervalo de confianza
μ	Normal con σ conocida	$\dfrac{\overline{x}-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$	$\overline{x} \pm z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
μ	Normal con σ desconocida	$\dfrac{\overline{x}-\mu}{S/\sqrt{n}}$	t_{n-1}	$\overline{x} \pm t_{n-1,1-\alpha/2}\dfrac{S}{\sqrt{n}}$
μ	No normal con σ conocida ($n \geq 30$)	$\dfrac{\overline{x}-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$	$\overline{x} \pm z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
μ	No normal con σ desconocida ($n \geq 30$)	$\dfrac{\overline{x}-\mu}{S/\sqrt{n}}$	$N(0,1)$	$\overline{x} \pm z_{1-\alpha/2}\dfrac{S}{\sqrt{n}}$
p	Bernoulli ($n \geq 30$)	$\dfrac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$	$N(0,1)$	$\widehat{p} \pm z_{1-\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$
λ	Poisson ($n \geq 30$)	$\dfrac{\overline{x}-\lambda}{\sqrt{\lambda/n}}$	$N(0,1)$	$\overline{x} \pm z_{1-\alpha/2}\sqrt{\dfrac{\overline{x}}{n}}$
σ^2	Normal con μ desconocida	$\dfrac{(n-1)S^2}{\sigma^2}$	χ^2_{n-1}	$\dfrac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \dfrac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}$

1.c Inferencia estadística. Intervalos

Casos de dos muestras

Parámetros	Poblaciones	Estadístico	Distribución	Intervalo de confianza
$\frac{\sigma_1^2}{\sigma_2^2}$	Normales indep., μ_1 y μ_2 desconocidas	$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$	F_{n_1-1,n_2-1}	$\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}}$
$\mu_1-\mu_2$	Normales indep., σ_1 y σ_2 conocidas	$\frac{\overline{x}_1-\overline{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$N(0,1)$	$\overline{x}_1-\overline{x}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
$\mu_1-\mu_2$	Normales indep., $\sigma_1=\sigma_2$ desconocidas	$\frac{\overline{x}_1-\overline{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}$	$t_{n_1+n_2-2}$	$\overline{x}_1-\overline{x}_2 \pm t_{n_1+n_2-2,1-\alpha/2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$
$\mu_1-\mu_2$	Normales indep., $\sigma_1\neq\sigma_2$ desconocidas	$\frac{\overline{x}_1-\overline{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	t_m $\frac{1}{m}=\frac{c^2}{n_1-1}+\frac{(1-c)^2}{n_2-1}$ $c=\frac{S_1^2/n_1}{S_1^2/n_1+S_2^2/n_2}$	$\overline{x}_1-\overline{x}_2 \pm t_{m,1-\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$
$\mu_1-\mu_2$	No Normales indep., σ_1,σ_2 desconocidas $n_1>30, n_2>30$	$\frac{\overline{x}_1-\overline{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	aprox. $N(0,1)$	$\overline{x}_1-\overline{x}_2 \pm z_{1-\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$
$\mu_1-\mu_2$	Normales apareadas, $D=X_1-X_2$	$\frac{\overline{D}-\mu_D}{\frac{S_D}{\sqrt{n}}}$	t_{n-1}	$\overline{D}\pm t_{n-1,1-\alpha/2}\frac{S_D}{\sqrt{n}}$
p_1-p_2	Bernoulli, indep., $(n_1\geq 30,\ n_2\geq 30)$	$\frac{\widehat{p}_1-\widehat{p}_2-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_1}}}$	$N(0,1)$	$\widehat{p}_1-\widehat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

2 Contrastes de hipótesis

2. Contrastes de hipótesis

Definiciones

Hipótesis estadística es una declaración acerca de un parámetro poblacional, una afirmación respecto a una característica de la población.

El **contraste de hipótesis** es la evaluación estadística de una afirmación con respecto a una población. Es un método a través del cual se investiga la afirmación o negación de una hipótesis acerca de una característica de una población o de un conjunto de poblaciones.

Contrastar una hipótesis es comparar las predicciones que se deducen de ella con la realidad de forma que si hay coincidencia, dentro del margen de error admisible, mantendremos la hipótesis y, en caso contrario, la rechazaremos

Los procedimientos de contraste de hipótesis se basan en estadísticos muestrales y teoría de probabilidad para probar si una hipótesis es razonable sostenerla o no.

El contraste de hipótesis respecto a un parámetro está muy relacionado con la construcción de intervalos de confianza. Son **hipótesis simples** aquellas que especifican un único valor para el parámetro y **compuestas** las que especifican un intervalo de valores.

2. Contrastes de hipótesis

Definiciones

La **hipótesis nula** H_0 es la hipótesis que se desea contrastar. Nula debe entenderse en el sentido de neutra. Es la hipótesis que mantendremos a no ser que los datos indiquen su falsedad. Se elige habitualmente según el principio de simplicidad.

Si rechazamos H_0, implícitamente se está aceptando una hipótesis alternativa, H_1, que puede ser simplemente la negación de H_0. Lo más habitual es que H_0 sea simple y H_1 se tome de alguna de estas dos formas:

- Contraste **bilateral**: desconocemos en qué sentido H_0 es falsa.
- Contraste **unilateral**: conocemos en qué sentido H_0 puede ser falsa.

2. Contrastes de hipótesis

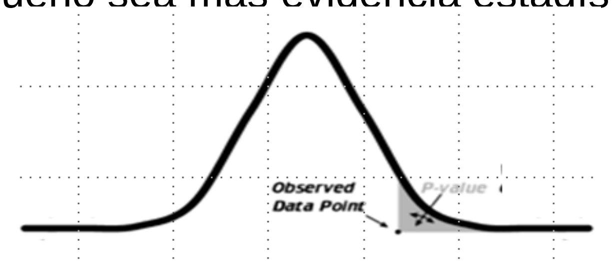
Definiciones

Para realizar el contraste se define una medida de discrepancia entre los datos muestrales y la hipótesis nula, que no dependa de las unidades de medida de la variable. Por tanto, lo más frecuente es:

$$\boxed{\text{discrepancia} = \frac{\text{estimador-parámetro}}{\text{error típico de estimación}}}$$

Hay que determinar qué discrepancias son inadmisibles bajo H_0. Esta decisión depende de la distribución de la medida de discrepancia bajo la hipótesis nula y de que el contraste sea unilateral o bilateral.

Llamamos **p-valor** del contraste a la probabilidad de obtener una discrepancia mayor que la observada. Se rechaza la hipótesis nula cuando el p-valor es menor que el nivel de significación α. Cuanto más pequeño sea más evidencia estadística a favor de H_1.

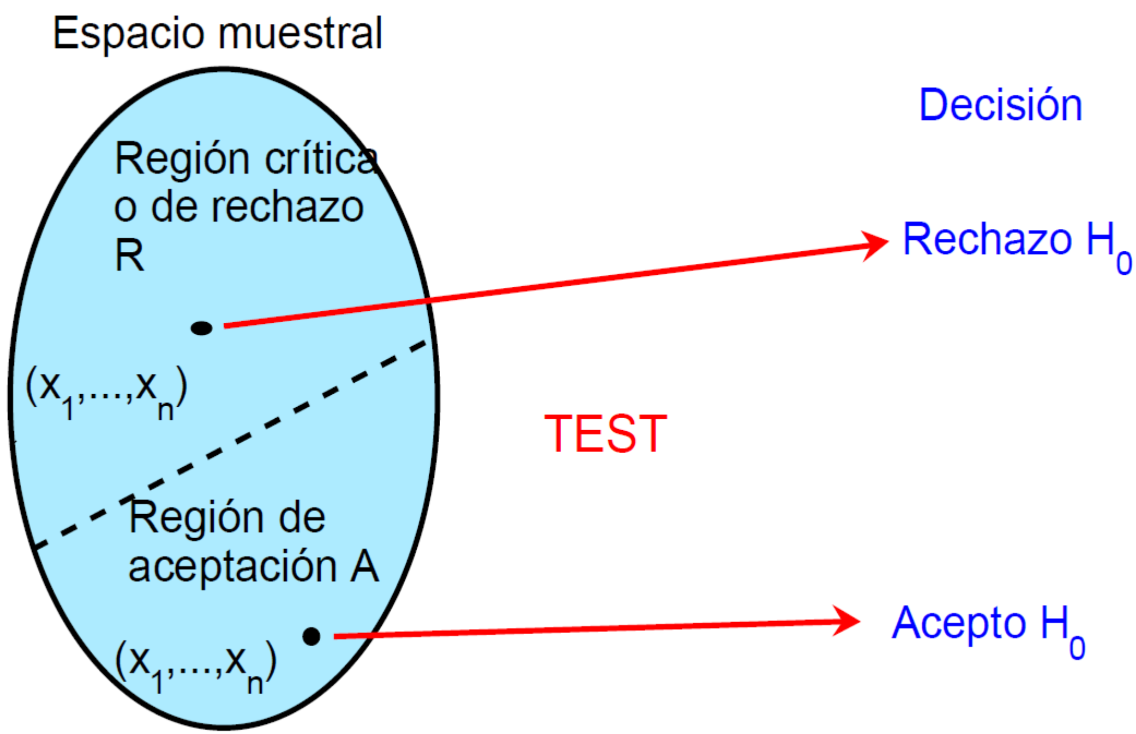


Para realizar el contraste se fija un **nivel de significación**, α, que representa la probabilidad de rechazar H_0 cuando es cierta. Este nivel permite definir una **región de rechazo**, R, y si la discrepancia está en esa región, rechazaremos la hipótesis nula.

2. Contrastes de hipótesis

Definiciones

Los contrastes habituales (no aleatorizados) se definen mediante una región crítica o región de rechazo $R \subset \mathbf{R}^n$, de tal manera que, cuando $(x_{1, \dots,} x_n) \in R$, se rechaza la hipótesis nula.



2. Contrastes de hipótesis

Definiciones

Hay que señalar que la metodología de contraste de hipótesis no demuestra la validez de la hipótesis que se acepta en cada caso (en el sentido de probar algo mediante un método deductivo).

Para interpretar correctamente los resultados hay que decir que “los datos disponibles proporcionan (o no proporcionan) suficiente evidencia estadística en contra de la hipótesis nula". En todo caso, la conclusión depende de información incompleta y aleatoria, procedente de una o varias muestras, y siempre existe la posibilidad de cometer un error aceptando una hipótesis equivocada.

2. Contrastes de hipótesis

Etapas de un contraste

Etapas del contraste de hipótesis

1. Definir la hipótesis
2. Seleccionar el estadístico de contraste
3. Especificar el nivel de significación
4. Definir la regla de decisión respecto a la hipótesis
5. Mediante la muestra calcular los estadísticos muestrales
6. Tomar una decisión referente a la hipótesis, basada en los resultados del contraste

2. Contrastes de hipótesis

Hipótesis nula y región de rechazo

Hipótesis nula (H_0): condición que cumple el parámetro, que el analista quiere contrastar.

Hipótesis alternativa (H_1): la negación de H_0. Conclusión a la que se llega si se rechaza la hipótesis nula.

Nivel de significación (α) : nivel de probabilidad de que haya diferencia entre los valores observados y esperados.

Sucesos con una probabilidad pequeña llevan a rechazar H_0. H_0 y H_1 no son simétricas.

Los valores del nivel de significación se suelen fijar en 10%, 5%, 1%.

Ejemplos:

H_0: $\mu=\mu_0$ o

H_0: $\mu\leq\mu_0$ o

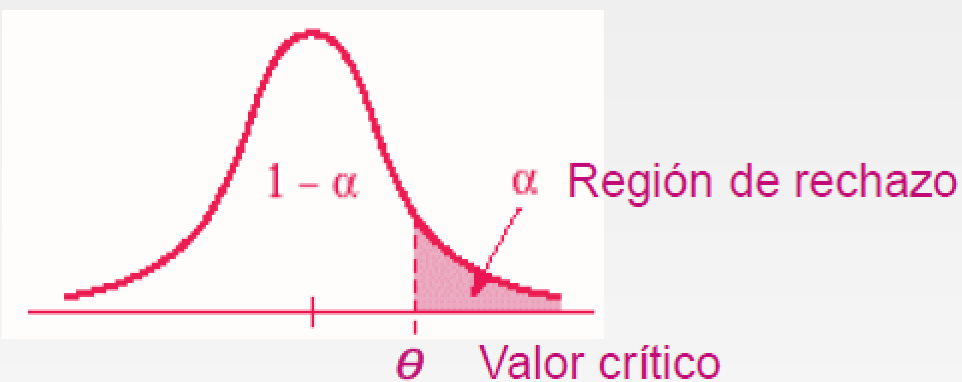
H_0: $\mu\geq\mu_0$, siendo μ la media poblacional, μ_0 la hipótesis sobre el valor de la media poblacional.

2. Contrastes de hipótesis

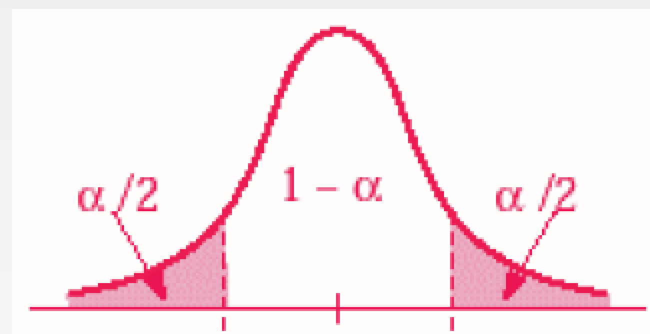
Hipótesis nula y región de rechazo

Valor muy bajo de la probabilidad del estadístico pivote (menor que α) $\rightarrow$ diferencia significativa (entre los valores de la muestra y los teóricos dados por la hipótesis nula H_0) $\rightarrow$ Rechaza H_0

$$H_0 : \theta \leq \theta_0$$



$$H_0 : \theta = \theta_0$$



Tipos de contrastes

Unilaterales

Bilaterales

2. Contrastes de hipótesis

Contrastes unilaterales o bilaterales

El uso depende de lo que desea probar, si se conoce o no la dirección en que la hipótesis nula es falsa.

Contraste bilateral:

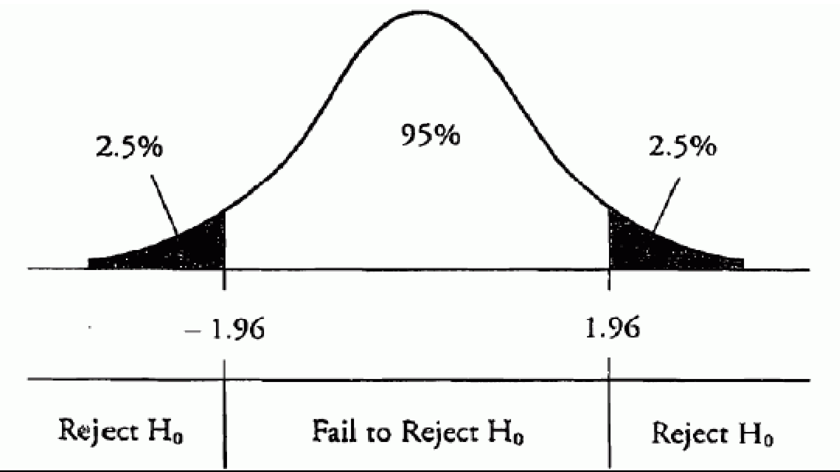
H_0: $\mu=\mu_0$ vs H_1: $\mu\neq\mu_0$

H_0 se rechaza si
- Estadístico de contraste> valor crítico superior o,
- Estadístico de contraste< valor crítico inferior

Regla de decisión

Ejemplo: Regla de decisión para un contraste bilateral normal con nivel de confianza 95% (α=5%):

H_0 se rechaza si
- Estadístico de contraste > 1.96, o
- Estadístico de contraste < -1.96



2. Contrastes de hipótesis

Contrastes unilaterales o bilaterales

Contraste unilateral:

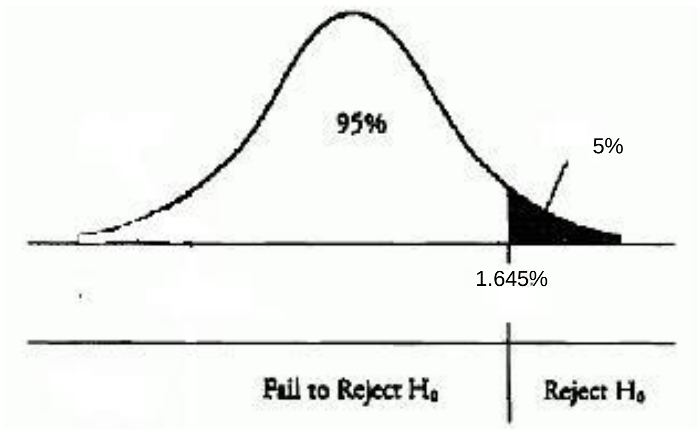
A: $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$, or
B: $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$,

H_0 se rechaza si
- A: Estadístico de contraste> valor crítico superior o,
- B: Estadístico de contraste< valor crítico inferior

Regla de decisión

Ejemplo: Regla de decisión para un contraste unilateral normal con nivel de confianza 95% (α=5%): (cola derecha de la distribución):

H_0 se rechaza si Estadístico de contraste > 1.645



2. Contrastes de hipótesis

Estadístico de contraste

Media muestral x=0.01

Estimador puntual del parámetro poblacional

$\mu_0=0$

Valor especificado en la hipótesis nula

$$Estadistico\ contraste = \frac{Estadistico\ muestral - valor\ de\ la\ hipotesis}{error\ estandar\ del\ estadistico\ muestral}$$

Error estándar del estadístico muestral o estimador puntual

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

o

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{n}}$$

El estadístico de contraste es una variable aleatoria. Las cuatro distribuciones principales son:

- **Distribución t_n**
- **Distribución normal, Z**
- **Distribución χ^2_n**
- **Distribución $F_{n,m}$**

2. Contrastes de hipótesis

Intervalos de confianza y contrastes de hipótesis

$$\text{Estadístico muestral} - \left[\text{Valor Crítico}\right]\left[\text{Error estándar}\right] < \text{Parámetro poblacional} < \text{Estadístico muestral} + \left[\text{Valor Crítico}\right]\left[\text{Error estándar}\right]$$

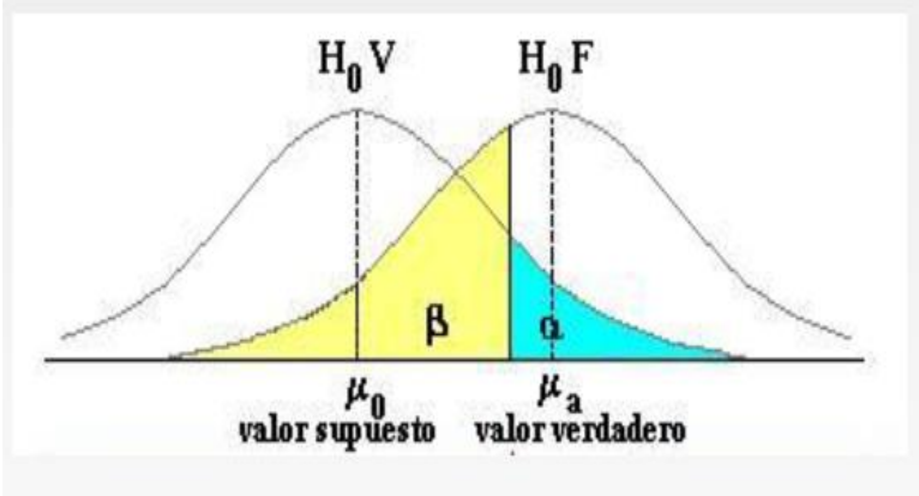
Los intervalos de confianza y los contrastes de hipótesis están relacionados mediante el valor crítico

Interpretación del intervalo de confianza: Para un nivel de confianza de, por ejemplo, el 95%, hay un 95% de probabilidad de que el verdadero parámetro de la población esté contenido en el intervalo

2. Contrastes de hipótesis

Hipótesis nula y región de rechazo

Valor muy bajo de la probabilidad del estadístico pivote (menor que α) $\rightarrow$ diferencia significativa (entre los valores de la muestra y los teóricos dados por la hipótesis nula H_0) $\rightarrow$ Rechaza H_0

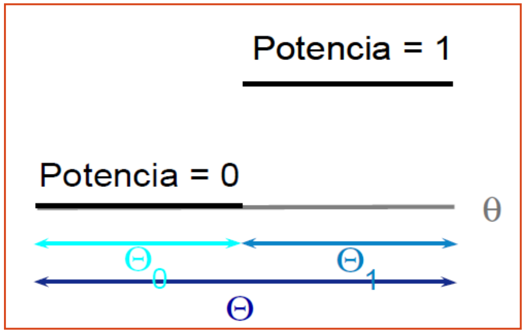


Error Tipo I:
α=P(rechazar H_0 | H_0 es correcta)
Error Tipo II:
β=P(aceptar H_0 | H_0 es incorrecta)

Se denomina **potencia** a la probabilidad de rechazar una hipótesis nula cuando es incorrecta

Potencia= 1-β=P(rechazar H_0 | H_0 es incorrecta)

Lo que nos gustaría:



2. Contrastes de hipótesis

Hipótesis nula y región de rechazo

Los contrastes de hipótesis están diseñados para controlar la probabilidad máxima de rechazar H_0 cuando es cierta, y por tanto, suelen ser “conservadores" con la hipótesis nula: hace falta mucha evidencia muestral para rechazar H_0.

Además, con los mismos datos, H_0 se puede rechazar para un nivel de significación α = 5% y se puede aceptar para α = 1%.

Lo que en realidad se suele hacer (teoría de Neyman-Pearson):

1. Acotar la máxima probabilidad de error de tipo I: se fija el nivel de significación y se define el tamaño de un contraste como la máxima probabilidad de error de tipo I.
2. Minimizar la probabilidad de error de tipo II.

En una primera aproximación, los problemas de contraste de hipótesis se pueden clasificar en problemas de una muestra (cuando hay una sola población de interés) y problemas de dos o más muestras (cuando se quiere comparar dos o más poblaciones y se dispone de una muestra de cada una de ellas).

2. Contrastes de hipótesis

Hipótesis nula y región de rechazo

Se considera mayor error al Error Tipo II.

Error Tipo I:
Se está rechazando una hipótesis cierta
Error Tipo II:
Se está aceptando una hipótesis falsa

Decisión	H_0 correcta	H_0 incorrecta
No rechazar H_0	Sin error $(1 - \alpha)$	Error Tipo II (β)
Rechazar H_0	Error Tipo I (α)	Sin error $(1 - \beta =$ potencia$)$

Ejemplo:

En un juicio se establece la hipótesis de que una persona es culpable.

El Error Tipo II se comete al determinar que el sujeto es culpable cuando en realidad es inocente.

El Error Tipo I se cometería en caso de declararle inocente cuando verdaderamente es culpable.

2.a Contrastes de hipótesis paramétricos

2.a Contrastes de hipótesis paramétricos

Contrastes para la media de una distribución

En cada caso se rechaza H_0 cuando $(x_1,\ldots,x_n) \in R$

Distribución normal con varianza conocida:
Sea $X_1,\ldots,X_n$ una muestra aleatoria de $X \sim N(\mu,\sigma)$ con σ conocida.

$$H_0 : \mu = \mu_0 \qquad R = \left\{ (x_1,\ldots,x_n) : |\bar{x} - \mu_0| \geq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

$$H_0 : \mu \leq \mu_0 \qquad R = \left\{ (x_1,\ldots,x_n) : \bar{x} - \mu_0 \geq z_{\alpha} \frac{\sigma}{\sqrt{n}} \right\}$$

$$H_0 : \mu \geq \mu_0 \qquad R = \left\{ (x_1,\ldots,x_n) : \bar{x} - \mu_0 \leq z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\}$$

donde z_β es tal que $\Phi(z) = 1 - \beta$ siendo Φ la función de distribución de la N(0,1)

2.a Contrastes de hipótesis paramétricos

Contrastes para la media de una distribución

Distribución normal con varianza desconocida:
Sea $X_1,\ldots,X_n$ una muestra aleatoria de $X \sim N(\mu,\sigma)$ con σ desconocida.

$$H_0 : \mu = \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : |\bar{x} - \mu_0| \geq t_{n-1;\alpha/2} \frac{s}{\sqrt{n}} \right\}$$

$$H_0 : \mu \leq \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : \bar{x} - \mu_0 \geq t_{n-1;\alpha} \frac{s}{\sqrt{n}} \right\}$$

$$H_0 : \mu \geq \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : \bar{x} - \mu_0 \leq t_{n-1;1-\alpha} \frac{s}{\sqrt{n}} \right\}$$

2.a Contrastes de hipótesis paramétricos

Contrastes para la media de una distribución

Contrastes de nivel aproximado α (muestras grandes) para el parámetro p en una Bernoulli:

Sea $X_1,\ldots,X_n$ una muestra aleatoria de X ~Bernoulli(p).

$$H_0 : p = p_0, \text{ frente a } H_1 : p \neq p_0.$$

El criterio de rechazo es $(x_1,\ldots,x_n)\in R$, siendo

$$R = \left\{ (x_1, \ldots, x_n) : \left| \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| > z_{\alpha/2} \right\}$$

y análogamente para los contrastes unilaterales.

2.a Contrastes de hipótesis paramétricos

Contrastes para la media de una distribución

Contrastes de nivel aproximado α (muestras grandes) para la media de cualquier distribución:

Sea $X_1,\ldots,X_n$ una muestra aleatoria de X con $E(X)=\mu<\infty$.

$H_0 : \mu = \mu_0$, frente a $H_1 : \mu \neq \mu_0$

$$R = \left\{ (x_1, \ldots, x_n) : \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > z_{\alpha/2} \right\}$$

$H_0 : \mu \leq \mu_0$, frente a $H_1 : \mu > \mu_0$

$$R = \left\{ (x_1, \ldots, x_n) : \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > z_{\alpha} \right\}$$

$H_0 : \mu \geq \mu_0$, frente a $H_1 : \mu < \mu_0$

$$R = \left\{ (x_1, \ldots, x_n) : \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -z_{\alpha} \right\}$$

2.a Contrastes de hipótesis paramétricos

Contrastes para la varianza de una normal

Contrastes para la varianza de una normal:

Sea $X_1,\ldots,X_n$ una muestra aleatoria de X~N(μ,σ), con σ desconocida.

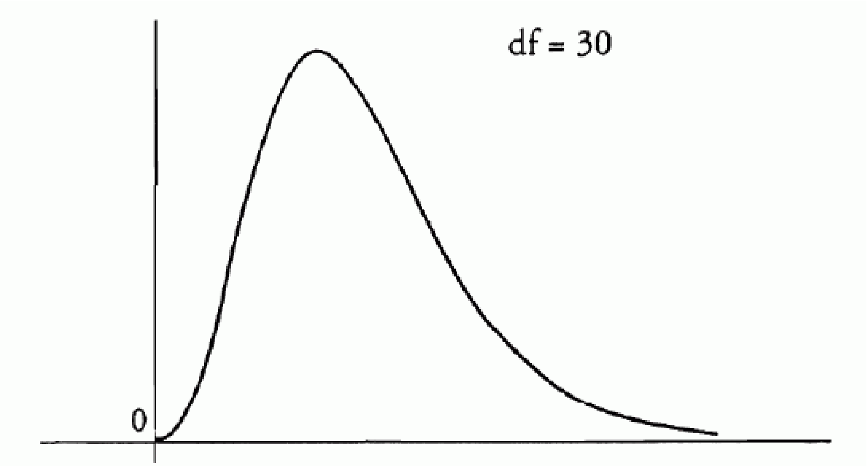
$$H_0 : \sigma = \sigma_0 \qquad R = \left\{ \frac{(n-1)s^2}{\sigma_0^2} \notin (\chi^2_{n-1;1-\alpha/2}, \chi^2_{n-1;\alpha/2}) \right\}$$

$$= \{\sigma_0^2 \notin \mathrm{IC}_{1-\alpha}(\sigma^2)\}$$

$$H_0 : \sigma \leq \sigma_0 \qquad R = \left\{ \frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{n-1;\alpha} \right\}$$

$$H_0 : \sigma \geq \sigma_0 \qquad R = \left\{ \frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{n-1;1-\alpha} \right\}$$

Es una distribución asimétrica, que tiende a la normal cuando los grados de libertad son muy altos



2.a Contrastes de hipótesis paramétricos

Contrastes para dos muestras

Caso de muestras independientes:

Se tienen dos muestras $X_1,\ldots,X_{n1}$ e $Y_1,\ldots,X_{n2}$ de dos v.a. X e Y. Ambas se suponen independientes entre sí. Se quiere contrastar hipótesis del tipo:

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 \leq \mu_2$$

$$H_0 : \sigma_1 = \sigma_2$$

Uno de los contrastes más usuales es el de igualdad de medias para dos poblaciones normales homocedásticas, es decir, con $\sigma_1 = \sigma_2$.

Se puede probar que, bajo $H_0 : \mu_1 = \mu_2$,

$$\frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

y, por tanto, una región crítica al nivel α es

$$R = \left\{ |\bar{x} - \bar{y}| > t_{n_1+n_2-2;\alpha/2}\ s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

Siendo

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

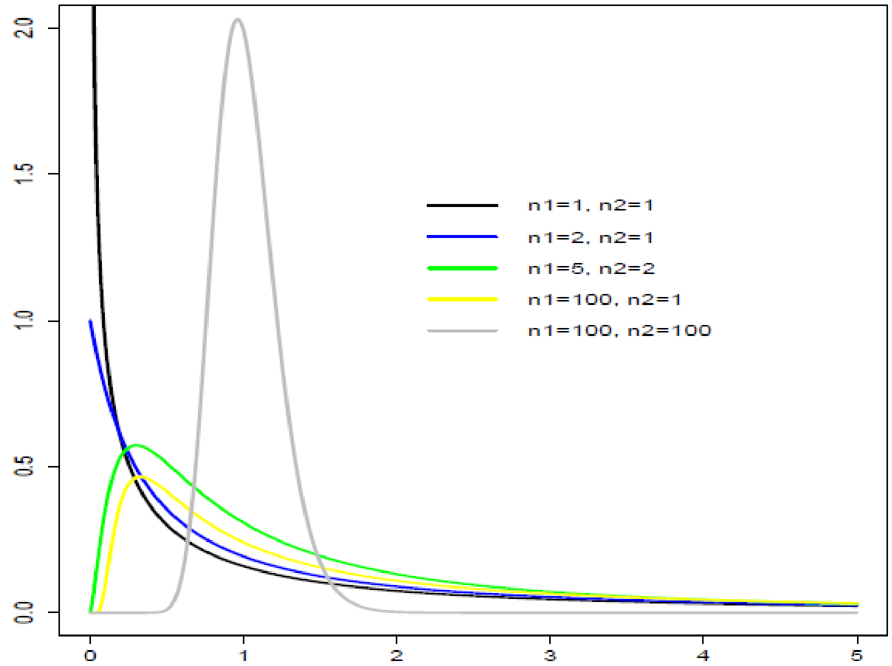
la varianza combinada (*pooled variance*).

2.a Contrastes de hipótesis paramétricos

Contrastes de igualdad de varianzas (hipótesis de homocedasticidad)

Caso de poblaciones normales:

Sean Q_1 y Q_2 dos v.a. independientes con distribuciones χ^2_{n1} y χ^2_{n2}, respectivamente. La distribución de $\frac{Q_1/n_1}{Q_2/n_2}$ se denomina F de Fisher-Snedecor con n_1 y n_2 grados de libertad, $F_{n1,n2}$.



2.a Contrastes de hipótesis paramétricos

Contrastes de igualdad de varianzas (hipótesis de homocedasticidad)

Caso de poblaciones normales:

Si s^2_1 , s^2_2 son las cuasi-varianzas de dos muestras independientes de tamaño n_1 y n_2 extraídas, respectivamente, de dos poblaciones $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$, se tiene

$$\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}, \quad \frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}.$$

Por tanto, bajo $H_0:\sigma_1=\sigma_2$,

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

De este resultado se derivan los contrastes para comparar σ_1 y σ_2.

Afi Escuela de Finanzas

2.a Contrastes de hipótesis paramétricos

Contrastes de igualdad de varianzas (hipótesis de homocedasticidad)

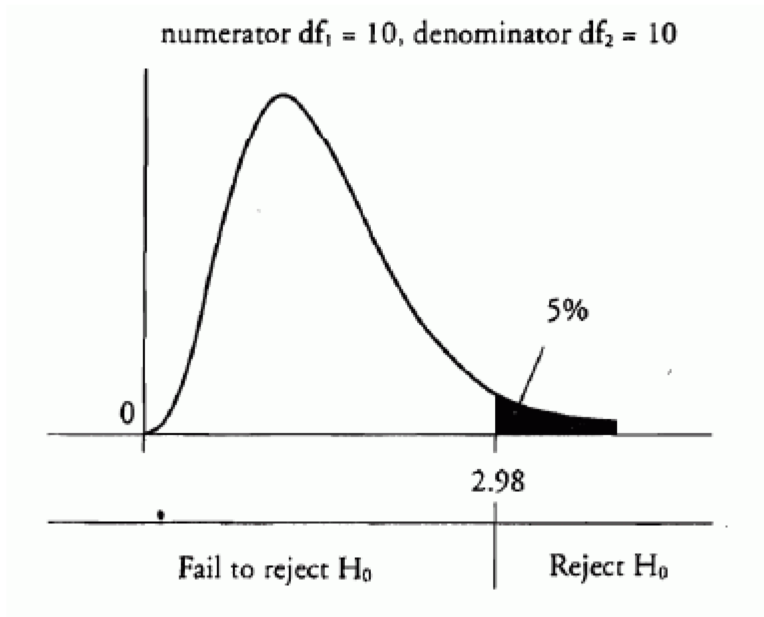
Caso de poblaciones normales:

$$H_0 : \sigma_1 = \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} \notin \left(F_{n_1-1;n_2-1;1-\alpha/2}, F_{n_1-1;n_2-1;\alpha/2}\right) \right\}$$

$$= \left\{ 1 \notin \mathrm{IC}_{1-\alpha}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) \right\}$$

$$H_0 : \sigma_1 \leq \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} > F_{n_1-1;n_2-1;\alpha} \right\}$$

$$H_0 : \sigma_1 \geq \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} < F_{n_1-1;n_2-1;1-\alpha} \right\}$$



2.a Contrastes de hipótesis paramétricos

Contrastes para dos muestras

Caso de muestras emparejadas:
Surge en aquellas situaciones con $n_1 = n_2$ en que X_i e Y_i no son independientes (porque corresponden a mediciones sobre la misma observación o individuo, por ejemplo, antes y después de una determinada acción como un tratamiento).

Se reducen a problemas de una muestra para la muestra de diferencias $D_i = X_i - Y_i$.

2.b Contrastes de hipótesis no paramétricos

2.b Contrastes de hipótesis no paramétricos

Introducción

Queremos decidir si es aceptable la hipótesis de que una cierta muestra está extraída de una distribución de probabilidad dada.

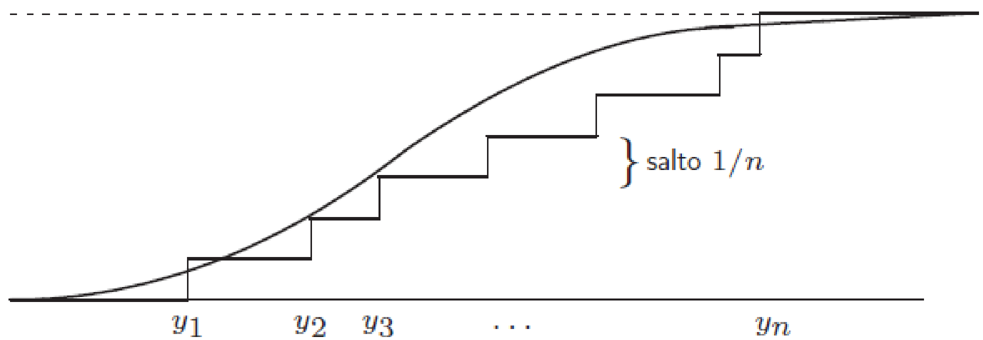
Presentamos a continuación dos contrastes sencillos y ampliamente utilizados:
- el contraste de Kolmogorov-Smirnov;
- el contraste de la χ^2

2.b Contrastes de hipótesis no paramétricos

Contraste de Kolmogorov-Smirnov

Contraste de Kolmogorov-Smirnov

(Sólo para distribuciones continuas). Tenemos una muestra de tamaño n y generamos la función de distribución empírica ordenando esas muestras de menor a mayor: digamos que $y_1, \ldots, y_n$ son los datos ya ordenados. Comparamos entonces con la función de distribución teórica. El tamaño de la muestra no suele ser grande.



La medida relevante es:

$$D_n = \max_{1 \leq j \leq n} |F(y_j) - F^{\text{emp}}(y_j)|$$

donde

$$F^{\text{emp}}(y_j) = \frac{j}{N}.$$

2.b Contrastes de hipótesis no paramétricos

Contraste de Kolmogorov-Smirnov

Contraste de Kolmogorov-Smirnov

Si D_n es mayor que un cierto valor crítico, rechazamos la hipótesis de partida (el que la muestra corresponda a una variable aleatoria con la función de distribución teórica F).

Los valores críticos dependen de n y de un nivel de significación, y están tabulados.

Distribución del estadístico de Kolmogorov-Smirnov (D_n).
Se tabula d tal que $P(D_n > d) = \alpha$.

			α						α		
n	0'2	0'1	0'05	0'02	0'01	n	0'2	0'1	0'05	0'02	0'01
1	0'900	0'950	0'975	0'990	0'995	21	0'226	0'259	0'287	0'321	0'344
2	0'684	0'776	0'842	0'900	0'929	22	0'221	0'253	0'281	0'314	0'337
3	0'565	0'636	0'780	0'785	0'829	23	0'216	0'247	0'275	0'307	0'330
4	0'493	0'565	0'624	0'689	0'734	24	0'212	0'242	0'269	0'301	0'323
5	0'447	0'509	0'563	0'627	0'669	25	0'208	0'238	0'264	0'295	0'317
6	0'410	0'468	0'519	0'577	0'617	26	0'204	0'233	0'259	0'290	0'311
7	0'381	0'436	0'483	0'538	0'576	27	0'200	0'229	0'254	0'284	0'305
8	0'358	0'410	0'454	0'507	0'542	28	0'197	0'225	0'250	0'279	0'300
9	0'339	0'387	0'430	0'480	0'513	29	0'193	0'221	0'246	0'275	0'295
10	0'323	0'369	0'409	0'457	0'489	30	0'190	0'218	0'242	0'270	0'290
11	0'308	0'352	0'391	0'437	0'468	31	0'187	0'214	0'238	0'266	0'285
12	0'296	0'338	0'375	0'419	0'449	32	0'184	0'211	0'234	0'262	0'281
13	0'285	0'325	0'361	0'404	0'432	33	0'182	0'208	0'231	0'258	0'277
14	0'275	0'314	0'349	0'390	0'418	34	0'179	0'205	0'227	0'254	0'273
15	0'266	0'304	0'338	0'377	0'404	35	0'177	0'202	0'224	0'251	0'269
16	0'258	0'295	0'327	0'366	0'392	36	0'174	0'199	0'221	0'247	0'265
17	0'250	0'286	0'318	0'355	0'381	37	0'172	0'196	0'218	0'244	0'262
18	0'244	0'279	0'309	0'346	0'371	38	0'170	0'194	0'215	0'241	0'258
19	0'237	0'271	0'301	0'337	0'361	39	0'168	0'191	0'213	0'238	0'255
20	0'232	0'265	0'294	0'329	0'352	40	0'165	0'189	0'21	0'235	0'252
						> 40	$\frac{1'07}{\sqrt{n}}$	$\frac{1'22}{\sqrt{n}}$	$\frac{1'36}{\sqrt{n}}$	$\frac{1'52}{\sqrt{n}}$	$\frac{1'63}{\sqrt{n}}$

2.b Contrastes de hipótesis no paramétricos

Contraste χ^2

Contraste χ^2

Tenemos una muestra $x_1, \ldots, x_n$. Agrupamos los datos en k clases. O_j es la frecuencia observada en la muestra de la clase j.

El modelo que estamos contrastando asigna probabilidad p_j a la clase j. La frecuencia esperada es $E_j = np_j$.

Calculamos

$$\text{discrepancia observada} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

que se distribuye aproximadamente como una χ^2 (con $k - 1$ grados de libertad) cuando el modelo es correcto.

Rechazaremos el modelo si la probabilidad de obtener una discrepancia mayor o igual que la observada sea suficientemente baja.

2.b Contrastes de hipótesis no paramétricos

Contraste χ^2

Ejemplo: Sea X el número de préstamos en default de una serie de carteras.

$$H_0: X \sim P(\lambda) \quad H_1: X \sim F(x)$$

Estimación del parámetro λ de la distribución de Poisson:

x_i	$n(x_i) = n_i$	$h(x_i)$	$x_i h(x_i)$
0	5	0,078125	0
1	10	0,15625	0,15625
2	13	0,203125	0,40625
3	19	0,296875	0,890625
4	11	0,171875	0,6875
5	5	0,078125	0,390625
6	0	0	0
7	1	0,015625	0,109375
	64	1	2,640625

$$\hat{\lambda} = \bar{x} = 2,64$$

2.b Contrastes de hipótesis no paramétricos

Contraste χ^2

Ejemplo: Sea X el número de préstamos en default de una serie de carteras.

$$H_0: X \sim P(\lambda) \quad H_1: X \sim F(x)$$

Reagrupación porque las frecuencias esperadas en la primera y tres últimas categorías son menores que 5.

x_i	n (xi) = ni	pi	n pi	$\frac{(n_i - np_i)^2}{np_i}$
0 y 1	15	0.2596	16.6168	0.1573
2	13	0.2486	15.9131	0.5333
3	19	0.2189	14.0068	1.7800
4	11	0.1445	9.2467	0.3324
5 y más	6	0.1284	8.2166	0.5980
	64	**1**	**64**	**3.4010**

$$\hat{\lambda} = \bar{x} = 3{,}40$$

$$\chi^2_4 = 3{,}4010 < \chi^2_{4,0.95} = 9{,}48773$$

Por tanto no hay evidencia suficiente para rechazar H_0

2.b Contrastes de hipótesis no paramétricos

Otros contrastes

Kolmogorov-Smirnov-Lilliefors: H_0: F=Normal H_0: X ~ N(μ,σ)

Compara ambas funciones, empírica y teórica, a lo largo de toda la curva.
Es necesario que la muestra esté ordenada

Shapiro-Wilk: H_0: F=Normal H_0: X ~ N(μ,σ)

Para muestras de menos de 30-50 datos es más potente.

Jarque-Bera: Bajo normalidad

$$\widehat{S} \sim N(0,6/T)$$
$$\widehat{K} \sim N(3,24/T)$$

Siendo los coeficientes de asimetría y curtosis muestrales, respectivamente.

$$\widehat{S} = \frac{1}{T\widehat{\sigma}^3}\sum_{i=1}^{T}(x_i - \widehat{\mu})^3 \qquad \widehat{K} = \frac{1}{T\widehat{\sigma}^4}\sum_{i=1}^{T}(x_i - \widehat{\mu})^4$$

Bajo normalidad,

$$JB = \widehat{S}^2 + \widehat{K}^2 \sim \chi^2_2$$

2.b Contrastes de hipótesis no paramétricos

Otros contrastes

Prueba de la mediana:

Contrasta la hipótesis nula de que las muestras proceden de *k* subpoblaciones en las que la probabilidad de obtener un resultado menor o igual que la mediana *Me* de la variable *X* sobre toda la población, es la misma en todas las subpoblaciones:

$$H_0: P_{x_1 \leq M_e} = \ldots = P_{x_k \leq M_e}$$

Análisis de la varianza de Kruskall-Wallis:

Contrastar la hipótesis nula de que las muestras proceden de *k* subpoblaciones en las que la distribución de *X* es la misma:

$$H_0 : F_1 = \ldots = F_k$$

Prueba de rachas:

Se utiliza para determinar la aleatoriedad en el orden de aparición de los valores de una variable. O bien, para determinar si una muestra se ha extraído de forma aleatoria o no.

3. Algunas funciones en R

Algunas funciones en R

Función	Comentarios
binom.test	Test exacto sobre el parámetro de una binomial
cor.test	Test de asociación entre muestras apareadas
wilcox.test	Test de suma de rangos de Wilcoxon para una y dos muestras
prop.test	Test de igualdad de proporciones
chisq.test	Test de la chi-cuadrado para datos de conteo
fisher.test	Test exacto de Fisher para datos de conteo
ks.test	Test de Kolmogorov-Smirnov para ajuste de datos a distribuciones dadas
shapiro.test	Test de Shapiro para comprobar ajuste de datos a una distribución normal
oneway.test	Test para comprobar la igualdad de medias entre varios grupos de datos
var.test	Test para comprobar la igualdad de varianzas entre dos grupos de datos

3 Tablas de contingencia

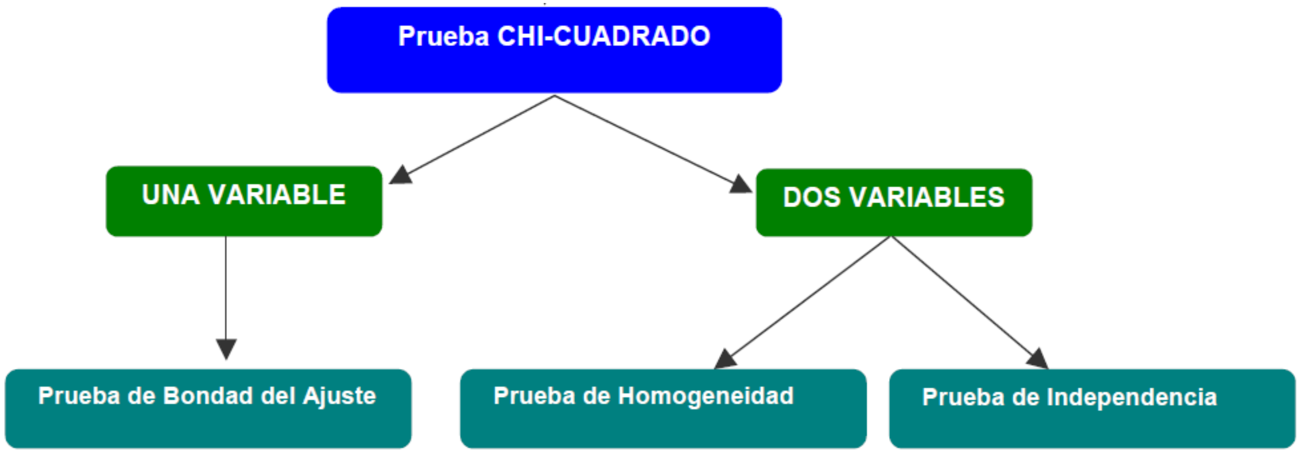
3. Tablas de contingencia

Introducción

Las tablas de contingencia permiten contrastes sobre la **independencia** de dos características diferentes de una población y contrastes de **homogeneidad**.

Estas características habitualmente son de naturaleza cualitativa (nominales u ordinales), o bien, cuantitativa en escala de razón o intervalo.

Resumen de pruebas χ^2



3. Tablas de contingencia

Contraste de independencia χ^2

Contraste χ^2

Se quiere determinar si existe relación entre dos características diferentes de una población, donde cada característica dispone de cierto número de categorías.

A \ B	B_1	B_2	...	B_j	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
...	...	...	...	...	...	...	...
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
...	...	...	...	...	...	...	...
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	$n_{..}$

H$_0$: A y B son independientes

$n_{i\cdot} = \sum_{i=1}^{r} n_{ij}$ *i=1,2,…,r* *Total de la i-ésima fila*

$n_{\cdot j} = \sum_{i=1}^{s} n_{ij}$ *j=1,2,…,s* *Total de j-ésima columna*

Estadístico*: diferencia entre las frecuencias observadas y las esperadas.*

Valores esperados*:* $e_{ij} = n_{i.}\, n_{.j} / n$

Estadístico de contraste*:*

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Sigue una χ^2 con (k-1)(r-1) grados de libertad

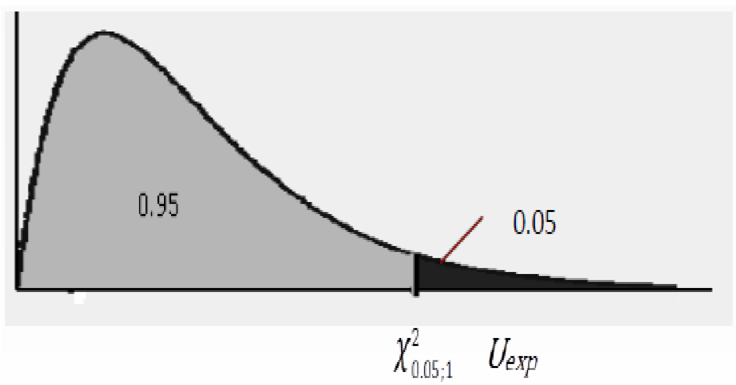
3. Tablas de contingencia

Contraste de independencia χ^2

Ejemplo χ^2

Se realiza una investigación para determinar si existe asociación aparente entre los análisis de riesgo de dos analistas diferentes (A, B) y el número de defaults.

Default	Si	No	Total
A	162 (170)	263 (255)	425
B	38 (30)	37 (45)	75
Total	200	300	500



Valores esperados:

$$e_{ij} = \frac{n_{i.}n_{.j}}{n} \qquad e_{11} = \frac{n_{1.}n_{.1}}{n} = \frac{425 \times 200}{500} = 170$$

Estadístico de contraste:

$$U_{exp} = \frac{(162-170)^2}{170} + \frac{(263-255)^2}{255} + \frac{(38-30)^2}{30} + \frac{(37-45)^2}{45} = 4.18$$

$$\chi^2_{0.05;1} = 3.84 < U_{exp} = 4.18$$

Rechazamos H_0

Sí existe asociación entre los defaults y los analistas (no siguen el mismo criterio)

3. Tablas de contingencia

Contraste de homogeneidad χ^2

Contraste χ^2

Se trata de contrastar si varias muestras proceden de una misma población, esto es que las muestras son homogéneas

Modalidades / Muestras	B_1	B_2	...	B_j	...	B_p	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1p}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2p}	$n_{2.}$
...	...	...	...	...	...	...	...
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ip}	$n_{i.}$
...	...	...	...	...	...	...	...
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rp}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.p}$	$n_{..}$

H_0: Las muestras son homogéneas

ni· = $\sum_{i=1}^{s} n_{ij}$ *i=1,2,…,r* *Total de la i-ésima fila*

n·j = $\sum_{i=1}^{p} n_{ij}$ *j=1,2,…,p* *Total de la j-ésima columna*

Estadístico: *diferencia entre las frecuencias observadas y las esperadas.*

Valores esperados: $e_{ij} = n_{i.}\, n_{.j} / n$

Estadístico de contraste:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{p} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Sigue una χ^2 con (r-1)(p-1) grados de libertad

3. Tablas de contingencia

Otras medidas

Medida de asociación	Tabla	Escala de medida	Observaciones
Coeficiente Phi	2 x 2	Nominales	Medida basada en el estadístico de χ^2.
Coeficiente de contingencia	r x c	Nominales	Toma valores comprendidos entre -1 y 1 que indican mínimo y máximo grado de asociación respectivamente.
V de Cramer	r x c	Nominales	Phi presenta el inconveniente de que puede alcanzar valores superiores a 1 en tablas r x c; el coeficiente de contingencia depende de una cota superior, y la V de Cramer tiende a subestimar la asociación. Además pueden tomar el mismo valor en muestras con tamaños muy diferentes. Son útiles para comparar grados de asociación entre pares de variables observadas sobre un mismo conjunto de individuos.

3. Tablas de contingencia

Otras medidas

Medida de asociación	Tabla	Escala de medida	Observaciones
Riesgo relativo	2 x 2	Nominales	Compara los dos grupos establecidos por los valores de una de las variables en términos de la frecuencia con que presentan cada uno de los valores de la otra.
Lambda	r x c	Nominales	Toma valores comprendidos entre 0 y 1 que indican mínimo y máximo grado de asociación respectivamente.
Coeficiente de incertidumbre	r x c	Nominales	Lambda es fácil de interpretar en términos de la proporción en que se reduce el error en la predicción del valor de una variable a partir de los valores de la otra; sin embargo, puede tomar el mínimo valor en tablas con asociación. El coeficiente de incertidumbre únicamente toma el valor cero en tablas con no asociación; sin embargo, su valor es más difícil de interpretar que el de Lambda.

3. Tablas de contingencia

Otras medidas

Medida de asociación	Tabla	Escala de medida	Observaciones
Kappa	r x r	Nominales	Los posibles valores de las dos variables son los mismos. Toma valores comprendidos entre -1 y 1 que indican, respectivamente, mínimo y máximo grado de acuerdo entre los valores de las dos variables.
Gamma	r x c	Ordinales	Toma valores comprendidos entre -1 y 1 que indican máximo grado de asociación negativa y positiva respectivamente.
Tau-b de Kendall	r x c	Ordinales	Gamma es fácil de interpretar, pero puede alcanzar valores extremos en tablas en las que la asociación no es total
Tau-c de Kendall	r x c	Ordinales	Tau-b únicamente alcanza valores extremos en tablas con asociación total; sin embargo, si r es distinto de c no puede alcanzarlos.
D de Somers	r x c	Ordinales	Tau-c puede alcanzar valores extremos aún en el caso de que r sea distinto de c; sin embargo, tiende a subestimar la asociación. D dispone de versión asimétrica; sin embargo, puede alcanzar valores extremos en tablas en las que la asociación no es total.

3. Tablas de contingencia

Otras medidas

Medida de asociación	Escala de medida	Observaciones
Eta	V.D.: intervalo V.I.: nominal	Los valores de la variable independiente establece grupos en la población. Toma valores entre 0 y 1. Cuanto más próximo a 1 sea su valor más diferenciados estarán los grupos en términos de la puntuaciones de la variable dependiente (mayor será la dependencia de las puntuaciones respecto de los grupos).
Correlación de Pearson	Intervalo	Son medidas del grado de asociación lineal entre las dos variables.
Correlación de Spearman	Intervalo u ordinal	Los coeficientes de correlación de Pearson y de Spearman toman valores comprendidos entre -1 y 1, que indican máximo grado de asociación lineal negativa y positiva respectivamente. La correlación de Spearman es la correlación de Pearson entre los rangos asignados a los valores ordenados.
Asociación lineal de Mantel-Haenszell	Intervalo	La medida de asociación lineal de Mantel-Haenszel se define como el cuadrado del coeficiente de correlación de Pearson multiplicado por (N-1), siendo N el tamaño muestral.

4 Anexo: Estimación de parámetros. Máxima verosimilitud

Anexo: Máxima verosimilitud

Estimación de parámetros. Máxima verosimilitud

El método de máxima verosimilitud es un procedimiento para ajustar los parámetros de un cierto modelo probabilístico a una serie de datos.

Disponemos de unos datos $z_1, z_2, \ldots, z_N$ y suponemos que se trata de una muestra de un modelo probabilístico N-dimensional (conocido) que depende de unos (en general, pocos) parámetros.

Es decir, fijados los valores de esos parámetros, sabemos calcular las probabilidades (conjuntas) con las que unas variables aleatorias $(X_1, X_2, \ldots, X_N)$ toman sus valores.

Por ejemplo, el modelo podría ser una normal N-dimensional, en la que los parámetros serían un vector de medias, uno de desviaciones típicas y una matriz de correlaciones.

En particular, sabemos calcular la probabilidad con la que el modelo multidimensional toma justamente los valores de la serie de datos:

$$\mathbf{P}(X_1 = z_1, X_2 = z_2, \ldots, X_N = z_N)$$

Anexo: Máxima verosimilitud

Esta probabilidad conjunta (que será el valor de la función de densidad conjunta, para el caso de modelos continuos) dependerá de los valores de los parámetros.

El método de máxima verosimilitud consiste en determinar los parámetros de la distribución que **maximizan** esa probabilidad, es decir, los que hacen que la muestra sea lo más "verosímil" posible.

Por supuesto, el método sólo tiene sentido si sabemos evaluar esas probabilidades conjuntas. El análisis comienza aplicando esperanza condicionada:

$$\mathbf{P}(X_1 = z_1, \ldots, X_N = z_N) =$$
$$= \mathbf{P}(X_N = z_N | X_{N-1} = z_{N-1}, \ldots, X_1 = z_1) \cdot \mathbf{P}(X_{N-1} = z_{N-1}, \ldots, X_1 = z_1)$$

Anexo: Máxima verosimilitud

Estimación de parámetros. Máxima verosimilitud

Ahora, repitiendo el proceso, conseguimos factorizar la probabilidad conjunta de la siguiente manera:

$$\begin{aligned}
&\mathbf{P}(X_1 = z_1, \ldots, X_N = z_N) = \\
&\qquad = \mathbf{P}(X_N = z_N | X_{N-1} = z_{N-1}, \ldots, X_1 = z_1) \\
&\qquad \quad \times \mathbf{P}(X_{N-1} = z_{N-1} | X_{N-2} = z_{N-2}, \ldots, X_1 = z_1) \\
&\qquad \quad \times \mathbf{P}(X_{N-2} = z_{N-2} | X_{N-3} = z_{N-3}, \ldots, X_1 = z_1) \\
&\qquad \quad \times \cdots \times \mathbf{P}(X_1 = z_1)
\end{aligned}$$

La expresión sigue siendo muy aparatosa. Pero, en algunas ocasiones, se simplifica considerablemente. Por ejemplo

- cuando las X_n sean **independientes**, porque entonces las probabilidades condicionadas son, simplemente, las probabilidades de que cada X_n valga el correspondiente z_n.
- cuando las X_n constituyan un **proceso estocástico** (con estructura sencilla). Porque entonces bastará condicionar al paso anterior (o a unos cuantos de los anteriores), en lugar de condicionar a toda la historia previa, lo que hace que la expresión sea más manejable.

Anexo: Máxima verosimilitud

Estimación de parámetros. Máxima verosimilitud. Muestras independientes

Partimos de unos datos z_1, z_2, . . . , z_N que son **muestras independientes** de una cierta variable aleatoria X con función de densidad $f_\theta(x)$. Hacemos explícito en la propia notación que la función de densidad depende de un parámetro (también podrían ser varios parámetros).

Buscamos el valor de θ que mejor se ajuste a los datos.

Poniéndolo en el contexto anterior, la serie completa es una única muestra de un vector (X_1, . . . ,X_N), donde las X_n son variables independientes, todas con la misma distribución que la X de referencia.

Por ser i.i.d., la función $h_\theta(x_1, . . . , x_N)$ de densidad conjunta se factoriza como

$$h_\theta(x_1, \ldots, x_N) = f_\theta(x_1) \cdot f_\theta(x_2) \cdots f_\theta(x_N)$$

Así que disponemos de una fórmula manejable para calcular probabilidades conjuntas.

Anexo: Máxima verosimilitud

Estimación de parámetros. Máxima verosimilitud. Muestras independientes

Procedimiento

Los datos son la serie de valores z_1, z_2, . . . , z_N y la función de densidad $f_\theta(x)$.

- Calculamos los valores dela función de densidad en cada uno de los datos,

$$f_\theta(z_1), f_\theta(z_2), \ldots, f_\theta(z_n)$$

- y formamos la función

$$\mathcal{L}_\theta(z_1, \ldots, z_n) = \prod_{j=1}^{n} f_\theta(z_j)$$

O mejor, su logaritmo:

$$\text{verosimilitud}(\theta) = \ln\left(\mathcal{L}_\theta(z_1, \ldots, z_n)\right) = \sum_{j=1}^{n} \ln\left(f_\theta(z_j)\right)$$

- Ahora buscamos (con algún método de búsqueda de extremos, usualmente numérico) el valor de θ que maximiza la función de verosimilitud (θ).

Anexo: Máxima verosimilitud

Estimación de parámetros. Máxima verosimilitud. Muestras independientes

Observaciones

- Nótese que maximizar una función o su logaritmo produce el mismo resultado (el logaritmo es una función continua y estrictamente creciente). Tomamos logaritmos para trabajar con sumas, en lugar de con productos.

- Es también habitual tomar el negativo de la función de verosimilitud, y entonces buscar su mínimo.

- En ocasiones (pocas habitualmente), el máximo se puede obtener analíticamente. En general, se debe calcular numéricamente.

- A veces puede haber varios máximos locales de la función de verosimilitud (o quizás ninguno), de manera que el procedimiento de optimización podría no determinar el máximo global.

- Puede producir estimadores sesgados de los parámetros.

Anexo: Máxima verosimilitud

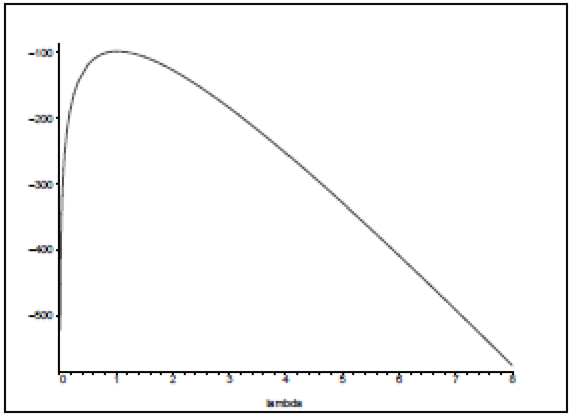
Estimación de parámetros. Máxima verosimilitud. Muestras independientes

Ejemplo: Ajuste de una exponencial

Digamos que los datos son $z_1, z_2, \ldots, z_N$ y que queremos determinar el parámetro λ de una exponencial, cuya función de densidad es $f_\lambda(x) = \lambda e^{-\lambda x}$.

Obsérvese que

$$\mathcal{L}_\lambda(z_1, \ldots, z_N) = \lambda^N \exp\Big(-\lambda \sum_{j=1}^{N} z_j \Big)$$



y por tanto

$$\ln\big(\mathcal{L}_\lambda(z_1, \ldots, z_N)\big) = N \ln(\lambda) - \lambda \sum_{j=1}^{N} z_j$$

Para localizar el máximo, derivamos con respecto a λ e igualamos a 0, para obtener:

$$0 = \frac{\partial \ln(\mathcal{L}_\lambda)}{\partial \lambda} = \frac{N}{\lambda} - \sum_{j=1}^{N} z_j \quad \Longrightarrow \quad \frac{1}{\lambda} = \frac{1}{N} \sum_{j=1}^{N} z_j$$

Así que el λ óptimo es el recíproco de la media muestral.

Referencias

An Introduction to Statistical Learning

http://fs2.american.edu/alberto/www/analytics/ISLRLectures.html

The Elements of Statistical Learning. Data Mining, Inference, and Prediction

https://web.stanford.edu/~hastie/ElemStatLearn//

Github

https://github.com/abhat222/Data-Science--Cheat-Sheet

Seeing theory

https://seeing-theory.brown.edu/index.html#firstPage