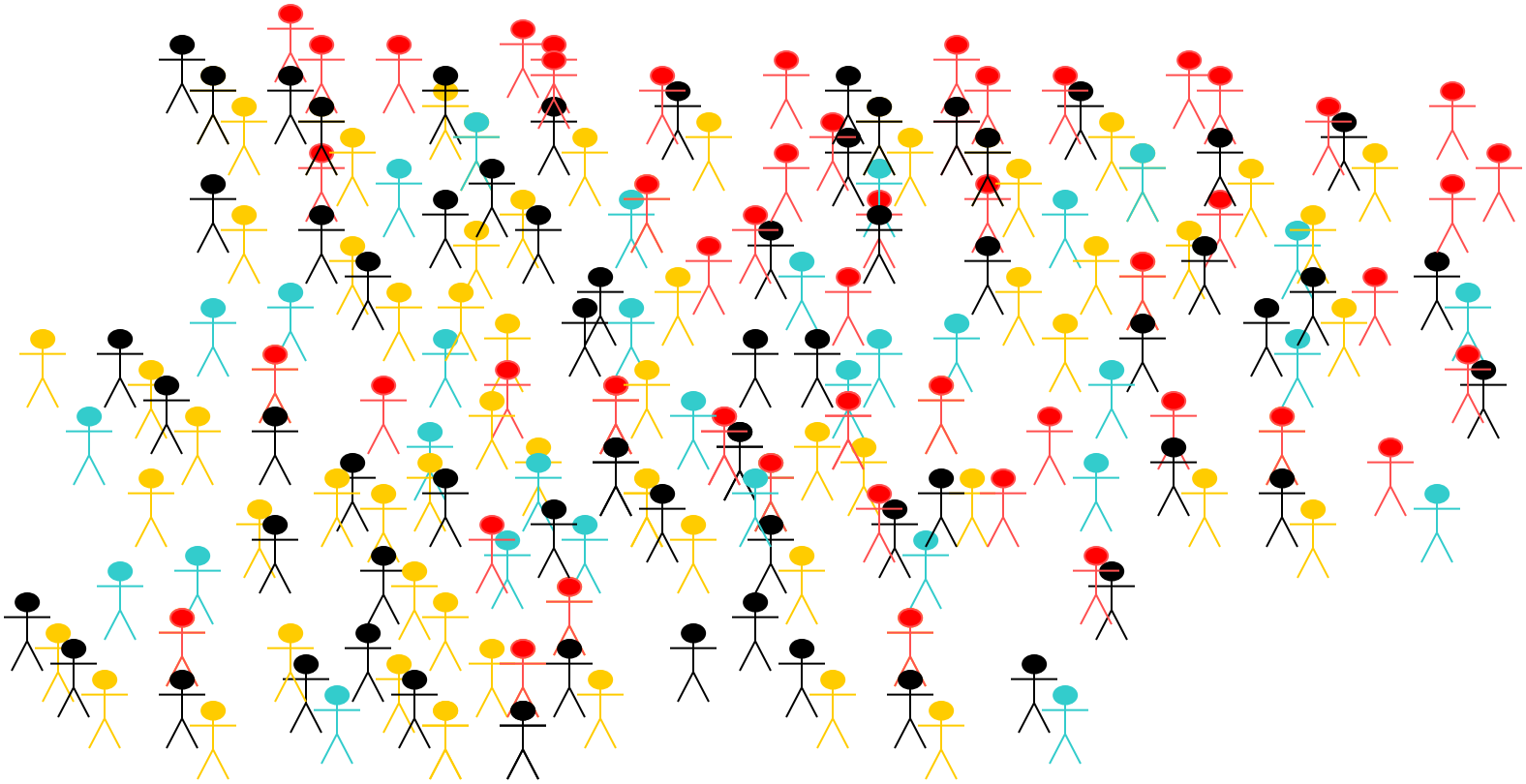
Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

La inferencia estadística. Población y muestra

Objetivo de la inferencia estadística:

- Aprender de la observación
- Generalizar lo que aprendemos de una muestra a toda la población



La inferencia estadística. Población y muestra

Objetivo de la inferencia estadística:

- Aprender de la observación
- Generalizar lo que aprendemos de una muestra a toda la población

La inferencia estadística. Población y muestra

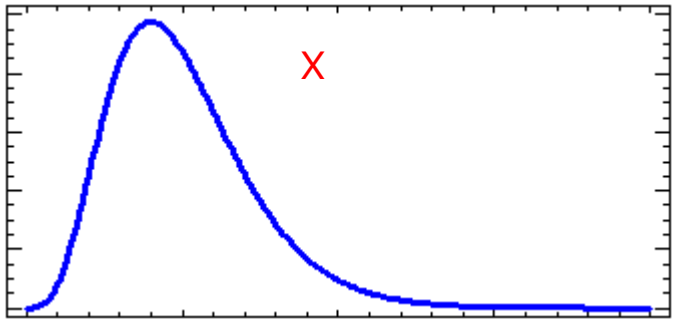
Objetivo de la inferencia estadística:

- Aprender de la observación
- Generalizar lo que aprendemos de una muestra a toda la población

INFERENCIA

MUESTRA DE **n** OBSERVACIONES → POBLACIÓN

$X_1, X_2, \ldots, X_n$

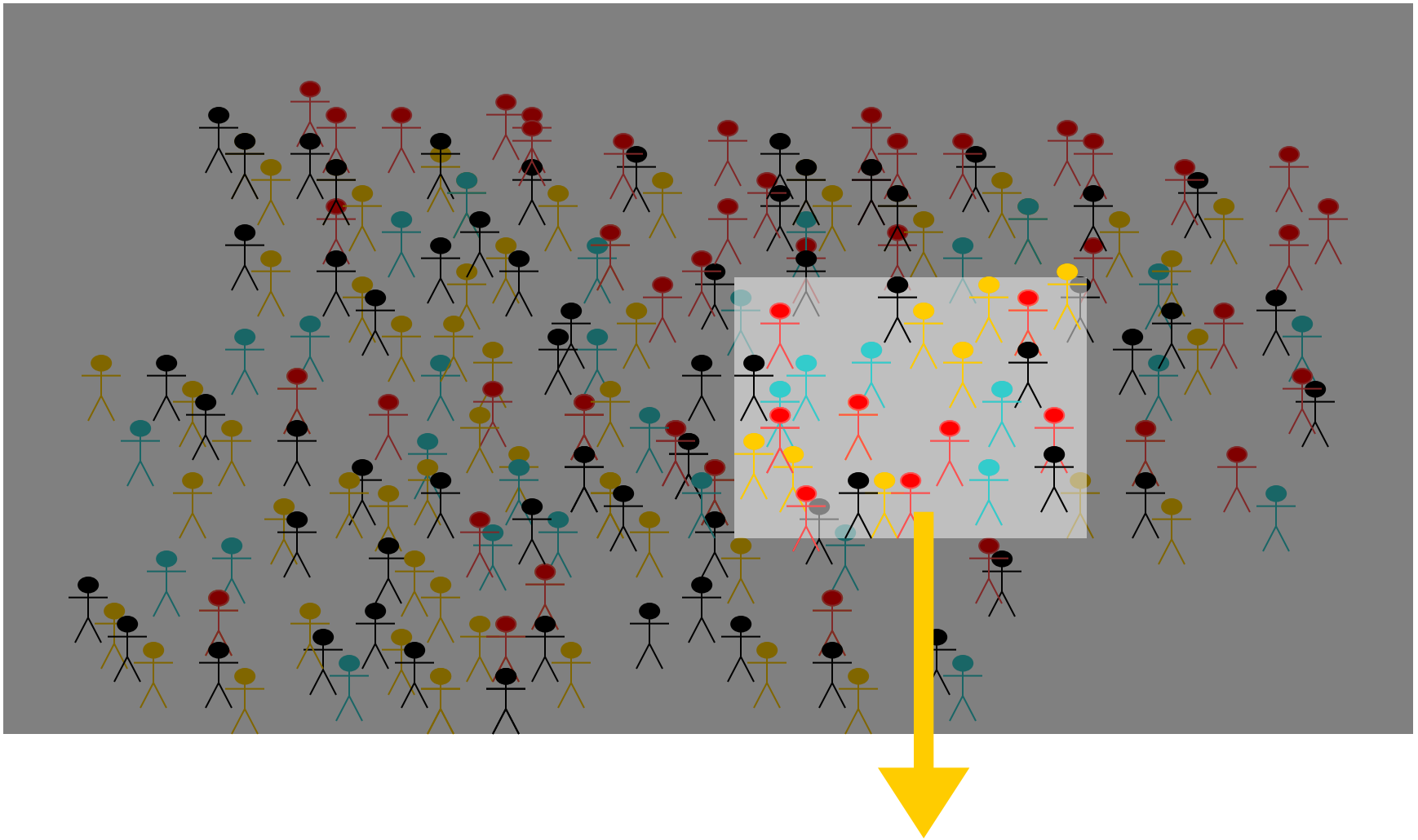


Muestra aleatoria simple:

- Todas las X_i tienen las mismas características que X
- Son independientes entre si

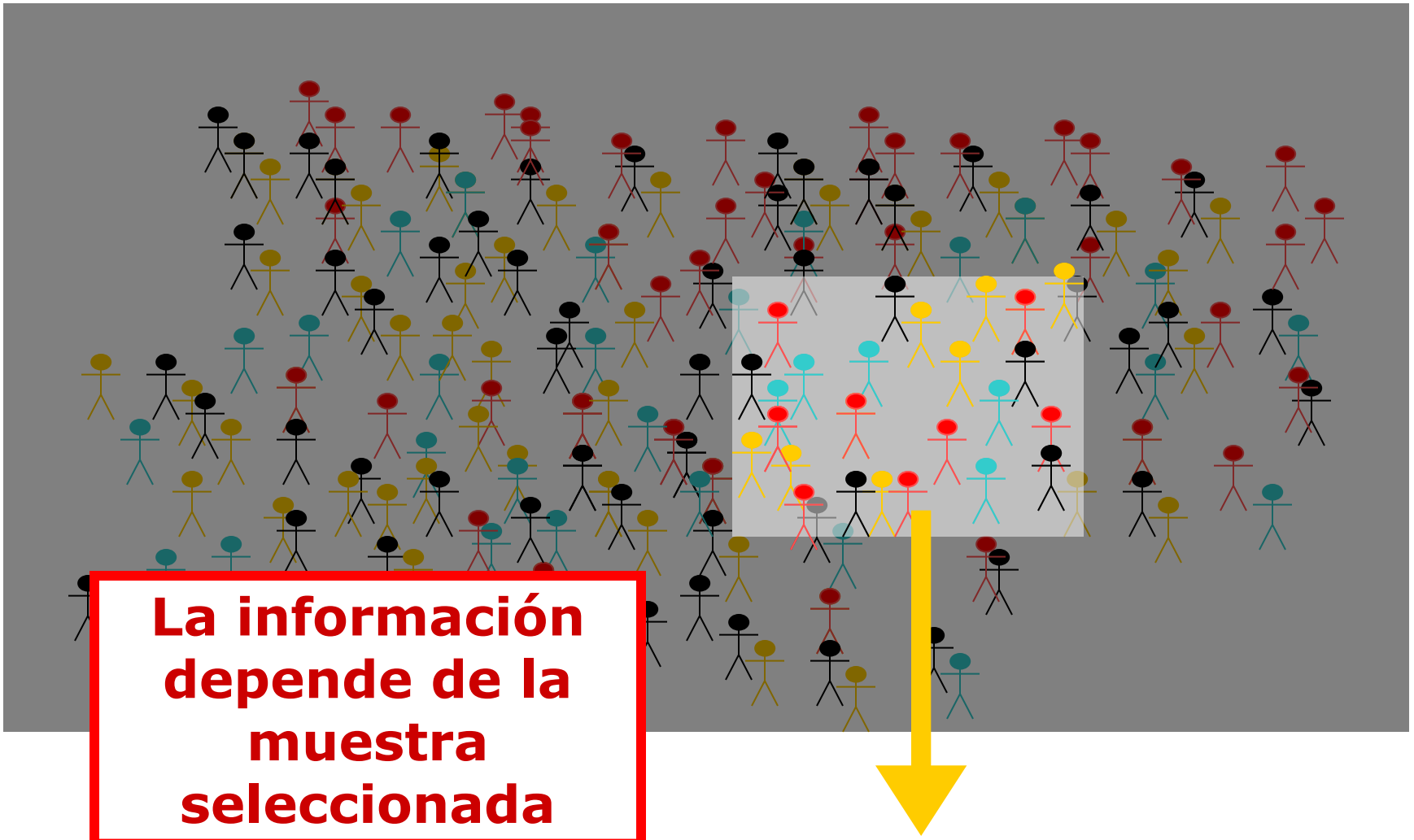
↓

$X_1, X_2, \ldots, X_n$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.)



Extraemos información de la muestra:

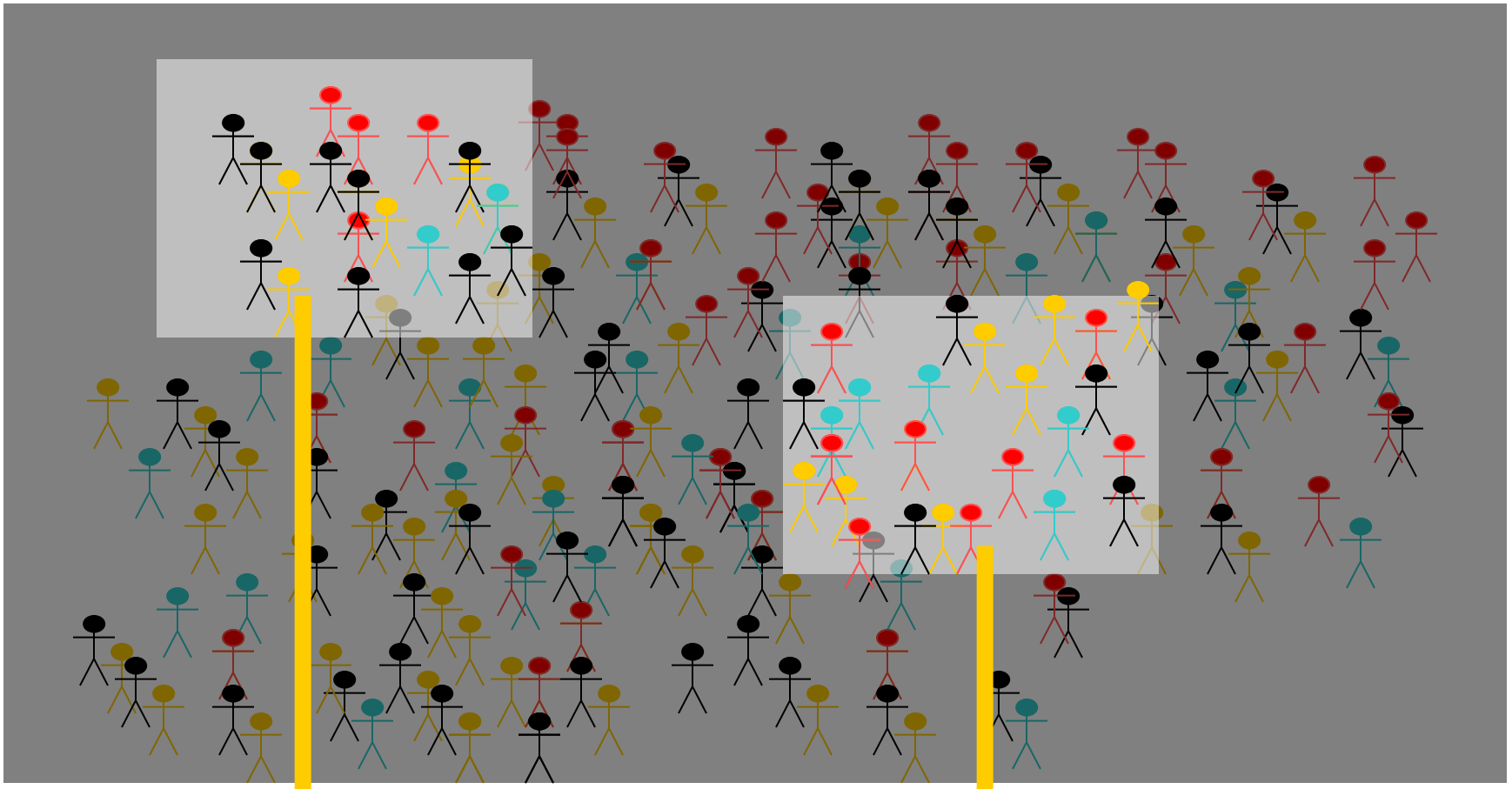
- Histograma
- Media muestral
- Varianza muestral ...

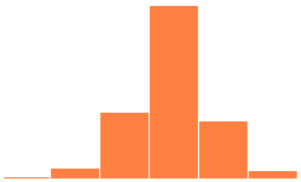


La información depende de la muestra seleccionada

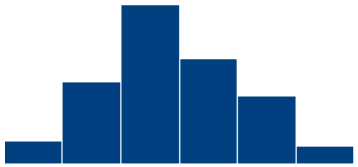
Extraemos información de la muestra:

- Histograma
- Media muestral
- Varianza muestral ...





$\bar{x} = 25.4$

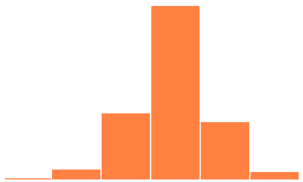


$\bar{x} = 23.2$

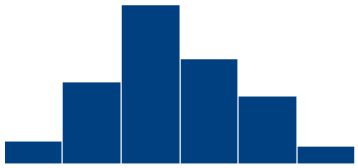
ESTADÍSTICO: función matemática realizada con una muestra

Ejemplo: media muestral, varianza muestral,…

El valor de un ESTADÍSTICO varía con la muestra



$\bar{x} = 25.4$



$\bar{x} = 23.2$

Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. **Estimación y estimadores**
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

Estimación y Estimadores
Distribución muestral de un estadístico

Estadístico: Cualquier función evaluada en una muestra

Ejemplo: Media muestral

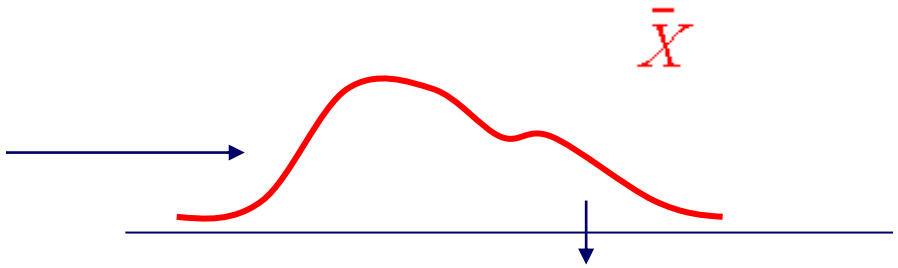
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Los elementos de la muestra son variables aleatorias. Su valor cambia de unas muestras a otras

Un estadístico es siempre una variable aleatoria. Su valor cambia de unas muestras a otras

Su distribución: distribución en el muestreo o **distribución muestral**

Depende de la operación que se realiza y de las propiedades de X



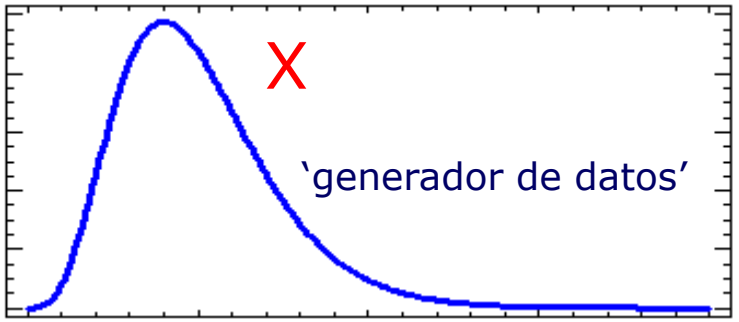
con una muestra tenemos sólo un valor concreto

La distribución de la media muestral

Estadístico: Media muestral

Distribución de X

$E(X) = \mu \quad \mathrm{Var}(X) = \sigma^2$



$X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_n$

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$\bar{X}$

?

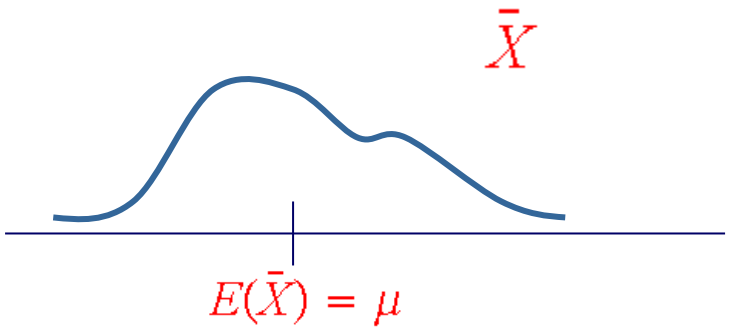
con una muestra tenemos sólo un valor concreto

- ¿Qué forma tiene la distribución de la media muestral? (la que se obtiene si cambiamos los elementos de la muestra)
- ¿Es la media muestral una buena aproximación a la media poblacional μ?

La distribución de la media muestral

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \frac{n\mu}{n} = \mu$$



La media poblacional está en el centro de las diferentes medias muestrales que podríamos haber obtenido con diferentes muestras diferentes

La distribución de la media muestral

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\mathrm{Var}(X_1 + X_2 + \cdots + X_n)}{n^2}$$

independientes

$$= \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)}{n^2} = \frac{n\sigma^2}{n} = \frac{\sigma^2}{n}$$

Disminuye con n

- Si **n es suficientemente grande**, la media muestral cambiaría poco de unas muestras a otras
- Si **n es suficientemente grande,** es muy poco probable que la media muestral dé un valor muy alejado de μ

La distribución de la media muestral

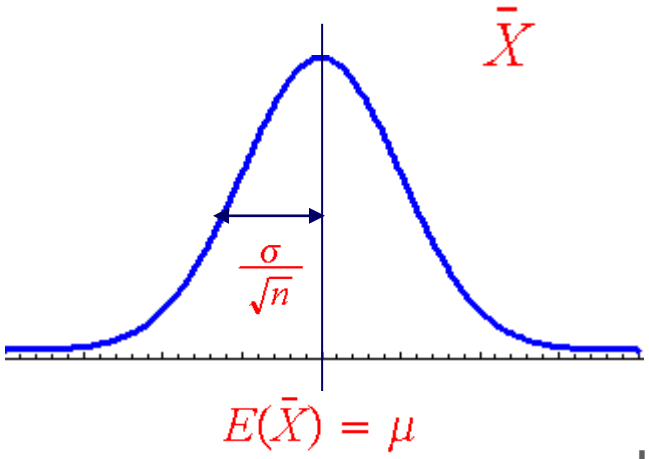
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \longrightarrow \bar{X} = \frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}$$

Estamos sumando variables aleatorias.

Por el **Teorema Central del Límite**, si n es grande (n>30)

NORMAL

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



- Simétrica
- Concentrada en μ
- Con n alto, es muy probable que cualquier muestra dé un valor próximo a μ

Estimación y estimadores

Parámetros → valores numéricos sobre características de la población:

μ , σ^2 , λ , Cuartiles,...

Si un parámetro es desconocido → le asignamos un valor a partir de una muestra de datos

ESTIMACIÓN DEL PARÁMETRO

ESTIMACIÓN: cálculo de un valor numérico a partir de una muestra, con el fin de asignar un valor a un parámetro desconocido

ESTIMADOR: estadístico que se emplea en la estimación de un parámetro (como es un estadístico, será una variable aleatoria)

Ejemplo: la media muestral se puede usar como estimador de la media poblacional

Estimación y estimadores

NOTACIÓN: El símbolo para denotar a un estimador será el mismo que el del parámetro pero con acento circunflejo ^

Estimador de la media poblacional μ → $\hat{\mu}$

Estimador de la varianza σ^2 → $\hat{\sigma}^2$

Estimador de un parámetro θ → $\hat{\theta}$

Se pueden proponer varios estimadores para un parámetro.

¿Cómo seleccionar el más adecuado?

¿Cómo seleccionar el estimador más adecuado?

¿Qué le vamos a pedir a un estimador de un parámetro θ ?

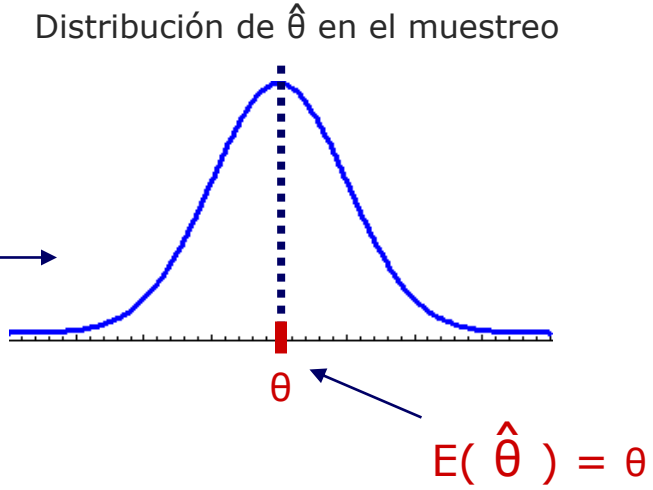
Que, aunque sea una variable aleatoria cuyo valor depende de la muestra empleada, dé un valor próximo al parámetro verdadero con mucha probabilidad

¿Cómo evaluarlo?

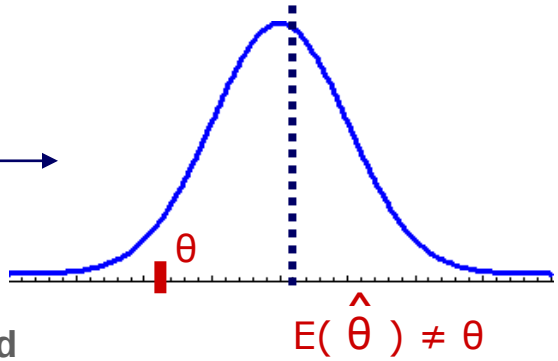
Varios criterios

1 Que $E(\hat{\theta})$ no se aleje mucho de θ

BUEN estimador del parámetro θ



MAL estimador del parámetro θ.
Tendencia a sobreestimar el valor



¿Cómo seleccionar el estimador más adecuado?

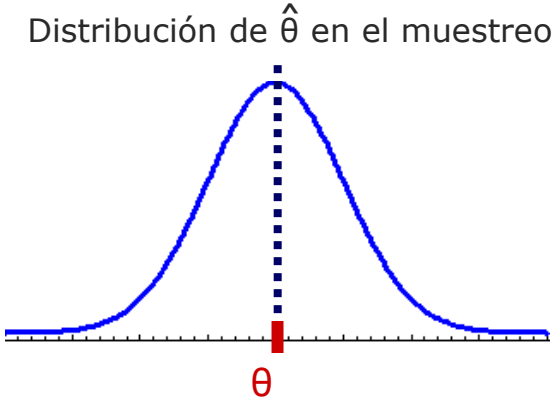
¿Qué le vamos a pedir a un estimador de un parámetro θ ?

Que, aunque sea una variable aleatoria cuyo valor depende de la muestra empleada, dé un valor próximo al parámetro verdadero con mucha probabilidad

¿Cómo evaluarlo?

Varios criterios

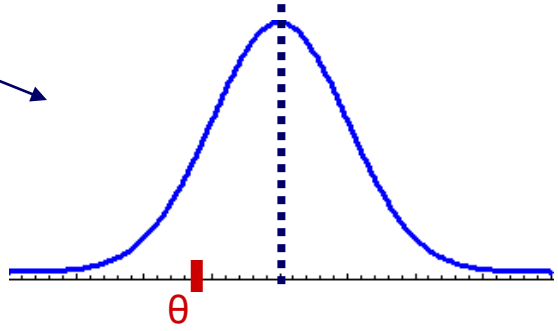
1 Que $E(\hat{\theta})$ no se aleje mucho de θ



insesgado o centrado

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Sesgado. Sesgo positivo

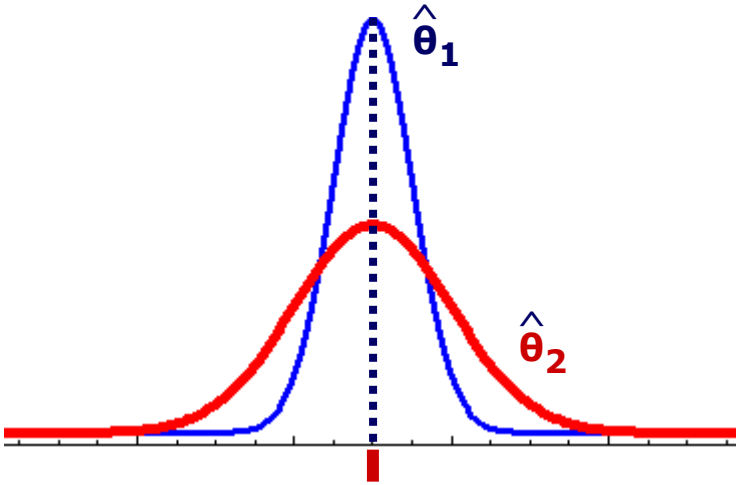


El estimador será mejor, cuanto menor sea su sesgo

¿Cómo seleccionar el estimador más adecuado?

1 Que $E(\hat{\theta})$ no se aleje mucho de θ

2 Que $\hat{\theta}$ tenga poca varianza

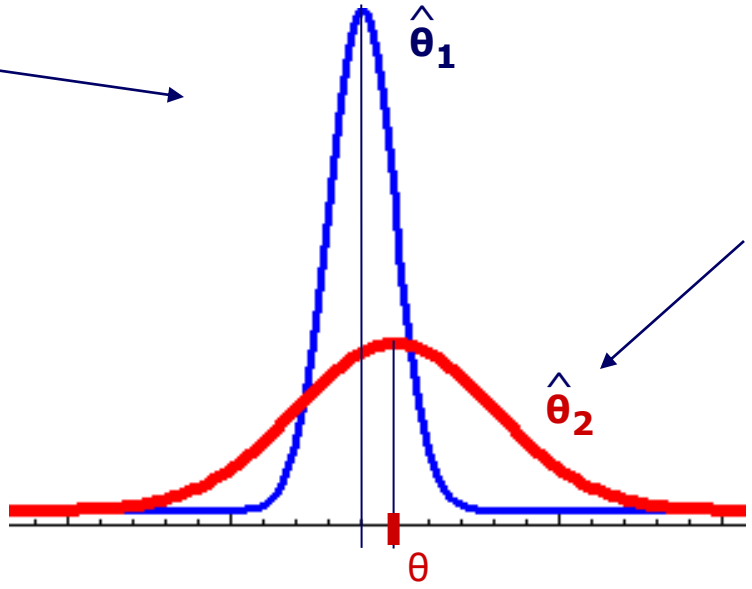


Aunque ambos son insesgados, el estimador $\hat{\theta}_2$ es peor que el $\hat{\theta}_1$, pues tiene mayor varianza. Es menos preciso.

¿Cómo seleccionar el estimador más adecuado?

1. Que $E(\hat{\theta})$ no se aleje mucho de θ
2. Que $\hat{\theta}$ tenga poca varianza
3. Si hay varios estimadores, con distinto sesgo y varianza. Es mejor el que tenga menor ERROR CUADRATICO MEDIO (ECM)

Es sesgado, pero de menor varianza



Es insesgado pero de mayor varianza

Calculamos el ECM de cada uno y elegimos el que tenga ECM menor

ESTIMADOR EFICIENTE

Es más probable que en una muestra dé un valor próximo al valor poblacional

Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

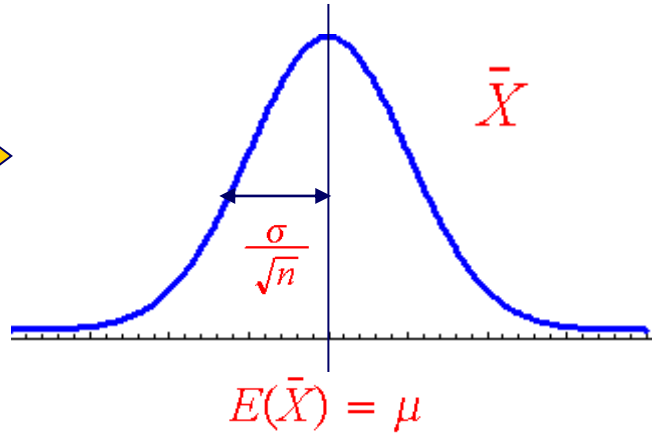
Intervalos de confianza para μ con muestras grandes

Sea X una v. aleatoria de interés con distribución **cualquiera** y con

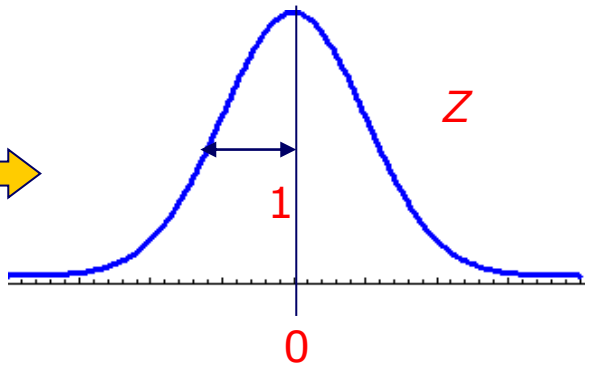
$$E(X) = \mu \qquad \mathrm{Var}(X) = \sigma^2$$

En el tema anterior vimos que si n es grande (n>30)

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



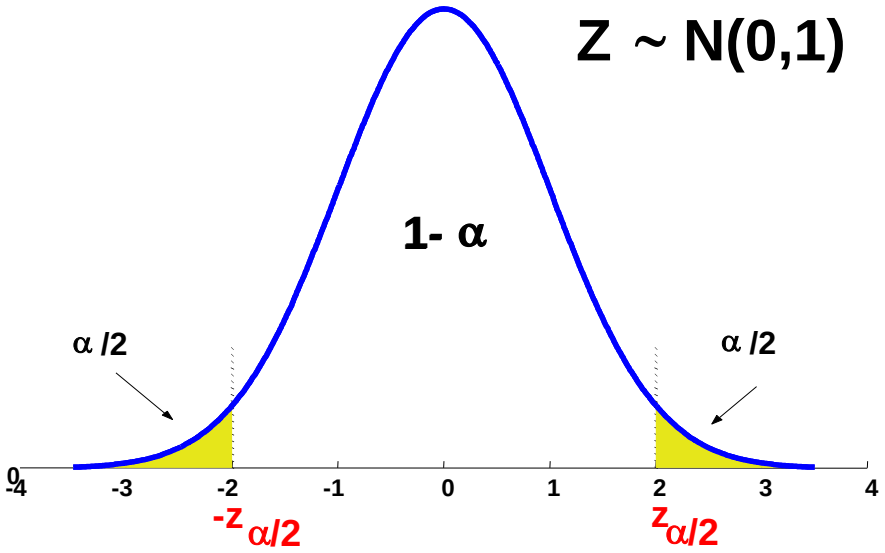
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \longrightarrow P(-z_{\alpha/2} < Z < z_{\alpha/2}) = (1-\alpha)$$

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

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

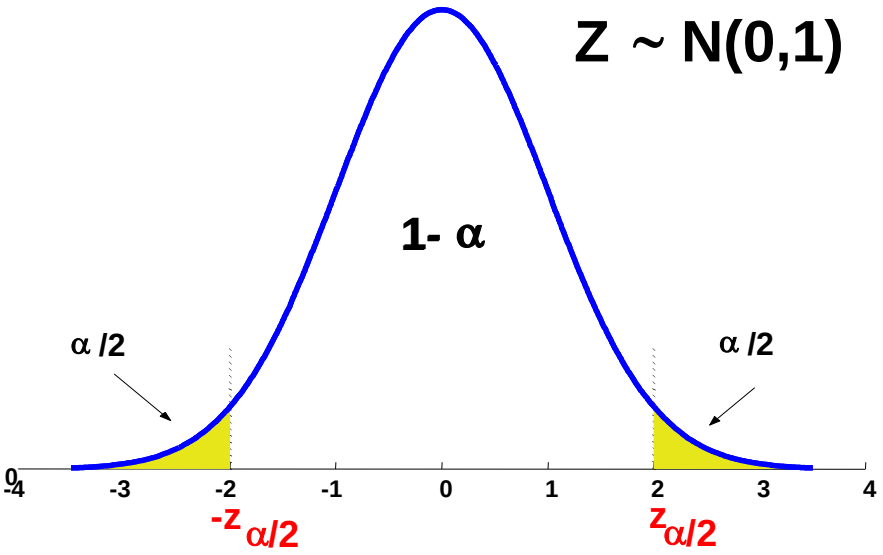


$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = (1-\alpha)$$

Si tomásemos infinitas muestras, y con cada una calculásemos el intervalo

$$\bar{X} \pm z_{a/2}\frac{s}{\sqrt{n}}$$

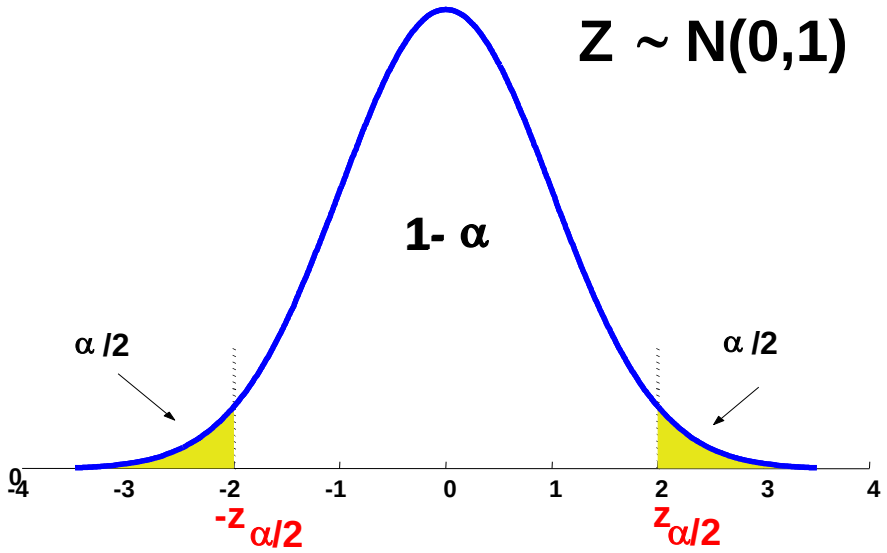
Entonces, el 100(1-α)% de esos intervalos contendría el valor de μ



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

En la práctica:

- ✓Sólo una muestra
- ✓Sólo un intervalo
- ✓El intervalo contendrá o no a μ
- ✓A la incertidumbre de si lo contendrá le llamaremos **confianza**



Intervalo de confianza de nivel de confianza $100\times(1-\alpha)\%$ para μ

$$IC(1-\alpha)\% = \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Ejemplo

Una muestra aleatoria extraída de una población con $\sigma^2=100$ de $n=144$ observaciones tiene una media muestral $=160$. se pide:

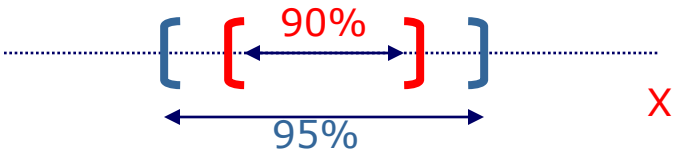
(a) Calcular un intervalo de confianza del 95% para μ.

(b) Calcular un intervalo de confianza del 90% para μ.

(a) $z_{\alpha/2} = z_{0.025} = 1.96 \longrightarrow IC(95\%) : \mu \in \left\{160 \pm 1.96\frac{10}{\sqrt{144}}\right\} \longrightarrow \mu \in [158.36, 161.63]$

(b) $z_{\alpha/2} = z_{0.05} = 1.65 \longrightarrow IC(90\%) : \mu \in \left\{160 \pm 1.65\frac{10}{\sqrt{144}}\right\} \longrightarrow \mu \in [158.625, 161.375]$

Mayor confianza = Mayor amplitud



$$IC(1-\alpha)\% = \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Es también un parámetro, y será desconocido

Lo sustituimos por un estimador

$$IC(1-\alpha)\% = \bar{x} \pm z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

¿Qué estimador usamos para σ^2?

¿Qué estimador usamos para σ^2 ?

Método de los momentos: varianza muestral →

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

Se puede demostrar que es SESGADO

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

$$\text{sesgo}(S^2) = \sigma^2 \frac{(n-1)}{n} - \sigma^2 = -\frac{\sigma^2}{n}$$

Subestima la verdadera varianza

¿Qué estimador usamos para σ^2 ?

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$ → es SESGADO

$$E(S^2) = \sigma^2 \frac{(n-1)}{n} \neq \sigma^2$$

$$E\left(S^2 \underbrace{\frac{n}{n-1}}\right) = \sigma^2, \longrightarrow S^2 \frac{n}{n-1} = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\right)\frac{n}{n-1} = \overbrace{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}^{\hat{S}^2}$$

Corregimos el sesgo

Nuestro estimador 'oficial' será el estimador insesgado → $$\hat{\sigma}^2 = \hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- Cuasivarianza
- Pseudo varianza
- Varianza corregida
- Varianza corregida por grados de libertad

Intervalo de confianza de nivel de confianza $100\times(1-\alpha)\%$ para μ

$$IC(1-\alpha)\% = \bar{x} \pm z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

Ejemplo

Se mide la duración de 200 componentes electrónicos hasta su avería. De esos 200 datos se tiene que la media muestral es 1300 horas y la cuasivarianza es 10.000 (horas al cuadrado).

Calcula un intervalo de confianza de μ de nivel de confianza 95%

$\bar{X} = 1300$

$\hat{S}^2 = 10.000$

$n = 200$

$a = 0.05$

$z_{0.025} = 1.96$

$$1300 \pm 1.96\frac{\sqrt{10000}}{\sqrt{200}} \rightarrow (1286.141, 1313.859)$$

Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. **Determinación del tamaño muestral**
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

Determinación del tamaño muestral

Acabamos de ver que...

Intervalo de confianza de nivel de confianza $100\times(1-\alpha)\%$ para μ

$$IC(1-\alpha)\% = \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$\bar{x} \pm L$$

¿Cuál debe ser n para conseguir un L determinado?

Se estima con alguna muestra piloto

$$L = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2}\sigma}{L}\right)^2$$

Ejemplo

Sea X el contenido de impurezas en un material obtenido en cierto proceso productivo (miligramos de impureza por kilogramo de producto obtenido).

Se toma una muestra aleatoria de 200 observaciones obteniéndose una media muestral del consumo de 120 mg/Kg y una desviación típica muestral 20 mg/Kg.

$$\bar{X} = 120$$

$$\hat{S} = 20$$

$$n_0 = 200$$

Estimar mediante un intervalo de un 95% de confianza el contenido medio de impurezas.

$$\mu \in \left[\bar{X} \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] = \left[120 \pm 1.96 \times \frac{20}{\sqrt{200}} \right] = [120 \pm 2.77]$$

¿Qué tamaño muestral sería necesario tomar para que L=1 mg?

$$n = \left(\frac{z_{\alpha/2} \hat{\sigma}}{L} \right)^2 = \left(\frac{1.96 \times 20}{1} \right)^2 \approx 1537$$

Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

Intervalo para una proporción p

Si np(1-p)>5, el intervalo de confianza es:

$$p \in \left\{ \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right\}$$

Intervalo para el parámetro de una Poisson

Si n y λ > 5 es grande:

$$IC(1-\alpha) : \lambda \in \left\{ \hat{\lambda} \pm z_{\alpha/2} \sqrt{\frac{\hat{\lambda}}{n}} \right\}$$

Intervalos sólo para poblaciones normales, pero con cualquier tamaño muestral

$$IC(1-\alpha) : \mu \in \left\{ \bar{x} \pm t_{n-1;\alpha/2} \sqrt{\frac{\hat{s}^2}{n}} \right\}$$

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

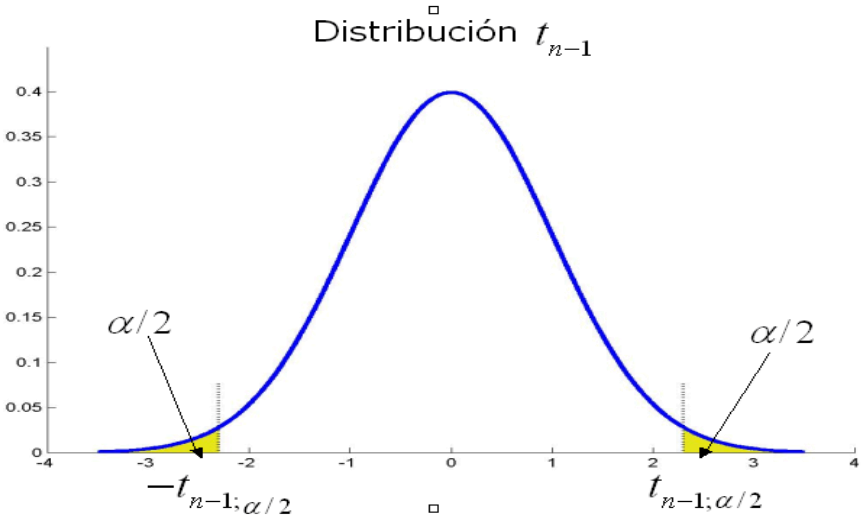
SÓLO SI LA POBLACION ES **NORMAL**:

en lugar de $z_{\alpha/2}$

$$IC(1-\alpha): \mu \in \left\{ \bar{X} \pm t_{n-1;\alpha/2} \frac{\hat{S}}{\sqrt{n}} \right\}$$

- La distribución t de Student es una variable aleatoria continua, simétrica, de media cero, y de perfil parecido a la normal estándar.
- Depende de un parámetro g que se denomina grados de libertad. Su notación habitual es t_g

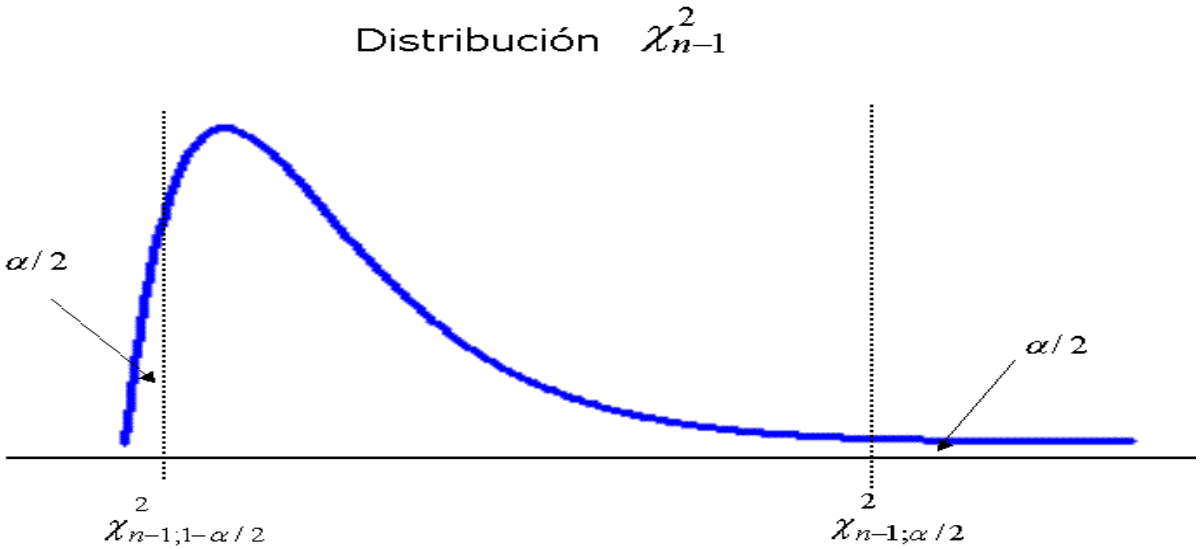
$$t_{n-1} \rightarrow N(0,1)$$



Con n grande los intervalos con la t de Student son aproximadamente iguales a los obtenidos con la N(0,1)

SÓLO SI LA POBLACION ES **NORMAL**:

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



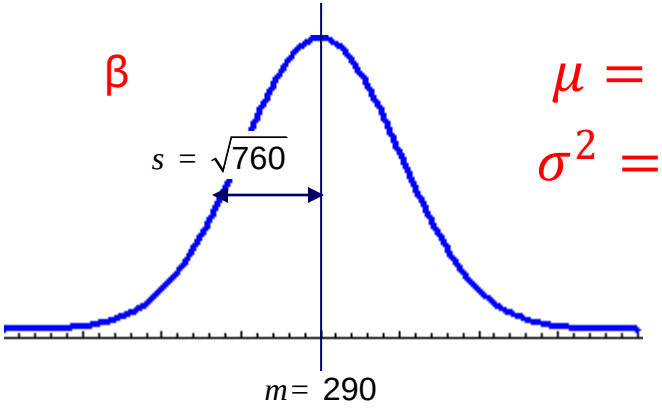
Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

Veamos la idea de contraste de hipótesis con un ejemplo

Ejemplo

Un fabricante de transistores sabe que cuando su producción se mantiene en los niveles de calidad deseables, el valor de la llamada ganancia en corriente de los transistores (conocida por **β**, adimensional) sigue una distribución normal de media 290 y varianza 760.



$$\mu = 290$$
$$\sigma^2 = 760$$

Son en realidad estimaciones con muchísimos datos históricos.

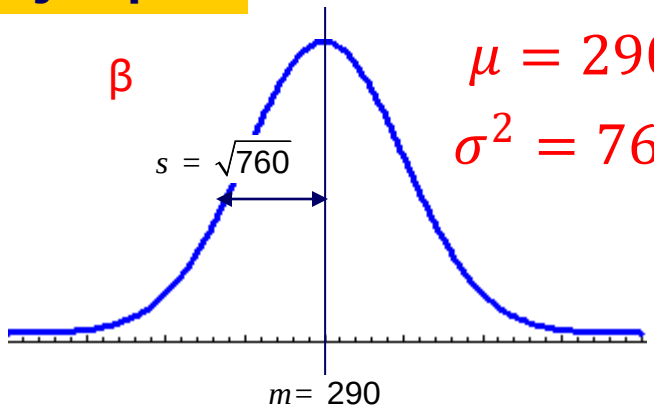
A efectos prácticos, los consideramos como si fuesen los valores poblacionales

¿Cómo puedo saber si el proceso se mantiene en los mismos parámetros?

¿Se mantiene la media?

¿Ha aumentado la variabilidad?

Ejemplo



$\mu = 290$

$\sigma^2 = 760$

¿Cómo puedo saber si se mantiene el proceso en los mismos parámetros?

¿Se mantiene la media?

¿Ha aumentado la variabilidad?

Son hipótesis que quiero comprobar

¿Cómo lo puedo hacer?

- Tomo una muestra de observaciones
- A la vista de los datos decido si mantengo o no la hipótesis (el objetivo no es estimar sino validar)

Si $\bar{x} >> 290$ → parece muy probable que la media haya cambiado

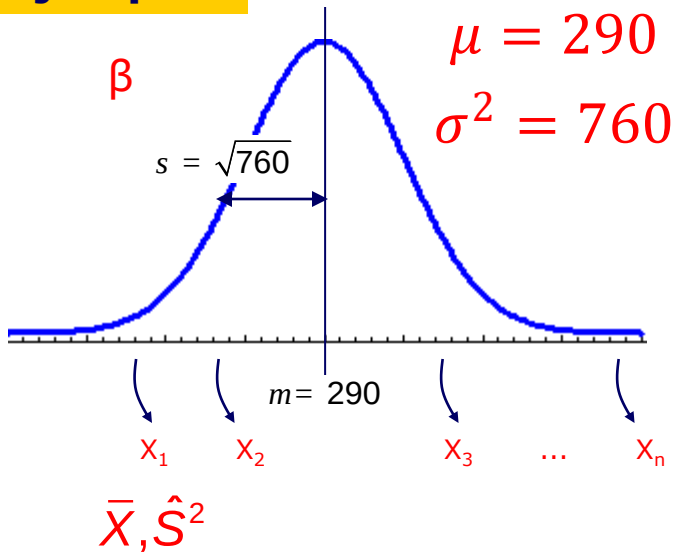
Si $\bar{x} \sim 290$ → parece probable que la media **NO** haya cambiado

A la vista de los datos, tomo la decisión que sea más plausible

(nunca estaré seguro al 100%)

¿Cómo me puede ayudar la estadística?

Ejemplo



Objetivo: Validar una hipótesis con los datos

↓

Contraste de hipótesis

↓

Las hipótesis serán restricciones sobre los parámetros

	Hipótesis nula H_0		**Hipótesis alternativa** H_1	
¿Ha variado la media? →	$\mu = 290$	ó	$\mu \neq 290$	**alternativa bilateral**
¿Ha aumentado la variabilidad? →	$\sigma^2 = 760$	ó	$\sigma^2 > 760$	**alternativa unilateral**

- Entre H_0 y H_1 está todo el rango de valores posibles.
- H_0 debe tener siempre el signo de igualdad.
- Se aceptará H_0 salvo que haya mucha evidencia en contra.

Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

Contraste de hipótesis de la media con muestras grandes

Para contrastar una hipótesis sobre la media μ seguimos los siguientes pasos:

PASO 1: Especificamos la hipótesis nula y la alternativa. Queremos contrastar alguna de estas hipótesis, donde μ_0 es un valor concreto

$$\begin{array}{c|c|c} H_0: \mu = \mu_0 & H_0: \mu = \mu_0 & H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 & H_1: \mu > \mu_0 & H_1: \mu < \mu_0 \end{array}$$

Ejemplo En el ejemplo de los transistores. Se desea saber si la población de transistores del proceso productivo mantiene la media en $\mu_0 = 290$

H_0	H_1
$\mu = 290$	$\mu \neq 290$

PASO 2: Hallamos una medida de la discrepancia entre los datos y H_0

Si la discrepancia es grande: Se rechaza H_0

Esa medida se denomina **estadístico de contraste**

¿Cómo se busca el **estadístico de contraste**, que resuma la información relevante para un contraste?

→ Usando las **propiedades de los estimadores**, e introduciendo la información de H_0

Sabemos que, para muestras grandes

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}} \sim N(0,1)$$

Estadístico de contraste

$$Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \qquad T_0 = \frac{\bar{X}-\mu_0}{\hat{S}/\sqrt{n}}$$

Ejemplo

En el ejemplo de los transistores. Se desea saber si la población de transistores del proceso productivo mantiene la media en $\mu_0 = 290$

H_0	H_1
$\mu = 290$	$\mu \neq 290$

Con 100 observaciones:

$$\bar{x} = 282.3; \hat{s} = 27.69;$$

$$t_0 = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \frac{282.3 - 290}{27.69/10} = -2.78$$

Resume en un número la información para decidir entre H_0 y H_1

PASO 3:

Para valorar el estadístico de contraste, buscamos una distribución de referencia que nos diga si es un valor grande o pequeño

La distribución de referencia es la del estadístico de contraste cuando $\mu=\mu_0$ → $N(0,1)$

PASO 4: Localizamos en qué zonas de la distribución de referencia rechazaremos H0.

Rechazamos H0 si los datos hacen lo que dice H1 de forma muy evidente.

Caso (a)

PASO 1:

$$H_0 : \mu = 290; H_1 : \mu \neq 290$$

PASO 2:

$$T_0 = \frac{\bar{X} - 290}{\hat{S}/\sqrt{n}}$$

PASO 3:

$T_0 \sim N(0,1)$

Rechazamos H0 si

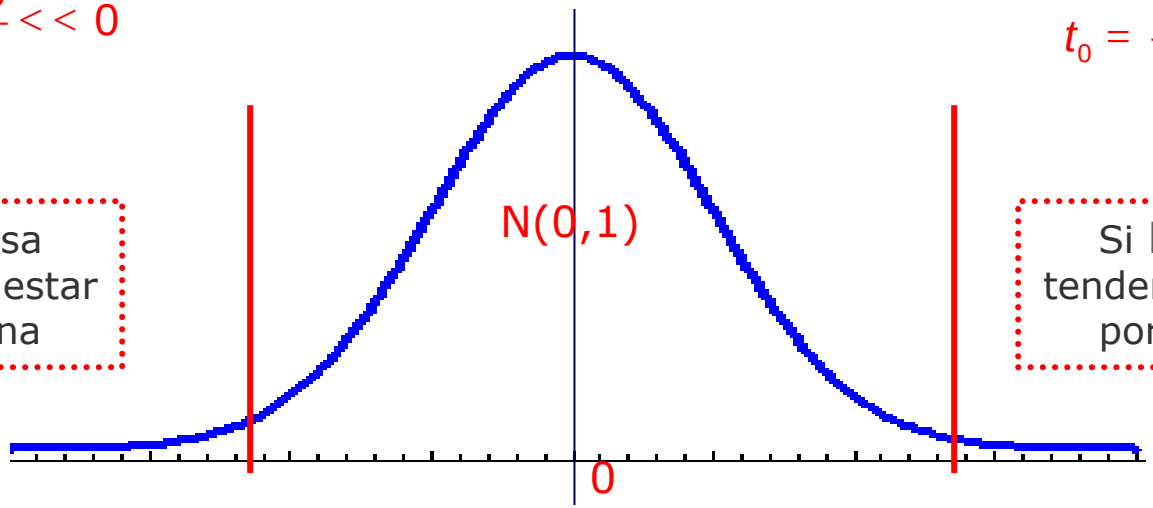
$\bar{x} << 290$

$$t_0 = \frac{\bar{x} - 290}{\hat{s}/\sqrt{n}} << 0$$

$\bar{x} >> 290$

$$t_0 = \frac{\bar{x} - 290}{\hat{s}/\sqrt{n}} >> 0$$

Si H0 es falsa tenderemos a estar por esta zona



Si H0 es falsa tenderemos a estar por esta zona

PASO 4: Localizamos en qué zonas de la distribución de referencia rechazaremos H_0

Rechazamos H_0 si los datos hacen lo que dice H_1 de forma muy evidente.

Caso (b)

PASO 1:

$$H_0 : \mu \leq 290; H_1 : \mu > 290$$

PASO 2:

$$T_0 = \frac{\bar{X} - 290}{\hat{S}/\sqrt{n}}$$

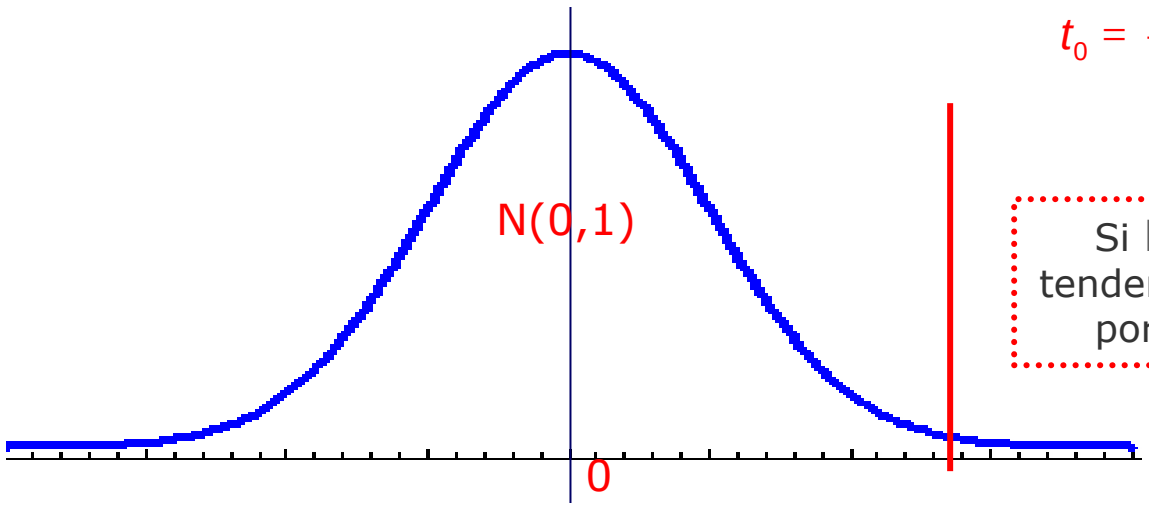
PASO 3:

$T_0 \sim N(0,1)$

Rechazamos H_0 si

$$\bar{x} >> 290$$

$$t_0 = \frac{\bar{x} - 290}{\hat{s}/\sqrt{n}} >> 0$$



Si H_0 es falsa tenderemos a estar por esta zona

PASO 4: Localizamos en qué zonas de la distribución de referencia rechazaremos H_0

Rechazamos H_0 si los datos hacen lo que dice H_1 de forma muy evidente.

Caso (c)

PASO 1:

$$H_0 : \mu \geq 290; H_1 : \mu < 290$$

PASO 2:

$$T_0 = \frac{\bar{X} - 290}{\hat{S}/\sqrt{n}}$$

PASO 3:

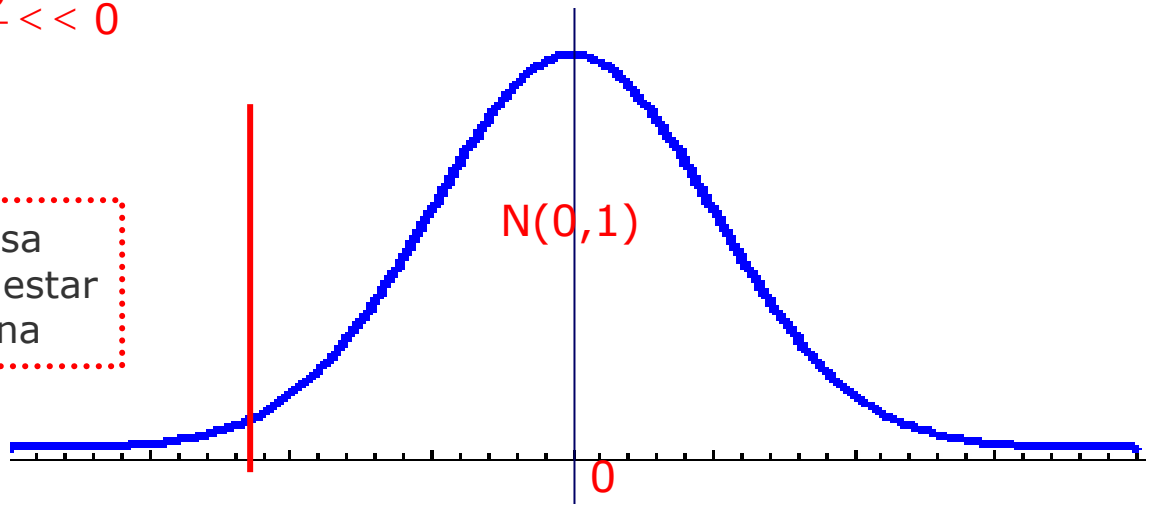
$T_0 \sim N(0,1)$

Rechazamos H_0 si

$$\bar{x} << 290$$

$$t_0 = \frac{\bar{x} - 290}{\hat{s}/\sqrt{n}} << 0$$

Si H_0 es falsa tenderemos a estar por esta zona



PASO 1:

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

(a)

$H_0 : \mu \leq \mu_0 ; H_1 : \mu > \mu_0$

(b)

$H_0 : \mu \geq \mu_0 ; H_1 : \mu < \mu_0$

(c)

PASO 2:

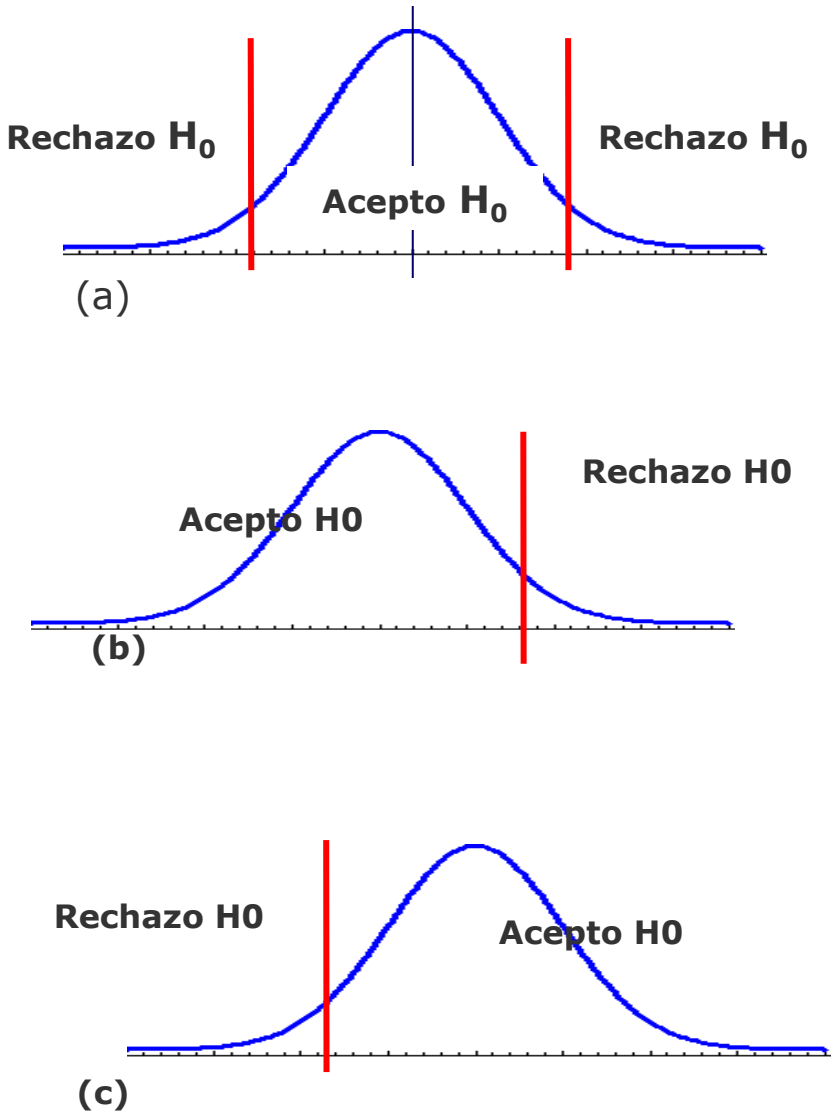
$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

$$T_0 = \frac{\bar{X} - \mu_0}{\hat{S} / \sqrt{n}}$$

PASO 3:

N(0,1)

PASO 4:



La región de rechazo está donde señala H_1

Metodología general para hacer un contraste de hipótesis

PASO 1: Especificamos la hipótesis nula y la alternativa.

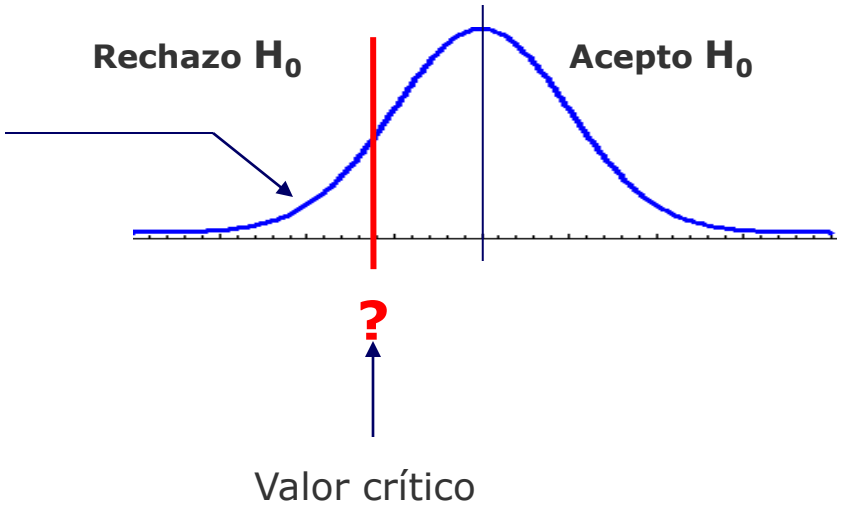
PASO 2: Estadístico de contraste

PASO 3: Distribución de referencia

PASO 4: Localizamos las zonas donde estará la región de rechazo

¿Qué área ocupa la región de rechazo?

- La región de rechazo ocupa un área pequeña
- Ese área se llama α=nivel de significación
- Su valor lo decide el analista
- Suele ser α=0.05, 0.10, 0.01



Ejemplo

En el ejemplo de los transistores. Se desea saber si la población de transistores del proceso productivo mantiene la media en $\mu_0 = 290$

H_0: $\mu = 290$ $\quad$ H_1: $\mu \neq 290$

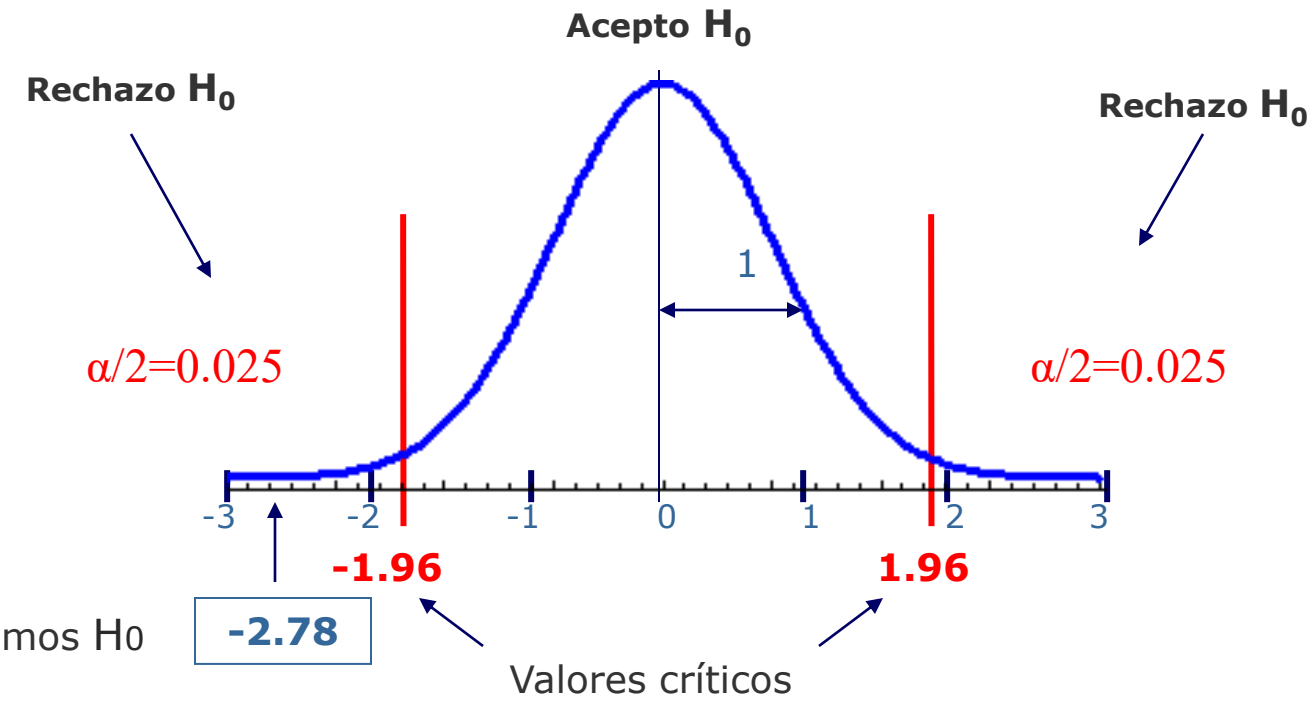
Con 100 observaciones:

$$\bar{x} = 282.3; \quad \hat{s} = 27.57;$$

$$t_0 = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \frac{282.3 - 290}{27.69/10} = -2.78$$

$T_0 \sim N(0,1)$

Nivel de significación, $\alpha = 0.05$



Ejemplo

En el ejemplo de los transistores. Se desea saber si la población de transistores del proceso productivo mantiene la media en $\mu_0 = 290$

H_0 $\quad$ H_1

$$\mu = 290 \qquad \mu \neq 290$$

Con 100 observaciones:

$$\bar{x} = 282.3; \quad \hat{s} = 27.57;$$

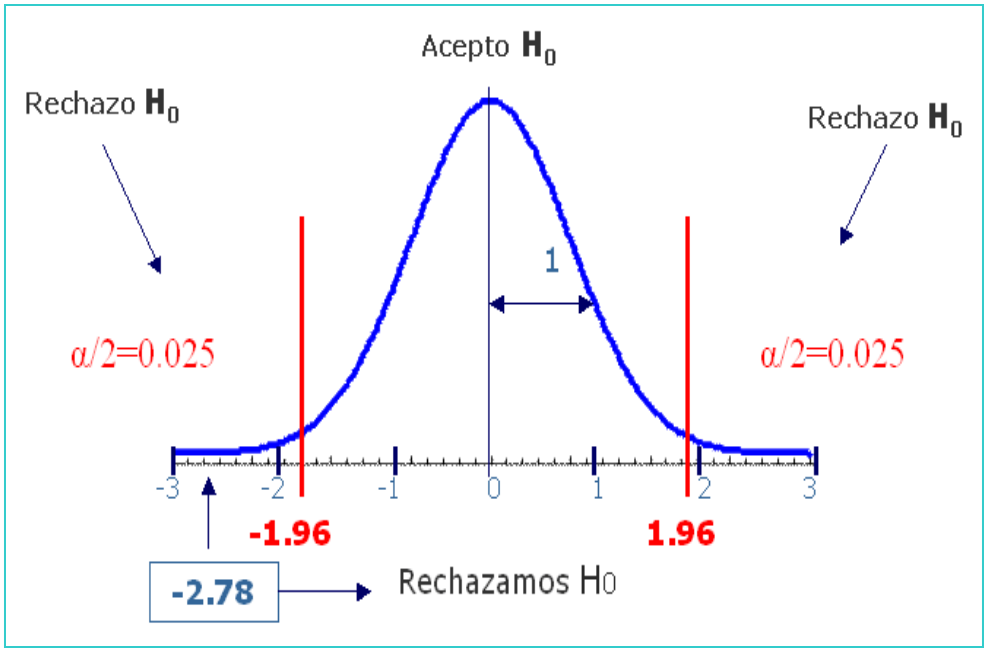
$$t_0 = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \frac{282.3 - 290}{27.69/10} = \underline{-2.78}$$

$T_0 \sim N(0,1)$

Nivel de significación, α=0.05

La diferencia entre la media de la muestra (282.3) y la de la hipótesis (290) **es significativa** (al 5%)

Concluimos, con un nivel de significación del 5%, que la media poblacional ha cambiado



Ejemplo

Según los estudios antropométricos, las personas en España con edades entre 18 y 25 años tienen una estatura media de $\mu_0 = 177$ cm.

Se toman las alturas de 50 personas en Madrid en ese rango de edad y resulta

$$\bar{x} = 175.9cm \qquad \hat{s} = 5.93cm$$

¿Hay evidencia suficiente para decir que en Madrid se tiene una estatura media **inferior** a la nacional?

PASO 1: Especificamos la hipótesis nula y la alternativa.

Dos opciones

- Estatura media inferior → $\mu < 177$
- Estatura media no inferior → $\mu \geq 177$

$$H_0 : \mu \geq 177$$

$$H_1 : \mu < 177$$

Ejemplo

Según los estudios antropométricos, las personas en España con edades entre 18 y 25 años tienen una estatura media de μ_0 =177 cm.

Se toman las alturas de 50 personas en Madrid en ese rango de edad y resulta

$$\bar{x} = 175.9cm \qquad \hat{s} = 5.93cm$$

¿Hay evidencia suficiente para decir que en Madrid se tiene una estatura media **inferior** a la nacional?

$$H_0 : \mu \geq 177$$
$$H_1 : \mu < 177$$

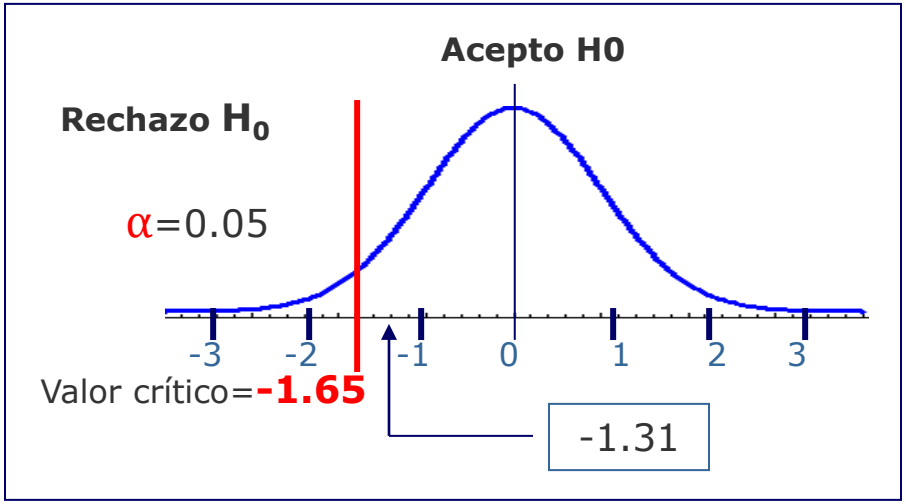
PASO 2: Estadístico de contraste → $t_o = \dfrac{\bar{x} - \mu_o}{\hat{s}/\sqrt{n}} = \dfrac{175.9 - 177}{5.93/\sqrt{50}} = -1.31$

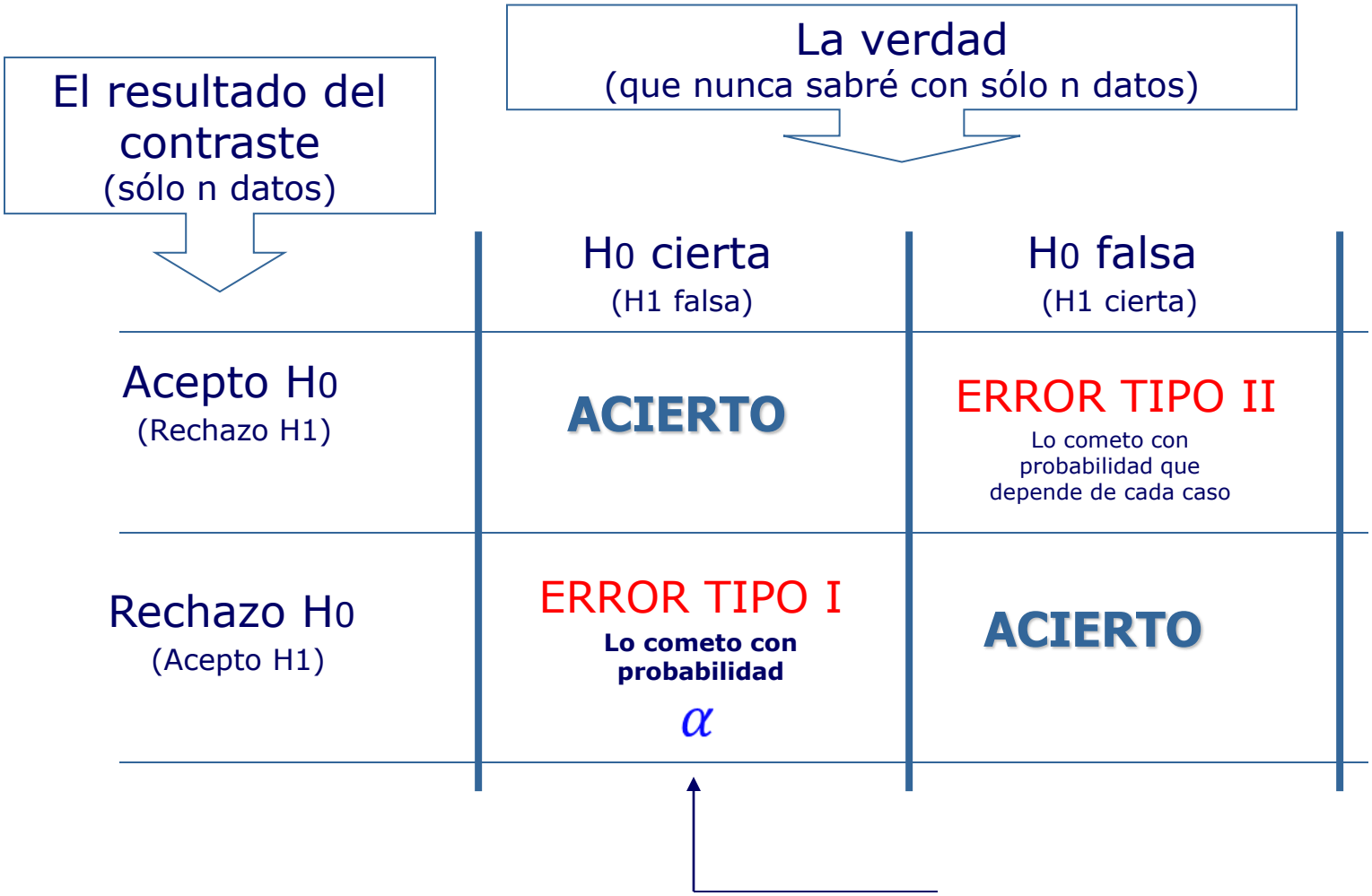
PASO 3: Distribución de referencia → **N(0,1)**

PASO 4: Localizamos las zonas donde estará la región de rechazo

La diferencia entre la media muestral (175.9) y la hipótesis nula **no es significativa** (al 5%)

La diferencia observada se atribuye, con un nivel de significatividad del 5%, a la variabilidad de la muestra y no a diferencias reales





Cuando demos la conclusión de un contraste debemos dar siempre el nivel de significación, para dar una medida de su precisión

Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

Interpretación de un contraste usando el p-valor

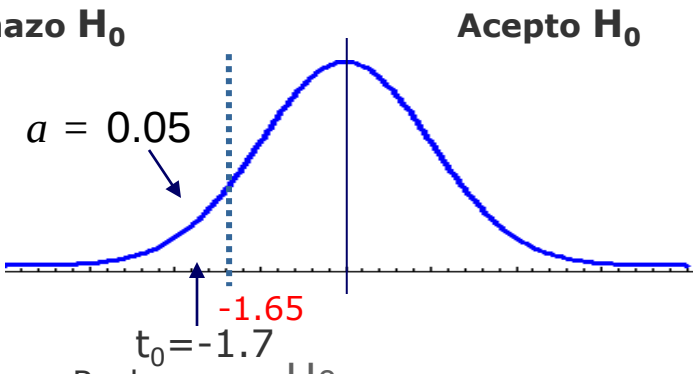
El resultado de un contraste tiene **dos** elementos:

1. Aceptamos o rechazamos H_0 → Conclusión del contraste
2. El nivel de significación α → Medida de su incertidumbre

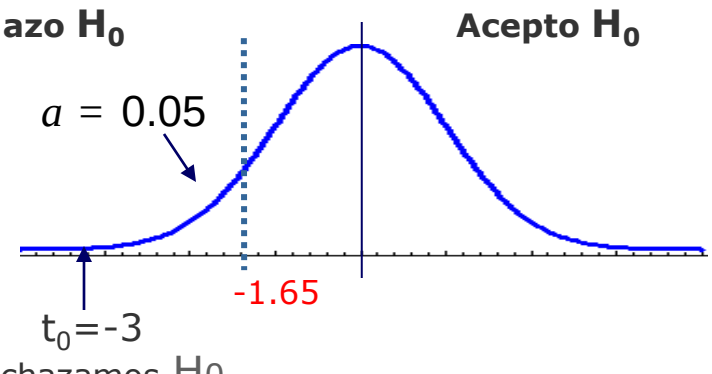
El nivel de significación es una medida de incertidumbre poco precisa

Ejemplo Hacemos el contraste $H_0 : \mu \geq \mu_0; H_1 : \mu < \mu_0$ con $a = 0.05$

Caso 1



Caso 2

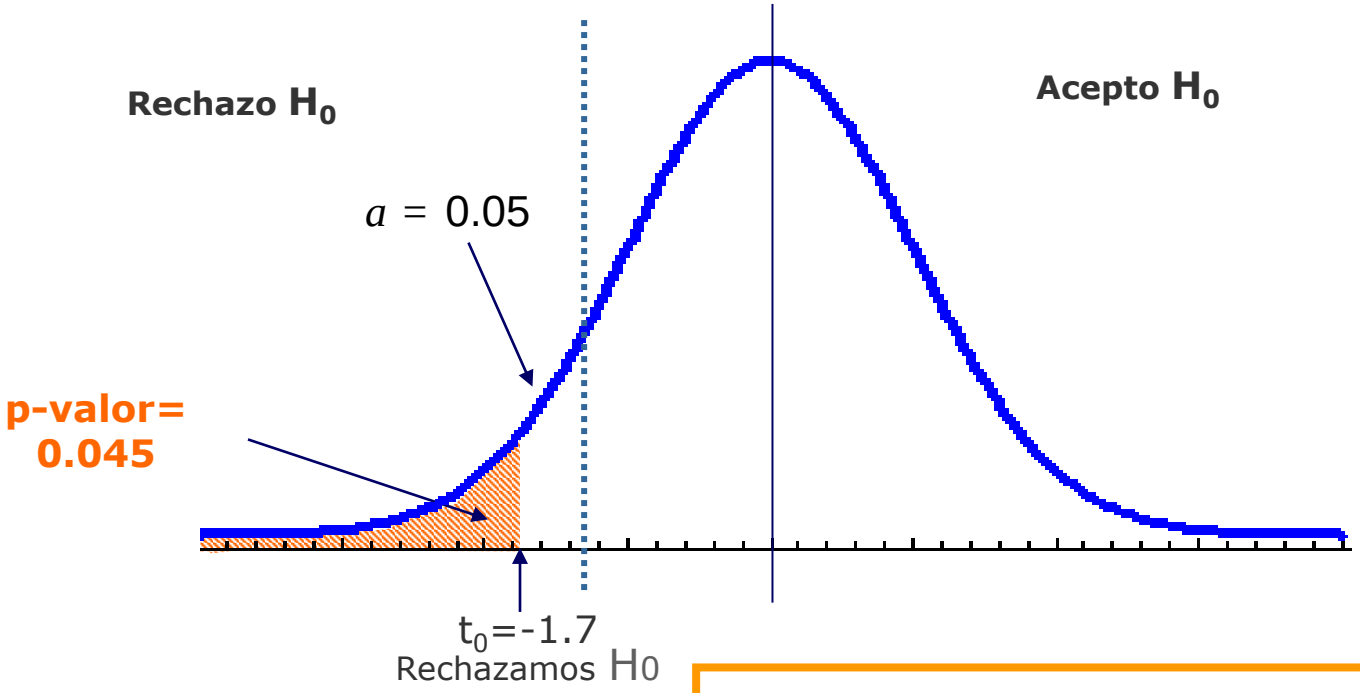


En ambos casos la conclusión sería la misma: Rechazamos con a=0.05

Sin embargo en el caso 2 estamos más seguros **¿Cómo expresarlo?**

El **p-valor** es el nivel de significación que deberíamos usar para dejar al valor del estadístico de contraste **justo en la frontera** de la región de rechazo

Caso 1

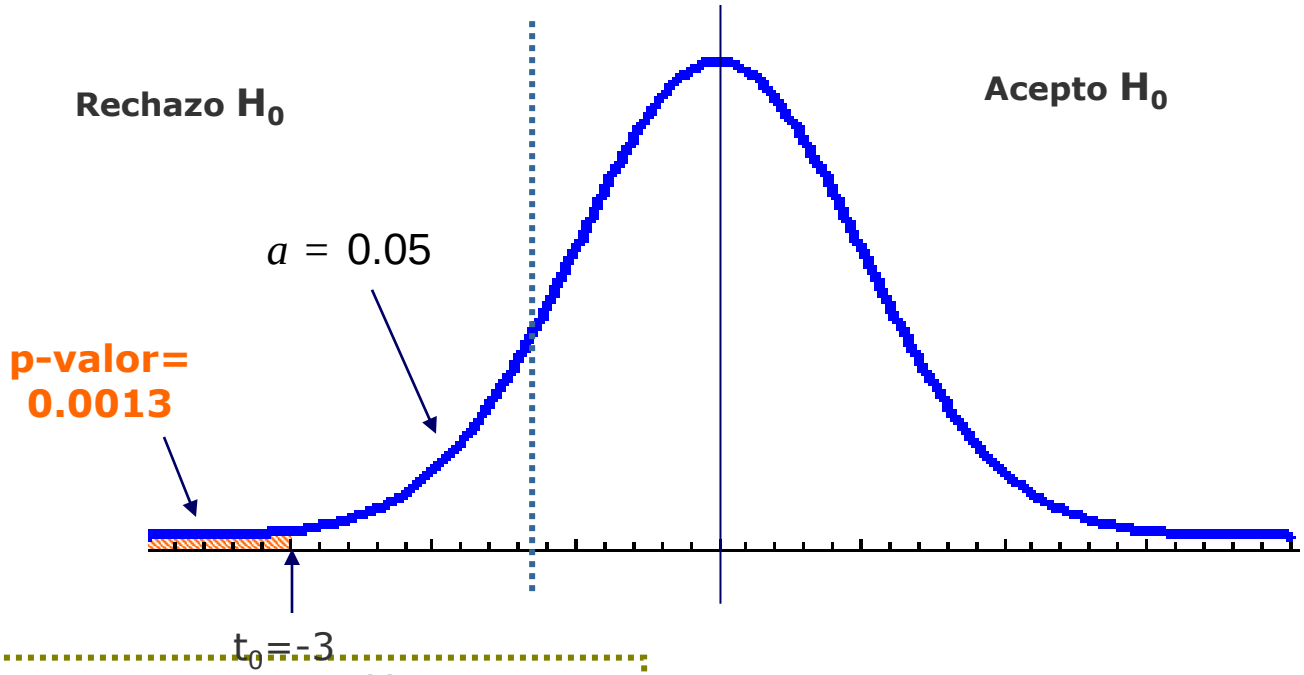


Como p-valor<a → Rechazamos H0

El p-valor es más informativo que el nivel de significación

El **p-valor** es el nivel de significación que deberíamos usar para dejar al valor del estadístico de contraste **justo en la frontera** de la región de rechazo

Caso 2

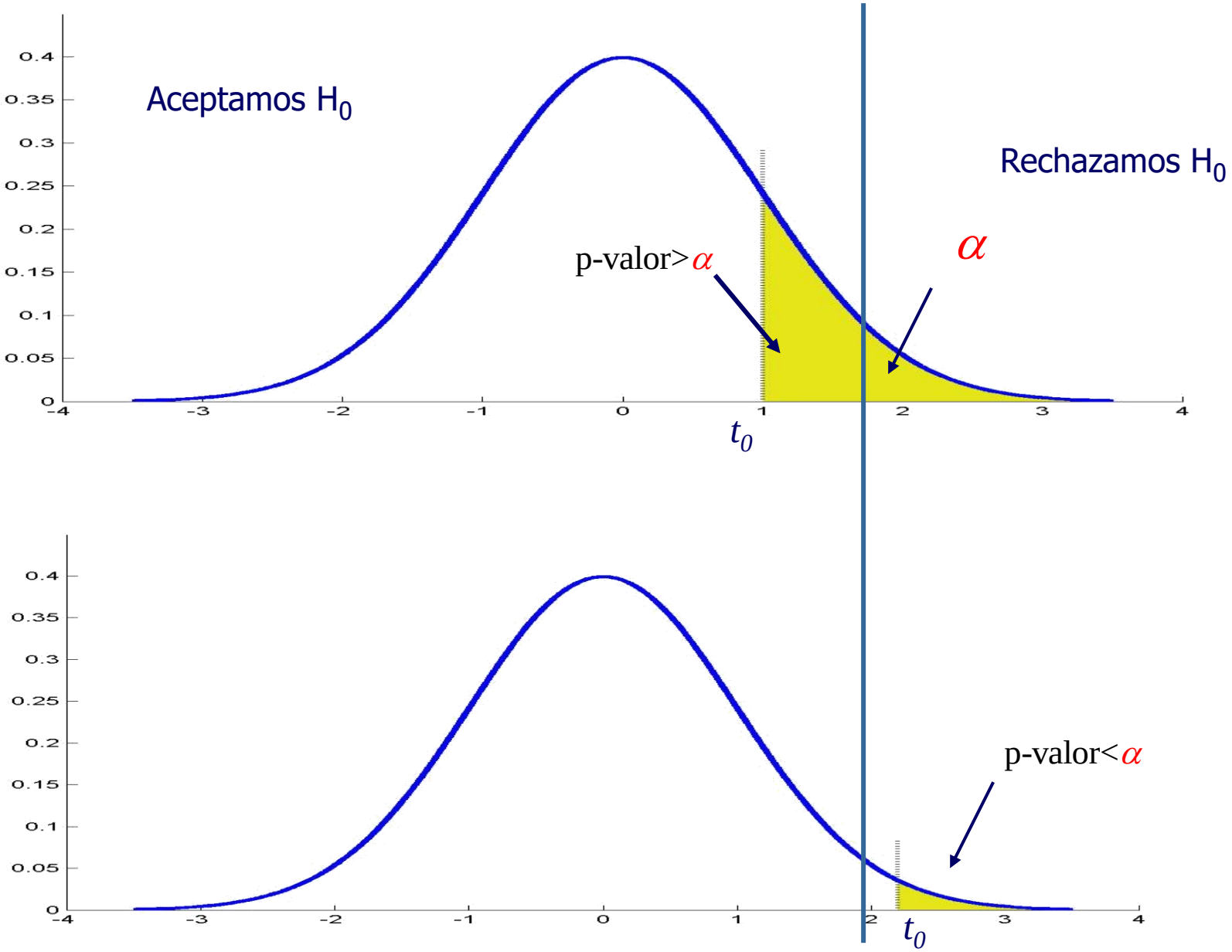


Rechazamos H0

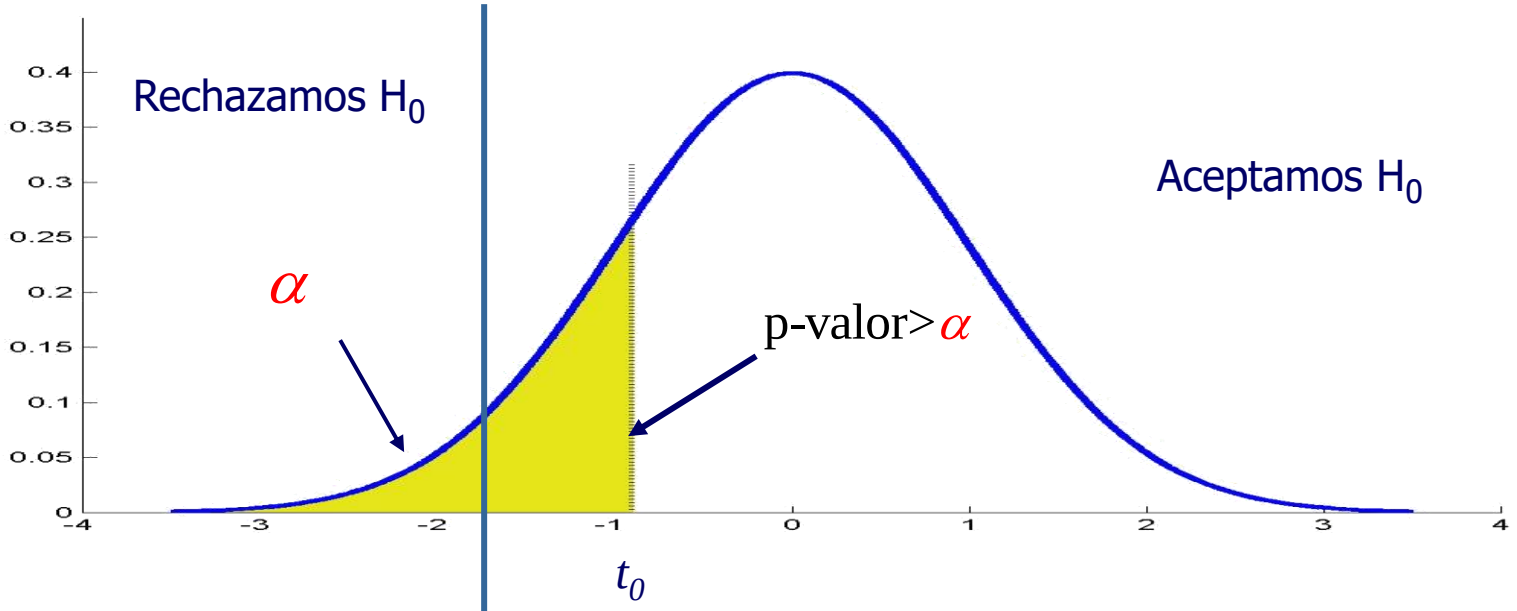
Como p-valor<<α → Rechazamos H0

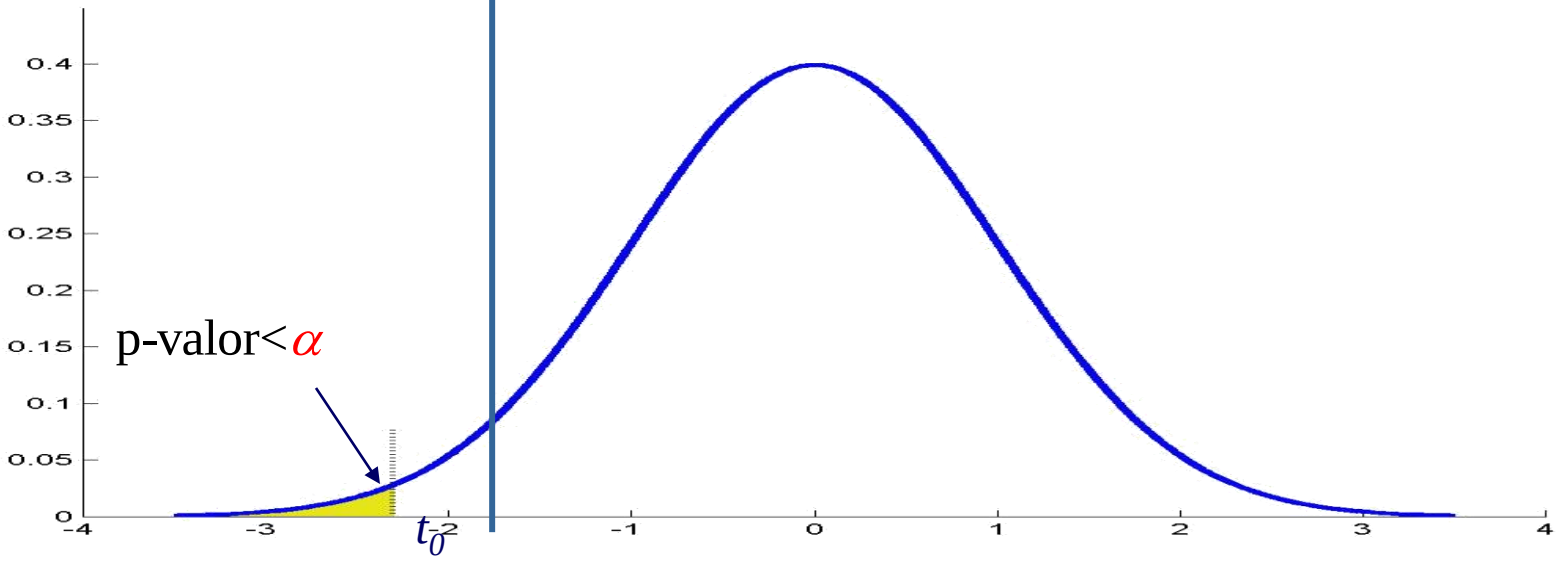
En este caso el p-valor es realmente pequeño. Estamos más seguros de nuestra conclusión

$$H_0 : \vartheta \leq \vartheta_0 ; H_1 : \vartheta > \vartheta_0$$

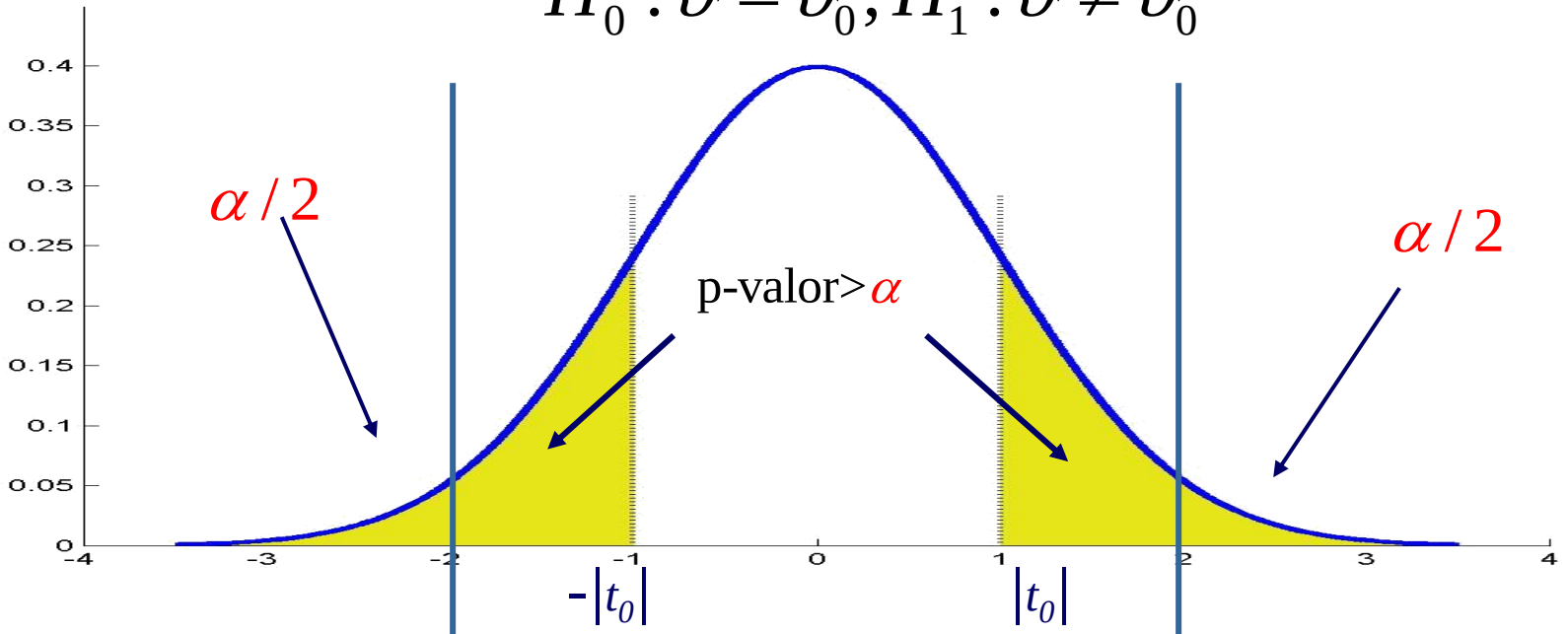


$$H_0 : \vartheta \geq \vartheta_0 ; H_1 : \vartheta < \vartheta_0$$

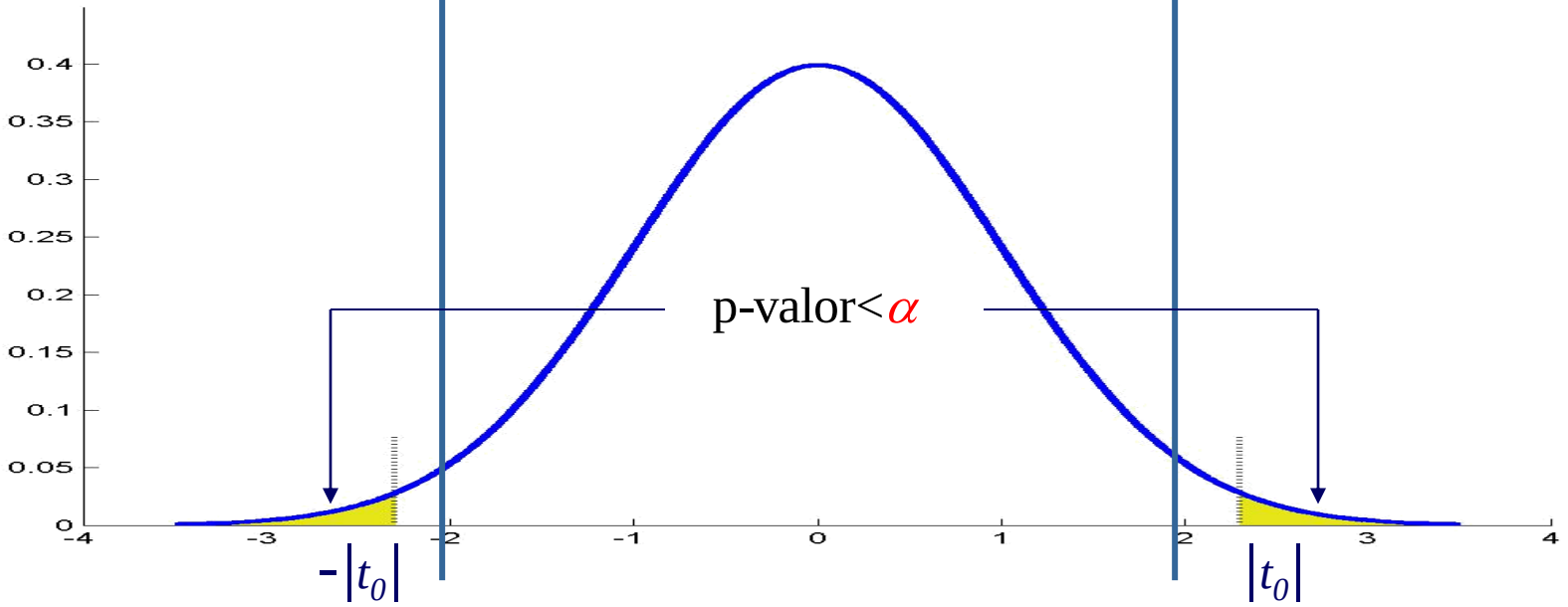




$$H_0 : \vartheta = \vartheta_0 ; H_1 : \vartheta \neq \vartheta_0$$



p-valor: es la suma de las dos áreas



Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

Diagnosis y crítica del modelo

Supongamos que el histograma de un conjunto de datos sugiere que éstos pueden seguir cierto modelo de probabilidad. ¿Cómo puedo comparar los datos con lo que predice el modelo?

Test de la chi-cuadrado

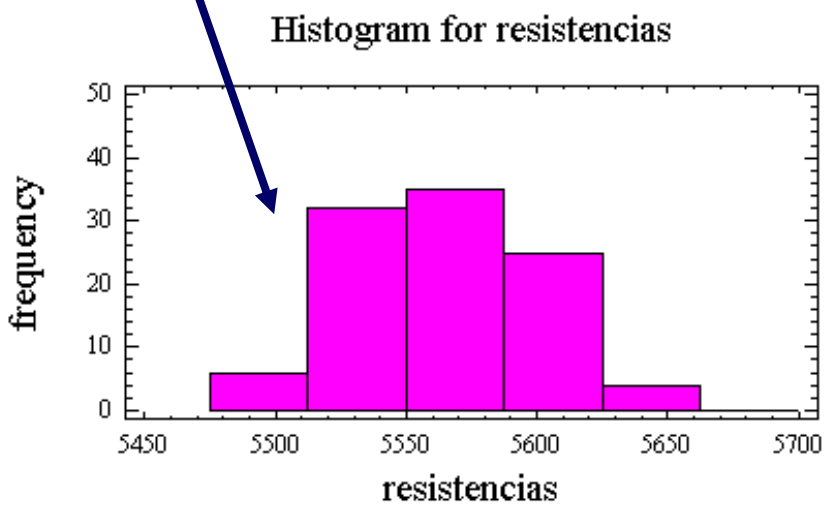
- Es un método para valorar la bondad del ajuste de un modelo.
- Es un método tanto para modelos continuos como discretos

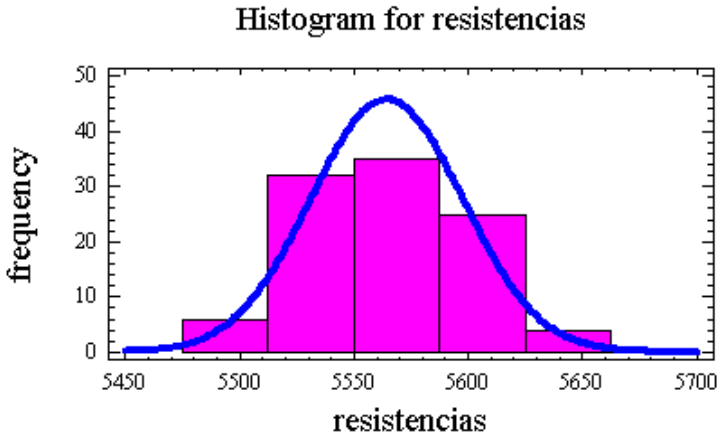
Estos datos sugieren una población normal

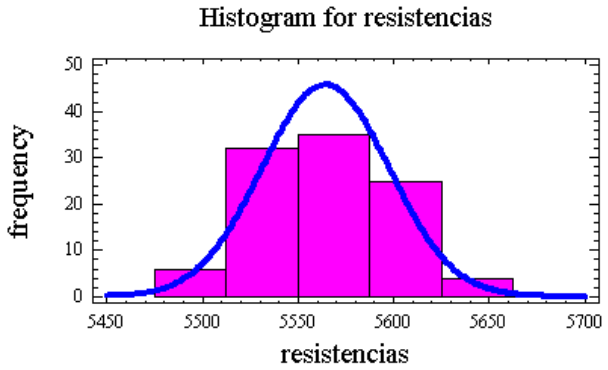
Con los datos se estiman μ y σ^2

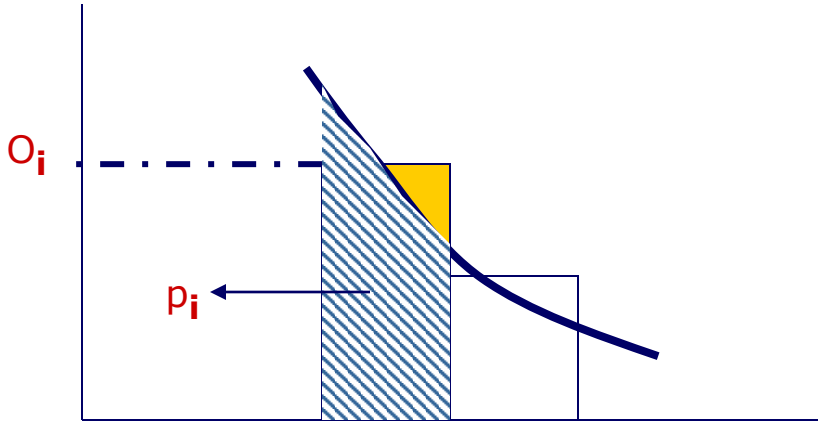
$$¿ N\left(\hat{\mu}, \hat{\sigma}^2\right) ?$$

Comparo el histograma con la curva de la normal estimada









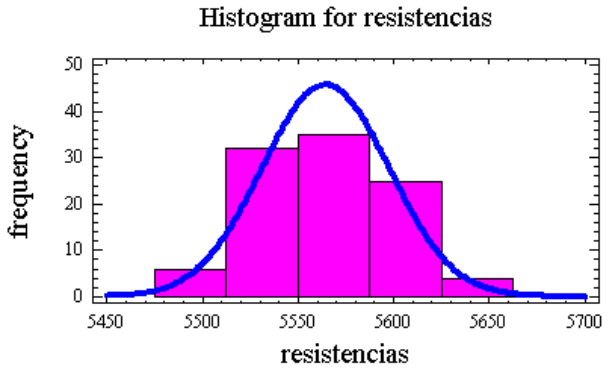
Para cada clase del histograma:

- Contamos el número de individuos observados en dicha clase: O_i
- Calculamos, según el modelo, la probabilidad de estar en esa clase: p_i
- Calculamos, el número esperado de individuos: $E_i = np_i$

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Si X_0^2 **es** muy grande: mucha discrepancia entre los datos y lo que dice el modelo.

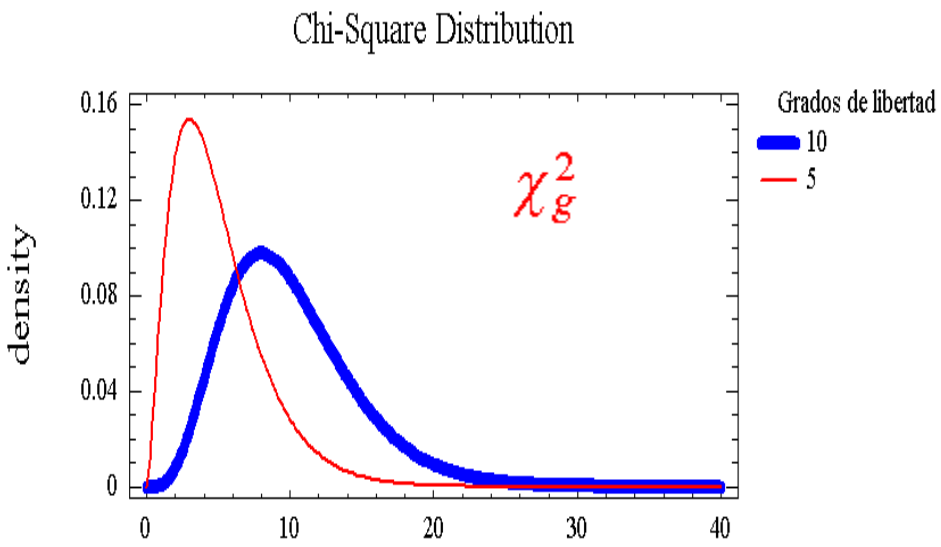
Rechazamos dicho modelo



$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Se demuestra que:

X_0^2 *sigue una distribución que se denomina chi-cuadrado*

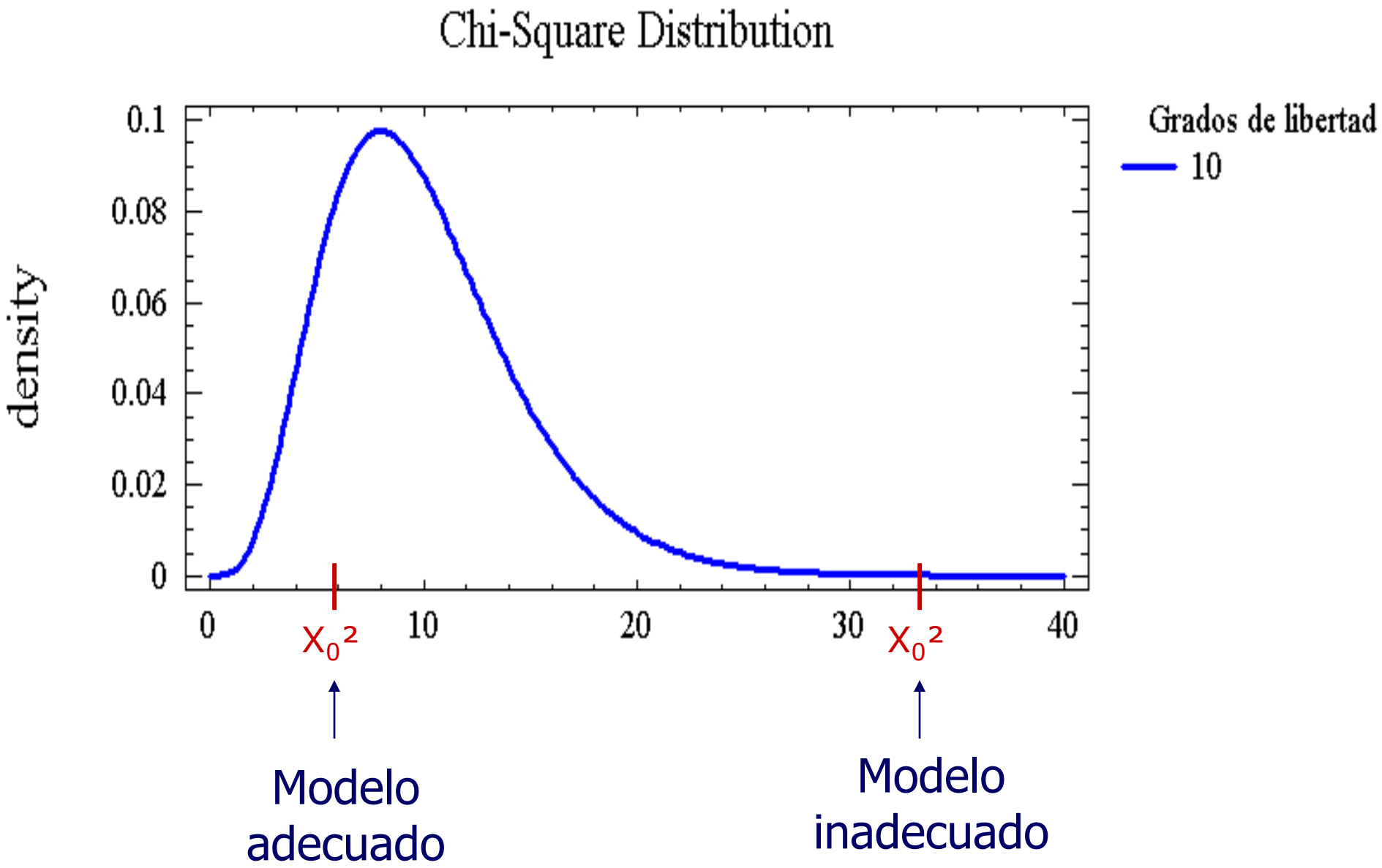


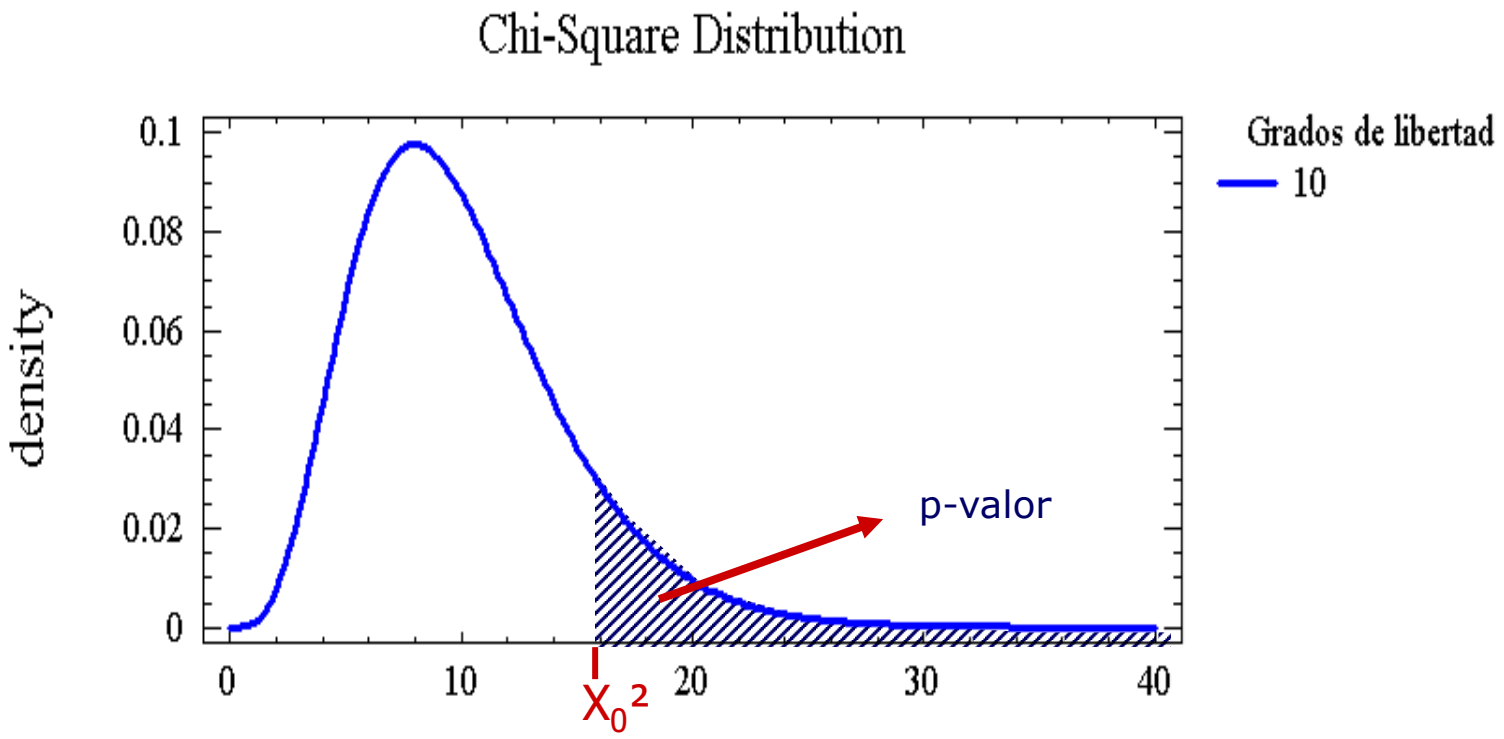
g= grados de libertad

$$g = k - v - 1$$

número de clases

parámetros estimados


Chi-Square Distribution
density
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
X_0^2
X_0^2
Grados de libertad
10
Modelo adecuado
Modelo inadecuado



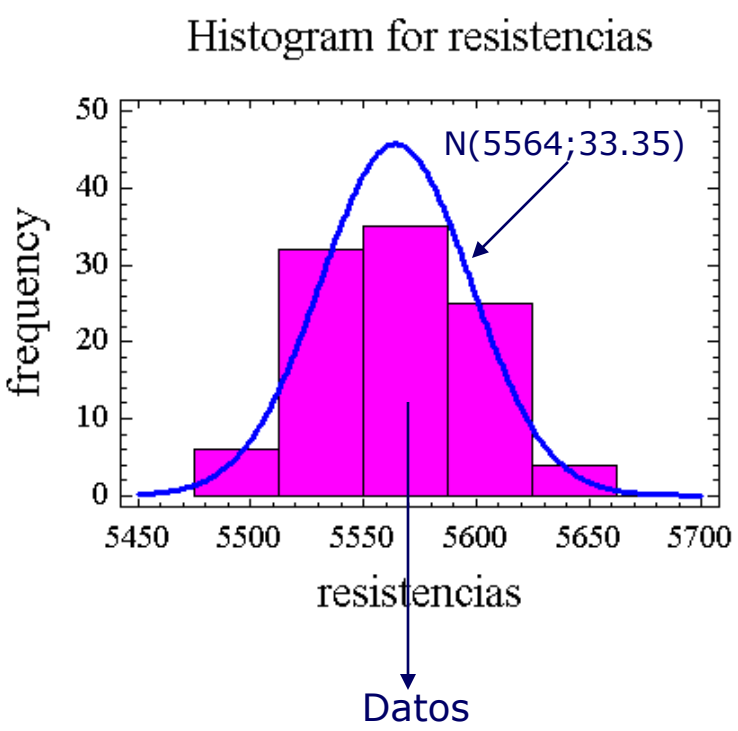
- Las funciones de R calculan el área a la derecha de X_0^2
- Ese área se llama p-valor (p-value)

En la práctica, si p-valor<0.05 descartamos el modelo

Ejemplo: Datos de resistividad de 102 resistencias similares

¿Ajuste a una normal?

Con los 102 datos: media muestral=5564. Desviación típica muestral=33.35



Goodness-of-Fit Tests for resistencias

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	5516.79	7	7.85	0.09
5516.79	5530.33	10	7.85	0.59
5530.33	5539.79	8	7.85	0.00
5539.79	5547.6	9	7.85	0.17
5547.6	5554.57	8	7.85	0.00
5554.57	5561.13	6	7.85	0.43
5561.13	5567.57	10	7.85	0.59
5567.57	5574.14	7	7.85	0.09
5574.14	5581.11	3	7.85	2.99
5581.11	5588.91	5	7.85	1.03
5588.91	5598.38	13	7.85	3.39
5598.38	5611.92	6	7.85	0.43
above 5611.92		10	7.85	0.59

Chi-Square = 10.4114 with 10 d.f. P-Value = 0.40517

O_i E_i $\frac{(O_i - E_i)^2}{E_i}$

X_0^2

p-valor

El p-valor es suficientemente grande. Podemos usar la normal para representar a la población de la que proceden estos datos

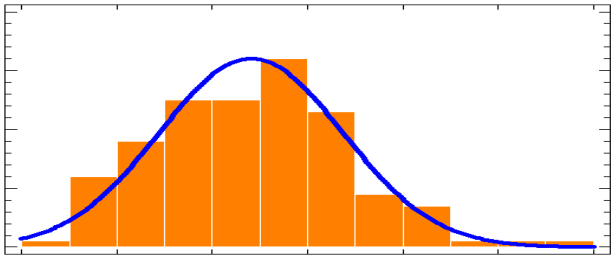
→ Resistividad ~ N(5564;33.35)

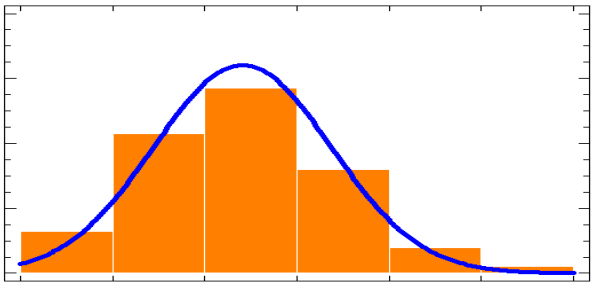
Tema 6: Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Estimación y estimadores
3. Intervalos de confianza para la media con muestras grandes
4. Determinación del tamaño muestral
5. Otros intervalos de confianza
6. Introducción al contraste de hipótesis
7. Contraste de hipótesis sobre la media con muestras grandes
8. Interpretación de un contraste usando el p-valor
9. Diagnosis y crítica del modelo
10. Transformaciones para aproximar a la normal

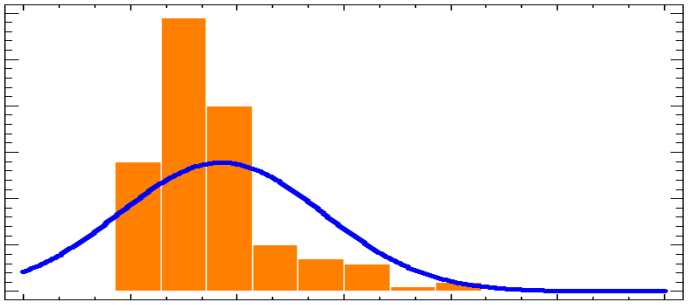
Transformaciones para aproximar a la normal

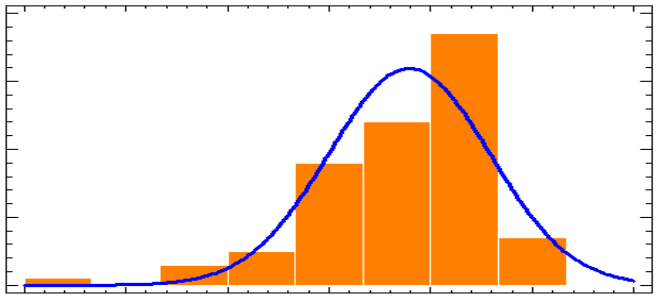
Muchos datos unimodales simétricos se ajustan a una distribución normal





Sin embargo es muy frecuente encontrar datos unimodales asimétricos





Transformaremos los datos de forma que los datos transformados sean unimodales y simétricos

Intentamos ajustar una normal a los datos transformados

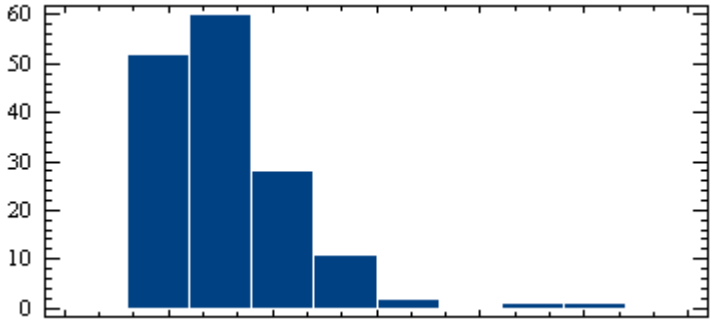
Transformaciones para aproximar a la normal

Dados un conjunto de datos x1, x2, x3, ..., xn con distribución UNIMODAL asimétrica

→

Buscamos una transformación y=h(x) tal que y sea más simétrica

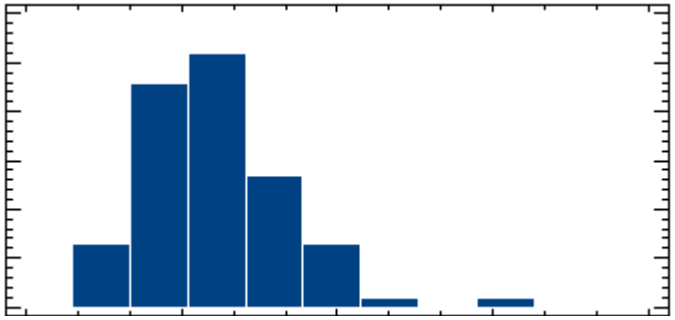
x



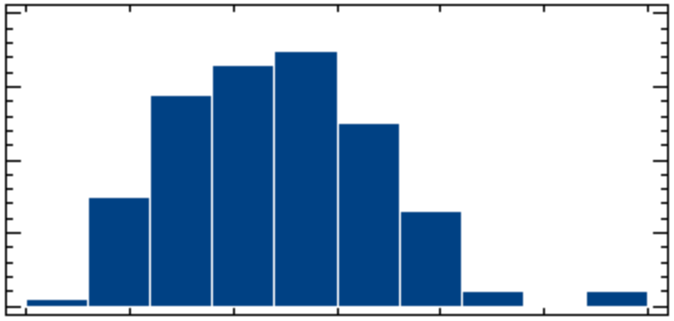
<u>Asimetrías positivas</u> (muy frecuentes)

- Transformaciones del tipo $y=x^c$,$c<1$
- $y=\ln(x)$
- Estas transformaciones comprimen mucho a los datos grandes y poco a los pequeños

$\sqrt{x}$



$\ln(x)$



- ln(x) puede interpretarse como el límite de la transformación $y=x^c/c$ cuando $c\to 0$
- Cuanto mayor sea la asimetría, necesitamos un c menor

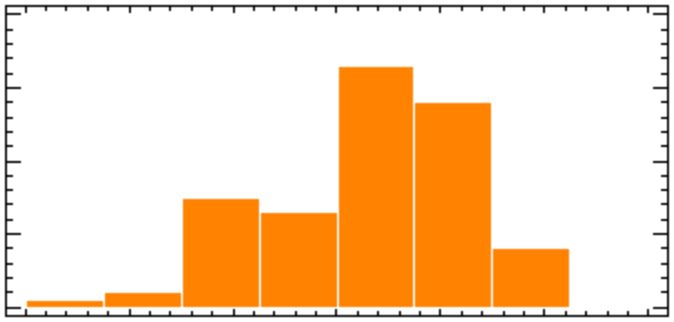
Transformaciones no lineales que mejoran la simetría

Dados un conjunto de datos x1, x2, x3, ..., xn con distribución UNIMODAL asimétrica

→

Buscamos una transformación $y=h(x)$ tal que y sea más simétrica

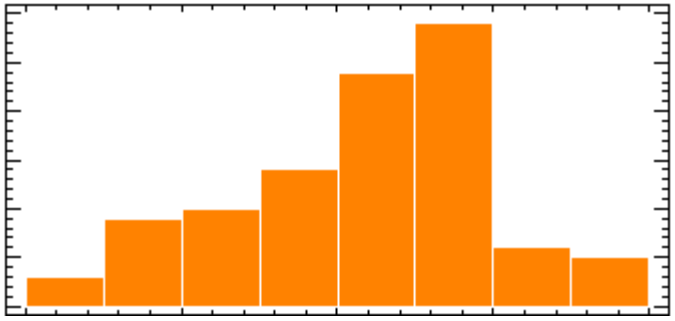
x



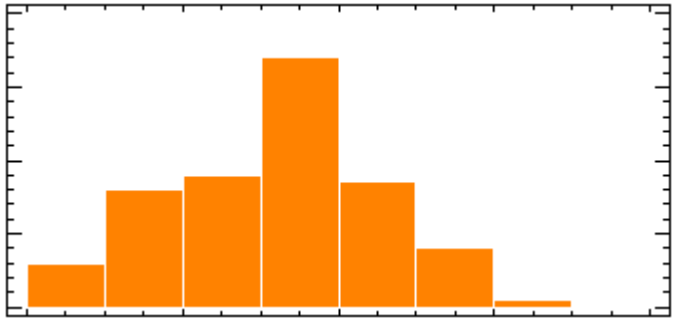
<u>Asimetrías negativas</u>

- Transformaciones del tipo $y=x^c, c>1$
- Estas transformaciones expanden mucho a los datos grandes y poco a los pequeños

$x^{1,5}$



x^2

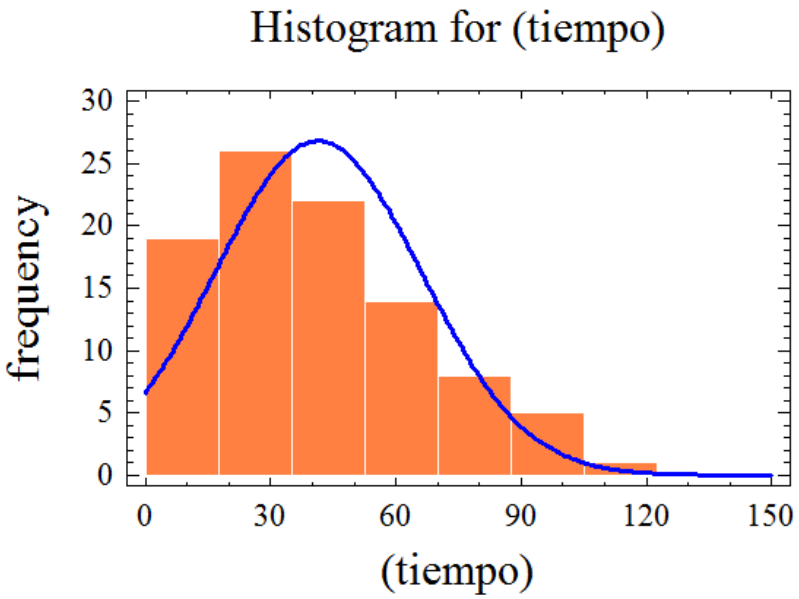


Cuanto mayor sea la asimetría, necesitamos un c mayor

Ejemplo

El fichero AlumnosIndustriales contiene una muestra de estudiantes de Ingeniería Industrial. La variable Tiempo tiene el tiempo (minutos) que tardan los alumnos de esta muestra en llegar a la universidad. ¿Cuál es la probabilidad de que un alumno tarde más de 60 minutos en llegar?

Nota: No queremos saber qué proporción de los alumnos encuestados tarda más de 60 minutos, sino generalizar a toda la población de estos estudiantes, es decir, hacer inferencia a toda la población



Goodness-of-Fit Tests for (tiempo)

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
	at or below	7,20844	4	7,92	1,94
	7,20844	17,4889	15	7,92	6,34
	17,4889	24,7355	8	7,92	0,00
	24,7355	30,7657	13	7,92	3,26
	30,7657	36,2154	5	7,92	1,07
	36,2154	41,4211	6	7,92	0,46
	41,4211	46,6267	11	7,92	1,20
	46,6267	52,0764	5	7,92	1,07
	52,0764	58,1066	2	7,92	4,42
	58,1066	65,3532	11	7,92	1,20
	65,3532	75,6337	8	7,92	0,00
above	75,6337		7	7,92	0,11

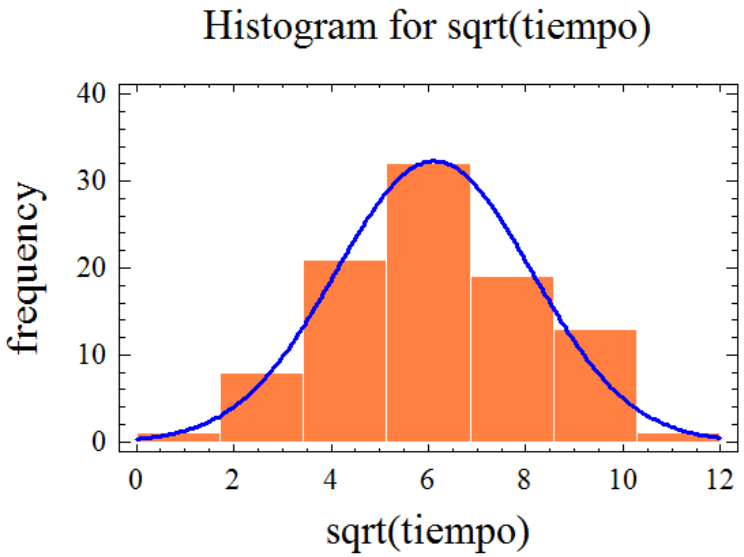
Chi-Square = 21,0842 with 9 d.f. P-Value = 0,0122818

La asimetría positiva es un problema para que se pueda ajustar a una normal.

Podemos 'probar suerte' con una transformación del tipo x^c, con $c<1$

Ejemplo

El fichero AlumnosIndustriales contiene una muestra de estudiantes de Ingeniería Industrial. La variable Tiempo tiene el tiempo (minutos) que tardan los alumnos de esta muestra en llegar a la universidad. ¿Cuál es la probabilidad de que un alumno tarde más de 60 minutos en llegar?



Analysis Summary

Data variable: sqrt(tiempo)

95 values ranging from 1,0 to 10,9545

Fitted normal distribution:
mean = 6,11693
standard deviation = 2,01167

Goodness-of-Fit Tests for sqrt(tiempo)

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below		3,3348	9	7,92	0,15
	3,3348	4,1708	10	7,92	0,55
	4,1708	4,76008	8	7,92	0,00
	4,76008	5,25045	3	7,92	3,05
	5,25045	5,69362	10	7,92	0,55
	5,69362	6,11693	5	7,92	1,07
	6,11693	6,54025	6	7,92	0,46
	6,54025	6,98341	11	7,92	1,20
	6,98341	7,47378	7	7,92	0,11
	7,47378	8,06307	11	7,92	1,20
	8,06307	8,89906	8	7,92	0,00
above	8,89906		7	7,92	0,11

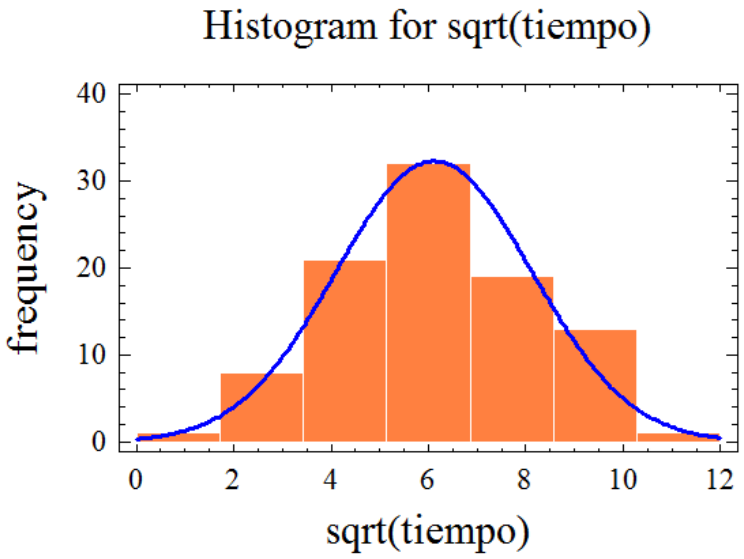
Chi-Square = 8,45304 with 9 d.f. P-Value = 0,489212

La transformación $x^{0.5}$ ha funcionado. El modelo que se ajusta a la normal es

$$X^{0.5} \sim N(6.12; 2.01^2)$$

Ejemplo

El fichero AlumnosIndustriales contiene una muestra de estudiantes de Ingeniería Industrial. La variable Tiempo tiene el tiempo (minutos) que tardan los alumnos de esta muestra en llegar a la universidad. ¿Cuál es la probabilidad de que un alumno tarde más de 60 minutos en llegar?



$$X^{0.5} \sim N(6.12; 2.01^2)$$

$$P(X > 60) = P(X^{0.5} > 60^{0.5})$$

$$= P(X^{0.5} > 7.746) = 0.21$$

Estimamos que el 21% de los alumnos de esa titulación tarda más de una hora en llegar a la Universidad.