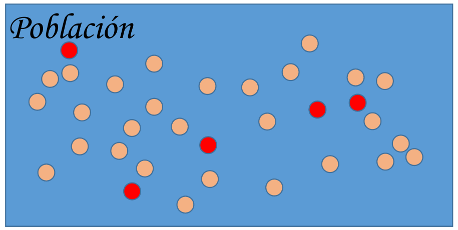
Tema 6
Introducción a la inferencia estadística

Carlos Montes – uc3m

1. Introducción

Proceso de inducción por el cual
a partir de una muestra
intentamos predecir
cómo será el resto de la población
que no se ha observado
(variable aleatoria).

1. Introducción



$X_1, X_2, \ldots X_n$

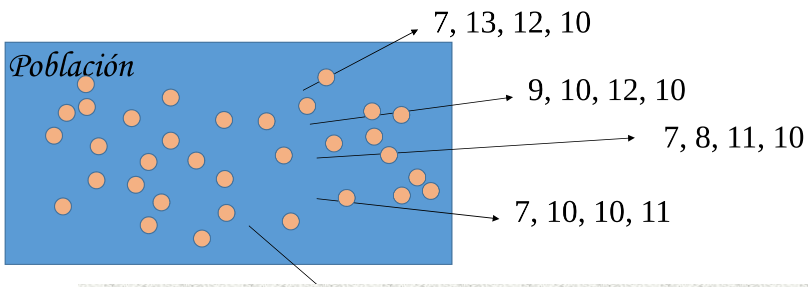
variables aleatorias independientes e idénticas

Muestra aleatoria simple

- Elementos de la muestra independientes entre sí.
- Elementos con las mismas características que la población.

Es una variable aleatoria ⇒ Cada muestra será diferente.

2.1. Distribución muestral de estimadores. Concepto



Cualquier estimador podrá tomar diferentes valores,
cada uno con diferente probabilidad dependiendo de la muestra.

50, 300, 1250, 689 ?

2.1. Distribución muestral de estimadores. Concepto

Distribución
del estadístico en el muestreo,
o distribución muestral del estadístico.

Función de los valores muestrales

Dependerá de:

-Distribución de la población base
-Tamaño de la muestra (n)

2.2. Distribución muestral de la media

Sea una variable aleatoria cualquiera,
de media μ y desviación típica σ:

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

$$=\frac{n\mu}{n}=\mu$$

2.2. Distribución muestral de la media

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

$$var\left(\bar{X}\right)=\frac{var(X_1)+var(X_2)+\ldots+var(X_n)}{n^2}=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

Al tomar una muestra de tamaño n
con media μ**,**varianza σ^2
y distribución cualquiera,
la distribución muestral de la media verifica:

$$E(\bar{X}) = \mu$$
$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

Cuando n es grande (n>30), la distribución de la media es asintóticamente normal, por el Teorema Central del Límite.

$$E(\bar{X}) = \mu$$
$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

Pero para cualquier *n*, **si *x* es N(μ,σ)**

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Muy importante para inferencia

Los parámetros (μ, σ, λ…) son valores numéricos de la población (constantes de valor desconocido)

Estimador es un estadístico
que nos da con cierta exactitud
el valor de los caracteres de la población
que se pretenden inferir.

$$\hat{\theta}$$

3.1. Estimación. Concepto

Para denotar un estimador usamos la misma letra que el parámetro, con el acento circunflejo (^) sobre él.

$$\hat{\mu}$$

3.2. Estimación. Propiedades

En general serán preferibles aquellos estimadores que verifiquen que:

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

(estimadores *insesgados* o *centrados*)

3.2. Estimación. Propiedades

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

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

La desviación típica de un estimador suele denominarse *error estándar* del estimador

$$var(\hat{\mu}) = \frac{\sigma^2}{n} \qquad e(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

3.2. Estimación. Propiedades

Error cuadrático medio del estimador (ECM), desviación cuadrática media o acuracidad.

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

Puede demostrarse que:

$$ECM(\hat{\theta}) = \left[sesgo(\hat{\theta})\right]^2 + var(\hat{\theta})$$

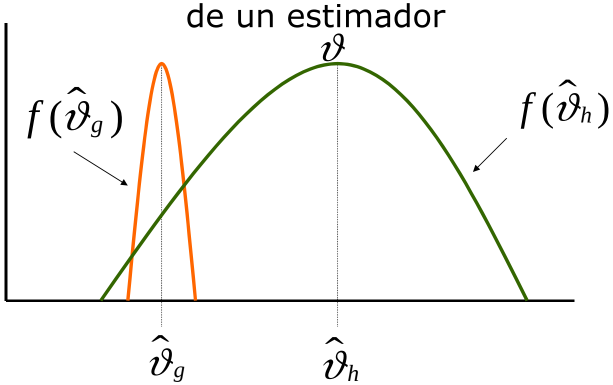
Así, un ECM mínimo supone:

sesgo mínimo → *insesgo*

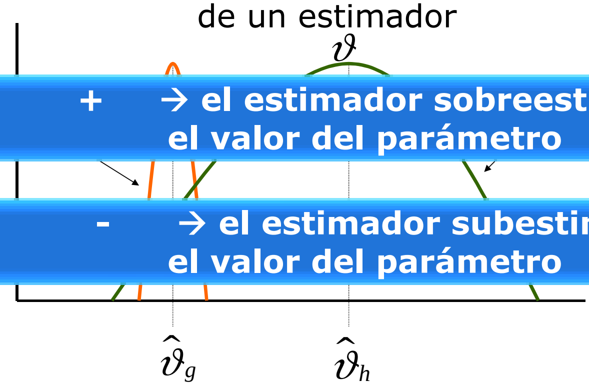
varianza mínima → *eficiencia o precisión*

$$efic(\hat{\theta}) = \frac{1}{var(\hat{\theta})}$$

Nos mide el error sistemático de un estimador



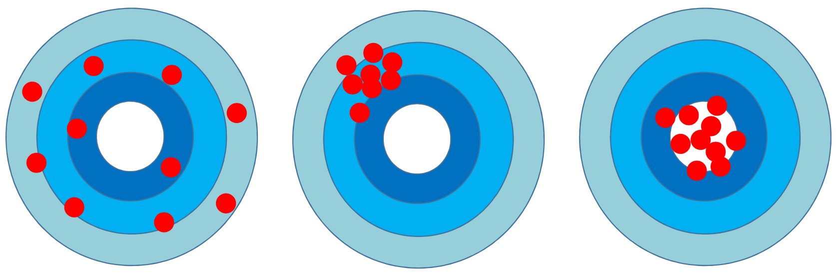
Nos mide el error sistemático de un estimador



+ → el estimador sobreestima el valor del parámetro

- → el estimador subestima el valor del parámetro

3.2. Estimación. Propiedades



insesgado impreciso | sesgado preciso | insesgado preciso

Carlos Montes – uc3m

En muestras aleatorias simples de tamaño n=3 de una variable aleatoria normal de media μ y varianza conocida $\sigma^2=1$, se consideran los siguientes estimadores de μ:

$$\widehat{\mu_1} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$$
$$\widehat{\mu_2} = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3$$
$$\widehat{\mu_3} = \frac{1}{8}X_1 + \frac{3}{8}X_2 + \frac{1}{2}X_3$$

Donde X_1, X_2 y X_3 son observaciones. Comprobar que son estimadores insesgados y estudiar su error cuadrático medio.

$$E(\hat{\mu}_1) = \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(X_3) = \frac{1}{3}\cdot 3\mu = \mu$$

$$E(\hat{\mu}_2) = \frac{1}{4}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{4}E(X_3) = \frac{\mu + 2\mu + \mu}{4} = \mu$$

$$E(\hat{\mu}_3) = \frac{1}{8}E(X_1) + \frac{3}{8}E(X_2) + \frac{1}{2}E(X_3) = \frac{\mu + 3\mu + 4\mu}{8} = \mu$$

Efectivamente, son insesgados.

Estudiamos su error cuadrático medio.

$$ECM(\hat{\theta}) = \cancel{\left[sesgo(\hat{\theta})\right]^2} + var\,(\hat{\theta})$$

$$var(\hat{\mu}_1) = var\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) = \frac{1}{9}var(X_1) + \frac{1}{9}var(X_2) + \frac{1}{9}var(X_3) = \frac{3}{9}\cdot\sigma^2 = \frac{1}{3}$$

$$var(\hat{\mu}_2) = var\left(\frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3\right) = \frac{1}{16}var(X_1) + \frac{1}{4}var(X_2) + \frac{1}{16}var(X_3) =$$
$$= \frac{\sigma^2 + 4\sigma^2 + \sigma^2}{16} = \frac{6}{16}\sigma^2 = \frac{3}{8}$$

$$var(\hat{\mu}_3) = var\left(\frac{1}{8}X_1 + \frac{3}{8}X_2 + \frac{1}{2}X_3\right) = \frac{1}{64}var(X_1) + \frac{9}{64}var(X_2) + \frac{1}{4}var(X_3) =$$
$$\frac{\sigma^2 + 9\sigma^2 + 16\sigma^2}{64} = \frac{26}{64}\sigma^2 = \frac{13}{32}$$

4. Método de los momentos

* Método sencillo de construcción de estimadores.
* Consiste en estimar una característica poblacional con la respectiva característica muestral.

Media poblacional = media muestral
Varianza poblacional = varianza muestral
...

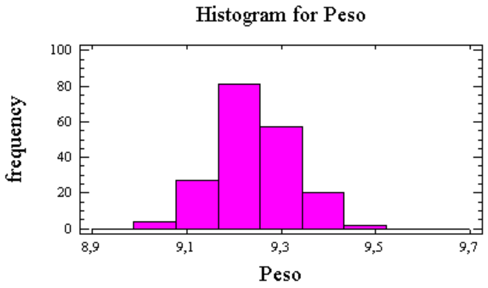
La duración de un sistema hasta que se produce un fallo por causas fortuitas se puede modelizar con una distribución exp(λ). Durante un tiempo se anota el tiempo que ha estado el sistema funcionando hasta que se produjo un fallo. Se obtienen así los siguientes valores de duraciones en horas: 18, 94, 22 143, 114. Estime el parámetro λ de la exponencial utilizando el método de los momentos.

$$E(X) = \frac{1}{\lambda}$$

$$\bar{x} = \frac{18 + 94 + 22 + 143 + 114}{5} = 78.2\ horas$$

$$78.2 = \frac{1}{\lambda} \qquad \lambda = 0.013\ fallos/hora$$

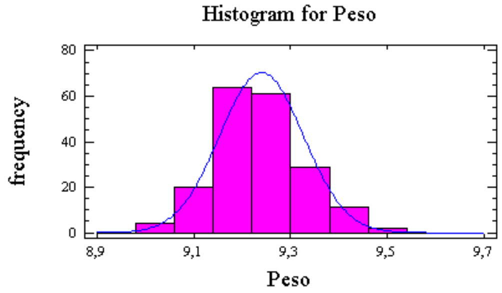
5. Diagnosis y crítica del modelo



Peso de 191 monedas

$$\bar{x} = 9{,}23g$$

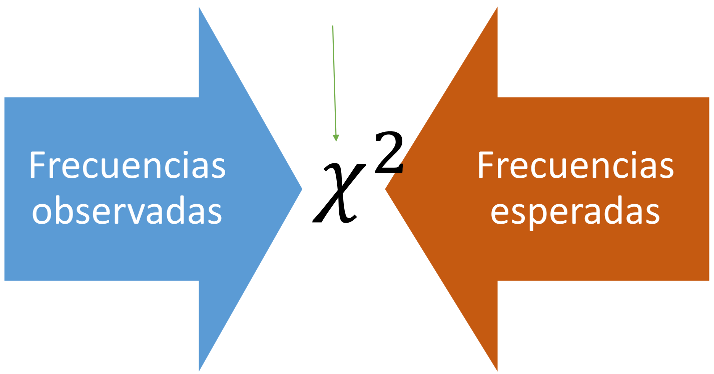
$$s^2 = 0{,}0075g^2$$



Peso de 191 monedas

$$N(9.23,\ 0.0075) \quad ?$$

Test de la chi-cuadrado



Procedimiento

- Se toma una muestra de tamaño ***n≥25***.
- Se agrupan los datos en k clases (***k≥5***) de tamaño **homogéneo**, con al menos ***3 datos en cada clase***.
- Calculamos la discrepancia entre las frecuencias observadas de cada clase O_i y las previstas por el modelo, E_i.

Se calcula el siguiente estadístico:

Frecuencias observadas

$$\chi_0^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

Frecuencias esperadas según el modelo

Resume la discrepancia entre datos y modelo:

- Discrepancia alta: rechazamos el modelo.
- Discrepancia baja: aceptamos el modelo.

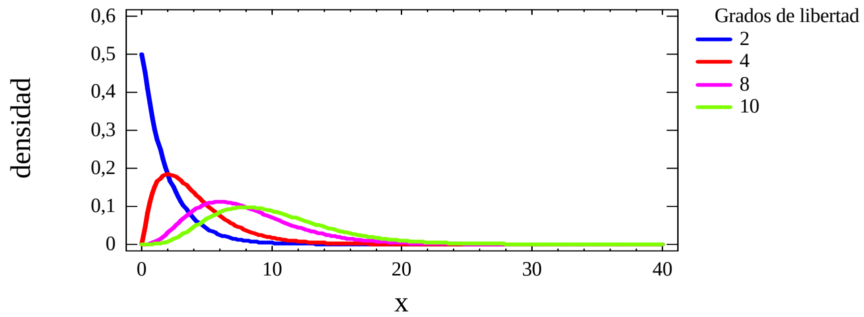
Llamamos discrepancia alta
a la que tiene
muy poca probabilidad de ocurrir
si el modelo es correcto.

¿Cómo valorar el modelo?

El estadístico calculado seguirá una distribución:

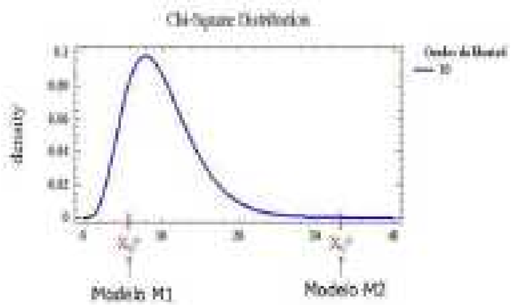
$$\chi^2$$

Depende de sus *g* grados de libertad
(si conocemos los parámetros del modelo, = *k -1*.
Si hay que estimar *v* parámetros, = *k-v-1*)



Para n>30 es prácticamente una normal

5. Diagnosis y crítica del modelo



* La probabilidad de obtener ese valor de χ_0^2 si M2 es cierto, es muy baja.

* M1 es adecuado, M2 no lo es.

5. Diagnosis y crítica del modelo

* Los programas informáticos proporcionan el área que queda a la derecha de χ_0^2 en la distribución ***(p-valor)***.

* En general, si el valor de χ_0^2 está en la zona de la cola de la derecha, el modelo no es adecuado (área bajo la curva pequeña ⇨ probabilidad pequeña de ocurrir si el modelo es cierto).

5. Diagnosis y crítica del modelo

*** Rechazaremos un modelo si el p-valor <0,05**

Peso de monedas

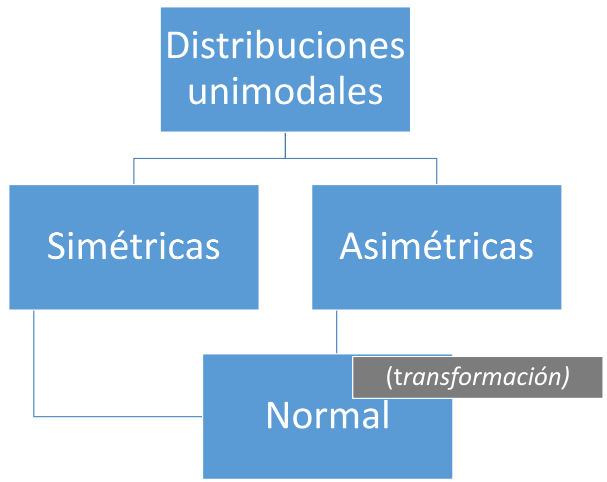
Goodness-of-Fit Tests for Peso

Chi-Square Test

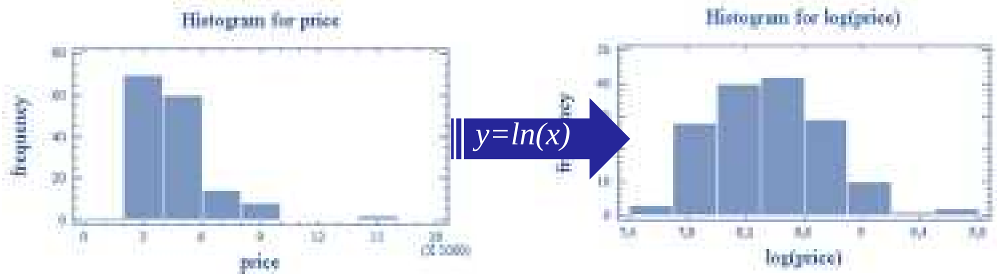
	Lower Limit	*Upper Limit*	*Observed Frequency*	*Expected Frequency*	*Chi-Square*
at or below		9,06	4	3,63	0,04
	9,06	9,14	20	20,27	0,00
	9,14	9,22	64	54,59	1,62
	9,22	9,3	61	66,26	0,42
	9,3	9,38	29	36,28	1,46
above	9,38		13	9,97	0,92

Chi-Square = 4,46353 with 3 d.f.   P-Value = 0,215564

6. Transformaciones para mejorar la normalidad



Datos con asimetría positiva



6. Transformaciones para mejorar la normalidad

+ efecto

$$y = \sqrt{x}$$
$$y = \ln x$$
$$y = \frac{1}{x}$$

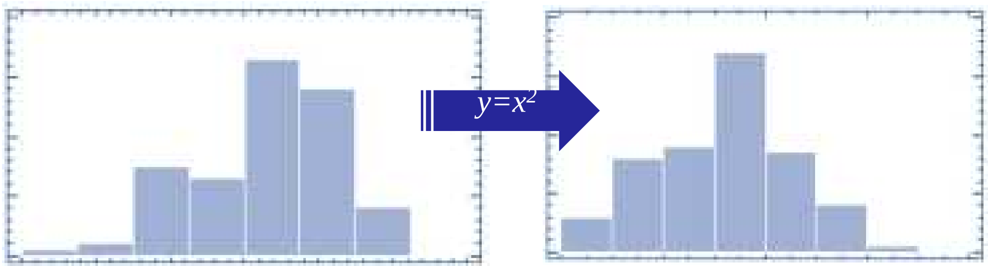
Comprimen la escala
en los valores altos
y la expanden
en los valores bajos.

En general, transformaciones del tipo:

$$y = x^c \quad c < 1$$

6. Transformaciones para mejorar la normalidad

Datos con asimetría negativa



6. Transformaciones para mejorar la normalidad

$$y = x^2$$

Comprime la escala
en los valores bajos
y la expande
en los valores altos.

En general, transformaciones del tipo:

$$y = x^c \quad c > 1$$