

Grado en Ingeniería Informática
2019-2020

Apuntes
Estadística

Jorge Rodríguez Fraile¹



Esta obra se encuentra sujeta a la licencia Creative Commons
Reconocimiento - No Comercial - Sin Obra Derivada

¹Universidad: 100405951@alumnos.uc3m.es | Personal: jrf1616@gmail.com

ÍNDICE GENERAL

I Tema 1. Estadística descriptiva univariable	3
II Tema 2. Estadística descriptiva bivariante	21
III Tema 3. Probabilidad	33
IV Tema 4. Variable aleatoria	45
V Tema 5. Modelos de probabilidad	57
VI Tema 6. Introducción a la inferencia estadística	77
VII Tema 7. Inferencia con muestras grandes	91
VIII Tema 8. Comparación de población	111
IX Tema 9. Regresión múltiple	125
X Recursos	141

Parte I

Tema 1. Estadística descriptiva univariable

Tema 1

Estadística descriptiva univariante

Carlos Montes – uc3m

1. Introducción
2. Análisis básico
 - 2.1. Generalidades
 - 2.2. Gráficos para variables cualitativas
 - 2.3. Variables cuantitativas
 - 2.4. Gráficos para variables cuantitativas
3. Medidas características
 - 3.1. Generalidades
 - 3.2. Medidas de tendencia central
 - 3.3. Medidas de dispersión
 - 3.4. Medidas de forma
4. Diagrama de Caja

1. Introducción

¿Qué es la Estadística?

Es una herramienta de aprendizaje
a partir de la observación.

Nos ayuda a extraer conclusiones
generalizables a partir de un conjunto de
datos observados ⇔ *inducción o inferencia*.

1. Introducción

DATOS (MUESTRA)
realizaciones de una variable

↓
CONCLUSIONES
sobre el fenómeno que los ha originado

<p>1. Introducción</p> <p>* Según su naturaleza, los datos pueden ser:</p> <p><i>Datos cuantitativos.</i> Toman valores numéricos ❖ Discretos: toman valores finitos. ❖ Continuos: toman valores en un intervalo.</p> <p><i>Datos cualitativos, categóricos o atributos.</i> No toman valores numéricos Su realización concreta es una cualidad o modalidad.</p> <p>Carlos Montes – uc3m</p>	<p>1. Introducción</p> <p>La cantidad de información aportada por ambos tipos de variables es muy distinta:</p> <p>- Variables cualitativas sin orden</p> <p>+ Variables cuantitativas orden</p>
<p>1. Introducción</p> <p>OBJETIVO: inferir cómo será la población de la variable de interés a partir de la información limitada que nos aporta la muestra.</p>	<p>2.1. Análisis básico. Generalidades</p> <p>A la hora de enfrentarse a un conjunto de datos hay que comenzar realizando dos operaciones básicas.</p> <p>ORDENAR RESUMIR</p>

2.1. Análisis básico. Generalidades

- **Frecuencia**
 - *absoluta (f)*: el número de veces que aparece cada dato de la variable.
 - *total (n)*: número total de datos de la variable (suma de frecuencias absolutas).
 - *relativa (fr)*: cociente entre frecuencia absoluta y frecuencia total.

Carlos Montes – uc3m

2.1. Análisis básico. Generalidades

- **acumulada**: supuesta la ordenación de los datos de menor a mayor, la frecuencia acumulada de x_i es la suma de frecuencias hasta el valor x_i .

- **Absoluta (F)**
- **Relativa (Fr)**



Tabla de distribución de frecuencias

2.1. Análisis básico. Generalidades

Value	Frequency	Relative Frequency			Cumulative Frequency	Cum. Rel.
		Cumulative Frequency	Cum. Rel.	0,4842 + 0,3789 + 0,1263		
1	46	46	0,4842		46	0,4842
2	36	46 + 36 = 82	0,3789		82	0,8632
3	12	46 + 36 + 12 = 94	0,1263		94	0,9895
4	1		0,0105		95	1,0000

2.2. Gráficos para variables cualitativas

Diagrama de barras

Eje 1: valor o categoría de la variable.
Eje 2: altura proporcional a la frecuencia.

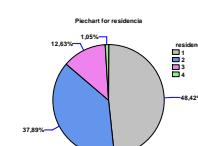
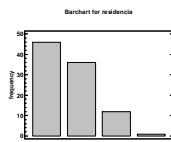
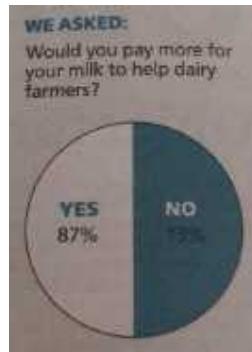


Diagrama de tarta
círculo dividido en sectores proporcionales a la frecuencia de cada valor.

2.2. Gráficos para variables cualitativas

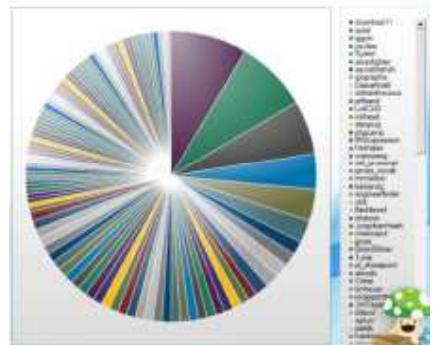
Encuesta en un periódico local



Carlos Montes – uc3m

2.2. Gráficos para variables cualitativas

Los 100 usuarios de Twitter más activos



Más de 4 o 5 sectores dificultan la lectura del diagrama.

2.3. Variables cuantitativas

En variables cuantitativas el análisis de frecuencias se realiza de la misma manera que en variables cualitativas.

- ✓ Absolutas
- ✓ Relativas
- ✓ Absolutas acumuladas
- ✓ Relativas acumuladas

Muchos valores diferentes



valores en clases o intervalos
(generalmente de la misma longitud)

2.3. Variables cuantitativas

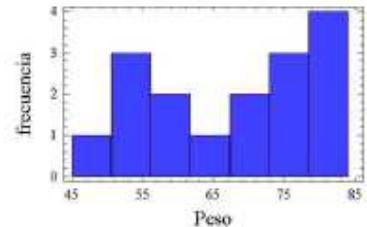
No confundir con el rango intercuartílico.

- *Rango o recorrido de una variable*: diferencia entre el mayor y el menor valor de ésta.
- *Amplitud de un intervalo*: diferencia entre el extremo superior e inferior del mismo.
- *Marca de clase (m)*: punto medio de cada intervalo o clase, valor representativo de todos los datos del intervalo.

El número de clases r debe oscilar entre 5 y 20; a menudo se escoge el entero más próximo a \sqrt{n}

2.4. Gráficos para variables cuantitativas

El **histograma** es una representación para variables agrupadas en intervalos.



- Abscisas: intervalo de valor de la variable.
- Ordenadas: altura proporcional a la frecuencia, de manera que las áreas de los rectángulos sean proporcionales a las frecuencias.

Carlos Montes – uc3m

2.4. Gráficos para variables cuantitativas

Muestra las tendencias generales de los datos:

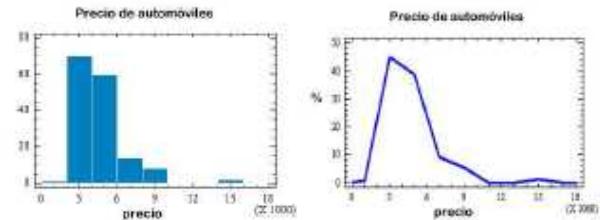
- Concentraciones: más de una concentración \Rightarrow datos heterogéneos.
- Huecos: indicio de que los datos proceden de poblaciones diferentes.
- Valores atípicos: aquellos que se separan mucho del patrón general que siguen los datos.

2.4. Gráficos para variables cuantitativas

- Asimetrías: tendencia de los datos cuando nos alejamos de las zonas de concentración.
 - Cola de la distribución de los datos hacia $+\infty$, \Rightarrow asimetría positiva.
 - Cola de la distribución de los datos hacia $-\infty$ \Rightarrow asimetría negativa.

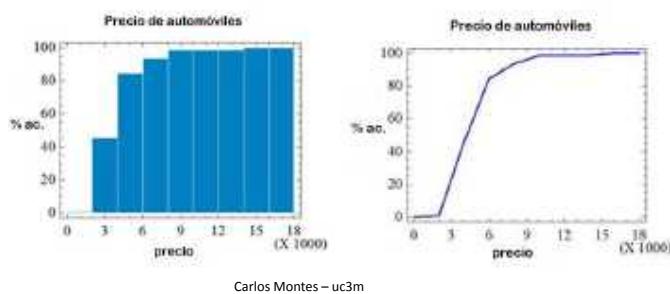
2.4. Gráficos para variables cuantitativas

* El polígono de frecuencias es una línea poligonal que resulta al unir los puntos centrales de la parte superior del histograma.



2.4. Gráficos para variables cuantitativas

- * Ambos pueden construirse a partir de las frecuencias acumuladas.



3.1. Medidas características. Generalidades

Podemos distinguir:

- ♦ Tendencia central o centralización: indican el valor medio de los datos.
- ♦ Dispersión: indican la variabilidad de los datos.
- ♦ Forma:
 - ♦ Simetría
 - ♦ Apuntamiento

3.1. Medidas características. Generalidades

- * Son aquellas que nos permiten resumir con un solo número los rasgos fundamentales de la distribución.
- * Deben acompañarse de herramientas gráficas para evitar errores.

3.2. Medidas de tendencia central

Media aritmética

$$\bar{x} = \frac{\sum_{j=1}^n x_j f(x_j)}{n}$$

$$\bar{x} = \frac{\sum_{j=1}^n m_j f(m_j)}{n} \quad \Leftrightarrow \text{Error de agrupamiento}$$

3.2. Medidas de tendencia central

Propiedades de la media aritmética

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

$$1) \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0$$

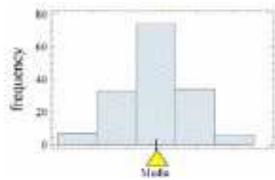
$$2) y = x + k$$

$$3) y = kx \quad \bar{y} = \frac{\sum_{i=1}^n kx_i}{n} = k \frac{\sum_{i=1}^n x_i}{n} = k\bar{x}$$

Carlos Montes – uc3m

3.2. Medidas de tendencia central

Es el centro de gravedad de los datos.



Si la distribución es asimétrica, se desplaza respecto a la clase más frecuente, y deja de ser una buena medida de centralización.

3.2. Medidas de tendencia central

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Std. skewness	-1,20518
Std. kurtosis	-1,70142

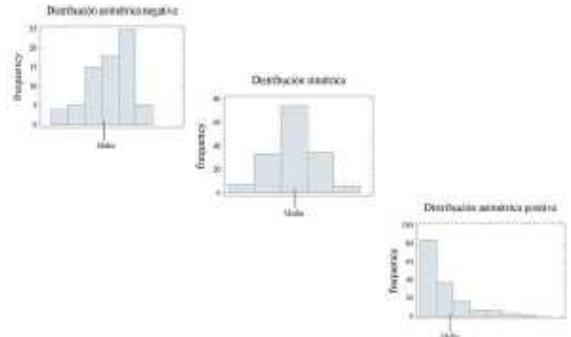
Es muy sensible a los datos atípicos.

$$1, 2, 4, 5, 7, 9, 11, 13 \quad \bar{x} = 6,5$$

$$1, 2, 4, 5, 7, 9, 11, 130 \quad \bar{x} = 21,125$$

Para muestras muy asimétricas o con muchos datos atípicos, la mediana es mejor medida de tendencia central.

3.2. Medidas de tendencia central



3.2. Medidas de tendencia central

Mediana

Valor de la muestra que la divide en dos partes iguales.

- * Para calcular la mediana se ordenan los datos de menor a mayor:
 - nº impar de datos: valor central.

$$2, 3, 4, \textcircled{5} 7, 7, 9$$

Carlos Montes – uc3m

3.2. Medidas de tendencia central

- nº par de datos: media aritmética de los valores centrales.

$$2, 3, 4, \textcircled{5}, 7, 9, 11$$

$$\frac{5+7}{2} = 6$$

3.2. Medidas de tendencia central

- Si tenemos los datos organizados en forma de tabla.

Accidentes mortales	n	f	F	N
0	7	0,039	0,039	7
1	26	0,144	0,183	33
2	33	0,182	0,365	66
3	38	0,210	0,575	104
4	29	0,160	0,735	133
5	20	0,110	0,846	153
6	15	0,083	0,929	168
7	9	0,050	0,978	177
8	2	0,011	0,989	179
9	2	0,011	1,000	181
10	0	0,000	1,000	181
>10	0	0,000	1,000	181
Total	181			

La mediana es el primer valor donde se alcanza la frecuencia relativa acumulada 0,5.

3.2. Medidas de tendencia central

La mediana NO es sensible a datos atípicos.

Robustez

Summary Statistics for altura	
Count	195
Average	174,621
Median	177,0
Mode	180,0
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

$$2, 3, 4, \textcircled{5}, 7, 7, 9$$

$$2, 3, 4, \textcircled{5}, 7, 7, 87$$

3.2. Medidas de tendencia central

Moda

Es el valor más frecuente de la distribución.

- Es apropiada para datos cualitativos o cuantitativos discretos.
- Pueden existir una o varias modas.
- En una muestra continua solo podemos hablar de un intervalo modal (el de mayor densidad de frecuencia)

Carlos Montes – uc3m

3.2. Medidas de tendencia central

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Std. skewness	-1,20518
Std. kurtosis	-1,70142

En variables continuas puede que no se repita ningún valor.

Pueden existir distribuciones con más de una moda.

3.3. Medidas de dispersión

Medidas de la separación de los datos (generalmente, respecto a la media).

medida
+ representativa



- dispersión

3.3. Medidas de dispersión

Varianza

$$s_x^2 = \frac{\sum_n (x_j - \bar{x})^2 f(x_j)}{n}$$

3.3. Medidas de dispersión

Propiedades de la varianza

- 1) Es una cantidad acotada y positiva
- 2) La varianza NO se ve afectada por los cambios de origen (transformaciones aditivas)

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$
$$y = x + k$$
$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^n (x_i + k - \bar{x} - k)^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = s_x^2$$

Carlos Montes - uc3m

3.3. Medidas de dispersión

- 3) La varianza SÍ se ve afectada por los cambios de escala (transformaciones multiplicativas)

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$
$$y = kx$$
$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^n (kx_i - k\bar{x})^2}{n} = \frac{k^2 \sum_{i=1}^n (x_i - \bar{x})^2}{n} = k^2 s_x^2$$
$$S_y^2 = k^2 \cdot S_x^2$$

3.3. Medidas de dispersión

Fórmula de cálculo

$$s_x^2 = \frac{\sum_{j=1}^n x_j^2 f(x_j)}{n} - \bar{x}^2$$

3.3. Medidas de dispersión

Una medida alternativa es la **cuasivarianza**

$$\hat{s}_x^2 = \frac{\sum_{j=1}^n (x_j - \bar{x})^2 f(x_j)}{n - 1}$$

La mayoría de los programas estadísticos calculan la cuasivarianza en lugar de la varianza, y la llaman varianza.

3.3. Medidas de dispersión

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Std. skewness	-1,20548
Std. kurtosis	-1,70142

No aparece por defecto
en el programa.

desviación típica

Carlos Montes – uc3m

- La varianza mide el promedio de las desviaciones (al cuadrado) de las observaciones respecto a la media.
- Al ser un cuadrado, siempre es positiva.
- Es muy sensible a datos atípicos.
- Problema: unidades $67,68 \text{ cm}^2$



3.3. Medidas de dispersión

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Std. skewness	-1,20548
Std. kurtosis	-1,70142

Desviación típica

- Toma siempre valores no negativos.
- Ventaja: tiene las mismas unidades que la variable.

8,22 cm

- Inconveniente: raíz cuadrada. La varianza es más fácil de usar en operaciones matemáticas al evitar la raíz.

3.3. Medidas de dispersión

Desviación típica

Es la raíz cuadrada positiva de la varianza.

$$s_x = \sqrt{\frac{\sum_n (x_j - \bar{x})^2 \cdot f(x_j)}{n}}$$

3.3. Medidas de dispersión

3.3. Medidas de dispersión

Cuasidesviación típica

$$\hat{s}_x = \sqrt{\frac{\sum_n (x_j - \bar{x})^2 f(x_j)}{n-1}}$$

- Para tamaños de muestra grande, casi no hay diferencia.

3.3. Medidas de dispersión

Coefficiente de variación

Es una medida de dispersión relativa.

$$CV = \frac{s}{|\bar{x}|} \cdot 100 \quad \bar{x} \neq 0$$

Carlos Montes – uc3m

3.3. Medidas de dispersión

Cuantiles

Son los valores de la variable que dividen la distribución en c partes iguales.

- **Cuartiles (Q)** $c=4$
- **Quintiles (K)** $c=5$
- **Percentiles (p)** $c=100$

3.3. Medidas de dispersión

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

Nos permite:

- 1) Comparar la dispersión entre distribuciones.
- 2) Evaluar la representatividad de la media.

3.3. Medidas de dispersión

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

3.3. Medidas de dispersión

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

Rango intercuartílico (RI)

Es la diferencia entre los percentiles 75 y 25 (o entre los cuartiles 3 y 1)

Carlos Montes – uc3m

3.3. Medidas de dispersión

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

3.4. Medidas de forma

Coeficiente de asimetría de Fisher

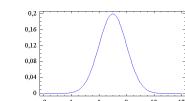


Ronald Aylmer Fisher
(1890-1962)

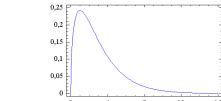
$$CA = \gamma_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{ns^3}$$

3.4. Medidas de forma

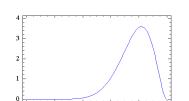
$\gamma_1 = 0 \Rightarrow$ Distribución simétrica



$\gamma_1 > 0 \Rightarrow$ Distribución asimétrica positiva o asimétrica a derechas



$\gamma_1 < 0 \Rightarrow$ Distribución asimétrica negativa o asimétrica a izquierdas



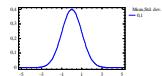
3.4. Medidas de forma

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Skewness	-0,302876
Stnd. skewness	-1,20518
Kurtosis	-0,855173
Stnd. kurtosis	-1,70142

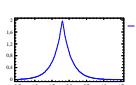
Carlos Montes – uc3m

3.4. Medidas de forma

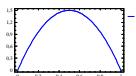
CAp=0: *mesocúrtica*



CAp>0: *leptocúrtica*



CAp<0: *platicúrtica*

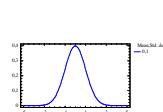


3.4. Medidas de forma

Coeficiente de apuntamiento o curtosis

Indica el mayor o menor agrupamiento de los datos en torno a la media.

Como referencia se toma el apuntamiento de la distribución normal, que cumple:



$$CA_p = \frac{\sum (x_i - \bar{x})^4}{ns^4} = 3$$

$$CA_p = \frac{\sum (x_i - \bar{x})^4}{ns^4} - 3$$

(Exceso de curtosis)

3.4. Medidas de forma

Summary Statistics for altura

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Skewness	-0,302876
Stnd. skewness	-1,20518
Kurtosis	-0,855173
Stnd. kurtosis	-1,70142

4. Diagrama de caja

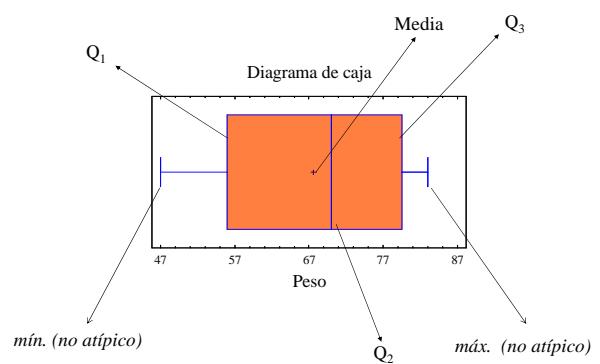
Representación gráfica de una distribución, construida para mostrar sus características principales y señalar los posibles datos atípicos.

Mínimo Máximo Cuartiles

$$LI = Q_1 - 1,5(Q_3 - Q_1) \quad LS = Q_3 + 1,5(Q_3 - Q_1)$$
$$LIE = Q_1 - 3(Q_3 - Q_1) \quad LSE = Q_3 + 3(Q_3 - Q_1)$$

Carlos Montes – uc3m

4. Diagrama de caja



Parte II

Tema 2. Estadística descriptiva bivariante

Tema 2

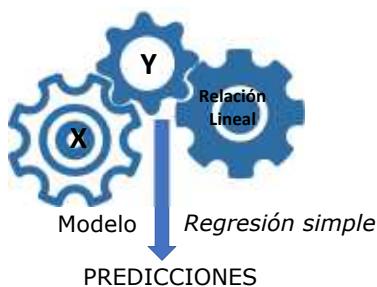
Descripción estadística de variables bidimensionales

Carlos Montes – uc3m

1. Introducción
2. Definiciones
3. Representación gráfica
4. Covariación
 - 4.1. Tipos
 - 4.2. Covarianza
 - 4.3. Coeficiente de correlación
 - 4.4. Matriz de covarianzas
5. Regresión simple
 - 5.1. Recta de regresión
 - 5.2. Interpretación de los coeficientes
 - 5.3. Evaluación del modelo
 - 5.4. Bondad del ajuste
6. Transformaciones

1. Introducción

Estudio de 2 caracteres simultáneos en cada elemento de la población.



2. Definiciones

- **Distribución conjunta de frecuencias de dos variables**

valores observados
y las frecuencias (relativas o absolutas)
de aparición de cada par.

Variable cualitativa \Rightarrow tabla de contingencia

$$\sum_i \sum_j fr(x_i, y_j) = 1$$

2. Definiciones

Distribución de frecuencias conjunta para las variables “número de hermanos” (columnas) y sexo (filas) de 95 estudiantes

	0	1	2	3	4	5	9	Row Total
0	3	13	11	2	2	0	1	32
	3,16%	13,68%	11,58%	2,11%	2,11%	0,00%	1,05%	33,68%
1	6	22	26	7	0	2	0	63
	6,32%	23,16%	27,37%	7,37%	0,00%	2,11%	0,00%	66,32%
Column Total	9	35	37	9	2	2	1	95
	9,47%	36,84%	38,95%	9,47%	2,11%	2,11%	1,05%	100,00%

Chicos

Chicas

Carlos Montes – uc3m

2. Definiciones

- **Distribución marginal**

Distribución de cada una de las variables, consideradas por separado (distribución de los valores de una sin tener en cuenta los de la otra).

$$f(x_i) = \sum_j f(x_i, x_j)$$

$$f(y_j) = \sum_i f(x_i, y_j)$$

Aparece en los márgenes de la tabla.

2. Definiciones

Frequency Table for sexo by hermanos

	0	1	2	3	4	5	9	Row Total
0	3	13	11	2	2	0	1	32
	3,16%	13,68%	11,58%	2,11%	2,11%	0,00%	1,05%	33,68%
1	6	22	26	7	0	2	0	63
	6,32%	23,16%	27,37%	7,37%	0,00%	2,11%	0,00%	66,32%
Column Total	9	35	37	9	2	2	1	95
	9,47%	36,84%	38,95%	9,47%	2,11%	2,11%	1,05%	100,00%

Alumnos con 2 hermanos

2. Definiciones

- **Distribución condicionada** de y para $x=x_i$ es la distribución que se obtiene imponiendo la condición $x = x_i$

$$f_r(y_j|x=x_i) = \frac{f(x_i, y_j)}{f(x_i)}$$

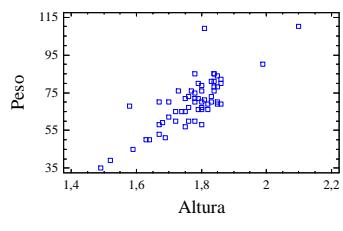
	0	1	2	3	4	5	9
0	3	13	11	2	2	0	1
1	6	22	26	7	0	2	0
Column Total	9	35	37	9	2	2	1

0	11/95 = 0,116
1	26/95 = 0,274
	0,39

0	11/37 = 0,298
1	26/37 = 0,702
	1

3. Representación gráfica

- *Diagrama de dispersión o nube de puntos*



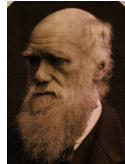
Carlos Montes – uc3m

4.1. Covariacion. Tipos

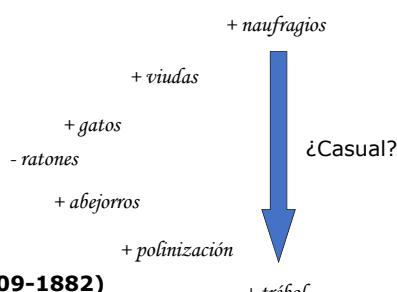
Variación conjunta o relación de dependencia entre las variables estudiadas (X, Y).

- **Dependencia causal unilateral:**
 X influye en Y , pero no la inversa.
- **Interdependencia:**
 X influye en Y , y viceversa.
- **Dependencia indirecta:**
Las variables muestran una covariación a través de una tercera variable que influye en ellas.
- **Concordancia.**
- **Covariación casual.**

4.1. Covariacion. Tipos



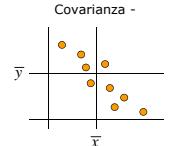
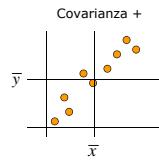
Charles Darwin (1809-1882)



4.2. Covarianza

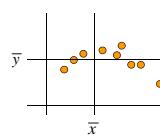
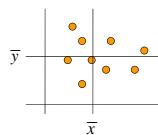
La covarianza es una medida descriptiva de la relación **lineal** entre cada par de variables.

$$\text{cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$



4.2. Covarianza

Covarianza 0



Carlos Montes – uc3m

4.2. Covarianza

* La covarianza tiene el inconveniente de depender de las unidades de medida.

* Para evitarlo, se emplea el *coeficiente de correlación lineal r*.

4.3. Coeficiente de correlación



Sir Francis Galton
(1822-1911)

$$r = \frac{s_{xy}}{s_x s_y}$$



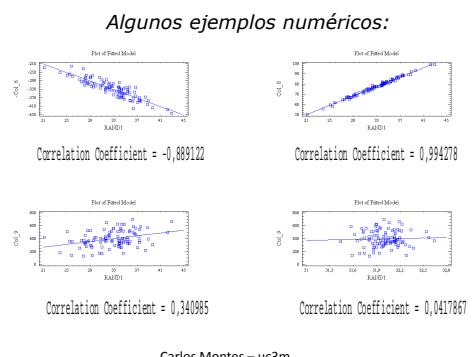
Karl Pearson
(1857-1936)

4.3. Coeficiente de correlación

r varía entre -1 y 1

- $r = -1$ Correlación lineal perfecta e inversa.
La nube de puntos es una recta de pendiente negativa.
- $r = 1$ Correlación lineal perfecta y directa.
La nube de puntos es una recta de pendiente positiva.
- $r = 0$ No existe correlación,
o bien existe una relación no lineal entre las variables.

4.3. Coeficiente de correlación



4.4. Matriz de covarianzas

* Las medidas de dependencia lineal de un conjunto de datos bidimensionales pueden presentarse en forma de matriz.

$$M = \begin{pmatrix} s_x^2 & \text{cov}(x, y) \\ \text{cov}(y, x) & s_y^2 \end{pmatrix}$$

Matriz de covarianzas muestrales

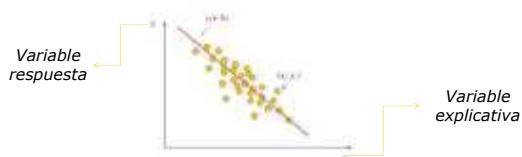
$$R = \begin{pmatrix} 1 & \text{corr}(x, y) \\ \text{corr}(y, x) & 1 \end{pmatrix}$$

Matriz de correlaciones muestrales

$\text{corr}(x, x) = \text{corr}(y, y)$

5.1. Recta de regresión

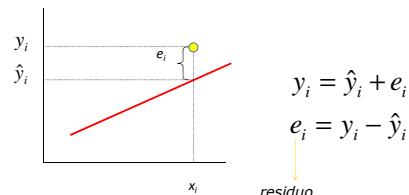
- * Recta que refleja, de la manera más aproximada posible, la evolución conjunta de dos variables.
- * Cuanto más próximo a ± 1 esté el coeficiente de correlación, mayor será la capacidad de explicación de la recta.



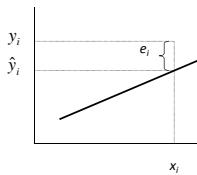
5.1. Recta de regresión

Para cada x_i tendremos

- ordenada real y_i
- ordenada sobre la recta de regresión \hat{y}_i



5.1. Recta de regresión



Possible criterio de construcción:
minimizar la suma de los errores.

$$e_i = y_i - \hat{y}_i$$

$$\min \sum_{i=1}^n e_i = \min \sum_{i=1}^n (y_i - \hat{y}_i)$$

Para evitar la influencia de los signos:

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Método de los mínimos cuadrados
Carl Friedrich Gauss (1777-1855)

Carlos Montes – uc3m



5.1. Recta de regresión

Si lo que queremos ajustar es una recta:

$$\hat{y}_i = a + b x_i$$

Minimizando llegamos a:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x})$$

(recta de regresión de Y sobre X)

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

(recta de regresión de X sobre Y)

5.2. Interpretación de los coeficientes

Como $y = a + bx$ $\frac{dy}{dx} = b$

b es la pendiente de la recta:
incremento de y cuando x aumenta en una unidad.

$$\begin{aligned}\Delta \hat{y} &= \hat{y}(x_i + 1) - \hat{y}(x_i) = \\ &= [a + b(x_i + 1)] - [a + bx_i] = b\end{aligned}$$

a es el valor de la recta cuando $x=0$

5.3. Evaluación del modelo

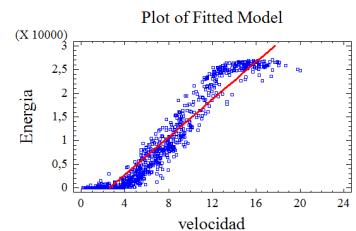


5.3. Evaluación del modelo



5.3. Evaluación del modelo

$$r = 0,96$$

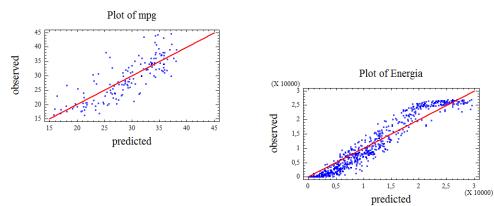


No hay relación lineal a pesar del elevado r .

5.3. Evaluación del modelo

Gráfico de valores previstos frente a valores observados

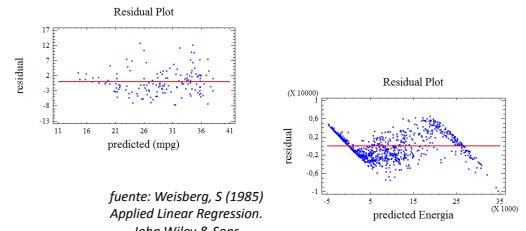
Linealidad \Rightarrow puntos distribuidos linealmente alrededor de la recta.



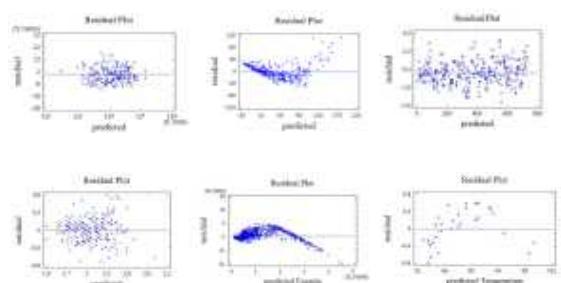
5.3. Evaluación del modelo

Gráfico de residuos frente a valores previstos

Linealidad \Rightarrow puntos distribuidos al azar

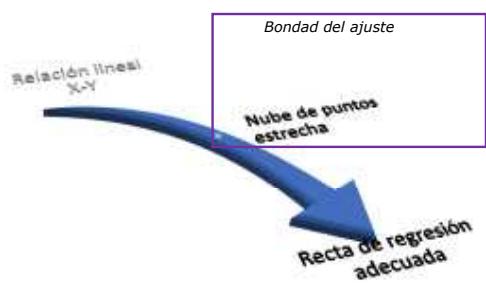


5.3. Evaluación del modelo



Carlos Montes – uc3m

5.4. Bondad del ajuste



5.4. Bondad del ajuste

- * La regresión simple será tanto mejor cuanto más estrecha sea la nube de puntos alrededor de la muestra.
- * La dispersión viene cuantificada por el coeficiente de correlación, o por el coeficiente de determinación R^2 , que varía entre 0 y 1:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = r^2$$

Variabilidad de los residuos
Variabilidad de los datos

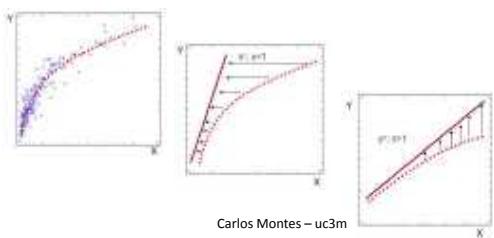
5.4. Bondad del ajuste

Cuanto más explicativa sea la regresión, menor será la variabilidad que queda en los residuos respecto a la de los datos, y R^2 será mayor.

Nos indica la proporción de la dispersión de la variable respuesta y que es capaz de explicar la recta de regresión.

6. Transformaciones

Cuando las hipótesis del modelo no se cumplen es necesario transformar los datos, de manera que los datos transformados cumplan las hipótesis.



6. Transformaciones

Las más utilizadas son:

■ Logaritmo

$$y = \ln x \quad x = e^y$$

■ Potencia

$$y = x^c \quad x = y^{\frac{1}{c}}$$

■ Inversa

$$y = I/x \quad x = I/y$$

■ Raíz cuadrada

$$y = \sqrt{x} \quad x = y^2$$

Parte III

Tema 3. Probabilidad

Tema 3 Probabilidad

Carlos Montes – uc3m

1. Definiciones y notación
2. Concepto de probabilidad
 - 2.1. Definición clásica
 - 2.2. Definición frecuentista
3. Propiedades fundamentales
4. Probabilidad condicionada
5. Independencia e incompatibilidad
6. Teorema de la Probabilidad Total
7. Teorema de Bayes

1. Definiciones y notación

- **Experimento aleatorio**

observación de una propiedad de interés que proporciona distintos resultados, sin que pueda precisarse cuál de ellos aparecerá.

Para su análisis debe conocerse el conjunto de todos los resultados posibles: espacio muestral (E, Ω)

1. Definiciones y notación

- **Suceso elemental (A, B)**

cada uno de los resultados posibles de un experimento aleatorio que verifican:

- Siempre ocurre alguno de ellos.
- Son mutuamente excluyentes.

- **Suceso compuesto**

aquel construido a partir de uniones de sucesos elementales.

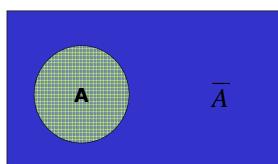
1. Definiciones y notación

• **Espacio muestral (E)**

unión de todos los sucesos elementales.

• **Suceso contrario o complementario de A (\bar{A})**

suceso que ocurre cuando no ocurre A.



1. Definiciones y notación

• **Suceso seguro**

el que siempre se observa.

• **Suceso imposible (\emptyset)**

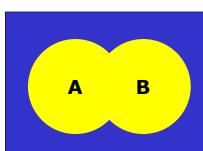
suceso que nunca se puede observar
(está fuera del espacio muestral)

1. Definiciones y notación

• **Suceso unión de A y B**

es el que se observa si ocurre
el suceso A o el suceso B.

$$A \cup B$$

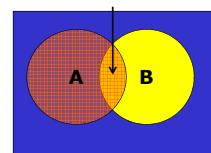


1. Definiciones y notación

• **Suceso intersección de A y B**

es el que se observa si sucede
A y B a la vez.

$$A \cap B$$

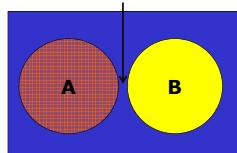


1. Definiciones y notación

- **Sucesos mutuamente excluyentes o disjuntos**

Sucesos sin elementos comunes

$$A \cap B = \emptyset$$



1. Definiciones y notación

La unión e intersección verifican las siguientes propiedades:

- *Commutativa*
- *Asociativa*
- *Distributiva*
- *Idempotente*
- *Elemento neutro*
- *Simplificación*
- *Absorción*

ÁLGEBRA
DE
BOOLE

1. Definiciones y notación

Leyes de De Morgan

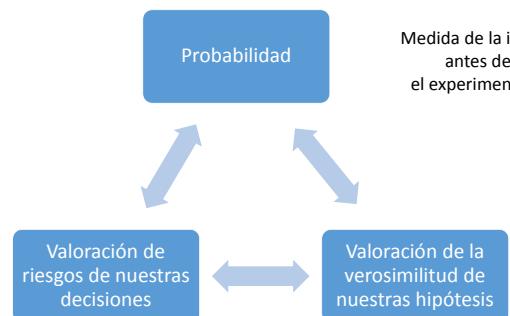
$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$



Augustus De Morgan
(1806-1871)

2.1. Probabilidad. Definición clásica



Medida de la incertidumbre
antes de realizar
el experimento aleatorio.

2.1. Probabilidad. Definición clásica

$$P(A) = \frac{n^o \text{ de casos favorables}}{n^o \text{ de casos posibles}}$$

si todos los casos son igualmente posibles (equiprobables)
(1812)



Pierre Simon Laplace
(1749-1827)

2.1. Probabilidad. Definición clásica

Limitaciones:

- Requiere que todos los casos sean igualmente probables.
- Requiere que el número de casos posibles sea finito.

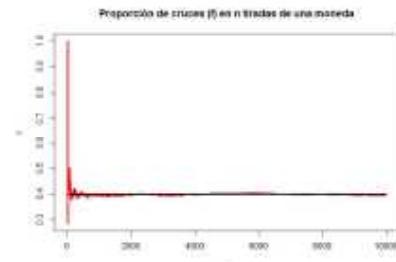
2.2. Probabilidad. Definición frecuentista

Probabilidad de un suceso es su frecuencia relativa de aparición si repetimos indefinidamente el experimento.



Richard von Mises
(1883-1953)

2.2. Probabilidad. Definición frecuentista



Pero la experimentación “indefinida” es imposible...

3. Propiedades fundamentales

- 1) $0 \leq P(A) \leq 1$
- 2) $P(E) = 1$
- 3) $P(\emptyset) = 0$
- 4) $P(\bar{A}) = 1 - P(A)$

3. Propiedades fundamentales

- 5) Si $A \cap B = \emptyset$
 $P(A \cup B) = P(A) + P(B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

4. Probabilidad condicionada

Es la probabilidad de un suceso sabiendo (condicionada a) la ocurrencia de otro suceso.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

ej. 7

Sean A y B dos sucesos tales que la probabilidad de que ocurra A es 0,6, la de que ocurra A o B es igual a 0,8 y, si se sabe que ha ocurrido B , la probabilidad de que ocurra A es igual a 0,5. ¿Puede determinarse de forma única el valor de $P(B)$? En caso afirmativo, determíñese dicho valor, y, en caso negativo, justificar por qué no es posible su determinación única.

$$\begin{aligned}
 P(A) &= 0.6 & P(A \cup B) &= 0.8 & P(A|B) &= 0.5 \\
 P(A \cup B) &= 0.8 = P(A) + P(B) - P(A \cap B) \\
 P(A|B) &= 0.5 = \frac{P(A \cap B)}{P(B)} & P(A \cap B) &= 0.5 \cdot P(B) \\
 0.8 &= 0.6 + P(B) - 0.5 \cdot P(B) \\
 0.2 &= 0.5 \cdot P(B) & P(B) &= \frac{0.2}{0.5} = 0.4
 \end{aligned}$$

5. Independencia e incompatibilidad

Dos sucesos A y B son **independientes** si el conocimiento de la ocurrencia de uno no modifica la probabilidad del otro.

$$P(A|B) = P(A), \quad P(B|A) = P(B)$$

5. Independencia e incompatibilidad

$$\begin{aligned}
 P(A|B) &= P(A) = \frac{P(A \cap B)}{P(B)} \\
 P(B|A) &= P(B) = \frac{P(A \cap B)}{P(A)}
 \end{aligned}$$

Si son independientes:

$$P(A \cap B) = P(A) \cdot P(B)$$

condición de independencia

5. Independencia e incompatibilidad

Dos sucesos A y B son incompatibles si no pueden verificarse simultáneamente.

$$A \cap B = \emptyset$$

LOS SUCESOS INCOMPATIBLES
NO SON INDEPENDIENTES

5. Independencia e incompatibilidad

Sean A y B incompatibles:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0}{P(A)} = 0$$

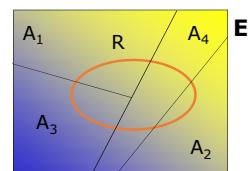
Si uno se da,
el otro no puede darse.

6. Teorema de la probabilidad total

Sea A_1, A_2, \dots, A_k una partición del espacio muestral E:

$$A_1, A_2, \dots, A_k | A_1 \cup A_2 \cup \dots \cup A_k = E$$

$$A_i \cap A_j = \emptyset \quad \forall i \neq j$$

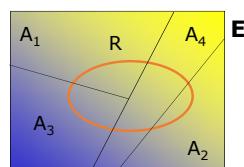


Sea R un suceso cualquiera de ese espacio muestral:

6. Teorema de la probabilidad total

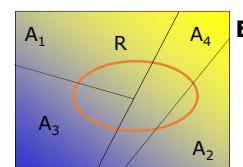
$$P(R) = P(R \cap A_1) + P(R \cap A_2) + \dots + P(R \cap A_k)$$

$$P(R) = P(R|A_1)P(A_1) + P(R|A_2)P(A_2) + \dots + P(R|A_k)P(A_k)$$



6. Teorema de la probabilidad total

$$P(R) = \sum_1^k P(R|A_i) \cdot P(A_i)$$



ej. 9

Una empresa ha adquirido memorias USB a dos proveedores diferentes, Al proveedor A se le compran 1000 memorias con un porcentaje de defectuosos del 4%. Al proveedor B se le compran 500 memorias con un porcentaje de defectuosos del 1%. Con estas condiciones, se pide:

a) **Probabilidad de elegir una memoria defectuosa entre todas las unidades adquiridas.**

b) Se ha elegido una memoria que resulta defectuosa. Calcule la probabilidad de que sea del proveedor B.

ej. 7

D: "memoria defectuosa"

A: "memoria del proveedor A"

B: "memoria del proveedor B"

$$P(A) = \frac{1000}{1500} = 0.6666$$

$$P(B) = \frac{500}{1500} = 0.3333$$

$$P(D|A) = 0.04$$

$$P(D|B) = 0.01$$

$$P(D) = P(D|A)P(A) + P(D|B)P(B) =$$

$$= 0.04 \frac{1000}{1500} + 0.01 \frac{500}{1500} = 0.03$$

7. Teorema de Bayes

Se aplica para calcular la probabilidad de cada una de las posibles causas, una vez observado el efecto.



$$P(A_i | R) = \frac{P(R | A_i) \cdot P(A_i)}{P(R)}$$

7. Teorema de Bayes

$$P(A_i | R) = \frac{P(R \cap A_i)}{P(R)} \quad P(R | A_i) = \frac{P(R \cap A_i)}{P(A_i)}$$

$$P(A_i | R) \cdot P(R) = P(R | A_i) \cdot P(A_i)$$

$$P(A_i | R) = \frac{P(R | A_i) \cdot P(A_i)}{P(R)}$$

ej. 7

Una empresa ha adquirido memorias USB a dos proveedores diferentes. Al proveedor A se le compran 1000 memorias con un porcentaje de defectuosos del 4%. Al proveedor B se le compran 500 memorias con un porcentaje de defectuosos del 1%. Con estas condiciones, se pide:

a) Probabilidad de elegir una memoria defectuosa entre todas las unidades adquiridas.

b) Se ha elegido una memoria que resulta defectuosa. Calcule la probabilidad de que sea del proveedor B.

ej. 7

$$P(B|D) = \frac{P(D|B)P(B)}{P(D|A)P(A) + P(D|B)P(B)}$$

$P(D)$

$$= \frac{0.01 \cdot \frac{500}{1500}}{0.03} = 0.11$$

***EN EL FONDO,
LA TEORÍA DE LA PROBABILIDAD
ES SOLO SENTIDO COMÚN
EXPRESADO CON NÚMEROS.***

Pierre Simon Laplace

Parte IV

Tema 4. Variable aleatoria

Tema 4 Variable aleatoria

Carlos Montes – uc3m

1. Concepto
2. Distribución de probabilidad
 - 2.1. Función de probabilidad
 - 2.2. Función de distribución
 - 2.3. Variables aleatorias discretas y continuas
 - 2.4. Función de densidad
3. Medidas características de una variable aleatoria
 - 3.1. Medidas de tendencia central
 - 3.2. Medidas de dispersión
4. Covarianza y correlación
5. Transformaciones y medidas características

1. Concepto

a) Ortodoxo

- Variable cuyo valor numérico está determinado por el resultado de un experimento aleatorio.

b) Ligeramente heterodoxo

- Variable que cuantifica la magnitud de interés, y cuya realización numérica concreta depende del azar(cada valor o intervalo de valores tendrá una probabilidad de aparición).

1. Concepto

Lanzamos 2 dados no trucados.

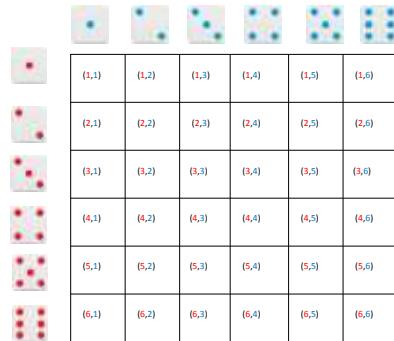
*Nos interesa estudiar el experimento aleatorio:
"suma de puntuaciones".*

¿Cómo definimos la variable aleatoria?

1. Concepto

Probabilidad de cada resultado:

$$1/36$$



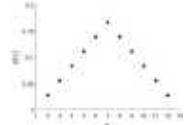
Carlos Montes – uc3m

1. Concepto

X: "Suma de puntos obtenidos al lanzar dos dados"

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

x	p(x)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36



1. Concepto

Suma de puntuaciones de los dos dados.

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

¿Cómo definimos la variable aleatoria?

X: "Suma de puntos obtenidos al lanzar dos dados".

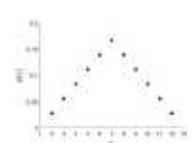
1. Concepto

X: "Suma de puntos obtenidos al lanzar dos dados"

Probabilidad de obtener 8 puntos:
5/36

Probabilidad de obtener menos de 6 puntos:
 $1/36 + 2/36 + 3/36 + 4/36 = 10/36$

x	p(x)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36



1. Concepto

- ~~Fallo de una maquinaria.~~
- Número de artículos defectuosos en un lote.
- Número de bits transmitidos correctamente.
- Distancia recorrida con determinada cantidad de combustible.
- Número de averías de un sistema.
- Número de clientes que llegan a un puesto de servicio por unidad de tiempo.

Carlos Montes – uc3m

1. Concepto



2.1. Función de probabilidad

Función de probabilidad, de cuantía o de masa de una variable aleatoria

Es la función $p(x)$ de una variable **discreta** X que asigna a cada valor diferente de X : x_1, x_2, \dots, x_k la probabilidad de ser obtenido en el experimento aleatorio.

$$p(x_0) = P(X = x_0)$$
$$\sum_E p(x_i) = 1$$

ej. 17

Sea una variable aleatoria discreta que toma los valores $X=\{a, 1, 2, 3\}$ y con función de probabilidad $p(x)=x/10$. ¿Qué valor debe tomar a?

<p>$X = \{a, 1, 2, 3\}$ $p(x) = x/10$</p> <p>Los valores 1,2,3 suman una probabilidad de: $\frac{1+2+3}{10} = \frac{6}{10}$</p> <p>Para que la probabilidad total sea 1.</p> <p>$a = 4$</p> <p>Carlos Montes – uc3m</p>	<p>2.2. Función de distribución</p> <p>Función de distribución F(x) (Cumulative probability function)</p> <p>Función de distribución de la variable aleatoria X en el punto $x = x_0$ es la probabilidad de que X tome un valor menor o igual que x_0.</p> $F(x_0) = P(x \leq x_0)$ $F(x_0) = P(-\infty < x \leq x_0)$ $F(x_0) = P(-\infty, x_0]$
<p>2.2. Función de distribución</p> <p><i>Propiedades</i></p> <p>1) $F(+\infty) = 1$ $F(-\infty) = 0$</p> <p>2) $P(x_0, x_{0+h}) = F(x_{0+h}) - F(x_0)$</p> <p>3) Es una función monótona no decreciente: $F(x_0) \leq F(x_{0+h})$</p>	<p>2.3. Variables aleatorias discretas y continuas</p> <p>Variable aleatoria discreta: toma un número de valores cuantitativos discretos</p> $F(x_0) = p(X \leq x_0) = \sum_{x_i \leq x_0} p(X = x_i)$ <p>En una distribución discreta, la probabilidad se concentra en los puntos de discontinuidad x_i</p>

2.3. Variables aleatorias discretas y continuas

Variable aleatoria continua:

puede tomar cualquier valor en un intervalo

$$p(X = x_i) = 0$$

$$p(X \leq x_i) = p(X < x_i)$$

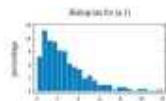
La probabilidad de cada punto concreto es nula.

Carlos Montes – uc3m

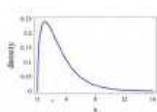
2.4. Función de densidad

Función que describe la **densidad de probabilidad** en cualquier intervalo.

$$f(x) = \frac{P(x_0 < X < x_0 + \Delta x)}{\Delta x}$$



Haciendo las clases del histograma cada vez más pequeñas, éste tenderá a una curva $f(x)$, capaz de describir el comportamiento de la variable.

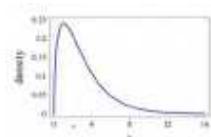


2.4. Función de densidad



$$f(x) = \frac{P(x_0 < X < x_0 + \Delta x)}{\Delta x}$$

La probabilidad de cualquier intervalo vendrá dada por el área que $f(x)$ encierra en ese intervalo.



2.4. Función de densidad

$$f(x) = \frac{P(x_0 < X < x_0 + \Delta x)}{\Delta x} = \frac{F(x_0 + \Delta x) - F(x_0)}{\Delta x}$$

Tomando un intervalo tan pequeño como queramos ($\Delta x \rightarrow 0$)

$$\lim_{\Delta x \rightarrow 0} \frac{F(x_0 + \Delta x) - F(x_0)}{\Delta x} = F'(x_0)$$

$f(x) = F'(x)$

2.4. Función de densidad

ej. 22

Propiedades

$$1) f(x) \geq 0 \quad \forall x \in D_x$$

$$2) f(x) = 0 \quad \forall x \notin D_x$$

$$3) P(x \leq x_0) = \int_{-\infty}^{x_0} f(x) dx$$

$$4) \int_{-\infty}^{+\infty} f(x) dx = [F(x)]_{-\infty}^{+\infty} = F(+\infty) - F(-\infty) = 1$$

$$5) P(a < x \leq b) = F(b) - F(a)$$

$$= \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx = \int_a^b f(x) dx$$

Carlos Montes – uc3m

La duración de la batería de un iPad mini (medida en días) viene dada por una variable aleatoria con función de densidad

$$f(x) = \begin{cases} (2 + kx)/6 & \text{si } 0 < x < 2 \\ 0 & \text{en el resto} \end{cases}$$

Calcule:

a) *El valor de k para que f(x) sea función de densidad*

b) *La duración media de la batería.*

c) *Se considera admisible un iPad cuya batería tenga una duración superior a 1.5 días. Sabiendo que la batería ha durado más de un día ¿cuál es la probabilidad de que ese iPad sea admisible?*

$$a) \int_0^2 \frac{(2 + kx)}{6} dx = 1$$

$$\frac{1}{6} \left[2x + \frac{kx^2}{2} \right]_0^2 = 1 \quad \frac{1}{6} [4 + 2k] = 1 \quad [4 + 2k] = 6 \quad k = 1$$

3.1. Medidas de tendencia central

Media o esperanza matemática: $\mu, E(x)$

$$\text{Caso discreto: } \mu = E(x) = \sum_i x_i P(x_i)$$

$$\text{Caso continuo: } \mu = E(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Mediana

Caso discreto: el m más pequeño que satisfaga:

$$F(m) \geq 0.5$$

Caso continuo: el m tal que:

$$F(m) = 0.5$$

3.1. Medidas de tendencia central

Moda

Es el valor de mayor probabilidad o densidad.

Carlos Montes – uc3m

3.2. Medidas de dispersión

Varianza: σ^2 , $var(x)$

$$\text{Caso discreto: } \sigma^2 = var(X) = \sum_{i=1}^n (x_i - \mu)^2 P(x_i)$$

$$\text{Caso continuo: } \sigma^2 = var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

$$\text{Equivale a } E(X - \mu)^2$$

3.2. Medidas de dispersión

$$\begin{aligned}\sigma^2 &= E[(X - \mu)^2] = E(X^2 - 2X\mu + \mu^2) = \\ &= E(X^2) - 2\mu E(X) + E(\mu^2) = \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2\end{aligned}$$

Fórmula de cálculo: $\sigma^2 = E(X^2) - [E(X)]^2$

3.2. Medidas de dispersión

Percentil p

Es el valor x_p que verifica:

$$F(x_p) = p$$

ej. 22

La duración de la batería de un iPad mini (medida en días) viene dada por una variable aleatoria con función de densidad

$$f(x) = \begin{cases} (2 + kx)/6 & \text{si } 0 < x < 2 \\ 0 & \text{en el resto} \end{cases}$$

Calcule:

- a) *El valor de k para que f(x) sea función de densidad*
- b) **La duración media de la batería.**
- c) *Se considera admisible un iPad cuya batería tenga una duración superior a 1.5 días. Sabiendo que la batería ha durado más de un día ¿cuál es la probabilidad de que ese iPad sea admisible?*

$$b) \int_0^2 x \cdot \frac{(2+x)}{6} dx = \frac{1}{6} \left[\frac{2x^2}{2} + \frac{x^3}{3} \right]_0^2 =$$

$$= \frac{1}{6} \left[4 + \frac{8}{3} \right] = \frac{20}{18} = 1.11$$

ej. 22

La duración de la batería de un iPad mini (medida en días) viene dada por una variable aleatoria con función de densidad

$$f(x) = \begin{cases} (2 + kx)/6 & \text{si } 0 < x < 2 \\ 0 & \text{en el resto} \end{cases}$$

Calcule:

- a) *El valor de k para que f(x) sea función de densidad*
- b) **La duración media de la batería.**
- c) *Se considera admisible un iPad cuya batería tenga una duración superior a 1.5 días. Sabiendo que la batería ha durado más de un día ¿cuál es la probabilidad de que ese iPad sea admisible?*

$$c) P(X > 1.5 | X > 1) = \frac{P[(X > 1.5) \cap (X > 1)]}{P(X > 1)} = \frac{P(X > 1.5)}{P(X > 1)}$$

$$P(X > 1.5) = \int_{1.5}^2 \frac{(2+x)}{6} dx = \frac{1}{6} \left[2x + \frac{x^2}{2} \right]_{1.5}^2 = 0.3125$$

$$P(X > 1) = \frac{1}{6} \left[2x + \frac{x^2}{2} \right]_1^2 = 0.5833$$

$$\frac{0.3125}{0.5833} = 0.5357$$

4. Covarianza y correlación

La covarianza y correlación poblacionales tienen la misma interpretación que las muestrales.

$$\text{cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$

$$\rho = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}}$$

Carlos Montes – uc3m

5. Transformaciones y medidas características

En el caso de la media:

$$E(a + bX) = a + bE(X)$$

Para dos variables:

$$E(aX + bY) = aE(X) + bE(Y)$$

5. Transformaciones y medidas características

En el caso de la varianza:

$$\text{var}(a + bX) = b^2 \text{var}(X)$$

Para dos variables:

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2abcov(X, Y)$$

Si X e Y están incorreladas:

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y)$$

5. Transformaciones y medidas características

De la misma manera:

$$\text{var}(aX - bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) - 2ab\text{cov}(X, Y)$$

Si X e Y están incorreladas:

$$\text{var}(aX - bY) = a^2 \text{var}(X) + b^2 \text{var}(Y)$$

**En ambos casos
es la SUMA de las varianzas.**

5. Transformaciones y medidas características

independencia \Rightarrow incorrelación

incorrelación $\not\Rightarrow$ independencia

(salvo en una normal bidimensional)

Carlos Montes – uc3m

Parte V

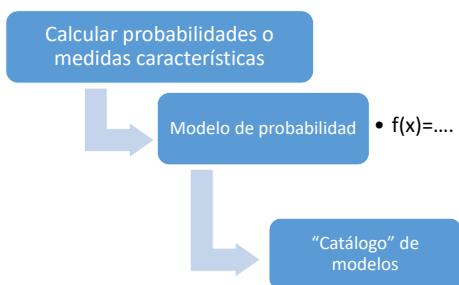
Tema 5. Modelos de probabilidad

Tema 5 Modelos de probabilidad

Carlos Montes – uc3m

1. Introducción
2. El proceso de Bernoulli
 - 2.1. Definición
 - 2.2. Ley binomial (1,0) o de Bernoulli
 - 2.3. Distribución binomial
3. El proceso de Poisson
 - 3.1. Definición
 - 3.2. Distribución de Poisson
 - 3.3. Distribución exponencial
4. Distribución uniforme
5. Distribución normal
6. Distribución lognormal
7. Teorema central del límite
 - 7.1. Definición
 - 7.2. Aplicaciones
8. Modelo de regresión lineal simple

1. Concepto



1. Proceso de Bernoulli. Definición.

- Fenómeno aleatorio dicotómico
- La observación consiste en la clasificación del resultado obtenido en una de 2 categorías posibles:
Éxito
Fracaso
- La proporción de cada una de las categorías en la población es constante:
 p : probabilidad de éxito
 $q = 1-p$: probabilidad de fracaso
- Las observaciones son independientes entre sí.

2.1. Variable de Bernoulli

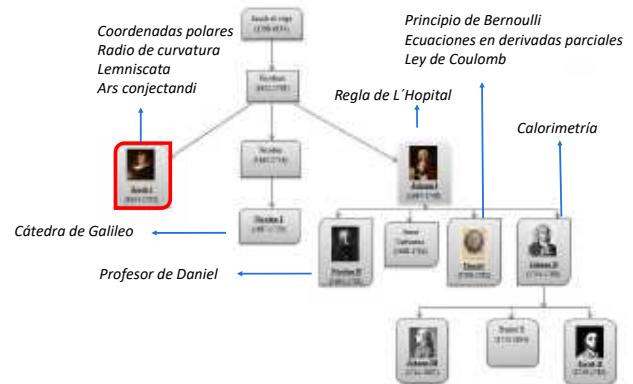
$$x = \begin{cases} 0 & \text{si obtenemos un fracaso} \\ 1 & \text{si obtenemos un éxito} \end{cases}$$



Jakob Bernoulli (1654-1715)

Carlos Montes – uc3m

2.1. Variable de Bernoulli



1. Ley binomial (1,0) o de Bernoulli

Características

Función de probabilidad

$$P(x) = p^x q^{1-x}; \quad x=0,1$$

$$P(x=1) = p$$

$$P(x=0) = q$$

Esperanza

$$E(X) = 0 \cdot q + 1 \cdot p = p$$

1. Ley binomial (1,0) o de Bernoulli

Varianza

$$\begin{aligned} var(X) &= \sum_{i=1}^n (x_i - \mu)^2 \cdot P(X = x_i) = \\ &= (0-p)^2 \cdot q + (1-p)^2 \cdot p = pq(p+q) = pq \end{aligned}$$

La varianza será máxima si:

$$\frac{d(pq)}{dp} = \frac{d[p(1-p)]}{dp} = 1 - 2p = 0; \quad p = 0,5$$

2.3. Distribución binomial

Modeliza una serie de fenómenos dicotómicos independientes entre sí.

(nº de veces que ha aparecido el suceso S en n repeticiones independientes del experimento, con $P(S)=p$)

Carlos Montes – uc3m

2.2. Distribución binomial



¿Cuántas caras saldrán en n tiradas de la moneda?

Función de probabilidad, $P(X=r)$

$$\overbrace{\text{E, E, E, E...E, F, F...}}^r \rightarrow r \quad \overbrace{\text{F, F...}}^{n-r} \rightarrow n-r$$

$$p^r (1-p)^{n-r}$$

(si los sucesos son independientes)

2.2. Distribución binomial

Órdenes posibles:

$$C_n^r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r} \quad r = 0, 1, \dots, n$$

2.3. Distribución binomial

Esperanza

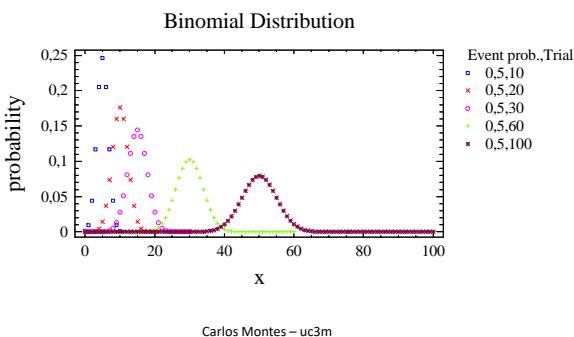
$$E(X) = E\left(\sum_{j=1}^n x_j\right) = \sum_{j=1}^n E(x_j) = np$$

Varianza

$$var(X) = var\left(\sum_{j=1}^n x_j\right) = \sum_{j=1}^n var(x_j) = npq$$

Desviación típica: \sqrt{npq}

2.3. Distribución binomial



La probabilidad de encontrar una persona zurda es de 0,1. En una clase de 20 alumnos hay 3 pupitres para zurdos. Calcule la probabilidad de que no haya suficientes pupitres

X : número de personas zurdas en la clase $\sim B(20, 0.10)$

$$P(X > 3) = 1 - P(X \leq 3) = \\ = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$P(X = 0) = \binom{20}{0} 0.10^0 \cdot 0.90^{20} = 0.12158 \\ \binom{20}{0} = \frac{20!}{0! 20!} = 1$$

$$P(X = 1) = \binom{20}{1} 0.10^1 \cdot 0.90^{19} = 0.27017 \\ \binom{20}{1} = \frac{20!}{1! 19!} = \frac{20 \cdot 19!}{1 \cdot 19!} = 20$$

$$P(X = 2) = \binom{20}{2} 0.10^2 \cdot 0.90^{18} = 0.28518$$

$$\binom{20}{2} = \frac{20!}{2! 18!} = \frac{20 \cdot 19 \cdot 18!}{2 \cdot 18!} = 190$$

$$P(X = 3) = \binom{20}{3} 0.10^3 \cdot 0.90^{17} = 0.190120$$

$$\binom{20}{3} = \frac{20!}{3! 17!} = \frac{20 \cdot 19 \cdot 18 \cdot 17!}{3 \cdot 2 \cdot 17!} = 1140$$

$$P(X > 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

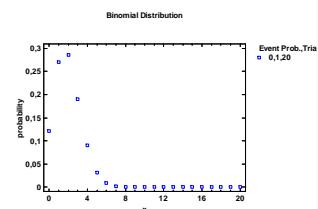
$$\begin{aligned} P(X > 3) &= 1 - [0.12158 + 0.27017 + 0.28518 + 0.190120] = \\ &= 0.13295 \end{aligned}$$

Carlos Montes – uc3m

Cumulative Distribution Distribution Result					
Lower Tail Area (\leq)					
Variable	Distr. 1	Distr. 2	Distr. 3	Distr. 4	Distr. 5
3	0.676927				
Probability Mass ($=$)					
Variable	Distr. 1	Distr. 2	Distr. 3	Distr. 4	Distr. 5
3	0.190120				
Upper Tail Area (\geq)					
Variable	Distr. 1	Distr. 2	Distr. 3	Distr. 4	Distr. 5
3	0.132953				

$$P(X > 3) = 0.132953$$

$$P(X > 3) = 1 - (0.676927 + 0.19012) = 0.132953$$



3.1. Proceso de Poisson. Definición

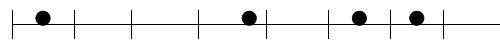
Aparición de sucesos puntuales sobre un soporte continuo, suponiendo que el proceso generador de estos sucesos:

- es estable.
- produce sucesos independientes.

3.2. Distribución de Poisson

Modeliza la aparición de cierto número de sucesos sobre un soporte continuo en un intervalo de longitud fija.

X : número de sucesos en un intervalo de longitud fija T



p pequeña \Rightarrow probabilidad despreciable de aparición de 2 o más sucesos en uno de los n segmentos.

Observamos si aparece o no el suceso estudiado en cada segmento.

3.2. Distribución de Poisson

Es una distribución binomial en la cual:

$$n \rightarrow \infty$$

$$p \rightarrow 0$$

$$np \rightarrow \lambda$$



Simeon Denis Poisson
(1781-1840)

Carlos Montes – uc3m

3.2. Distribución de Poisson

Características

Función de probabilidad

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

$$P(X = r) \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r} \quad r = 0, 1, 2 \dots$$

Tomando límites: $P(X = r) = \frac{\lambda^r}{r!} e^{-\lambda} \quad r = 0, 1, 2 \dots$

(1838)

3.2. Distribución de Poisson

Esperanza

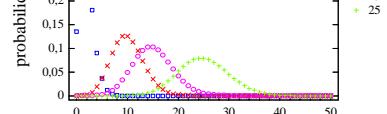
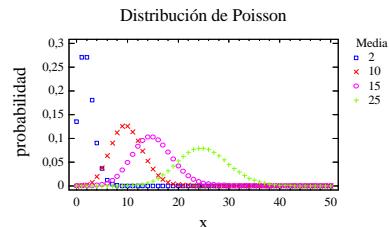
$$E(X) = \lambda$$

Varianza

$$\text{var}(X) = \lambda$$

3.2. Distribución de Poisson

Es una distribución asimétrica, que tiende a la simetría al aumentar λ



3.2. Distribución de Poisson

La suma de varias variables de Poisson independientes, también es una variable de Poisson.

Sea $X_i \sim P(\lambda_i)$, $i = 1, \dots, k$, un conjunto de variables de Poisson independientes.

$$Y = \sum_{i=1}^k X_i \rightarrow P(\lambda^*)$$

$$\lambda^* = \sum_{i=1}^k \lambda_i$$

Carlos Montes – uc3m

Un servidor de una pequeña red recibe una media de 7 accesos por minuto. Suponiendo que los accesos a dicho servidor suceden de forma independiente y con ritmo medio constante, se quiere calcular la probabilidad de que reciba más de 10 accesos en un minuto, porque el servidor tendría entonces un rendimiento deficiente.

X : número de accesos en un minuto

$X \sim P(\lambda = 7)$

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \sum_{r=0}^{10} \frac{7^r}{r!} e^{-7} = 0.09852$$

Cumulative Distribution
Distribution: Poisson

Lower Tail Area (≤)					
Variable	Distr. 1	Distr. 2	Distr. 3	Distr. 4	Distr. 5
10	0.830496				

Probability Mass (≥)					
Variable	Distr. 1	Distr. 2	Distr. 3	Distr. 4	Distr. 5
10	0.0709833				

Upper Tail Area (≥)					
Variable	Distr. 1	Distr. 2	Distr. 3	Distr. 4	Distr. 5
10	0.0985218				

3. Distribución exponencial

Modela el tiempo entre la ocurrencia de dos sucesos consecutivos, siendo estos independientes y estables.

Características

Función de distribución

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}$$

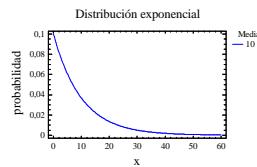
Con λ número medio de sucesos por unidad de tiempo.

3.3. Distribución exponencial

$$f(t) = \frac{dF(t)}{dt} = \lambda e^{-\lambda t}$$

Esperanza

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



Varianza

$$\text{var}(T) = \frac{1}{\lambda^2}$$

Carlos Montes – uc3m

Considerando la red anterior, se pide:

a) Calcular el tiempo medio que transcurre entre dos accesos consecutivos.

$$\lambda = 7 \text{ accesos/minuto}$$

$$E(T) = \frac{1}{\lambda} = \frac{1}{7} = 0.143 \text{ minutos/acceso}$$

b) Probabilidad de que entre dos accesos consecutivos transcurran más de 15 segundos.

$$\begin{aligned} P(T > 0.25 \text{ minutos}) &= 1 - P(T < 0.25) = \\ 1 - (1 - e^{-\lambda t}) &= e^{-\lambda t} = e^{-7 \cdot 0.25} = 0.17 \end{aligned}$$

ej. 35

La duración de ciertos componentes electrónicos sigue una distribución exponencial de media 100 días.

a) ¿Cuál es la probabilidad de que uno de los componentes anteriores dure más de 50 días?

X: duración de un componente → $\exp(\lambda = 1/100)$

$$\begin{aligned} P(X > 50) &= 1 - P(X \leq 50) = 1 - (1 - e^{-\lambda t}) \\ &= 1 - (1 - e^{-50/100}) = e^{-\frac{1}{2}} = 0.607 \end{aligned}$$

b) Un equipo electrónico está formado por 5 componentes de los anteriores, que trabajan de manera independiente, y funciona mientras funcionen correctamente al menos dos de ellos, ¿cuál es la probabilidad de que dicho equipo dure más de 50 días?

Y : Número de componentes que funcionan más de 50 días

$$Y \rightarrow B(5, 0.607)$$

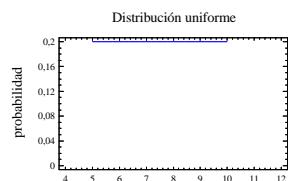
$$P(\text{equipo funcione}) = P(Y \geq 2) = 1 - P(Y = 0) - P(Y = 1) = 1 - \binom{5}{0} 0.607^0 \cdot 0.393^5 - \binom{5}{1} 0.607^1 \cdot 0.393^4 = 0.9182$$

Carlos Montes – uc3m

4. Distribución uniforme

Características

$$f(x) = \frac{1}{b-a} \quad a \leq x \leq b$$

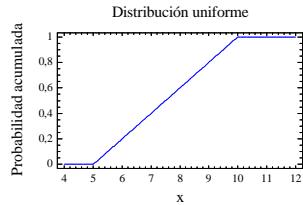


4. Distribución uniforme

Función de distribución

$$F(x) = P(X \leq x) = \int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a}$$

$$a \leq x \leq b$$



4. Distribución uniforme

Esperanza

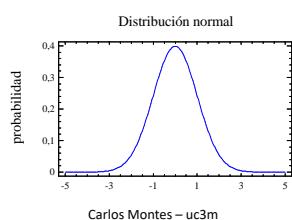
$$E(X) = \frac{a+b}{2}$$

$$\text{Varianza} \quad \text{var}(X) = \frac{(b-a)^2}{12}$$

4. Distribución normal

Es la distribución con función de densidad:

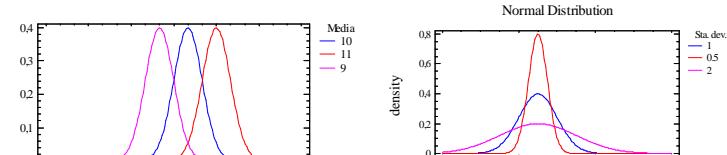
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad -\infty \leq x \leq \infty$$



4. Distribución normal

La distribución normal depende de dos parámetros:

- ✓ Media μ
- ✓ Desviación típica σ

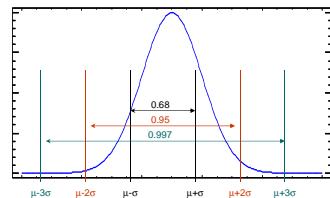


4. Distribución normal

- Es uno de los modelos más frecuentes para describir variables reales continuas.
- Simétrica, centrada en la media μ que es su mediana y su moda.
- Forma de campana.
- Coeficiente de apuntamiento igual a 3.

4. Distribución normal

Regla empírica



4. Distribución normal

- Se ajusta a lo observado en muchos procesos de medición, si no influyen los errores sistemáticos.
- La normal de media 0 y desviación típica 1 se denomina:
 - ❖ normal tipificada (Z)
 - ❖ normal estándar
 - ❖ normal (0,1)y su función de distribución está tabulada.

Carlos Montes – uc3m

4. Distribución normal

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$

↓

$$Z \rightarrow N(0,1)$$

La longitud L en milímetros, de las piezas fabricadas en un proceso es una variable aleatoria que se distribuye según una $N(32, 0.3^2)$, considerándose aceptables aquellas cuya medida se encuentra dentro del intervalo $(31.1, 32.6)$. Calcule la probabilidad de que una pieza elegida al azar sea aceptable.

$$L \rightarrow N(32, 0.3^2)$$

$$P(\text{acceptable}) = P(31.1 < L < 32.6)$$

$$P\left(\frac{31.1 - 32}{0.3} < Z < \frac{32.6 - 32}{0.3}\right) = P(-3 < Z < 2)$$

$$\begin{aligned}F(2) - F(-3) &= F(2) - [1 - F(3)] = F(2) - 1 + F(3) = \\&= 0.9772 - 1 + 0.9987 = 0.9759\end{aligned}$$

4. Distribución normal



$$F(2) - 1 + F(3)$$

0.9772 0.9987

0.9759

6. Distribución lognormal

Se llama *lognormal* a la variable aleatoria cuyo logaritmo neperiano es normal:

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

$$E(X) = e^{\mu + \frac{1}{2}\sigma^2} \quad \text{var}(X) = \left(e^{\sigma^2} - 1\right) \cdot e^{2\mu + \sigma^2}$$

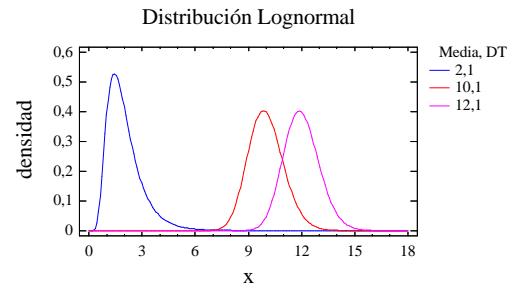
6. Distribución lognormal

Si la variable y puede considerarse como un producto de variables aleatorias:

$$y = x_1 \cdot x_2 \cdot x_3 \dots x_n$$

su logaritmo neperiano seguirá una distribución normal (distribución lognormal o de Mac Alister).

6. Distribución lognormal



7.1. Teorema central del Límite. Definición

La suma de un conjunto de variables aleatorias se aproxima, al aumentar el número de variables, a una variable aleatoria **normal** independientemente de cual sea la distribución de esas variables.

Carlos Montes – uc3m

7.1. Teorema central del Límite. Definición

Sean x_1, x_2, \dots, x_n v.a. independientes con media μ_i , desviación típica σ_i y distribución CUALQUIERA

$$Y = x_1 + x_2 + \dots + x_n \quad \text{Al crecer } n:$$
$$\frac{Y - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} \rightarrow N(0,1) \quad Y \rightarrow N\left(\sum \mu_i, \sqrt{\sum \sigma_i^2}\right)$$

7.1. Teorema central del Límite. Definición

$$Y \rightarrow N\left(\sum \mu_i, \sqrt{\sum \sigma_i^2}\right)$$

La desviación típica es la raíz cuadrada de la suma de varianzas, NO la suma de las desviaciones típicas.

7.2. Teorema central del Límite. Aplicaciones

Normal

$$E(x_i) = p \quad Var(x_i) = pq$$

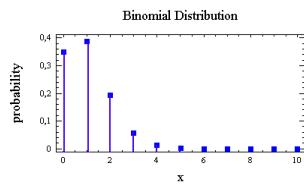
Binomial

Aplicando el Teorema Central del Límite:

$$Y \rightarrow N\left(np, \sqrt{npq}\right)$$

$$n > 30 \\ npq > 5$$

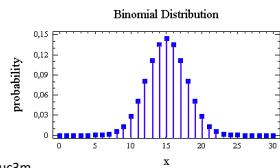
7.2. Teorema central del Límite. Aplicaciones



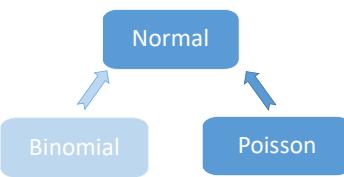
$$npq = 0,9$$

$$npq = 7,5$$

Carlos Montes – uc3m



7.2. Teorema central del Límite. Aplicaciones



Sea $Y(0,T)$ una variable de Poisson que cuenta el nº de sucesos en el intervalo $(0,T)$

$$Y(0,T) = x_1(0,t_1) + x_2(t_1,t_2) + \dots + x_n(t_{n-1},T)$$

Para $\lambda > 5$
podemos aproximar esta variable mediante una distribución Normal

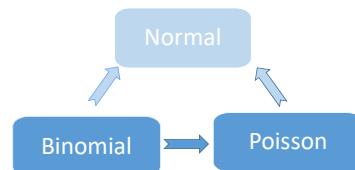
7.2. Teorema central del Límite. Aplicaciones

Para $\lambda > 5$
podemos aproximar esta variable mediante una distribución Normal

$$x_p \rightarrow P(\lambda) \quad \lambda > 5$$

$$x_n \rightarrow N(\lambda, \sqrt{\lambda})$$

7.2. Teorema central del Límite. Aplicaciones

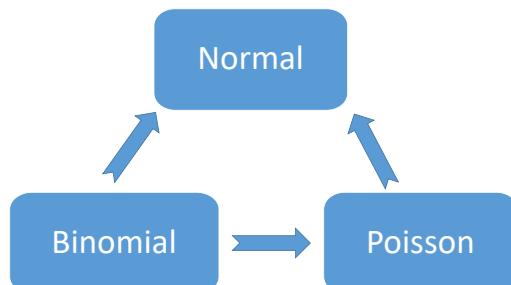


La aproximación es válida cuando:

$$\begin{aligned} np &> 1 \\ p &< 0,1 \end{aligned}$$

$$\lambda = np$$

7.2. Teorema central del Límite. Aplicaciones



Carlos Montes – uc3m

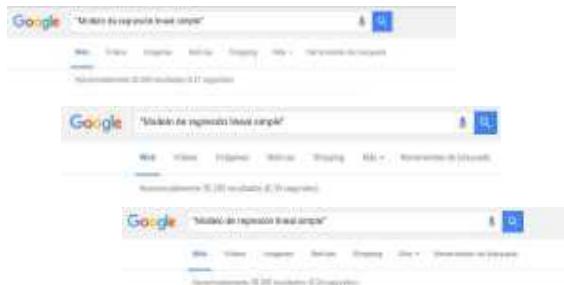
8. Modelo de regresión lineal simple

La recta de regresión de y sobre x es de la forma:

$$y_i = a + bx_i + e_i \rightarrow \begin{array}{l} \text{Término de error} \\ \downarrow \\ \text{Valor observado de la variable } y \text{ para el individuo } i\text{-ésimo} \end{array} \quad \begin{array}{l} \text{Valor observado de la variable } x \text{ para el individuo } i\text{-ésimo} \end{array}$$

Si queremos definir un tipo de relación válido para toda la población, encontramos numerosos factores que no controlamos.

8. Modelo de regresión lineal simple



8. Modelo de regresión lineal simple

Si fijamos el valor de la variable X en $X = x_i$ repitiendo el experimento, observaremos valores diferentes debidos al efecto de las variables recogidas en el término e .

$$y_i = a + bx_i + e_i \quad \begin{array}{l} \downarrow \text{aleatorio (ruido)} \\ \downarrow \text{fijo} \end{array}$$

8. Modelo de regresión lineal simple

Si asumimos que todas las variables que influyen sobre Y lo hacen de forma lineal (aditiva):

$$Y = a + bX + (c_1Z_1 + c_2Z_2 + c_3Z_3 + \dots)$$

e

Por el Teorema del Límite Central, e seguirá una distribución normal.

Carlos Montes – uc3m

8. Modelo de regresión lineal simple

$$E(e) = 0$$

$$\text{var}(e) = \sigma^2 \text{ (constante)}$$

$$e \rightarrow N(0, \sigma^2)$$

$$E(Y \setminus X = x_i) = E(a + bx_i + e) = a + bx_i + E(e) = a + bx_i$$

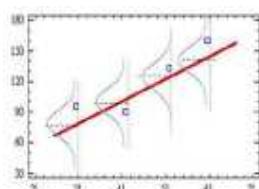
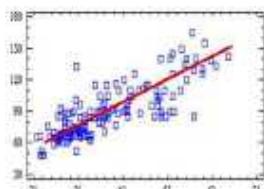
$$\text{var}(Y \setminus X = x_i) = \text{var}(a + bx_i + e) = \text{var}(e)$$

$$Y \rightarrow N(a + bx_i, \sigma^2)$$

8. Modelo de regresión lineal simple

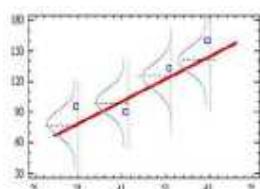
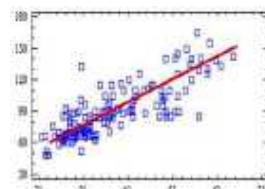
Cada punto y_i que observamos se interpreta como un valor al azar de la normal.

$$y_i \rightarrow N(a + bx_i, \sigma^2)$$



8. Modelo de regresión lineal simple

Suponemos que el "ruido" es homogéneo a lo largo de la recta (varianza constante u *homocedasticidad*).



Sea el modelo de regresión simple $Y=50 + 2X + e$, siendo $e = N(0, 5^2)$. Calcule la probabilidad de que Y sea mayor que 160 para los casos siguientes:

- a) $X=60$
- b) $X=50$

Carlos Montes – uc3m

$$P(Y > 160) \quad Y=50 + 2X + e,$$

a) Necesitamos el modelo que sigue Y

$$Y \rightarrow N(a + bx_i; \sigma^2)$$

$$Y = 50 + 2X + e = 50 + 2 \cdot 60 + e = 170 + e$$

$$Y \rightarrow N(170; 5^2)$$

$$P(Y > 160) = P\left(Z > \frac{160 - 170}{5}\right) =$$

$$= P(Z > -2) = P(Z < 2) = 0.9772$$

$$P(Y > 160) \quad Y=50 + 2X + e,$$

$$b) \quad Y \rightarrow N(a + bx_i; \sigma^2)$$

$$Y = 50 + 2X + e = 50 + 2 \cdot 50 + e = 150 + e$$

$$Y \rightarrow N(150; 5^2)$$

$$P(Y > 160) = P\left(Z > \frac{160 - 150}{5}\right) =$$

$$= P(Z > 2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228$$

Parte VI

Tema 6. Introducción a la inferencia estadística

Tema 6 Introducción a la inferencia estadística

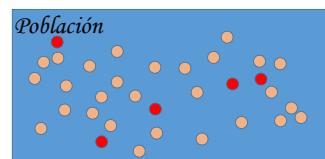
Carlos Montes – uc3m

1. Introducción
2. Distribución muestral de estimadores
 - 2.1. Concepto
 - 2.2. Distribución muestral de la media
3. Estimación
 - 3.1. Concepto
 - 3.2. Propiedades
4. Método de los momentos
5. Diagnóstico y crítica del modelo
6. Transformaciones para mejorar la normalidad

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



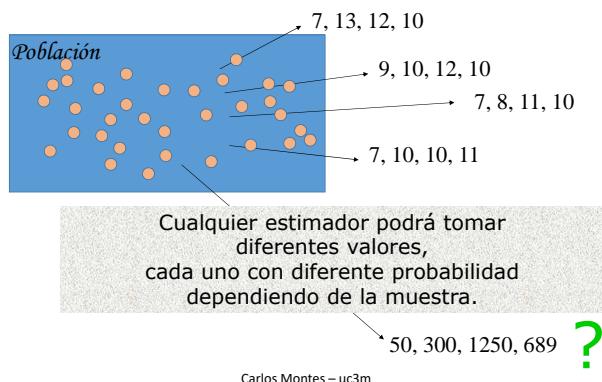
X_1, X_2, \dots, X_n
variables aleatorias independientes e idénticas

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



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 :

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \dots + E(X_n)}{n} = \\ &= \frac{n\mu}{n} = \mu \end{aligned}$$

2.2. Distribución muestral de la media

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

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

2.2. Distribución muestral de la media

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

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

Carlos Montes – uc3m

2.2. Distribución muestral de la media

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

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

2.2. Distribución muestral de la media

Pero para cualquier n , **si x es $N(\mu, \sigma)$**

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

Muy importante para inferencia

3.1. Estimación. Concepto

Los parámetros ($\mu, \sigma, \lambda, \dots$) 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.

$$\hat{\mu}$$

Carlos Montes – uc3m

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

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

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

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

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

3.2. Estimación. Propiedades

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

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

3.2. Estimación. Propiedades

Puede demostrarse que:

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

Carlos Montes – uc3m

3.2. Estimación. Propiedades

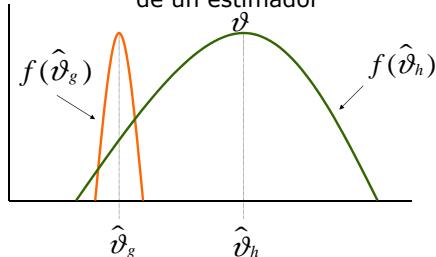
Así, un ECM mínimo supone:

sesgo mínimo \Rightarrow insesgo
varianza mínima \Rightarrow eficiencia o precisión

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

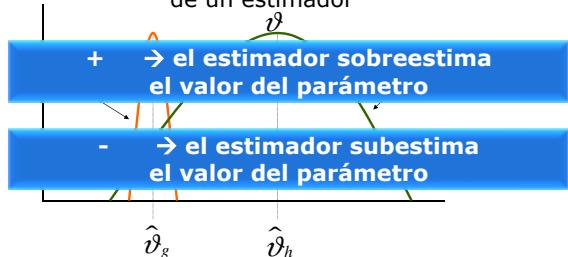
3.2. Estimación. Propiedades

Nos mide el error sistemático de un estimador

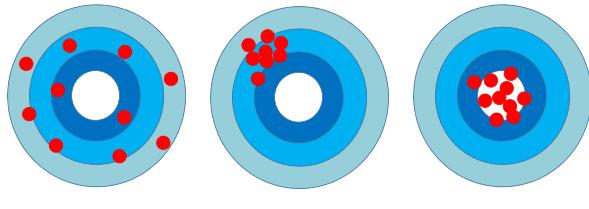


3.2. Estimación. Propiedades

Nos mide el error sistemático de un estimador



3.2. Estimación. Propiedades



inesesgado
impreciso

sesgado
preciso

inesesgado
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 :

$$\begin{aligned}\widehat{\mu}_1 &= \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3 \\ \widehat{\mu}_2 &= \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3 \\ \widehat{\mu}_3 &= \frac{1}{8}X_1 + \frac{3}{8}X_2 + \frac{1}{2}X_3\end{aligned}$$

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

Estudiamos su error cuadrático medio.

$$\begin{aligned}E(\widehat{\mu}_1) &= \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(X_3) = \frac{1}{3} \cdot 3\mu = \mu \\ E(\widehat{\mu}_2) &= \frac{1}{4}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{4}E(X_3) = \frac{\mu + 2\mu + \mu}{4} = \mu \\ E(\widehat{\mu}_3) &= \frac{1}{8}E(X_1) + \frac{3}{8}E(X_2) + \frac{1}{2}E(X_3) = \frac{\mu + 3\mu + 4\mu}{8} = \mu\end{aligned}$$

Efectivamente, son inexactos.

$$\begin{aligned}ECM(\widehat{\theta}) &= \boxed{[sesgado(\widehat{\theta})]^2 + var(\widehat{\theta})} \\ var(\widehat{\mu}_1) &= var\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) = \frac{1}{9}var(X_1) + \frac{1}{9}var(X_2) + \frac{1}{9}var(X_3) = \frac{3}{9} \cdot \sigma^2 = \frac{1}{3} \\ var(\widehat{\mu}_2) &= var\left(\frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3\right) = \frac{1}{16}var(X_1) + \frac{1}{4}var(X_2) + \frac{1}{16}var(X_3) = \\ &= \frac{\sigma^2 + 4\sigma^2 + \sigma^2}{16} = \frac{6}{16}\sigma^2 = \frac{3}{8} \\ var(\widehat{\mu}_3) &= var\left(\frac{1}{8}X_1 + \frac{3}{8}X_2 + \frac{1}{2}X_3\right) = \frac{1}{64}var(X_1) + \frac{9}{64}var(X_2) + \frac{1}{4}var(X_3) = \\ &= \frac{\sigma^2 + 9\sigma^2 + 16\sigma^2}{64} = \frac{26}{64}\sigma^2 = \frac{13}{32}\end{aligned}$$

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
 ...*

Carlos Montes – uc3m

La duración de un sistema hasta que se produce un fallo por causas fortuitas se puede modelizar con una distribución $\exp(\lambda)$. 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.

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

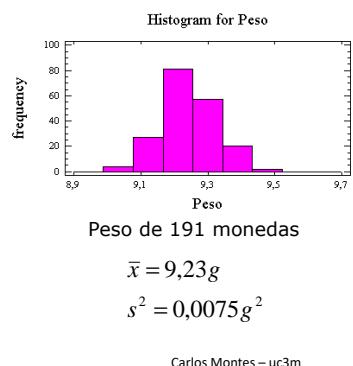
$$\bar{x} = \frac{18 + 94 + 22 + 143 + 114}{5} = 78.2 \text{ horas}$$

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

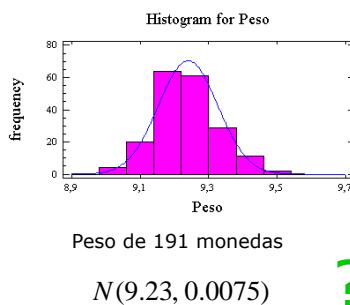
5. Diagnosis y crítica del modelo



5. Diagnosis y crítica del modelo



5. Diagnosis y crítica del modelo



5. Diagnosis y crítica del modelo

Test de la chi-cuadrado



5. Diagnosis y crítica del modelo

Procedimiento

- Se toma una muestra de tamaño $n \geq 25$.
- Se agrupan los datos en k clases ($k \geq 5$) 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 .

5. Diagnosis y crítica del modelo

Se calcula el siguiente estadístico:

$$\chi^2_0 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \begin{array}{l} \text{Frecuencias observadas} \\ \text{Frecuencias esperadas} \\ \text{según el modelo} \end{array}$$

Resume la discrepancia entre datos y modelo:

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

Carlos Montes – uc3m

5. Diagnosis y crítica del modelo

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

5. Diagnosis y crítica del modelo

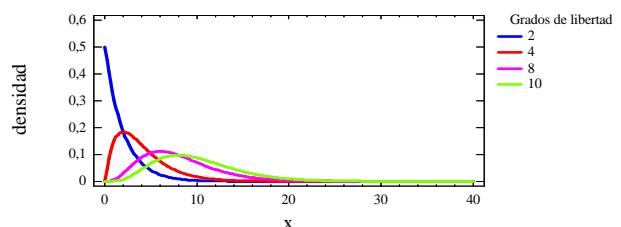
¿Cómo valorar el modelo?

El estadístico calculado seguirá una distribución:

$$\chi^2$$

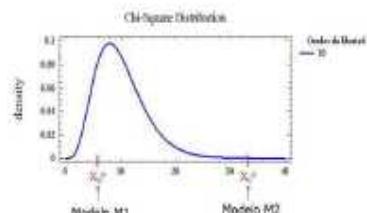
Depende de sus g grados de libertad
(si conocemos los parámetros del modelo, $= k - 1$.
Si hay que estimar v parámetros, $= k - v - 1$)

5. Diagnosis y crítica del modelo



Para $n > 30$ es prácticamente una normal

5. Diagnosis y crítica del modelo



* La probabilidad de obtener ese valor de χ^2_0 si M2 es cierto, es muy baja.

* M1 es adecuado, M2 no lo es.

Carlos Montes – uc3m

5. Diagnosis y crítica del modelo

* Los programas informáticos proporcionan el área que queda a la derecha de χ^2_0 en la distribución (**p-valor**).

* En general, si el valor de χ^2_0 está en la zona de la cola de la derecha, el modelo no es adecuado (área bajo la curva pequeña \Rightarrow probabilidad pequeña de ocurrir si el modelo es cierto).

5. Diagnosis y crítica del modelo

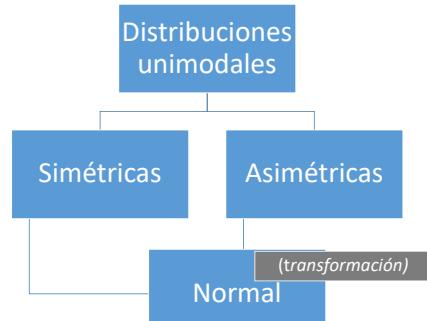
*** Rechazaremos un modelo
si el p-valor <0,05**

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

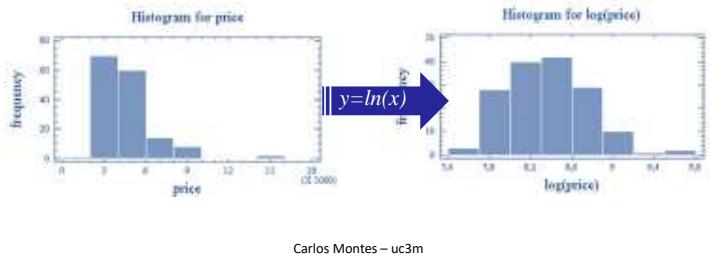
Chi-Square = 4,46353 with 3 d.f. P-Value = **0,215564**

6. Transformaciones para mejorar la normalidad



6. Transformaciones para mejorar la normalidad

Datos con asimetría positiva



6. Transformaciones para mejorar la normalidad

+ efecto

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

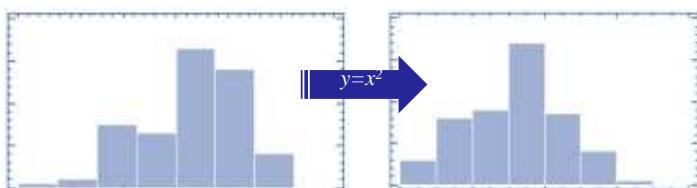
Comprimen la escala en los valores altos y la expanden en los valores bajos.

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:

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

Parte VII

Tema 7. Inferencia con muestras grandes

Tema 7

Inferencia con muestras grandes

Carlos Montes – uc3m

1. Concepto de intervalo de confianza
2. Intervalo de confianza para la media
 - 2.1. Varianza poblacional conocida
 - 2.2. Varianza poblacional desconocida
3. T de Student
4. Otros intervalos de confianza
 - 4.1. Para la varianza de poblaciones normales
 - 4.2. Para una proporción
 - 4.3. Para la λ de Poisson
5. Determinación del tamaño muestral
6. Contraste de hipótesis
 - 6.1. Concepto
 - 6.2. Método
 - 6.3. p-valor
7. Contraste para la media
8. Contrastos e intervalos

1. Concepto de intervalo de confianza

Intervalo de confianza para el parámetro θ con nivel de confianza $(1-\alpha)$ es el intervalo $[\theta_1(X), \theta_2(X)]$ tal que:

$$P[\theta_1(X) \leq \theta \leq \theta_2(X)] = 1 - \alpha$$

La probabilidad de que el **intervalo aleatorio** (θ_1, θ_2) contenga al verdadero valor del parámetro es $1-\alpha$

1. Concepto de intervalo de confianza



De cada 100 intervalos construidos a partir de 100 muestras tendrán el $(1-\alpha)100\%$ contendrán el verdadero valor del parámetro

1. Concepto de intervalo de confianza

Dado un α , o nivel de significación se trata de encontrar un intervalo centrado en el parámetro que contenga su verdadero valor el $(1-\alpha)100\%$ de las veces.

Carlos Montes – uc3m

2.1. Intervalo de confianza para la media con σ conocida

Por el teorema del límite central sabemos que:

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

Tipificando:

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

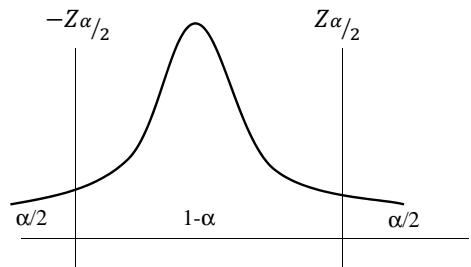
2.1. Intervalo de confianza para la media con σ conocida

Definimos $Z_{\alpha/2}$ como el valor de Z tal que:

$$P(Z > Z_{\alpha/2}) = \frac{\alpha}{2}$$

$$P(Z < -Z_{\alpha/2}) = \frac{\alpha}{2}$$

2.1. Intervalo de confianza para la media con σ conocida



$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = (1 - \alpha)$$

2.1. Intervalo de confianza para la media con σ conocida

$$P\left(-Z_{\alpha/2} < Z < Z_{\alpha/2}\right) = 1 - \alpha$$

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

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

Carlos Montes – uc3m

2.1. Intervalo de confianza para la media con σ conocida

$$P(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}) = 1 - \alpha$$

$$P(\bar{X} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \bar{X} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}) = 1 - \alpha$$

$$\mu \in \left(\bar{X} \pm Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right)$$

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

Se pide:

a) Calcular un intervalo de confianza del 95% para la media poblacional μ .

b) Calcular un intervalo de confianza del 90% para la media poblacional μ .

$$\mu \in \left(\bar{X} \pm Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right)$$

$$\mu \in \left(160 \pm Z_{\alpha/2} \cdot 10/12\right)$$

a) $1 - \alpha = 0.95 \quad \alpha = 0.05$

$$P(Z > Z_{\alpha/2}) = \frac{0.05}{2} = 0.025$$

$$P(Z < Z_{\alpha/2}) = 1 - 0.025 = 0.975$$

P(Z<z)	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,963	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,970	0,971	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,977	0,977

Carlos Montes – uc3m

$$Z\alpha_{/2} = 1.96$$

$$\mu \in (160 \pm 1.96 \cdot 10/12) = (158.36; 161.63)$$

$$\mu \in (160 \pm Z\alpha_{/2} \cdot 10/12)$$

$$b) \quad 1 - \alpha = 0.90 \quad \alpha = 0.10$$

$$P(Z > Z\alpha_{/2}) = \frac{0.10}{2} = 0.05$$

$$P(Z < Z\alpha_{/2}) = 1 - 0.05 = 0.95$$

P(Z<z)	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,963	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977

$$Z_{\alpha/2} = 1.64$$

$$\mu \in (160 \pm 1.64 \cdot \frac{10}{\sqrt{12}}) = (158.63; 161.36)$$

(158.63; 161.36)

Carlos Montes – uc3m

2.2. Intervalo de confianza para la media con σ desconocida

σ^2 suele ser desconocido, con lo que lo sustituimos por una estimación.

$$\hat{\sigma}^2 = s^2$$

Pero es un estimador sesgado, pudiéndose demostrar que:

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

2.2. Intervalo de confianza para la media con σ desconocida

$$\text{sesgo}(s^2) = E(s^2) - \sigma^2$$

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

El sesgo es negativo $\Leftrightarrow s^2$ subestima la verdadera varianza.

2.2. Intervalo de confianza para la media con σ desconocida

Para corregir el sesgo:

$$E \left(s^2 \cdot \frac{n}{n-1} \right) = \frac{n}{n-1} \cdot E(s^2) = \frac{n}{n-1} \cdot \sigma^2 \cdot \frac{n-1}{n} = \sigma^2$$

Pero ¿qué es $(s^2 \cdot \frac{n}{n-1})$?

$$(s^2 \cdot \frac{n}{n-1}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \hat{s}^2$$

2.2. Intervalo de confianza para la media con σ desconocida

Así, si n es grande, el intervalo de confianza para la media poblacional μ es:

$$\bar{x} \pm Z_{\alpha/2} \frac{\hat{s}}{\sqrt{n}}$$

\hat{s}^2

*cuasivarianza
varianza corregida*

Carlos Montes – uc3m

2.2. Intervalo de confianza para la media con σ desconocida

Si la muestra no es grande, pero la población es normal:

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

t de Student

2.2. Intervalo de confianza para la media con σ desconocida

Statgraphics calcula siempre estos intervalos :

- Para muestras grandes valen para cualquier población.
- Para muestras pequeñas, solo para poblaciones normales.

3. T de Student

William Gosset "Student" (1876-1937)

Definida en 1908, es la distribución:

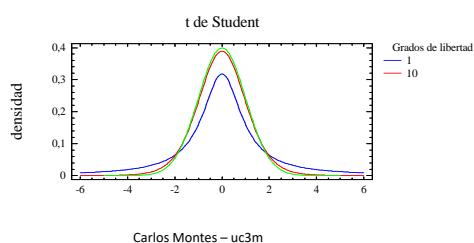


$$t_n = \frac{Z}{\sqrt{\frac{1}{n} \chi_n^2}}$$

Con $Z \rightarrow N(0,1)$

3. T de Student

Es una variable simétrica, con mayor dispersión que la normal estándar, a la que tiende rápidamente al aumentar n .



4.1. Intervalo de confianza para la varianza de poblaciones normales

$$\frac{\hat{s}^2}{\sigma^2} \rightarrow \chi_{n-1}^2 \quad \text{Lema de Fisher-Cochran}$$



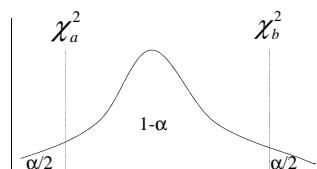
$$\frac{(n-1)\hat{s}^2}{\sigma^2} \rightarrow \chi_{n-1}^2$$

(1890-1962) (1909-1980)

4.1. Intervalo de confianza para la varianza de poblaciones normales

Llamando χ_a^2 y χ_b^2 a los valores χ_{n-1}^2 que dejan entre sí el $1-\alpha$ de la distribución:

$$P\left(\chi_a^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_b^2\right) = 1 - \alpha$$



4.1. Intervalo de confianza para la varianza de poblaciones normales

El intervalo de confianza buscado queda:

$$\frac{(n-1)\hat{s}^2}{\chi_a^2} \geq \sigma^2 \geq \frac{(n-1)\hat{s}^2}{\chi_b^2}$$

Para simplificar, se suele tomar un intervalo simétrico:

$$P(\chi_{n-1}^2 \geq \chi_b^2) = \frac{\alpha}{2}$$

$$P(\chi_{n-1}^2 \leq \chi_a^2) = \frac{\alpha}{2}$$

4.1. Intervalo de confianza para la varianza de poblaciones normales

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

Carlos Montes – uc3m

4.2. Intervalo de confianza para una proporción

Se desea estimar la proporción p de elementos de la población que poseen un atributo determinado.

$$\bar{x} = p \quad \bar{x} \pm Z_{\alpha/2} \frac{\hat{s}}{\sqrt{n}} \quad \hat{s}^2 = \hat{p}\hat{q}$$

$$p \in \left(\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

ej. 48

Para estimar el porcentaje de tornillos defectuosos producidos por una máquina, se extrae una muestra de 400 unidades, de las cuales 8 resultan ser defectuosas. Construya un intervalo de confianza al 95% para el porcentaje de tornillos defectuosos fabricados.

$$n = 400 \quad \hat{p} = \frac{8}{400} = 0.02 \quad \hat{q} = 1 - 0.02 = 0.98$$

El intervalo de confianza para el porcentaje de tornillos defectuosos es:

$$p \in \left(\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) \quad p \in \left(0.02 \pm 1.96 \sqrt{\frac{0.02 \cdot 0.98}{400}} \right)$$

$$Z_{\alpha/2} = 1.96 \quad (6.28 \cdot 10^{-3}, 0.03372)$$

4.3. Intervalo de confianza para la λ de Poisson

Sean X_1, X_2, \dots, X_n m.a.s de una distribución de Poisson.

$$\frac{\bar{x} - \lambda}{\sqrt{s/\sqrt{n}}} \rightarrow N(0,1)$$

Como $\lambda = \text{var}(X) = E(X)$

$$\frac{\bar{x} - \lambda}{\sqrt{\bar{x}/\sqrt{n}}} \rightarrow N(0,1)$$

Carlos Montes – uc3m

4.3. Intervalo de confianza para la λ de Poisson

Un intervalo de confianza $(1-\alpha)$ será aquel para el cual:

$$P\left(-Z_{\alpha/2} \leq \frac{\bar{x} - \lambda}{\sqrt{\bar{x}/\sqrt{n}}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

El intervalo de confianza queda:

$$\lambda \in \left(\bar{x} \pm Z_{\alpha/2} \sqrt{\frac{\bar{x}}{n}} \right)$$

5. Determinación del tamaño muestral

Función de la precisión que se quiera conseguir.

a) Tamaño muestral para la estimación de la media

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \bar{x} \pm L$$

$$L = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \frac{Z_{\alpha/2}^2 \sigma^2}{L^2}$$

2.1. Intervalo de confianza para la media con σ conocida

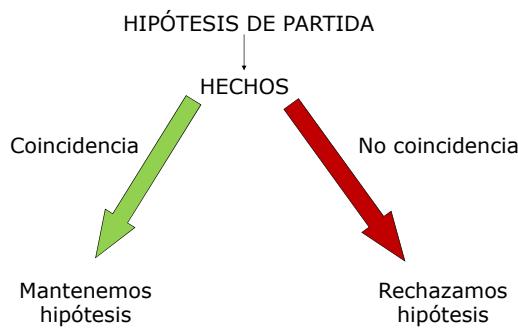
b) Tamaño muestral para la estimación de una proporción

$$\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad \hat{p} \pm L \quad L = Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Como p es desconocido, tomamos el caso más desfavorable: $p=0.5$; $q=0.5$

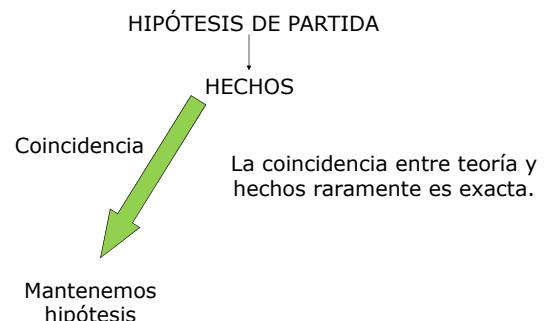
$$n = \frac{Z_{\alpha/2}^2}{4L^2}$$

6.1. Contraste de hipótesis. Concepto

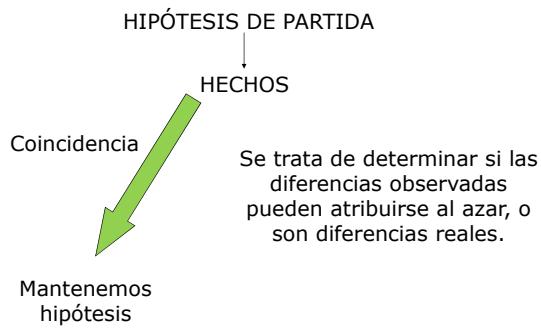


Carlos Montes – uc3m

6.1. Contraste de hipótesis. Concepto



6.1. Contraste de hipótesis. Concepto



6.1. Contraste de hipótesis. Concepto

En un contraste de hipótesis se comparan 2 alternativas:

H_0 : hipótesis nula o "neutra"

- La que se contrasta (la más simple).
- La que mantendremos a no ser que los datos indiquen su falsedad.
- Aceptarla no equivale a probarla, sino a no rebatirla.

6.1. Contraste de hipótesis. Concepto

H₁: hipótesis alternativa

- La que aceptamos si los datos parecen incompatibles con la hipótesis nula.

Carlos Montes – uc3m

6.1. Contraste de hipótesis. Concepto

$$\begin{aligned} H_0 : \theta = \theta_0 & \quad \text{La hipótesis nula debe tener siempre el signo =} \\ H_1 : \theta \neq \theta_0 & \quad \text{Contraste bilateral} \\ H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \\ \text{o bien:} \\ H_1 : \theta < \theta_0 & \quad \text{La hipótesis alternativa es lo que nos preguntamos.} \\ & \quad \left. \begin{array}{l} H_1 : \theta > \theta_0 \\ \text{o bien:} \\ H_1 : \theta < \theta_0 \end{array} \right\} \quad \text{Contrastes unilaterales} \end{aligned}$$

6.2. Contraste de hipótesis. Método

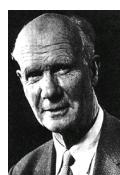
Fisher, Neyman y Pearson (1920-33)



(1890-1962)



(1894-1981)



(1895-1980)

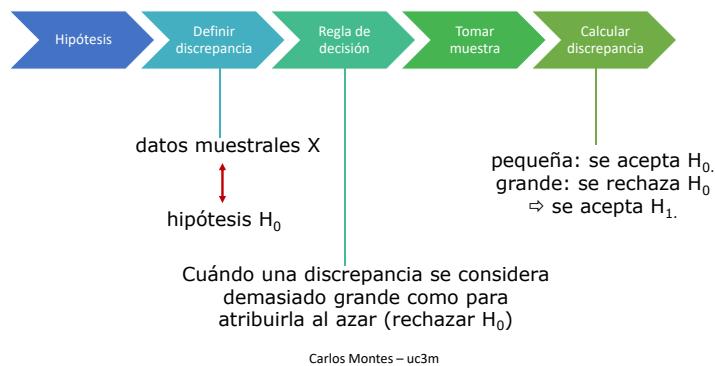
6.2. Contraste de hipótesis. Método

Su lógica es similar
a la de un juicio penal:

el acusado es inocente
si no se demuestra lo contrario.

la hipótesis nula es verdadera
si no se demuestra lo contrario.

6.2. Contraste de hipótesis. Método



6.2. Contraste de hipótesis. Método

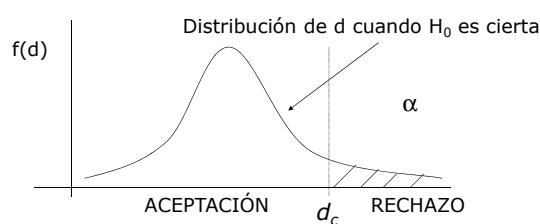
Se consideran discrepancias "demasiado grandes" aquellas que tienen una probabilidad "demasiado pequeña" de ocurrir si H_0 es cierta.

Si la discrepancia se considera demasiado grande como para atribuirlo al azar (rechazar H_0) se acepta H_1 . Si la discrepancia es pequeña: se rechaza H_0 y se acepta H_0 .

6.2. Contraste de hipótesis. Método

Si la discrepancia observada en la muestra, \hat{d} cae dentro de la región de rechazo, rechazaremos H_0 .

En caso contrario, la aceptaremos.



6.2. Contraste de hipótesis. Método

Podemos cometer dos tipos de errores

	H_0 cierta	H_0 falsa
Rechazamos H_0	Error tipo I (α)	
Aceptamos H_0		Error tipo II

$$\alpha = P\left(\text{rechazar } H_0 \middle/ H_0 \text{ cierta}\right)$$

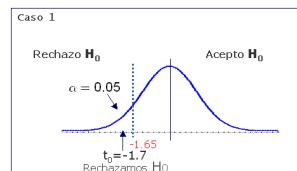
6.2. Contraste de hipótesis. Método

Objeciones:

- El resultado puede ser distinto al emplear α ligeramente diferentes.
- Al emplear un α prefijado no podemos saber el grado de evidencia que la muestra indica a favor o en contra de H_0 .

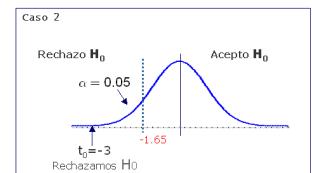
Carlos Montes – uc3m

6.2. Contraste de hipótesis. Método



... en el caso 2
tenemos
mayor
seguridad.

Aunque
rechazamos
 H_0 en los dos
casos...



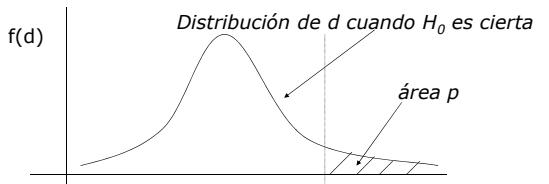
6.3. p-valor

**Para determinar la región de rechazo,
empleamos en lugar del nivel de
significación α
el nivel crítico p (p-value)**

Es la probabilidad de obtener
una discrepancia mayor o igual
que la observada en la muestra
cuando H_0 es cierta.

6.3. p-valor

- Es el mínimo nivel de significación que nos llevaría a rechazar la hipótesis nula.



6.3. p-valor

- $p < 0,05$ ($0,01$): existe muy poca evidencia en la muestra a favor de la hipótesis, luego RECHAZAMOS.

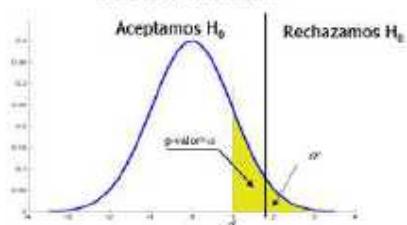
Probabilidad pequeña \Rightarrow discrepancia grande
(supuesta H_0 cierta)

Carlos Montes – uc3m

6.3. p-valor

$$H_0: \theta \leq \theta_0; H_1: \theta > \theta_0$$

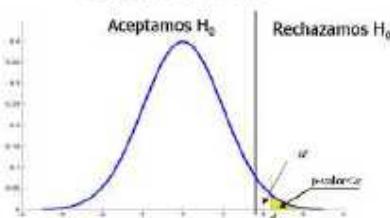
Distribución de referencia



6.3. p-valor

$$H_0: \theta \leq \theta_0; H_1: \theta > \theta_0$$

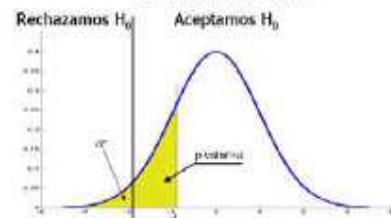
Distribución de referencia



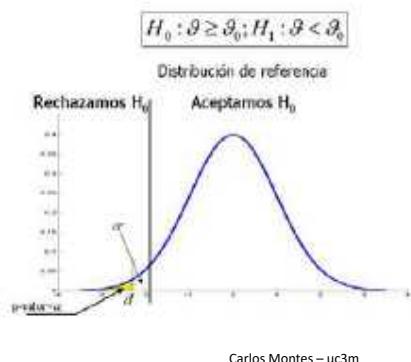
6.3. p-valor

$$H_0: \theta \geq \theta_0; H_1: \theta < \theta_0$$

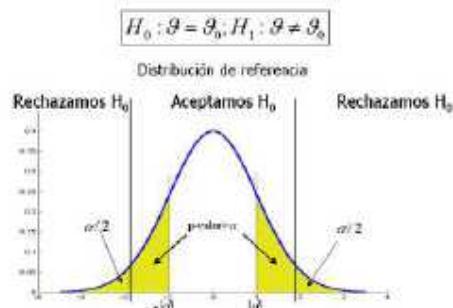
Distribución de referencia



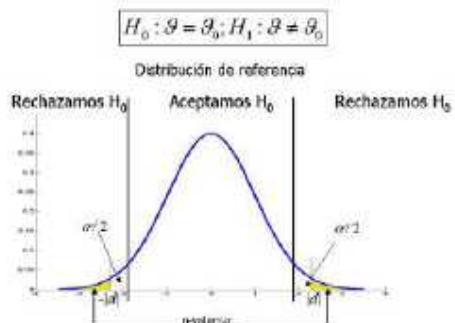
6.3. p-valor



6.3. p-valor



6.3. p-valor

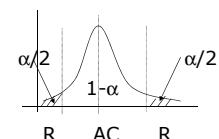


7. Contraste para la media

Se desea contrastar la hipótesis de que la media de una distribución normal es μ_0

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$



Contraste bilateral

7. Contraste para la media

Para cualquier variable X de media μ y varianza σ^2 , **si el tamaño muestral es suficientemente grande** se cumple que:

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

Y además:

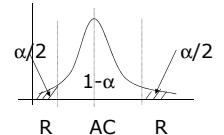
$$T = \frac{\bar{x} - \mu}{\hat{s} / \sqrt{n}} \rightarrow N(0, 1)$$

Carlos Montes – uc3m

7. Contraste para la media

Rechazamos H_0 si:

$$\frac{\bar{x} - \mu_0}{\hat{s} / \sqrt{n}} \begin{cases} > Z_{\alpha/2} \\ < -Z_{\alpha/2} \end{cases}$$



7. Contraste para la media

$$\begin{aligned} H_0: \mu &\leq \mu_0 \\ H_1: \mu &> \mu_0 \end{aligned}$$

Contraste unilateral

Rechazamos H_0 si:

$$\frac{\bar{x} - \mu_0}{\hat{s} / \sqrt{n}} > Z_\alpha$$

7. Contraste para la media

$$\begin{aligned} H_0: \mu &\geq \mu_0 \\ H_1: \mu &< \mu_0 \end{aligned}$$

Contraste unilateral

Rechazamos H_0 si:

$$\frac{\bar{x} - \mu_0}{\hat{s} / \sqrt{n}} < -Z_\alpha$$

7. Contraste para la media

Un intervalo de confianza
con nivel $1-\alpha$
y un contraste
usan la misma información.

La realización
de un contraste de hipótesis bilateral con
nivel de significación α
es equivalente a realizar
un intervalo de confianza de nivel $(1 - \alpha)$.

Carlos Montes – uc3m

Parte VIII

Tema 8. Comparación de población

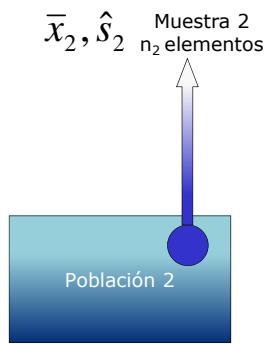
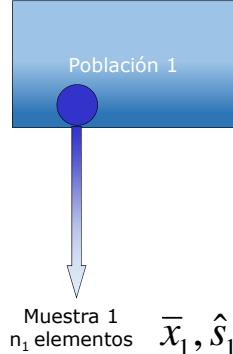
Tema 8

Comparación de poblaciones

Carlos Montes – uc3m

1. Introducción
2. Intervalo de confianza para la diferencia de medias
 - 2.1. Varianzas conocidas
 - 2.2. Varianzas estimadas
3. Contraste para la diferencia de medias
 - 3.1. Varianzas conocidas
 - 3.2. Varianzas estimadas
4. Contraste para la comparación de medias con muestras emparejadas
5. Intervalo de confianza para la diferencia de proporciones
6. Contraste para la comparación de proporciones
7. Intervalo de confianza para la razón de varianzas en poblaciones normales
8. Contraste para la igualdad de varianzas en poblaciones normales

1. Introducción



2.1. Intervalo de confianza para la diferencia de medias con σ conocidas

Nos interesa la distribución de la variable:

$$\bar{x}_1 - \bar{x}_2$$

Con **muestras grandes** o **poblaciones normales**:

$$\frac{\bar{x}_1 - \mu_1}{\sigma / \sqrt{n_1}} \rightarrow Z \quad \frac{\bar{x}_2 - \mu_2}{\sigma / \sqrt{n_2}} \rightarrow Z$$

$\bar{x}_1 - \bar{x}_2$ seguirá una ley Normal

2.1. Intervalo de confianza para la diferencia de medias con σ conocidas

Media: $\frac{\mu_1 - \mu_2}{\sigma^2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Varianza: $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$$\bar{x}_1 - \bar{x}_2 \rightarrow N\left[\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Carlos Montes – uc3m

2.1. Intervalo de confianza para la diferencia de medias con σ conocidas

Análogamente a como hicimos en el caso del intervalo de confianza para la media poblacional, podemos escribir:

$$P\left[-Z_{\alpha/2} \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z_{\alpha/2}\right] = 1 - \alpha$$

2.1. Intervalo de confianza para la diferencia de medias con σ conocidas

Entonces:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

2.1. Intervalo de confianza para la diferencia de medias con σ conocidas

Operando:

$$P\left((\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Con lo que el intervalo queda: $DT \text{ estimador}$

$$IC(1-\alpha) : \mu_1 - \mu_2 \in \left\{ (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\}$$

Parámetro Estimación Valor tablas

2.1. Intervalo de confianza para la diferencia de medias con σ conocidas

Si las varianzas son iguales: $(\sigma_1^2 = \sigma_2^2)$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow N(0,1)$$

Carlos Montes – uc3m

2.2. Intervalo de confianza para la diferencia de medias con σ estimada

Con **muestras grandes**, la aproximación a la normal sigue siendo válida si sustituimos la varianza por su estimador:

Varianzas poblacionales distintas:

$$IC(1-\alpha) : \mu_1 - \mu_2 \in \left\{ (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right\}$$

2.1. Intervalo de confianza para la diferencia de medias con σ conocidas

Con lo que el intervalo queda:

$$IC(1-\alpha) : \mu_1 - \mu_2 \in \left\{ (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

2.2. Intervalo de confianza para la diferencia de medias con σ estimada

Varianzas poblacionales iguales:

$$IC(1-\alpha) : \mu_1 - \mu_2 \in \left\{ (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

$$\hat{s}_T^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1 + n_2 - 2}$$

Varianza experimental

Combinación más precisa de estimadores para σ^2

2.2. Intervalo de confianza para la diferencia de medias con σ estimada

¿Y si la muestra es pequeña?

Necesitamos que la población se distribuya normalmente.

Varianzas iguales

$$IC(1-\alpha) : \mu_1 - \mu_2 \in \left\{ (\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, \alpha/2} \hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

$$\hat{s}_T^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2}$$

Carlos Montes – uc3m

2.2. Intervalo de confianza para la diferencia de medias con σ estimada

Varianzas distintas:

$$IC(1-\alpha) : \mu_1 - \mu_2 \in \left\{ (\bar{x}_1 - \bar{x}_2) \pm t_{v, \alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right\}$$

$$v = \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left(\frac{\hat{s}_1^2}{n_1} \right) + \frac{1}{n_2-1} \left(\frac{\hat{s}_2^2}{n_2} \right)}$$

Como, en general, no será entero, usamos el entero más próximo

3.1. Contraste para la diferencia de medias. Varianzas conocidas

Se desea contrastar la hipótesis de que las medias de dos poblaciones son iguales.

Muestras grandes, o poblaciones normales

Varianzas iguales

$$Z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow Z$$

3.1. Contraste para la diferencia de medias. Varianzas conocidas

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq Z_{\alpha/2} \\ H_1: \mu_1 \neq \mu_2 & \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq -Z_{\alpha/2} \end{array}$$

$$\begin{array}{ll} H_0: \mu_1 \leq \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq Z_\alpha \\ H_1: \mu_1 > \mu_2 & \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq -Z_\alpha \end{array}$$

$$\begin{array}{ll} H_0: \mu_1 \geq \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq -Z_\alpha \\ H_1: \mu_1 < \mu_2 & \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq Z_\alpha \end{array}$$

3.1. Contraste para la diferencia de medias. Varianzas conocidas

Varianzas distintas

$$Z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow Z$$

Carlos Montes – uc3m

3.1. Contraste para la diferencia de medias. Varianzas conocidas

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq Z_{\alpha/2} \\ H_1: \mu_1 \neq \mu_2 & \end{array}$$

$$\begin{array}{ll} H_0: \mu_1 \leq \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq Z_\alpha \\ H_1: \mu_1 > \mu_2 & \end{array}$$

$$\begin{array}{ll} H_0: \mu_1 \geq \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -Z_\alpha \\ H_1: \mu_1 < \mu_2 & \end{array}$$

3.2. Contraste para la diferencia de medias. Varianzas estimadas

Con **muestras grandes**, la aproximación a la normal sigue siendo válida si sustituimos la varianza por su estimador:

Varianzas poblacionales iguales $Z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow Z$

Varianzas poblacionales distintas $Z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \rightarrow Z$

3.2. Contraste para la diferencia de medias. Varianzas estimadas

Y los contrastes se realizan de la misma manera:

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq Z_{\alpha/2} \\ H_1: \mu_1 \neq \mu_2 & \end{array}$$

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 & \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \geq Z_{\alpha/2} \\ H_1: \mu_1 \neq \mu_2 & \end{array}$$

etc...

3.2. Contraste para la diferencia de medias. Varianzas estimadas

¿Y si la muestra es pequeña?

Necesitamos que la población se distribuya normalmente.

Varianzas poblacionales iguales

$$\hat{s}_T = \frac{(n_1 - 1)\hat{s}_1 + (n_2 - 1)\hat{s}_2}{n_1 + n_2 - 2}$$

$$T_0 = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow t_{n_1 + n_2 - 2}$$

Carlos Montes – uc3m

3.2. Contraste para la diferencia de medias. Varianzas estimadas

$$\begin{aligned} H_0: \mu_1 &= \mu_2 \\ H_1: \mu_1 &\neq \mu_2 \end{aligned} \quad \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{n_1 + n_2 - 2; \alpha/2}$$

$$\begin{aligned} H_0: \mu_1 &\leq \mu_2 \\ H_1: \mu_1 &> \mu_2 \end{aligned} \quad \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{n_1 + n_2 - 2; \alpha}$$

$$\begin{aligned} H_0: \mu_1 &\geq \mu_2 \\ H_1: \mu_1 &< \mu_2 \end{aligned} \quad \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{n_1 + n_2 - 2; \alpha}$$

3.2. Contraste para la diferencia de medias. Varianzas estimadas

Varianzas distintas

$$T_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \rightarrow t_v$$

$$v = \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{\hat{s}_1^2}{n_1} \right) + \frac{1}{n_2 - 1} \left(\frac{\hat{s}_2^2}{n_2} \right)}$$

3.2. Contraste para la diferencia de medias. Varianzas estimadas

$$\begin{aligned} H_0: \mu_1 &= \mu_2 \\ H_1: \mu_1 &\neq \mu_2 \end{aligned} \quad \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} > t_{v; \alpha/2}$$

$$\begin{aligned} H_0: \mu_1 &\leq \mu_2 \\ H_1: \mu_1 &> \mu_2 \end{aligned} \quad \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} > t_{v; \alpha}$$

$$\begin{aligned} H_0: \mu_1 &\geq \mu_2 \\ H_1: \mu_1 &< \mu_2 \end{aligned} \quad \text{Rechazamos si: } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} < -t_{v; \alpha}$$

4. Contraste para la comparación de medias con muestras emparejadas

Para detectar pequeñas diferencias entre medias es más conveniente tomar muestras por pares en condiciones semejantes.

Pares de datos de los mismos elementos.

- Antes /después de cierto cambio.
- Antes / después de un tratamiento.
- Una misma medición con distintos aparatos.

Carlos Montes – uc3m

4. Contraste para la comparación de medias con muestras emparejadas

Con las n observaciones de X_1 y X_2 construimos una nueva variable:

$$Y = X_1 - X_2$$

$$\mu_Y = E(Y) = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

$$H_0 = \mu_1 - \mu_2 = \mu_Y = 0$$

$$H_1 = \mu_1 - \mu_2 = \mu_Y \neq 0$$

4. Contraste para la comparación de medias con muestras emparejadas

El estadístico del contraste será como el estudiado para la media:

$$T = \frac{\bar{x} - \mu}{\hat{s}/\sqrt{n}} \rightarrow Z$$

$$T_0 = \frac{\bar{y} - 0}{\hat{s}_y/\sqrt{n}} \rightarrow Z$$

$$\hat{s}_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

4. Contraste para la comparación de medias con muestras emparejadas

Rechazamos H_0 si:

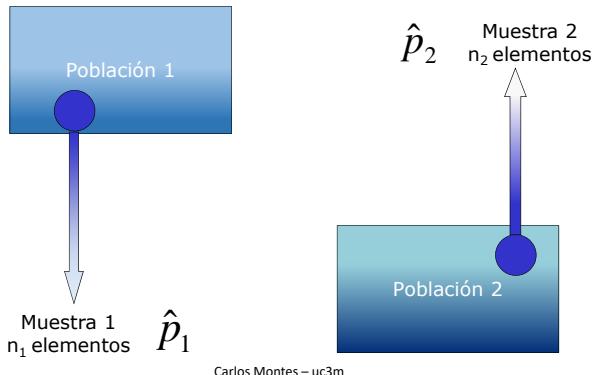
$$\frac{\bar{y}}{\hat{s}_y/\sqrt{n}} \geq t_{(n-1); \alpha/2} \quad o \quad Z_{\alpha/2}$$

$$\frac{\bar{y}}{\hat{s}_y/\sqrt{n}} \leq -t_{(n-1); \alpha/2} \quad o \quad -Z_{\alpha/2}$$

↓ ↓

*Muestra pequeña
(población normal)* *Muestra grande*

5. Intervalo de confianza para la diferencia de proporciones



5. Intervalo de confianza para la diferencia de proporciones

$$\hat{p}_1 \rightarrow N\left(p_1, \sqrt{\frac{p_1 q_1}{n_1}}\right) \quad \hat{p}_2 \rightarrow N\left(p_2, \sqrt{\frac{p_2 q_2}{n_2}}\right)$$

La diferencia de proporciones seguirá una distribución:

$$(\hat{p}_1 - \hat{p}_2) \rightarrow N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

5. Intervalo de confianza para la diferencia de proporciones

Luego el intervalo de confianza queda:

$$IC(1-\alpha) : (p_1 - p_2) \in \left\{ (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right\}$$

6. Contraste para la comparación de proporciones

Se desea contrastar la hipótesis de que la proporción de elementos con cierto atributo es la misma en dos poblaciones.

$$H_0 : p_1 = p_2 = p_0$$

$$H_1 : p_1 \neq p_0$$

Se toman muestras independientes de tamaño n₁ y n₂ de ambas poblaciones

$$\hat{p}_1 \quad \hat{p}_2$$

6. Contraste para la comparación de proporciones

Si H_0 es cierta, la estimación de mayor precisión para p_0 es:

$$\hat{p}_0 = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

$$\text{var}(\hat{p}_1) = \frac{\hat{p}_0 \hat{q}_0}{n_1} \quad \text{var}(\hat{p}_2) = \frac{\hat{p}_0 \hat{q}_0}{n_2}$$

Carlos Montes – uc3m

6. Contraste para la comparación de proporciones

Si la muestra es suficientemente grande:

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \rightarrow Z$$

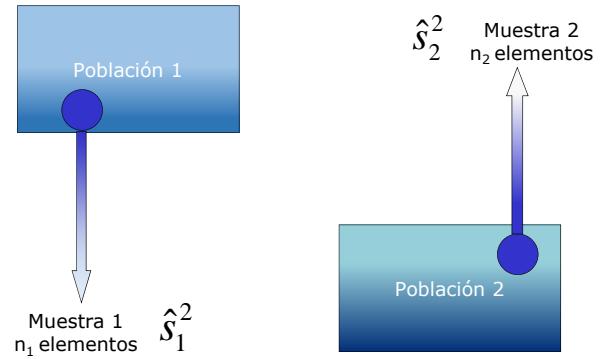
6. Contraste para la comparación de proporciones

$$\begin{array}{ll} H_0: p_1 = p_2 & \text{Rechazamos si: } \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} > Z_{\alpha/2} \\ H_1: p_1 \neq p_2 & \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} < -Z_{\alpha/2} \end{array}$$

$$\begin{array}{ll} H_0: p_1 \leq p_2 & \text{Rechazamos si: } \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} > Z_\alpha \\ H_1: p_1 > p_2 & \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} < -Z_\alpha \end{array}$$

$$\begin{array}{ll} H_0: p_1 \geq p_2 & \text{Rechazamos si: } \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} < -Z_\alpha \\ H_1: p_1 < p_2 & \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} > Z_\alpha \end{array}$$

7. Intervalo de confianza para la razón de varianzas en poblaciones normales



7. Intervalo de confianza para la razón de varianzas en poblaciones normales

Se demuestra que:

$$F = \frac{\hat{s}_1^2 / \sigma_1^2}{\hat{s}_2^2 / \sigma_2^2} \rightarrow F_{(n_1-1, n_2-1)} \quad (\text{F de Fisher})$$

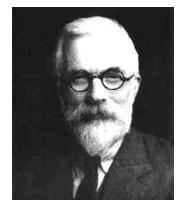
↓ ↓

Grados de libertad del numerador
Grados de libertad del denominador

Carlos Montes – uc3m

7. Intervalo de confianza para la razón de varianzas en poblaciones normales

Ronald Aylmer Fisher
(1890-1962)

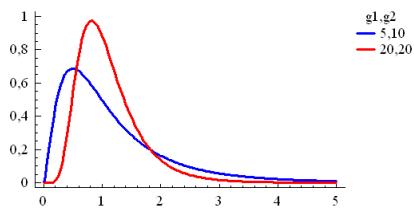


$$F_{n,m} = \frac{\chi_n^2 / n}{\chi_m^2 / m}$$

Compara la longitud de vectores aleatorios de variables normales independientes.

7. Intervalo de confianza para la razón de varianzas en poblaciones normales

Distribuciones F



Propiedad: $\frac{1}{F_{(n_1, n_2, a)}} = F_{(n_2, n_1, 1-a)}$

7. Intervalo de confianza para la razón de varianzas en poblaciones normales

$$F = \frac{\hat{s}_1^2 / \sigma_1^2}{\hat{s}_2^2 / \sigma_2^2} \rightarrow F_{(n_1-1, n_2-1)}$$

$$IC(1-\alpha) : \frac{\sigma_1^2}{\sigma_2^2} \in \left\{ \frac{\hat{s}_1^2}{\hat{s}_2^2} F_{n_1-1, n_2-1; 1-\alpha/2}, \frac{\hat{s}_1^2}{\hat{s}_2^2} F_{n_1-1, n_2-1; \alpha/2} \right\}$$

8. Contraste para la igualdad de varianzas en poblaciones normales

$$F = \frac{\hat{s}_1^2 / \sigma_1^2}{\hat{s}_2^2 / \sigma_2^2} \rightarrow F_{(n_1-1, n_2-1)}$$

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{Si } H_0 \text{ es cierta:}$$

$$F_0 = \frac{\hat{s}_1^2}{\hat{s}_2^2} \rightarrow F_{(n_1-1, n_2-1)}$$

Carlos Montes – uc3m

8. Contraste para la igualdad de varianzas en poblaciones normales

$$\begin{array}{lll} H_0: \sigma_1^2 = \sigma_2^2 & \text{Rechazamos si:} & \frac{\hat{s}_1^2}{\hat{s}_2^2} > F_{n_1-1, n_2-1; \alpha/2} \\ H_1: \sigma_1^2 \neq \sigma_2^2 & & \frac{\hat{s}_1^2}{\hat{s}_2^2} < F_{n_1-1, n_2-1; 1-\alpha/2} \end{array}$$

$$\begin{array}{lll} H_0: \sigma_1^2 \leq \sigma_2^2 & \text{Rechazamos si:} & \frac{\hat{s}_1^2}{\hat{s}_2^2} > F_{n_1-1, n_2-1; \alpha} \\ H_1: \sigma_1^2 > \sigma_2^2 & & \frac{\hat{s}_1^2}{\hat{s}_2^2} < F_{n_1-1, n_2-1; 1-\alpha} \end{array}$$

$$\begin{array}{lll} H_0: \sigma_1^2 \geq \sigma_2^2 & \text{Rechazamos si:} & \frac{\hat{s}_1^2}{\hat{s}_2^2} < F_{n_1-1, n_2-1; 1-\alpha} \\ H_1: \sigma_1^2 < \sigma_2^2 & & \end{array}$$

Parte IX

Tema 9. Regresión multiple

Tema 9 Regresión múltiple

Carlos Montes – uc3m

1. Introducción
2. Modelo lineal general
3. Estimación del modelo
 - 3.1. Definición
 - 3.2. Coeficiente de determinación corregido
4. Contraste sobre los parámetros
5. Diagnosis del modelo
6. Transformaciones
 - 6.1. Generalidades
 - 6.2. Gráfico de componentes
7. Regresión con variables binarias

1. Introducción

Modelo de regresión simple

$$y_i = a + bx_i + e_i$$

El valor que resulta de aplicar la recta $a+bx$ al valor $x=x_i$ es la predicción:

$$\hat{y}(x_i) \quad o \quad \hat{y}$$

1. Introducción

La recta que predice el valor de y cuando $x=x_i$ puede expresarse como:

$$\hat{y}_i = a + bx$$

Luego el residuo puede expresarse como:

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

1. Introducción

Los valores de la variable Y pueden dividirse en dos partes:

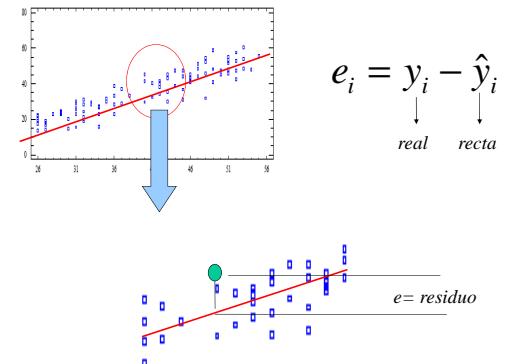
- **Parte lineal o determinista** (explicada por la variable X)
- **Parte aleatoria** (parte de Y no explicada linealmente por X)

$$Y = \underset{\text{real}}{a + bx} + \underset{\text{residuo}}{e}$$

Cuando en el modelo de regresión simple asumimos que e sigue una normal, le denominamos *modelo lineal general* (de regresión simple)

Carlos Montes – uc3m

1. Introducción



1. Introducción

El coeficiente R^2 (coeficiente de determinación) indica la proporción de Y que es explicada por X.

Es el coeficiente de correlación al cuadrado.

Analysis of Variance				
Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	120,475	1	120,475	20,23
Residual	247,225	8	30,900	
Total (Omn.)	367,699	9		

Correlation Coefficient = -0,24724
R-squared = 71,7447 percent
Standard Error of Est. = 5,0991

$R^2 = 71,76\%$

1. Introducción

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Analysis of Variance				
Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	120,475	1	120,475	20,23
Residual	247,225	8	30,900	
Total (Omn.)	367,699	9		

Correlation Coefficient = -0,24724
R-squared = 71,7447 percent
Standard Error of Est. = 5,0991

$R^2 = 71,76\%$

2. Modelo lineal general

En un modelo de regresión múltiple, queremos conocer el valor de una variable respuesta a partir de más de una variable explicativa.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + e_i$$

Fijo Variable

x_1, x_2, \dots, x_n son las variables independientes o explicativas, que pueden ser cualitativas o cuantitativas.

$$E(e_i) = 0$$

También se le denomina "recta de regresión" aunque no sea una recta, sino un hiperplano.

Carlos Montes – uc3m

2. Modelo lineal general

Observado x_i , el valor esperado de y_i es:

$$E(y_i | x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + E(e_i | x_i)$$



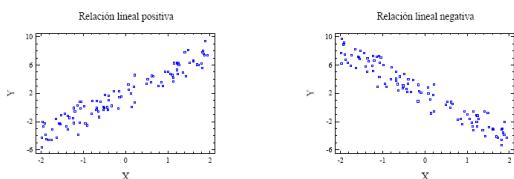
0

2. Modelo lineal general

El modelo se basa en una serie de hipótesis:

1) Linealidad

Las variables explicativas X influyen en Y según una combinación lineal.



2. Modelo lineal general

$$2) E(e)=0$$

$$3) \text{cov}(X, e) = 0$$

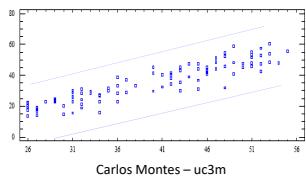
El término de error representa el resto de variables no contenidas en $X = (X_1, \dots, X_n)$, luego contiene información independiente de X .

2. Modelo lineal general

4) Homocedasticidad

La varianza de los errores es constante, y no depende del nivel de las variables.

La nube de puntos de los datos tiene una anchura más o menos constante a lo largo de la recta de regresión.



2. Modelo lineal general

$$\text{var}(e) = E[(e - E(e))^2] = E(e^2) - E(e)^2 = \sigma^2 \quad \forall i$$

$$\text{var}(y_i \setminus X = x_i) = \sigma^2$$

2. Modelo lineal general

5) Normalidad

Los errores se distribuyen según una distribución normal.

$$e_i \sim N(0, \sigma^2)$$

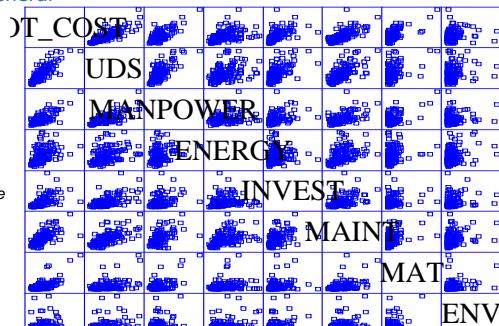
Es una hipótesis razonable por el Teorema Central del Límite: la suma de muchas causas (variables) pequeñas tiende a distribuirse normalmente.

2. Modelo lineal general

6) Independencia

La secuencia de valores de e_i es independiente.

2. Modelo lineal general



Carlos Montes – uc3m

3.1. Estimación del modelo. Definición.

1) Estimación de los coeficientes β

Para estimar los coeficientes β empleamos el método de los mínimos cuadrados ordinarios (MCO).

Dada una muestra de n datos:

$$\begin{aligned} & (y_1, x_{11}, \dots, x_{k1}) \\ & (y_2, x_{12}, \dots, x_{k2}) \\ & \dots \\ & (y_n, x_{1n}, \dots, x_{kn}) \end{aligned}$$

3.1. Estimación del modelo. Definición.

Los valores que asignamos a β_1, \dots, β_k son aquellos que minimizan los errores al cuadrado.

Buscamos el mínimo de la función:

$$S(\beta) = \sum_{i=1}^n \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki}) \right]^2 = \sum_{i=1}^n e^2_i$$

Puede demostrarse que el vector de estimadores de β que minimiza $S(\beta)$ es:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Modelo de regresión de altura (Y) sobre peso (X₁)

$$\hat{Y} = 138,4 + 0,53X_1 + e$$

La estatura de un individuo que pesa 80 kg es una variable aleatoria normal de media estimada:

$$138.4 + 0.53 \cdot 80 = 180.4 \text{ cm}$$

Los individuos que pesan 1 kg más tienden a medir

0.53 cm más

Modelo de regresión de altura (Y) sobre peso (X_1) y talla de zapato (X_2)

$$\hat{Y} = 77,7 + 0,13X_1 + 2,16X_2 + e$$

La estatura media de un individuo de 80 kg que calza un 37 es:

$$77.7 + 0.13 \cdot 80 + 2.16 \cdot 37 = 168.02 \text{ cm}$$

Si calza un 43:

$$77.7 + 0.13 \cdot 80 + 2.16 \cdot 43 = 181.98 \text{ cm}$$

Carlos Montes – uc3m

3.1. Estimación del modelo. Definición.

2) Estimación de σ^2

$$e_i = N(0, \sigma^2)$$

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki})$$

Como la media de e es 0, la varianza muestral de los residuos será:

$$\frac{\sum_{i=1}^n e_i^2}{n}$$

3.1. Estimación del modelo. Definición.

Pero para que sea insesgado, debemos usar la varianza residual:

$$\hat{S}_R^2 = \frac{\sum_{i=1}^n e_i^2}{n-p}$$

Donde p es el número de parámetros beta.

Analizamos datos de una muestra de $n=82$ automóviles. Entre las variables se encuentra **velmax**, que es la velocidad máxima (km/h) que puede alcanzar el vehículo. Queremos construir un modelo lineal que prediga esa velocidad máxima a partir de la variable **Potencia** (caballos de vapor -cv-) y la variable **Peso** (kg) del vehículo.

El modelo de regresión tal y como lo muestra Statgraphics es:

Multiple Regression Analysis				
Dependent variable: velmax				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	155,468	1,3399	116,027	0,0000
Potencia	0,519647	0,00966429	53,7698	0,0000
Peso	-0,0252839	0,00148786	-16,9935	0,0000

Analysis of Variance				
Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	40555,0	2	20277,5	2698,00
Residual	593,746	79	7,51577	
Total (Corr.)	41148,8	81		

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

$$velmax = 155,5 + 0,52 \cdot \text{Potencia} - 0,025 \cdot \text{Peso} + e$$

Carlos Montes – uc3m

$$velmax = 155,5 + 0,52 \cdot \text{Potencia} - 0,025 \cdot \text{Peso} + e$$

Para un mismo peso del vehículo, cada caballo de potencia adicional permite aumentar la velocidad máxima

0.52 km/h por término medio.

Para una misma potencia, cada kg adicional disminuye la velocidad máxima en 0.025 km/h.

$$velmax = 155,5 + 0,52 \cdot \text{Potencia} - 0,025 \cdot \text{Peso} + e$$

La velocidad máxima de un coche que pesa 1500 kg y tenga 100 CV de potencia es una variable aleatoria de media

$$155,5 + 0,52 \cdot 100 - 0,025 \cdot 1500 = 170 \text{ km/h}$$

Su varianza es la del término de error

Analysis of Variance				
Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	40555,0	2	20277,5	2698,00
Residual	593,746	79	7,51577	
Total (Corr.)	41148,8	81		

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

En la tabla: 7.51577

$$S_R^2 = \frac{593,746}{79} = 7,51577$$

$$S_R^2 = 2,74149^2 = 7,51577$$

3.2. Coeficiente de determinación corregido

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \text{corr}(\hat{y}, y)^2$$

A medida que aumenta el número de variables, el coeficiente puede aumentar aunque las variables no sean significativas.

Por ello se define el coeficiente de determinación corregido o ajustado:

3.2. Coeficiente de determinación corregido

$$\bar{R}^2 = 1 - \frac{\hat{s}_R^2}{\hat{s}_y^2} \longrightarrow \text{cuasivarianza de } Y$$

Aunque no es exactamente el % de la variabilidad de Y explicada por las variables independientes, se interpreta de manera informal de la misma manera.

El objetivo es conseguir el máximo número de variables explicativas que consigan aumentar el coeficiente.

Carlos Montes – uc3m

Multiple Regression Analysis				
Dependent variable: velmax				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	155,465	1,3399	116,027	0,0000
Potencia	0,519647	0,00966429	53,7698	0,0000
Peso	-0,0252839	0,00148786	-16,9938	0,0000

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	40565,0	2	20277,5	2698,00	0,0000
Residual	593,746	79	7,51577		
Total (Corr.)	41148,8	81			

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

4. Contraste sobre los parámetros

Los coeficientes: $\beta = (\beta_0, \beta_1, \dots, \beta_k)^T$

Son parámetros poblacionales, y por tanto desconocidos.

Para su estimación usamos: $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)^T$

4. Contraste sobre los parámetros

Basándonos en la distribución de β_i podemos hacer el contraste de si una variable es o no "significativa".

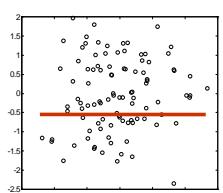
Variable significativa es aquella que aporta información sobre Y no incluida en el resto de las variables.

Por tanto, será relevante incluirla en la regresión.

4. Contraste sobre los parámetros

Una variable será no significativa si:

$$\beta_i = 0$$



Carlos Montes – uc3m

4. Contraste sobre los parámetros

Hacemos el contraste:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

Ya que el estadístico: $t = \frac{\hat{\beta}_i}{s_R \sqrt{\frac{1}{(n-1)s^2_{x_i}}}}$

Sigue una distribución Z en muestras grandes y una t_{n-p} en poblaciones normales.

4. Contraste sobre los parámetros

Multiple Regression Analysis				
Dependent Variable: velmax				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANTE	155,465	1,3399	116,027	0,0000
Potencia	0,519647	0,0096429	53,7698	0,0000
Peso	-0,0262839	0,00148786	-16,9386	0,0000

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	40555,0	2	20277,5	2698,00	0,0000
Residual	593,746	79	7,51577		
Total (Corr.)	41148,8	81			

R-squared = 99,5571 percent
R-squared (adjusted for d.f.) = 99,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

$$p < 0.05$$

Rechazamos H_0 . Las dos variables son significativas.

4. Contraste sobre los parámetros

Multiple Regression Analysis				
Dependent variable: velmax				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANTE	155,465	1,3399	116,027	0,0000
Potencia	0,519647	0,0096429	53,7698	0,0000
Peso	-0,0262839	0,00148786	-16,9386	0,0000

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	40555,0	2	20277,5	2698,00	0,0000
Residual	593,746	79	7,51577		
Total (Corr.)	41148,8	81			

R-squared = 99,5571 percent
R-squared (adjusted for d.f.) = 99,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

$$|t| > 2$$

Rechazamos H_0 . Las dos variables son significativas.

5. Diagnosis del modelo

Es la comprobación de las hipótesis del modelo, lo que garantiza que éste va a describir la verdadera relación entre variables:

- Linealidad.
 - Gráfico de residuos frente a valores predichos. Si el modelo está bien ajustado, no debe presentar ninguna estructura.
- Homocedasticidad.
 - Gráfico de residuos frente a valores ajustados.
 - Gráfico de residuos frente a X.
- Independencia. Existen contrastes específicos como el de Durbin-Watson.
- Normalidad.
 - Gráfico probabilístico normal de los residuos.

Carlos Montes – uc3m

5. Diagnosis del modelo

A la hora de analizar la normalidad:

- Puede ser suficiente con comprobar que los residuos sean unimodales y simétricos.
- Si queremos calcular probabilidades con el modelo normal, necesitamos asegurar la normalidad de los residuos mediante el test de la chi cuadrado.

6.1. Transformaciones. Generalidades.

Cuando las hipótesis del modelo no se cumplen es necesario transformar los datos, de manera que los datos transformados cumplan las hipótesis.

Hay que detectar las variables que no tienen un comportamiento lineal.

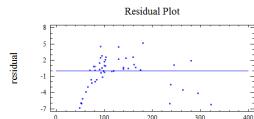
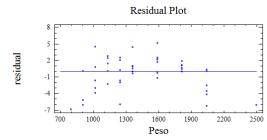


Gráfico de residuos frente a X



6.1. Transformaciones. Generalidades.

Como en la regresión simple, las más utilizadas son:

- Logaritmo
 $y' = \ln x$ $x' = \ln y$

Frecuente para evitar la falta de linealidad o heterocedasticidad

- Cuadrado
 $y' = y^2$ $x' = x^2$

- Inversa
 $y' = 1/y$ $x' = 1/x$

- Raíz cuadrada
 $y' = \sqrt{y}$ $x' = \sqrt{x}$

Muy útil cuando los datos proceden de una Poisson.

6.2. Transformaciones. Gráfico de componentes

El gráfico de componentes (*component effect plot*) representa:

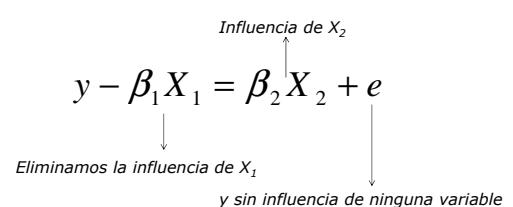
$$X_i \quad \text{frente a:} \quad e_i + \hat{\beta}_i(X_i - \bar{X}_i)$$

Puede interpretarse como la variable Y a la que se la ha eliminado la influencia de todas las variables, salvo la X_i ,

Carlos Montes – uc3m

6.2. Transformaciones. Gráfico de componentes

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$



Los programas informáticos permiten, además, la realización de otro tipo de regresiones.

7. Regresión con variables binarias

* Variable binaria o dicotómica es aquella que toma solo 2 valores (0 y 1).

* Se utilizan para describir la ausencia o presencia de una propiedad.

7. Regresión con variables binarias

Al estudiar la altura, podemos plantearnos en qué medida es explicada por el sexo.

Podemos separar el modelo en dos, uno para cada valor de la variable binaria.

$$E(\text{altura} | \text{chica}) = 165.313 + 14.0367 \times 0 = 165.313$$

$$E(\text{altura} | \text{chico}) = 165.313 + 14.0367 \times 1 = 179.3497$$

Multiple Regression Analysis				
Dependent variable: altura				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	165.313	0.056112	199.097	0.0000
sexo	14.0367	1.05129	13.3819	0.0000

7. Regresión con variables binarias

Podemos completar el modelo añadiendo otras variables explicativas, por ejemplo, peso.

$$\text{Altura} = \beta_0 + \beta_1 \cdot \text{sexo} + \beta_2 \cdot \text{peso} + e$$

Multiple Regression Analysis				
Dependent variable: altura				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	169,306	3,12145	48,1528	0,0000
peso	0,28133	0,0640419	4,39689	0,0000
sexo	9,28133	1,34214	6,91581	0,0000

Entre un chico y una chica del mismo peso, el chico tiene una altura 9,28 cm mayor.

Carlos Montes – uc3m

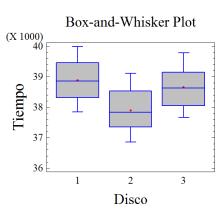
7. Regresión con variables binarias

* Para comparar G grupos podemos construir G variables binarias que denotaremos por D_g , donde cada una toma el valor 1 en aquellos elementos que pertenezcan al grupo g -ésimo y 0 al resto.

* Hay que introducir un máximo de $G-1$ variables binarias, porque si no la primera columna (de unos, correspondiente al término constante de la regresión) será combinación lineal de las otras.

Se quiere comparar el comportamiento de tres discos duros con el fin de ver cuál es el más rápido. Para ello se graba un fichero de 200 MB en cada uno de ellos y se cronometra el tiempo de descarga. Se repite ese experimento un número de veces con cada disco.

Los resultados se encuentran en el fichero Discosduros.sf3. ¿Cuál es el disco duro más rápido?



Queremos ver si el tiempo del disco 2 es significativamente mejor.

$$Y = \beta_0 + \beta_1 D_1 + \beta_3 D_3 + e$$

$$E(Y | \text{disco } 2) = \beta_0$$

$$E(Y | \text{disco } 1) = \beta_0 + \beta_1$$

$$E(Y | \text{disco } 3) = \beta_0 + \beta_3$$

$$Y = \beta_0 + \beta_1 D_1 + \beta_3 D_3 + e$$

Multiple Regression Analysis				
Dependent variable: Tiempo				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	37896,3	75,572	501,46	0,0000
Disco=1	978,018	107,235	9,12029	0,0000
Disco=3	747,922	118,785	6,29644	0,0000

El 2 es significativamente mejor

Carlos Montes – uc3m

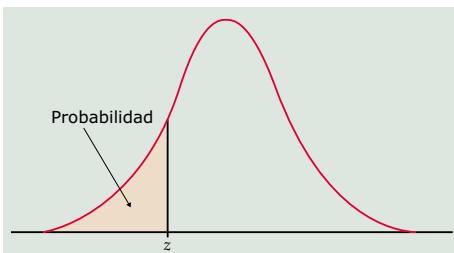
$$Y = \beta_0 + \beta_2 D_2 + \beta_3 D_3 + e$$

Multiple Regression Analysis				
Dependent variable: Tiempo				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	38874,4	76,0809	510,96	0,0000
Disco=2	-978,018	107,235	-9,12029	0,0000
Disco=3	-230,096	119,109	-1,93181	0,0548

- * El disco 2 tiene una duración media significativamente inferior en 978,018 unidades de tiempo.
- * La diferencia del disco 1 con el disco 3 no parece ser significativa al tener el p-valor mayor que 0,05.
- * No podemos asegurar que el 1 y el 3 sean diferentes entre si.

Parte X

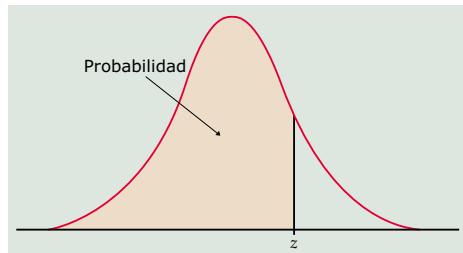
Recursos



El valor de la tabla para z es el área bajo la curva de la normal estándar a la izquierda de z

TABLA A: Probabilidades de la normal estandar

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641



El valor de la tabla para z es el área bajo la curva de la normal estándar a la izquierda de z .

TABLA A: Probabilidades de la normal estándar (cont.)

FORMULARIO DE ESTADÍSTICA

MODELOS DE PROBABILIDAD					
Nombre	Símbolo	$p(x) = P(X = x); F(x) = P(X \leq x); f(x) = \frac{\partial F(x)}{\partial x}$	Media μ	Varianza σ^2	
Bernoulli	$B(p)$	$p(x) = p^x q^{1-x}; x = 0, 1$	p	pq	
Binomial	$B(n, p)$	$p(x) = \binom{n}{x} p^x q^{n-x}; x = 0, 1, \dots, n$	np	npq	
Geométrica	$G(p)$	$p(x) = pq^{x-1}; x = 1, 2, \dots$	$\frac{1}{p}$	$\frac{q}{p^2}$	
Poisson	$P(\lambda)$	$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}; x = 0, 1, \dots$	λ	λ	
Uniforme continua en (a, b)	$U(a, b)$	$f(x) = 1/(b - a); a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Exponencial	$Exp(\lambda)$	$f(x) = \lambda e^{-\lambda x}; x > 0$ $F(x) = 1 - e^{-\lambda x}; x > 0$	$1/\lambda$	$1/\lambda^2$	
Normal	$N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; x \in \mathbb{R}$	μ	σ^2	

$$\begin{aligned} \text{REGRESIÓN SIMPLE} \\ \hat{y} = a + bx \\ a = \bar{y} - b\bar{x} \\ b = \frac{s_{xy}}{s_x^2} = \frac{\hat{s}_{xy}}{\hat{s}_x^2} \end{aligned}$$

INFERENCIA PARA UNA POBLACIÓN

Población	Contraste de hipótesis	Estadístico de contraste	Región de rechazo (p -valor < α)	Intervalo de confianza $IC_{1-\alpha}$
Cualquier v.a. X con $E[X] = \mu$, $\text{Var}[X] = \sigma^2$ y $n \rightarrow \infty$	(1) $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$ (2) $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$ (3) $H_0 : \mu \leq \mu_0$; $H_1 : \mu > \mu_0$	(a) $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ (b) $t_0 = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}}$	(1a) $ z_0 > z_{\alpha/2}$ (1b) $ t_0 > z_{\alpha/2}$ (2a) $z_0 < -z_{\alpha}$ (2b) $t_0 < -t_{n-1;\alpha}$ (3a) $z_0 > z_{\alpha}$ (3b) $t_0 > t_{n-1;\alpha}$	$\mu \in (\bar{x} \mp z_{\alpha/2}\sigma/\sqrt{n})$ $\mu \in (\bar{x} \mp z_{\alpha/2}\hat{s}/\sqrt{n})$
Normal $N(\mu, \sigma^2)$	(1) $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$ (2) $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$ (3) $H_0 : \mu \leq \mu_0$; $H_1 : \mu > \mu_0$	(a) $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ (b) $t_0 = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}}$	(1a) $ z_0 > z_{\alpha/2}$ (1b) $ t_0 > t_{n-1;\alpha/2}$ (2a) $z_0 < -z_{\alpha}$ (2b) $t_0 < -t_{n-1;\alpha}$ (3a) $z_0 > z_{\alpha}$ (3b) $t_0 > t_{n-1;\alpha}$	$\mu \in (\bar{x} \mp z_{\alpha/2}\sigma/\sqrt{n})$ $\mu \in (\bar{x} \mp t_{n-1;\alpha/2}\hat{s}/\sqrt{n})$
Bernoulli $B(p)$ con $n \rightarrow \infty$	(1) $H_0 : p = p_0$; $H_1 : p \neq p_0$ (2) $H_0 : p \geq p_0$; $H_1 : p < p_0$ (3) $H_0 : p \leq p_0$; $H_1 : p > p_0$	$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$	(1) $ z_0 > z_{\alpha/2}$ (2) $z_0 < -z_{\alpha}$ (3) $z_0 > z_{\alpha}$	$p \in (\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}})$
Normal $N(\mu, \sigma^2)$	(1) $H_0 : \sigma^2 = \sigma_0^2$; $H_1 : \sigma^2 \neq \sigma_0^2$ (2) $H_0 : \sigma^2 \geq \sigma_0^2$; $H_1 : \sigma^2 < \sigma_0^2$ (3) $H_0 : \sigma^2 \leq \sigma_0^2$; $H_1 : \sigma^2 > \sigma_0^2$	$\chi_0^2 = \frac{(n-1)\hat{s}^2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2}$	(1) $\chi_0^2 > \chi_{n-1;\alpha/2}^2$ ó $\chi_0^2 < \chi_{n-1;1-\alpha/2}^2$ (2) $\chi_0^2 < \chi_{n-1;1-\alpha}^2$ (3) $\chi_0^2 > \chi_{n-1;\alpha}^2$	$\sigma^2 \in \left(\frac{(n-1)\hat{s}^2}{\chi_{n-1;\alpha/2}^2}; \frac{(n-1)\hat{s}^2}{\chi_{n-1;1-\alpha/2}^2} \right)$
Cualquier v.a. X con $\hat{\theta}_{MV}$ y $n \rightarrow \infty$	(1) $H_0 : \theta = \theta_0$; $H_1 : \theta \neq \theta_0$ (2) $H_0 : \theta \geq \theta_0$; $H_1 : \theta < \theta_0$ (3) $H_0 : \theta \leq \theta_0$; $H_1 : \theta > \theta_0$	$t_0 = \frac{\hat{\theta}_{MV} - \theta_0}{\sqrt{\widehat{\text{Var}}(\hat{\theta}_{MV})}}$ $\widehat{\text{Var}}(\hat{\theta}_{MV}) = -\left(\frac{\partial^2 L(\hat{\theta}_{MV})}{\partial \theta^2}\right)^{-1}$	(1) $ t_0 > z_{\alpha/2}$ (2) $t_0 < -z_{\alpha}$ (3) $t_0 > z_{\alpha}$	$\theta \in \left(\hat{\theta}_{MV} \mp z_{\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\theta}_{MV})} \right)$

INFERENCIA PARA DOS POBLACIONES

Dos poblaciones o v.a. X_1, X_2	Contraste de hipótesis	Estadístico de contraste	Región de rechazo (p -valor < α)
Cualesquiera con $E(X_1) = \mu_1, E(X_2) = \mu_2$ $\text{Var}(X_1) = \sigma_1^2$ $\text{Var}(X_2) = \sigma_2^2$	(1) $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$ (2) $H_0 : \mu_1 \geq \mu_2; H_1 : \mu_1 < \mu_2$ (3) $H_0 : \mu_1 \leq \mu_2; H_1 : \mu_1 > \mu_2$	(a) $z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ (b) $t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$	(1a) $ z_0 > z_{\alpha/2}$ (1b) $ t_0 > z_{\alpha/2}$ (1b) $ t_0 > t_{v;\alpha/2}$ (2a) $z_0 < -z_\alpha$ (2b) $t_0 < -z_\alpha$ (2b) $t_0 < -t_{v;\alpha}$ (3a) $z_0 > z_\alpha$ (3b) $t_0 > z_\alpha$ (3b) $t_0 > t_{v;\alpha}$ Bajo normalidad Si $n_1, n_2 \rightarrow \infty$ Bajo normalidad o si $n_1, n_2 \rightarrow \infty$
Datos pareados $D = X_1 - X_2$	(1) $H_0 : \mu_D = 0; H_1 : \mu_D \neq 0$ (2) $H_0 : \mu_D \geq 0; H_1 : \mu_D < 0$ (3) $H_0 : \mu_D \leq 0; H_1 : \mu_D > 0$	(a) $z_0 = \frac{\bar{d}}{\sigma_D / \sqrt{n}}$ (b) $t_0 = \frac{\bar{d}}{\hat{s}_D / \sqrt{n}}$	(1a) $ z_0 > z_{\alpha/2}$ (1b) $ t_0 > z_{\alpha/2}$ (1b) $ t_0 > t_{n-1;\alpha/2}$ (2a) $z_0 < -z_\alpha$ (2b) $t_0 < -z_\alpha$ (2b) $t_0 < -t_{n-1;\alpha}$ (3a) $z_0 > z_\alpha$ (3b) $t_0 > z_\alpha$ (3b) $t_0 > t_{n-1;\alpha}$ Bajo normalidad Si $n \rightarrow \infty$ Bajo normalidad o si $n \rightarrow \infty$
Cualesquiera con $E(X_1) = \mu_1, E(X_2) = \mu_2$ $\text{Var}(X_1) = \text{Var}(X_2) = \sigma^2$	(1) $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$ (2) $H_0 : \mu_1 \geq \mu_2; H_1 : \mu_1 < \mu_2$ (3) $H_0 : \mu_1 \leq \mu_2; H_1 : \mu_1 > \mu_2$	(a) $z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ (b) $t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ con $\hat{s}_T^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2}$	(1a) $ z_0 > z_{\alpha/2}$ (1b) $ t_0 > z_{\alpha/2}$ (1b) $ t_0 > t_{n_1+n_2-2;\alpha/2}$ (2a) $z_0 < -z_\alpha$ (2b) $t_0 < -z_\alpha$ (2b) $t_0 < -t_{n_1+n_2-2;\alpha}$ (3a) $z_0 > z_\alpha$ (3b) $t_0 > z_\alpha$ (3b) $t_0 > t_{n_1+n_2-2;\alpha}$ Bajo normalidad Si $n_1, n_2 \rightarrow \infty$ Bajo normalidad o si $n_1, n_2 \rightarrow \infty$
v.a. de Bernoulli $X_1 \sim B(p_1), X_2 \sim B(p_2)$	(1) $H_0 : p_1 = p_2; H_1 : p_1 \neq p_2$ (2) $H_0 : p_1 \geq p_2; H_1 : p_1 < p_2$ (3) $H_0 : p_1 \leq p_2; H_1 : p_1 > p_2$	$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ con $\hat{p}_0 = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$	(1) $ z_0 > z_{\alpha/2}$ (2) $z_0 < -z_\alpha$ (3) $z_0 > z_\alpha$ Si $n_1, n_2 \rightarrow \infty$
v.a. Normales $X_1 \sim N(\mu_1, \sigma_1^2)$ $X_2 \sim N(\mu_2, \sigma_2^2)$	(1) $H_0 : \sigma_1^2 = \sigma_2^2; H_1 : \sigma_1^2 \neq \sigma_2^2$ (2) $H_0 : \sigma_1^2 \geq \sigma_2^2; H_1 : \sigma_1^2 < \sigma_2^2$ (3) $H_0 : \sigma_1^2 \leq \sigma_2^2; H_1 : \sigma_1^2 > \sigma_2^2$	$F_0 = \frac{\hat{s}_1^2}{\hat{s}_2^2}$	(1) $F_0 > F_{n_1-1; n_2-1; \alpha/2}$ ó $F_0 < F_{n_1-1; n_2-1; 1-\alpha/2}$ donde $F_{n_1-1; n_2-1; 1-\alpha/2} = 1/F_{n_2-1; n_1-1; \alpha/2}$ (2) $F_0 < F_{n_1-1; n_2-1; 1-\alpha}$ (3) $F_0 > F_{n_1-1; n_2-1; \alpha}$

Dos poblaciones o v.a. X_1, X_2	Parámetro	Intervalo de confianza IC $_{1-\alpha}$
Cualesquiera con $E(X_1) = \mu_1, E(X_2) = \mu_2$ $\text{Var}(X_1) = \sigma_1^2$ $\text{Var}(X_2) = \sigma_2^2$	$\mu_1 - \mu_2$	$\mu_1 - \mu_2 \in \left(\bar{x}_1 - \bar{x}_2 \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$ Bajo normalidad o si $n_1, n_2 \rightarrow \infty$ $\mu_1 - \mu_2 \in \left(\bar{x}_1 - \bar{x}_2 \mp z_{\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right)$ Si $n_1, n_2 \rightarrow \infty$ $\mu_1 - \mu_2 \in \left(\bar{x}_1 - \bar{x}_2 \mp t_{v;\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right)$ con $v \approx \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left(\frac{\hat{s}_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \left(\frac{\hat{s}_2^2}{n_2} \right)^2}$ Bajo normalidad
Cualesquiera con $E(X_1) = \mu_1, E(X_2) = \mu_2$ $\text{Var}(X_1) = \text{Var}(X_2) = \sigma^2$	$\mu_1 - \mu_2$	$\mu_1 - \mu_2 \in \left(\bar{x}_1 - \bar{x}_2 \mp z_{\alpha/2} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$ Bajo normalidad o si $n_1, n_2 \rightarrow \infty$ $\mu_1 - \mu_2 \in \left(\bar{x}_1 - \bar{x}_2 \mp z_{\alpha/2} \hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$ Si $n_1, n_2 \rightarrow \infty$ $\mu_1 - \mu_2 \in \left(\bar{x}_1 - \bar{x}_2 \mp t_{n_1+n_2-2;\alpha/2} \hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$ Bajo normalidad
v.a. de Bernoulli $X_1 \sim B(p_1), X_2 \sim B(p_2)$	$p_1 - p_2$	$p_1 - p_2 \in \left(\hat{p}_1 - \hat{p}_2 \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right)$ Si $n_1, n_2 \rightarrow \infty$
v.a. Normales $X_1 \sim N(\mu_1, \sigma_1^2)$ $X_2 \sim N(\mu_2, \sigma_2^2)$	$\frac{\sigma_1^2}{\sigma_2^2}$	$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{\hat{s}_1^2}{\hat{s}_2^2} F_{n_2-1; n_1-1; 1-\alpha/2}; \frac{\hat{s}_1^2}{\hat{s}_2^2} F_{n_2-1; n_1-1; \alpha/2} \right)$ donde $F_{n_2-1; n_1-1; 1-\alpha/2} = 1/F_{n_1-1; n_2-1; \alpha/2}$