PROPEDÉUTICO

Modulo: Bioestadística

INSTRUCCIONES PARA EL MÓDULO:

1. ESTE MÓDULO ESTÁ COMPUESTO POR CINCO SECCIONES Y CINCO GUÍAS DE ESTUDIO.
2. DENTRO DE CADA SECCIÓN ENCONTRARÁS EN PRIMER LUGAR SU GUÍA DE ESTUDIO Y SUBSECUENTEMENTE EL MATERIAL DE APOYO PARA RESOLVERLA.
3. UTILIZA EL MATERIAL DIDACTICO DE LAS SECCIONES Y EL LIBRO BIOSTATISTICAL ANALYSIS, ZAR, J. PRENTICE-HALL 1984 Ó 1999, Y RESUELVE CADA UNA DE LAS GUÍAS DE ESTUDIO PROPORCIONADAS.
4. LOS REACTIVOS DEL EXAMEN DE ADMISIÓN ESTARÁN BASADOS EN LOS PUNTOS EXPUESTOS EN CADA GUÍA.
5. ¡BUENA SUERTE!

PROPEDÉUTICO

Modulo: Introducción a la estadística

Guía de estudio para la Unidad 1: Introducción

UTILIZANDO LA INFORMACIÓN DE ESTA SECCIÓN ó DEL LIBRO BIOSTATISTICAL ANALYSIS, ZAR, J. PRENTICE-HALL 1984 Ó 1999 RESUELVE CADA UNO DE LOS INCISOS:

1. ¿Qué es la estadística?
2. ¿Cuáles son los pasos básicos en el método científico y la estadística?
3. Define población.
4. Define muestra.
5. Define qué es una variable ordinal.
6. Define qué es una tabla de frecuencias y cómo se relaciona con los métodos gráficos para presentar la información de los datos.
7. ¿Cuál es la utilidad de los diagramas de barras?
8. ¿Qué es una diagrama integral y que información aportan?

Introducción a la Estadística

Tema 1: Introducción

Temario

- Introducción.- conceptos, tipos de datos, presentación de datos
- Estadística descriptiva.- medidas de tendencia central, dispersión, posición forma
- Modelos probabilísticas.- distribuciones de probabilidades, distribución normal
- Inferencia estadística.- Teorema central del limite, estimación puntual y por intervalos
- Pruebas de hipótesis.- conceptos, pruebas sobre la media de una población

¿ Que es la estadística ?

- **Estadística (De Estadista)** f. Censo o recuento de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de un Estado, provincia pueblo etc. // Estudio de los hechos morales o físicos del mundo que se prestan a numeración o recuento y a comparación de las cifras a ellos referentes.

  - Real Academia Española, Diccionario de la lengua española.

¿ Que es la estadística ?

- A menudo la información de que se dispone es incompleta.
  - Existe incertidumbre en cualquier proceso en el que se extienden conclusiones que aquel que se tiene información.
  - El método de razonamiento que nos conduce a esta extensión es conocido como inductivo.

- La misma experiencia realizada repetidas veces arroja resultados diferentes.
  - Mediciones repetidas de una misma persona
  - Determinar el rendimiento de una variedad de maíz sembrándola varias veces
  - La variabilidad introduce un elemento de incertidumbre.

- **Actividad importante de la estadística** es cuantificar la incertidumbre

Definición

La Estadística es la Ciencia de la

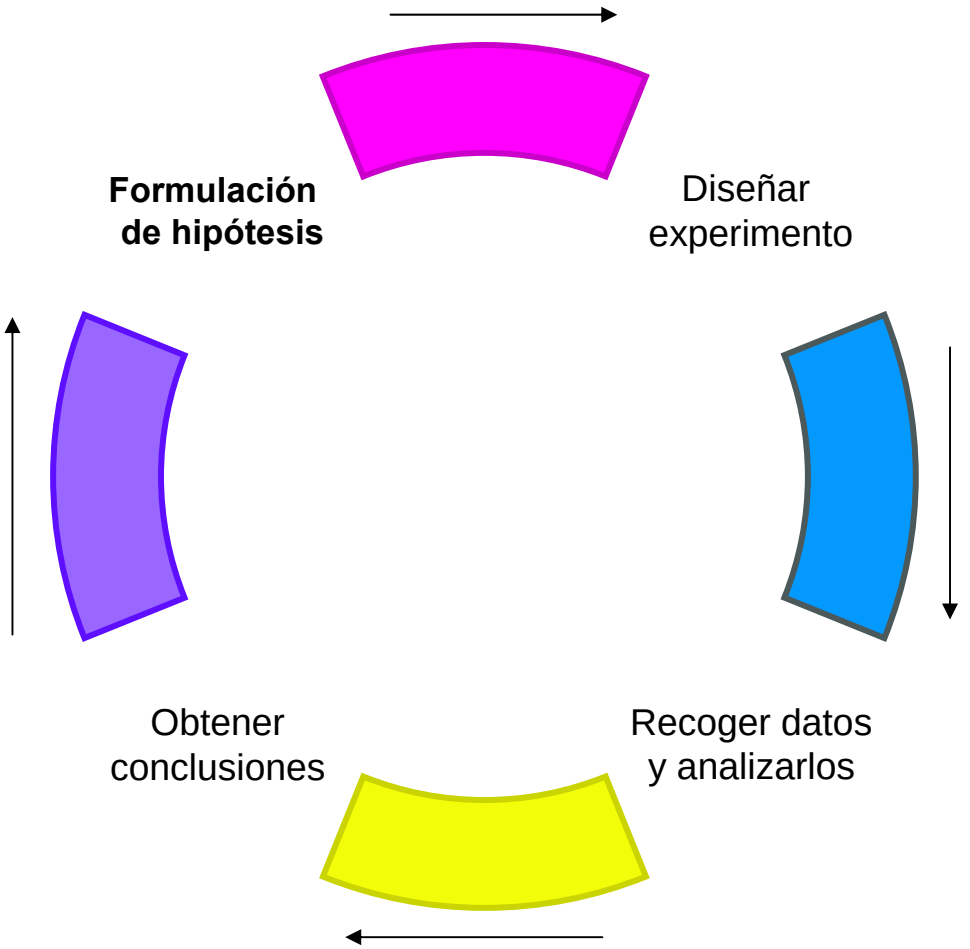
Descriptiva

- **Colección, manejo , descripción y presentación** de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico,

Inferencia

- y poder de esa forma hacer análisis sobre los mismos, para la toma de **decisiones** u obtener **conclusiones**.

Método científico y estadística

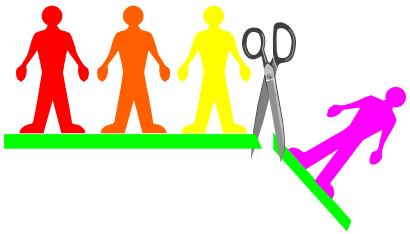


- La información sea relevante al problema.
- Las conclusiones que de ella se extraigan tengan un cierto grado de confiabilidad.

Población y muestra

- **Población** es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).
  - Normalmente es demasiado grande para poder abarcarlo.

- **Muestra** es un subconjunto de la poblaciónal que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)
  - Debería ser “representativo”
  - Esta formado por miembros “seleccionados” de la población (individuos, unidades experimentales).



Variables

- Una **variable** es una característica observable *que varía entre los diferentes individuos* de una población. La información que disponemos de cada individuo es resumida en **variables**.

- En los individuos de la *población* mexicana, de uno a otro ***es variable***:

  - El grupo sanguíneo
    - {A, B, AB, O} ← Var. Cualitativa
  - Su nivel de felicidad “declarado”
    - {Deprimido, Ni fu ni fa, feliz, Muy Feliz} ← Var. Ordinal
  - El número de hijos
    - {0,1,2,3,...} ← Var. Numérica discreta
  - La altura
    - {1.62 ; 1.74; ...} ← Var. Numérica continua

Tipos de variables

- Cualitativas
  Si sus valores (*modalidades*) no se pueden asociar naturalmente a un número (no se pueden hacer operaciones algebraicas con ellos)
  - Nominales: Si sus valores no se pueden ordenar
    - Sexo, tipo de cultivo, especie, Religión, Nacionalidad, Fumar (Sí/No)
    - La única relación aritmética que se admite es la de igualdad
    - La única estadística valida es la frecuencia de una clase
  - Ordinales: Si sus valores se pueden ordenar de menor a mayor
    - Mejoría a un tratamiento, Grado de satisfacción, Intensidad del dolor
    - No solo se admite la relación de igual, sino además de la mayor que y menor que
    - Frecuencias, mediana
- Cuantitativas o Numéricas
  Si sus valores son numéricos (tiene sentido hacer operaciones algebraicas con ellos)
  - Discretas: Si toma valores enteros
    - Número de hijos, Número de especies, Num. de “cumpleaños”
    - las anteriores relaciones más la suma
    - Media, varianza, coeficiente de variación
  - Continuas: Si entre dos valores, son posibles infinitos valores intermedios.
    - Peso de un individuo, rendimiento por ha de una planta, Dosis de medicamento administrado, edad
    - las anteriores relaciones más la suma
    - Media, varianza, coeficiente de variación

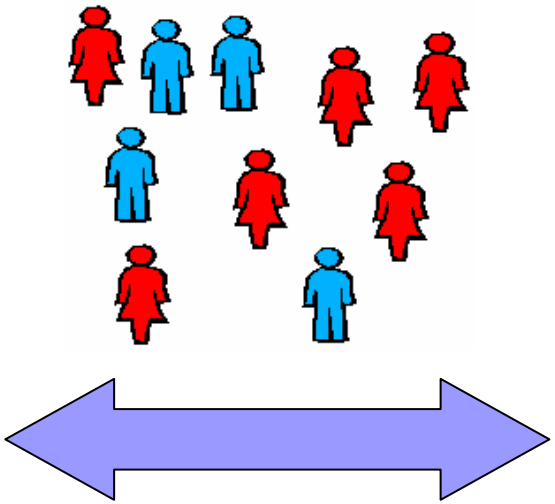
- Los posibles valores de una variable suelen denominarse modalidades.

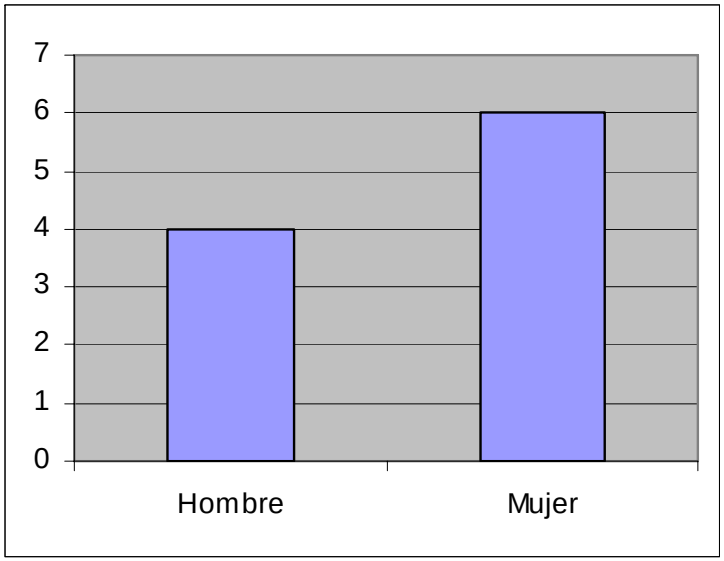
- Las modalidades pueden agruparse en clases (intervalos)
  - Edades:
    - Menos de 20 años, de 20 a 50 años, más de 50 años
  - Hijos:
    - Menos de 3 hijos, De 3 a 5, 6 o más hijos

- Las modalidades/clases deben forman un sistema exhaustivo y excluyente
  - Exhaustivo: No podemos olvidar ningún posible valor de la variable
    - **Mal: ¿Cuál es su color del pelo: (Rubio, Moreno)?**
    - Bien: ¿Cuál es su grupo sanguíneo?
  - Excluyente: Nadie puede presentar dos valores simultáneos de la variable
    - Estudio sobre el ocio
      - Mal: De los siguientes, qué le gusta: (deporte, cine)
      - Bien: Le gusta el deporte: (Sí, No)
      - Bien: Le gusta el cine: (Sí, No)
      - Mal: Cuántos hijos tiene: (Ninguno, Menos de 5, Más de 2)

Presentación ordenada de datos

Género	Frec.
Hombre	4
Mujer	6





- Las tablas de frecuencias y las representaciones gráficas son dos maneras ***equivalentes*** de presentar la información. Las dos exponen ordenadamente la información recogida en una muestra.

Tablas de frecuencia

- Exponen la información recogida en la muestra, de forma que no se pierda nada de información (o poca).
  - Frecuencias absolutas: Contabilizan el número de individuos de cada modalidad
  - Frecuencias relativas (porcentajes): Idem, pero dividido por el total
  - Frecuencias acumuladas: Sólo tienen sentido para variables ordinales y numéricas
    - Muy útiles para calcular cuantiles (ver más adelante)
      - ¿Qué porcentaje de individuos tiene menos de 3 hijos? Sol: 83,8
      - ¿Entre 4 y 6 hijos? Soluc 1ª: 8,4%+3,6%+1,6%= **13,6%**. Soluc 2ª: 97,3% - 83,8% = **13,5%**

Sexo del encuestado

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	Hombre	636	41,9	41,9
	Mujer	881	58,1	58,1
	Total	1517	100,0	100,0

Nivel de felicidad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Muy feliz	467	30,8	31,1	31,1
	Bastante feliz	872	57,5	58,0	89,0
	No demasiado feliz	165	10,9	11,0	100,0
	Total	1504	99,1	100,0	
Perdidos	No contesta	13	,9		
Total		1517	100,0		

Número de hijos

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	419	27,6	27,8	27,8
	1	255	16,8	16,9	44,7
	2	375	24,7	24,9	69,5
	3	215	14,2	14,2	83,8
	4	127	8,4	8,4	92,2
	5	54	3,6	3,6	95,8
	6	24	1,6	1,6	97,3
	7	23	1,5	1,5	98,9
	Ocho o más	17	1,1	1,1	100,0
	Total	1509	99,5	100,0	
Perdidos	No contesta	8	,5		
Total		1517	100,0		

Datos desordenados y ordenados en tablas

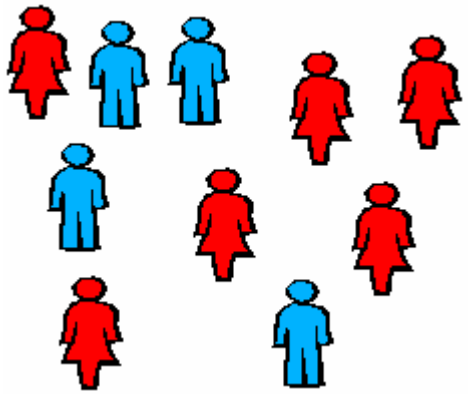
- Variable: Género
  - Modalidades:
    - H = Hombre
    - M = Mujer

Género	Frec.	Frec. relat. porcentaje
Hombre	4	4/10=0,4=40%
Mujer	6	6/10=0,6=60%
	10=tamaño muestral	

- Muestra:

  M H H M M H M M M H

  - equivale a
    HHHH MMMMMM

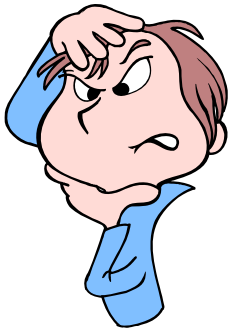


Ejemplo

- ¿Cuántos individuos tienen menos de 2 hijos?
  - frec. indiv. sin hijos
    +
    frec. indiv. con 1 hijo
    = 419 + 255
    = 674 individuos

- ¿Qué porcentaje de individuos tiene 6 hijos o menos?
  - 97,3%

- ¿Qué cantidad de hijos es tal que al menos el 50% de la población tiene una cantidad inferior o igual?
  - 2 hijos

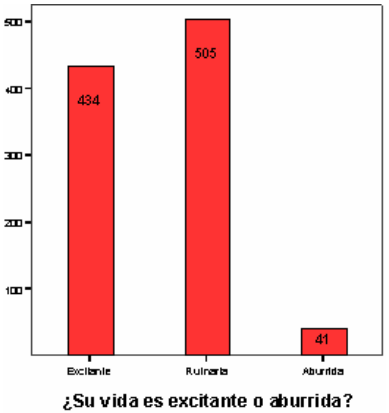


Número de hijos

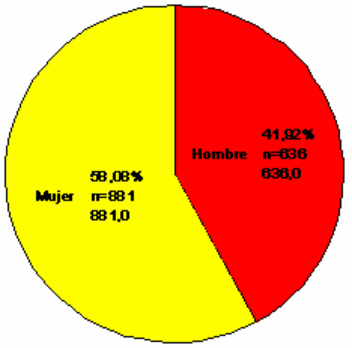
	Frec.	Porcent. (válido)	Porcent. acum.	
0	419	27,8	27,8	
1	255	16,9	44,7	
2	375	24,9	69,5	≥50%
3	215	14,2	83,8	
4	127	8,4	92,2	
5	54	3,6	95,8	
6	24	1,6	97,3	
7	23	1,5	98,9	
Ocho+	17	1,1	100,0	
Total	1509	100,0		

Gráficos para v. cualitativas

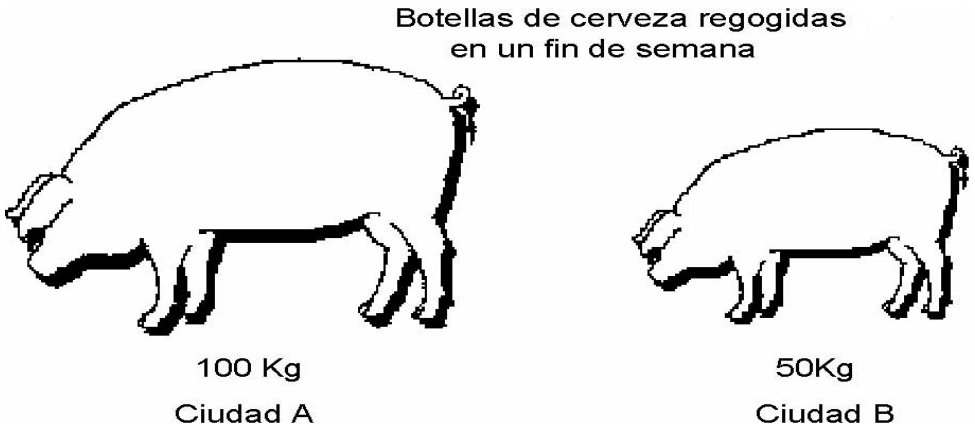
- Diagramas de barras
  - Alturas proporcionales a las frecuencias (abs. o rel.)
  - Se pueden aplicar también a variables discretas

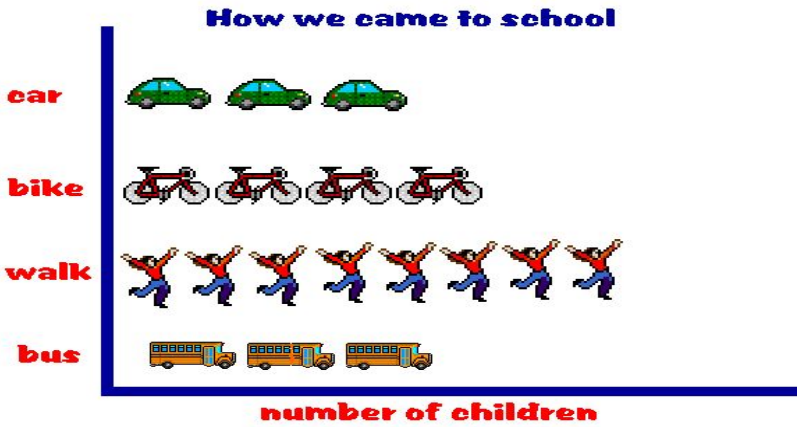


- Diagramas de sectores (pay, polares)
  - No usarlo con variables ordinales.
  - El área de cada sector es proporcional a su frecuencia (abs. o rel.)



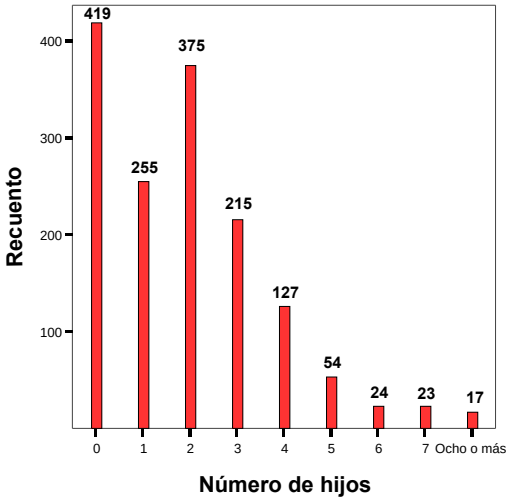
- Pictogramas
  - Fáciles de entender.
  - El área de cada modalidad debe ser proporcional a la frecuencia. ¿De los dos, cuál es incorrecto?.

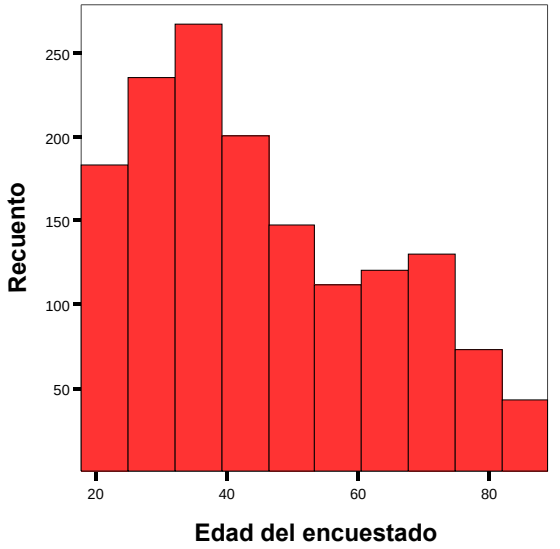




Gráficos diferenciales para variables numéricas

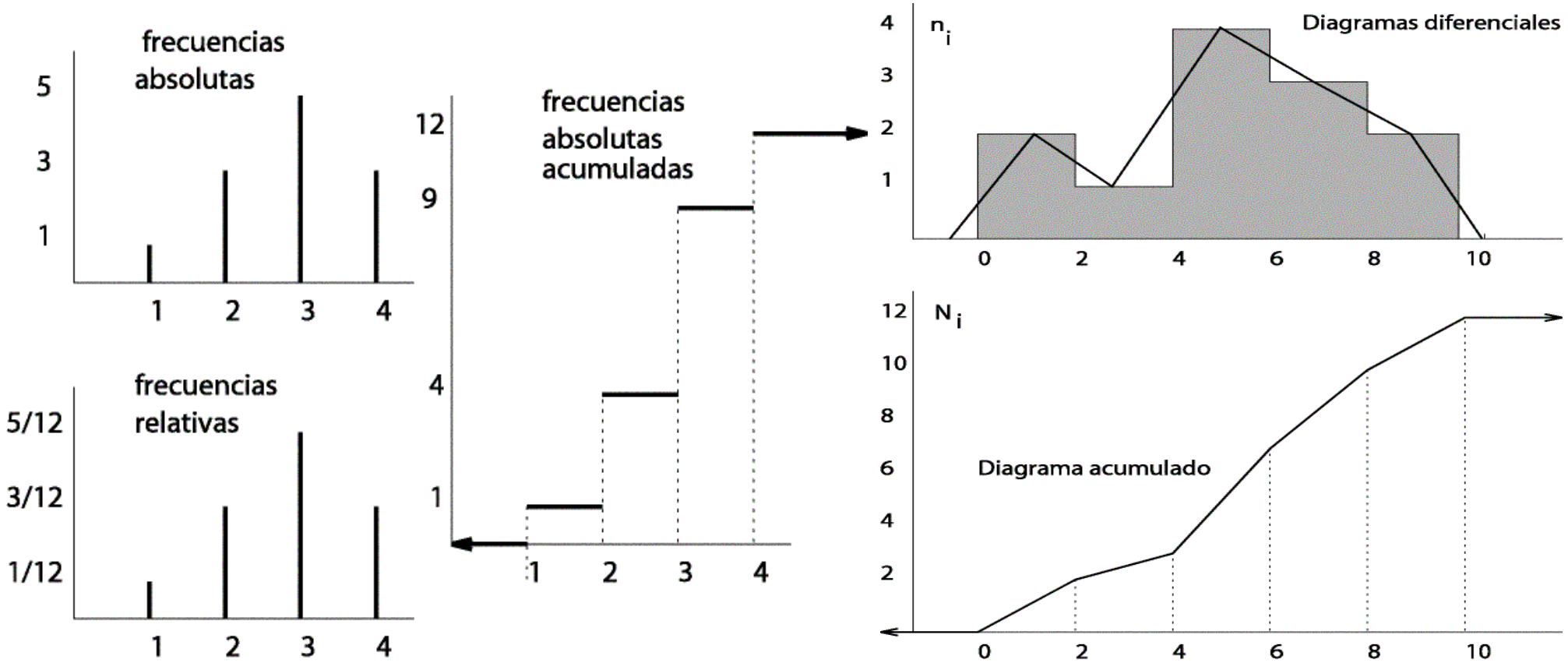
- Son diferentes en función de que las variables sean discretas o continuas. Valen con frec. absolutas o relativas.
  - **Diagramas barras para v. discretas**
    - Se deja un hueco entre barras para indicar los valores que no son posibles
  - **Histogramas para v. continuas**
    - El área que hay bajo el histograma entre dos puntos cualesquiera indica la cantidad (porcentaje o frecuencia) de individuos en el intervalo.





Diagramas integrales

- Cada uno de los anteriores diagramas tiene su correspondiente diagrama integral. Se realizan a partir de las frecuencias acumuladas. Indican, para cada valor de la variable, la cantidad (frecuencia) de individuos que poseen un valor inferior o igual al mismo.



PROPEDÉUTICO

Modulo: Introducción a la estadística

Guía de estudio para la Unidad 2: Estadística descriptiva

UTILIZANDO LA INFORMACIÓN DE ESTA SECCIÓN ó DEL LIBRO BIOSTATISTICAL ANALYSIS, ZAR, J. PRENTICE-HALL 1984 Ó 1999 RESUELVE CADA UNO DE LOS INCISOS:

1. ¿Qué son los cuartiles y que significada cada uno de ellos?

2. Define parámetro y estadístico.

3. Define rango intercuartílico y cuánta dispersión engloba.

4. ¿Qué es el coeficiente de variación?

5. ¿Qué es la varianza?

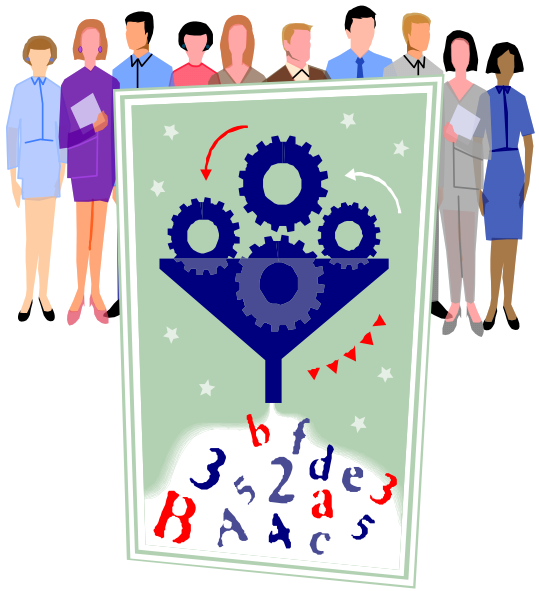
6. ¿Que son los cuantiles y cuál es su relación con los cuartiles?

Introducción a la Estadística

Tema 2: Estadística Descriptiva

Parámetros y estadísticos

- Parámetro: Es una cantidad numérica calculada sobre una población
  - La altura media de los individuos de un país
  - La idea es resumir toda la información que hay en la población en unos pocos números (parámetros).
- Estadístico: Ídem (cambiar población por muestra)
  - La altura media de los que estamos en este aula.
    - Somos una muestra (¿representativa?) de la población.
  - Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.



Normalmente nos interesa conocer un parámetro, pero por la dificultad que conlleva estudiar a ***TODA*** la población, calculamos un estimador sobre una muestra y "confiamos" en que sean próximos. Como elegir muestras para que el error sea "confiablemente" pequeño.

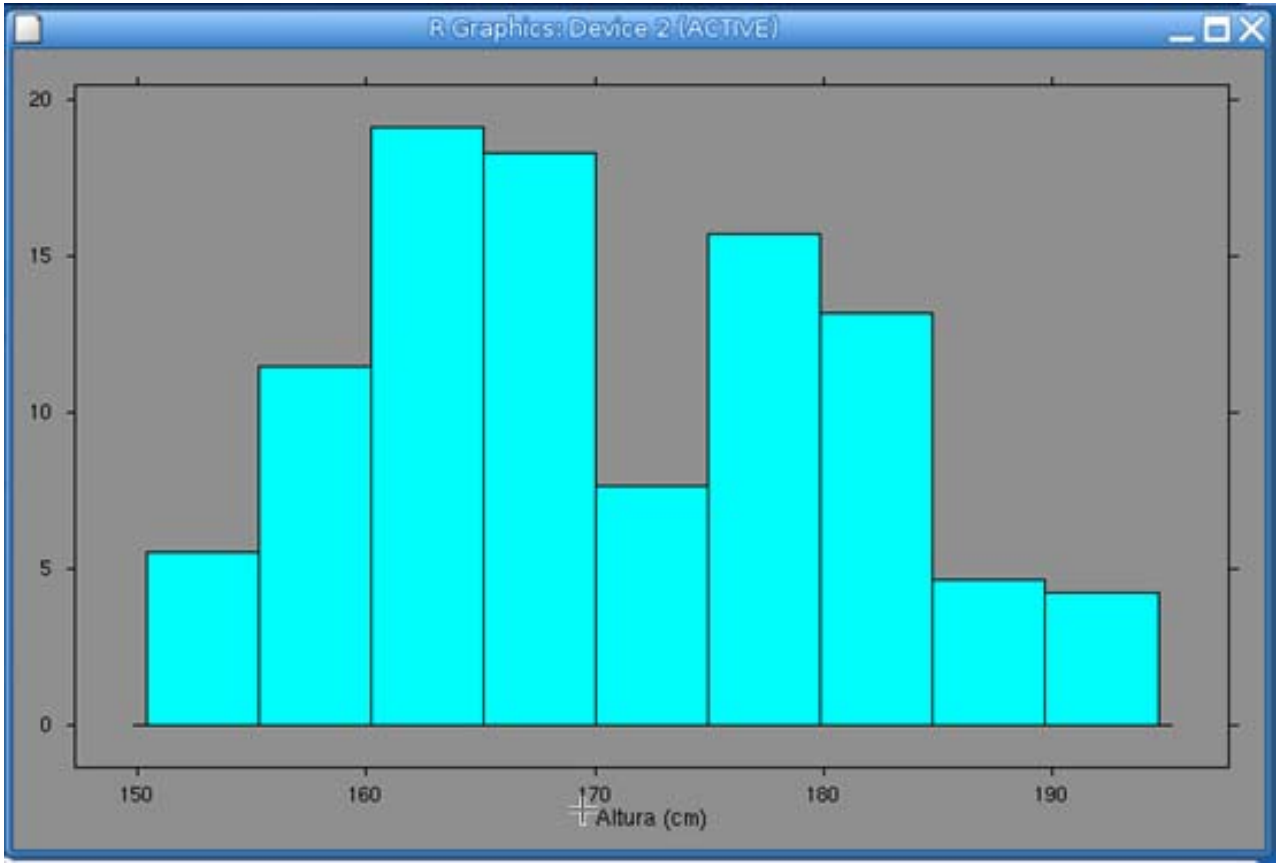
Importancia de los parámetros

Base de datos

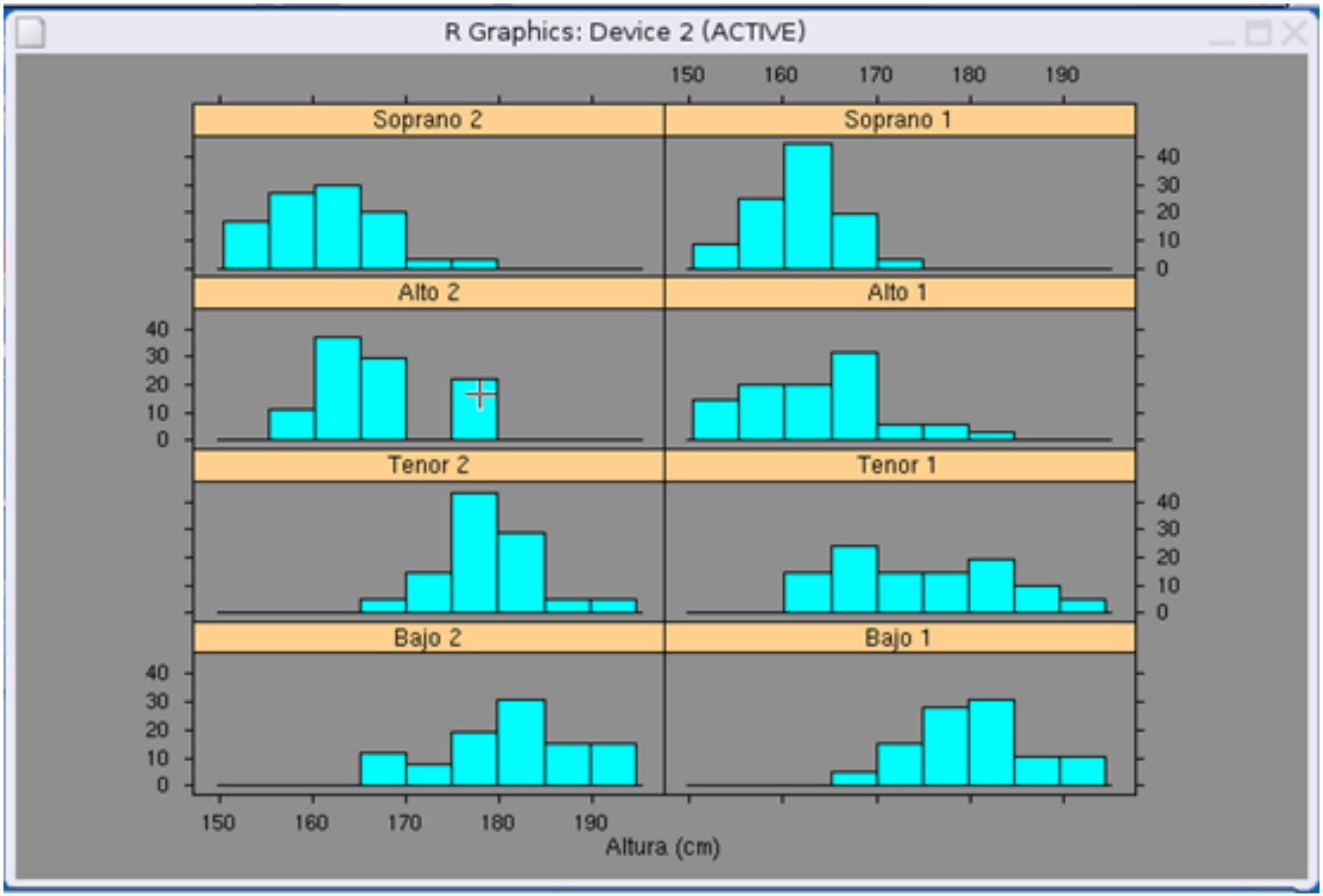
R DataEntryWindow

Copy Paste Quit

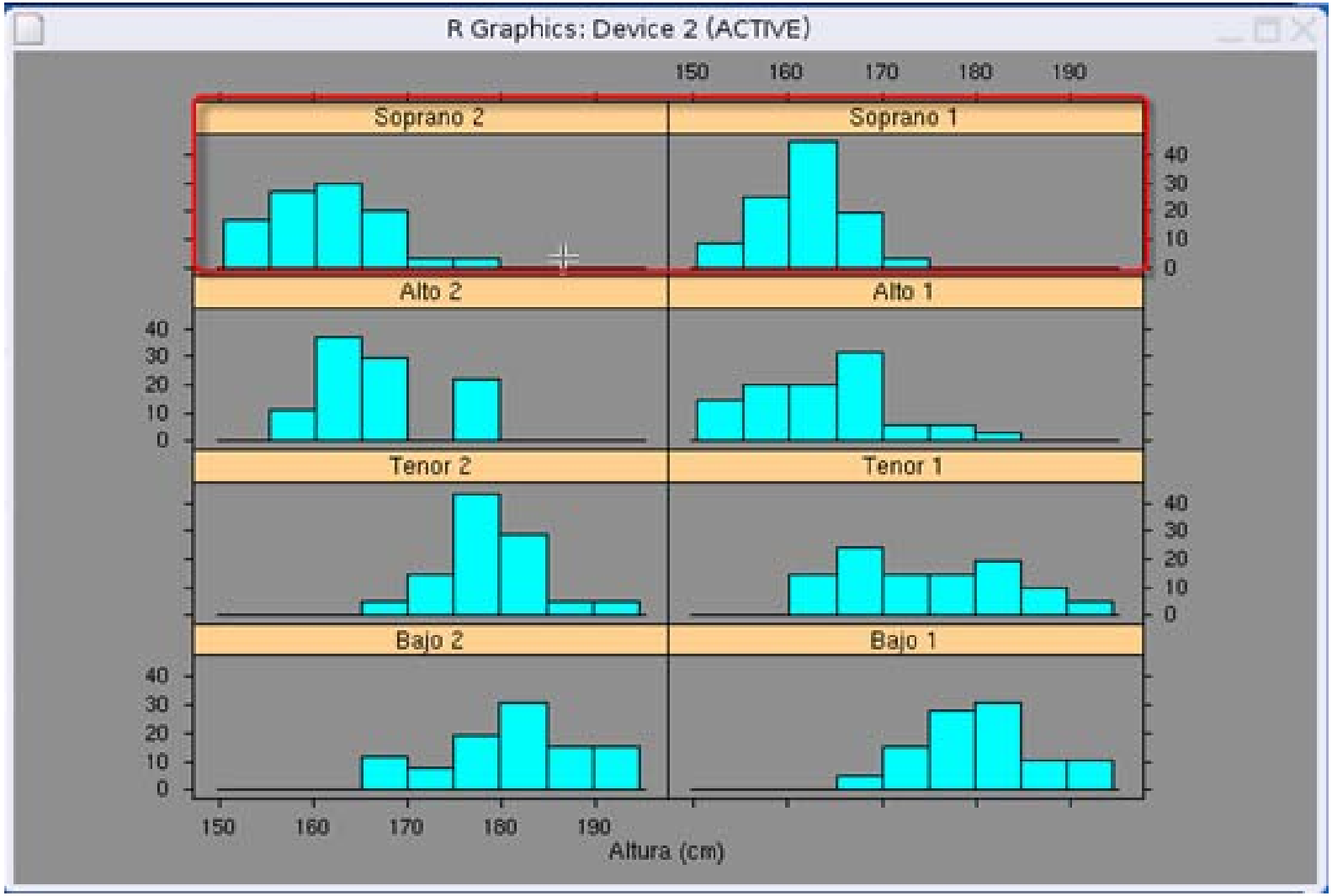
	altura	voz	var3
49	165	Soprano 2	
50	165	Soprano 2	
51	155	Soprano 2	
52	163	Soprano 2	
53	173	Soprano 2	
54	163	Soprano 2	
55	160	Soprano 2	
56	157	Soprano 2	
57	163	Soprano 2	
58	157	Soprano 2	
59	163	Soprano 2	
60	165	Soprano 2	
61	152	Soprano 2	
62	165	Soprano 2	
63	178	Soprano 2	
64	160	Soprano 2	
65	170	Soprano 2	
66	168	Soprano 2	
67	165	Alto 1	
68	157	Alto 1	
69	173	Alto 1	
70	170	Alto 1	
71	170	Alto 1	
72	160	Alto 1	
73	170	Alto 1	



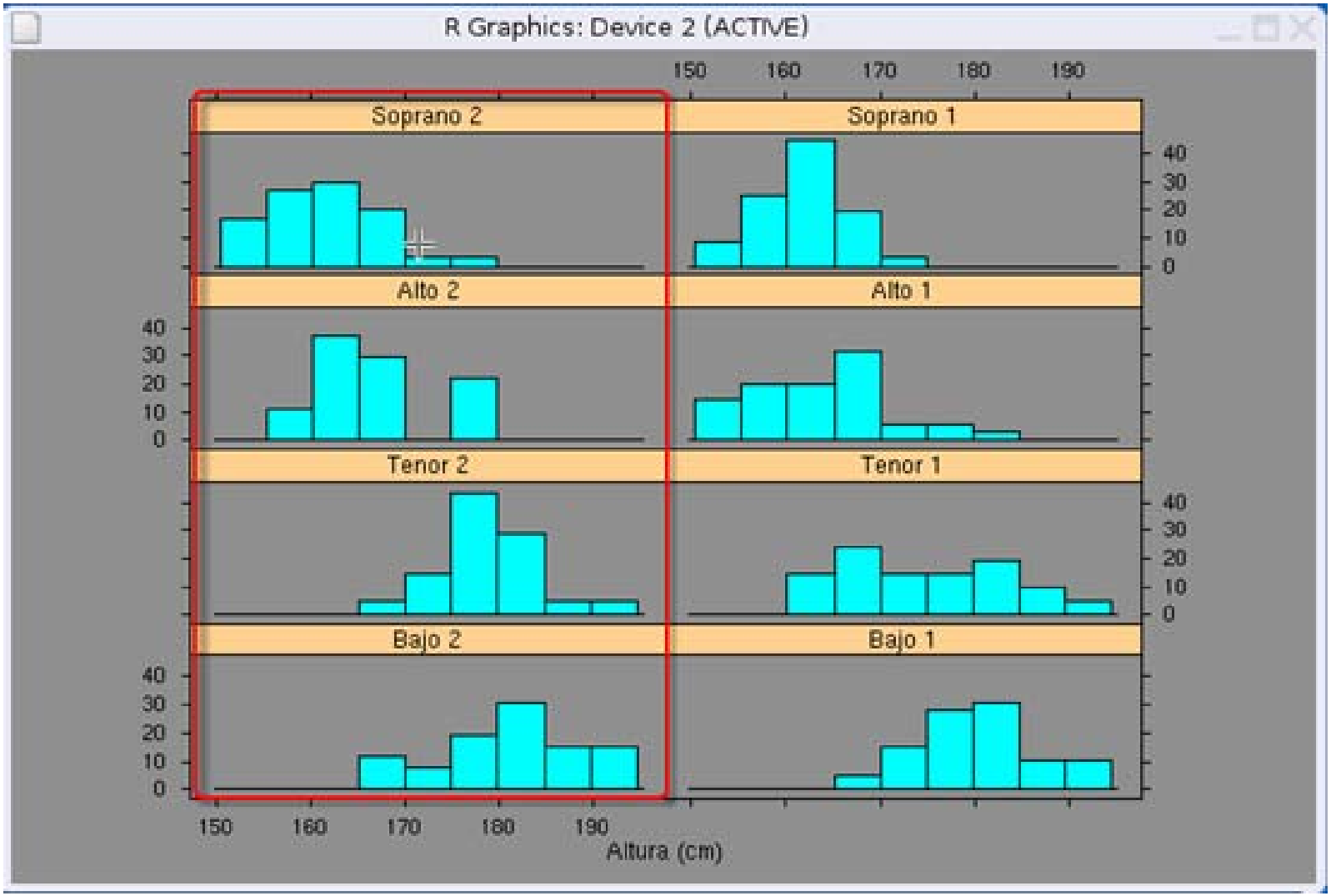
Importancia de los parámetros (continuación)



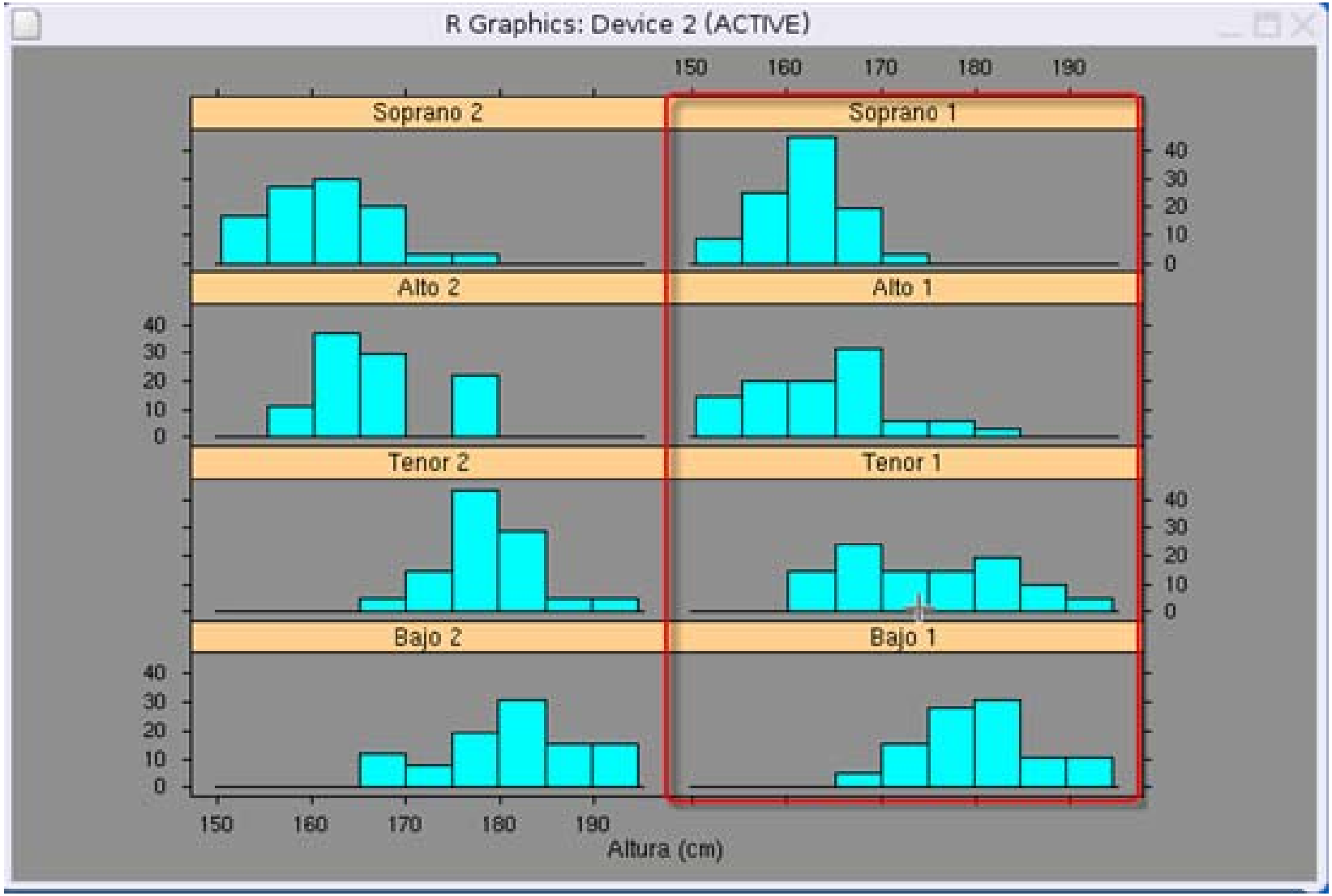
Importancia de los parámetros (continuación)



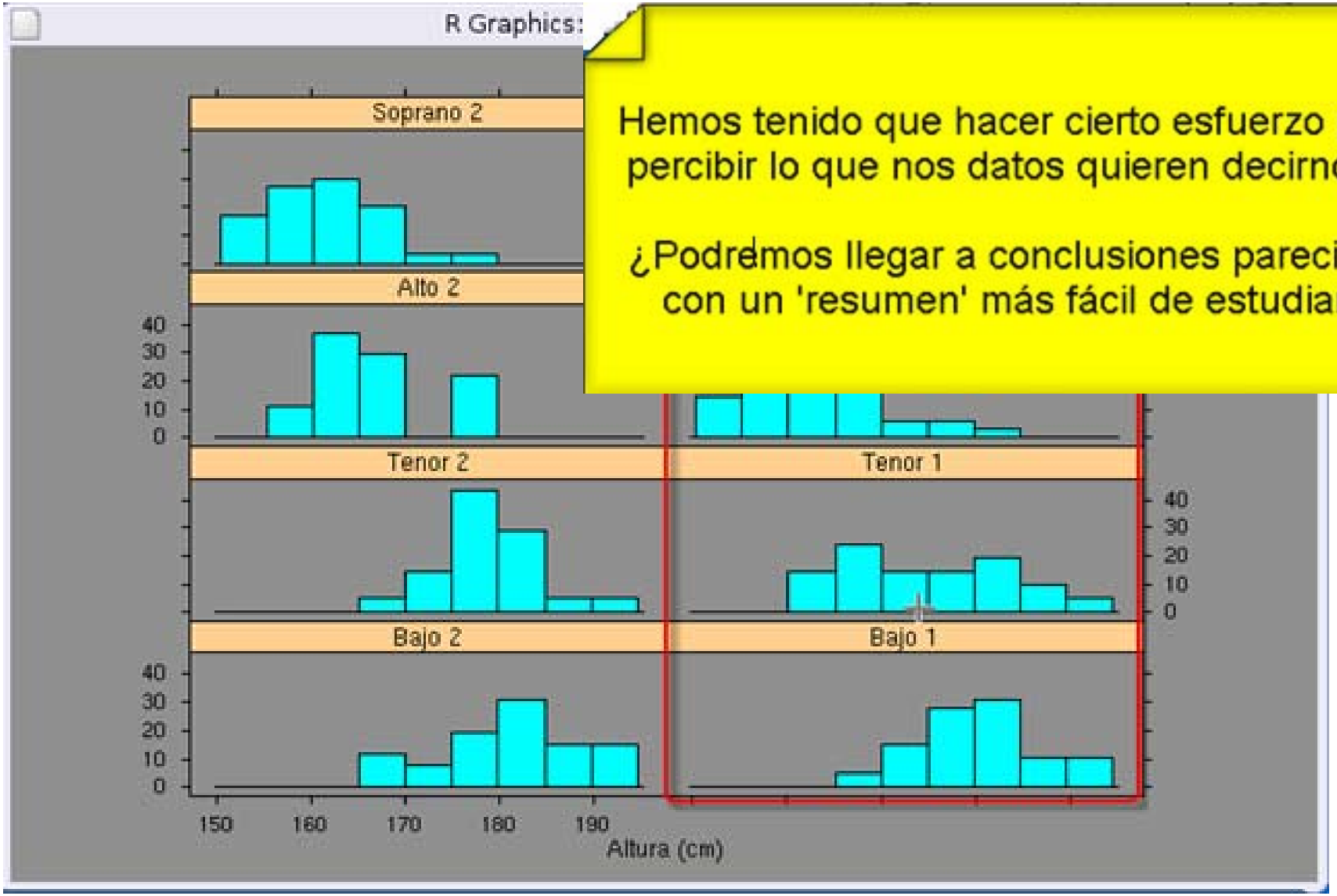
Importancia de los parámetros (continuación)



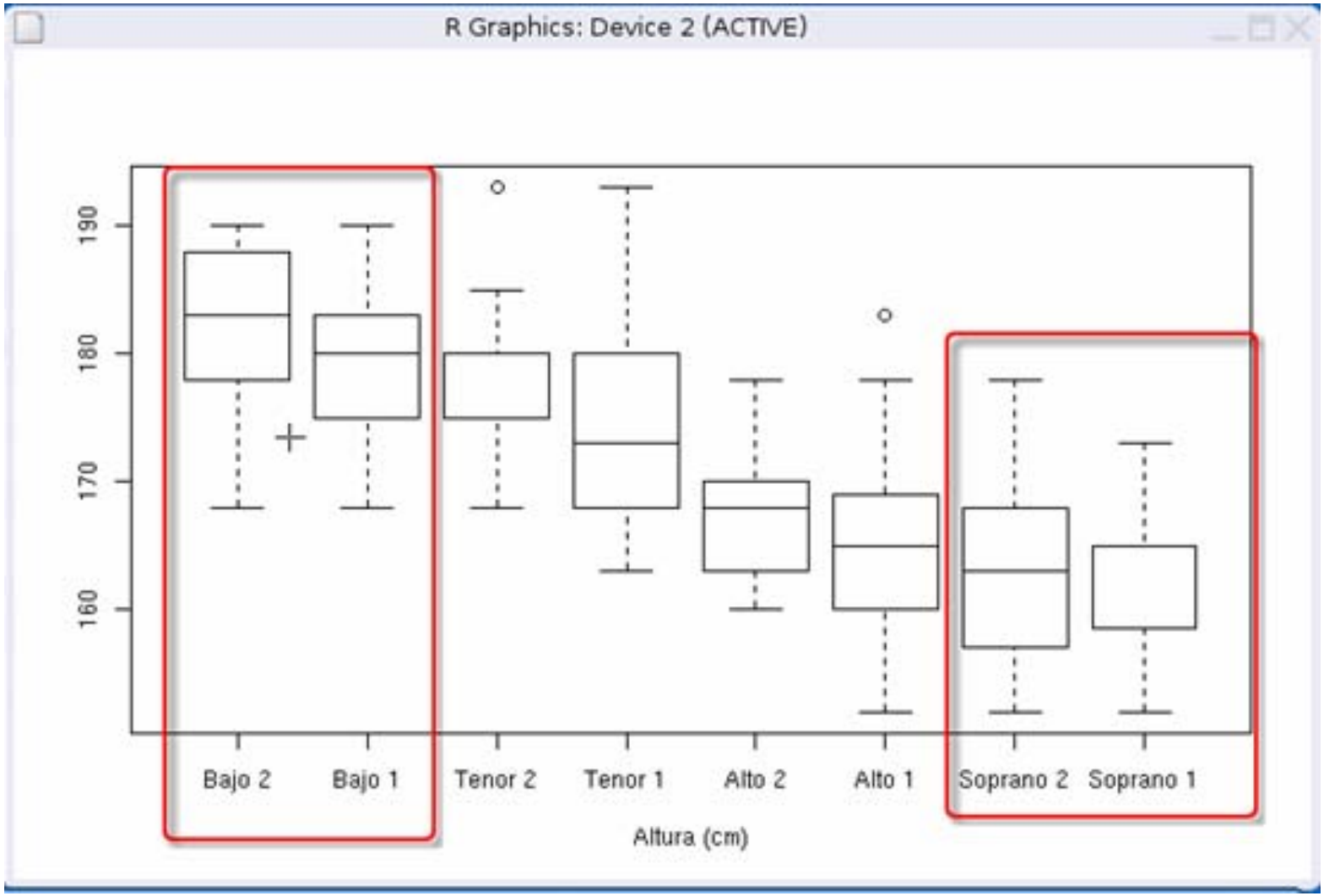
Importancia de los parámetros (continuación)

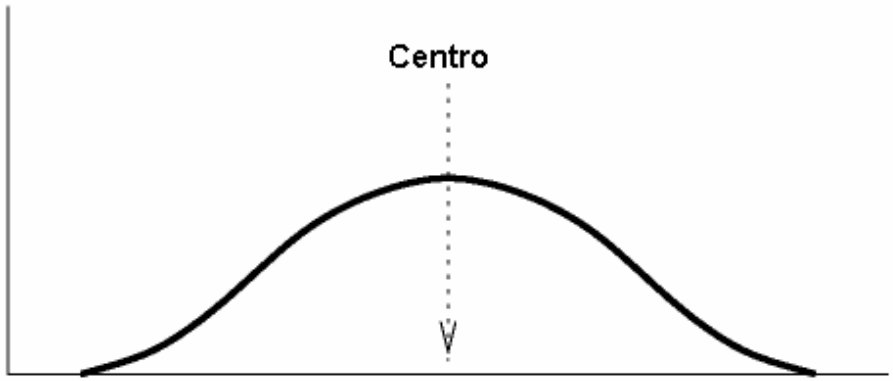


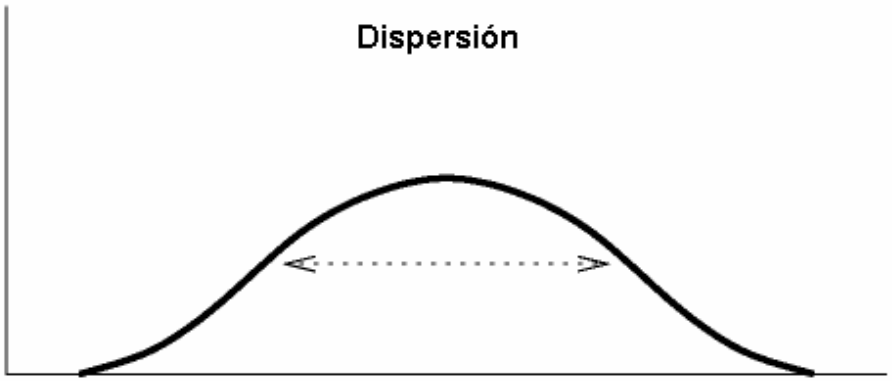
Importancia de los parámetros (continuación)

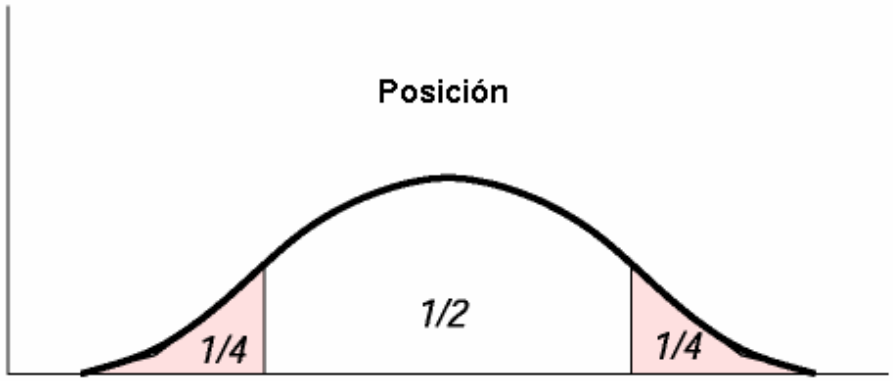


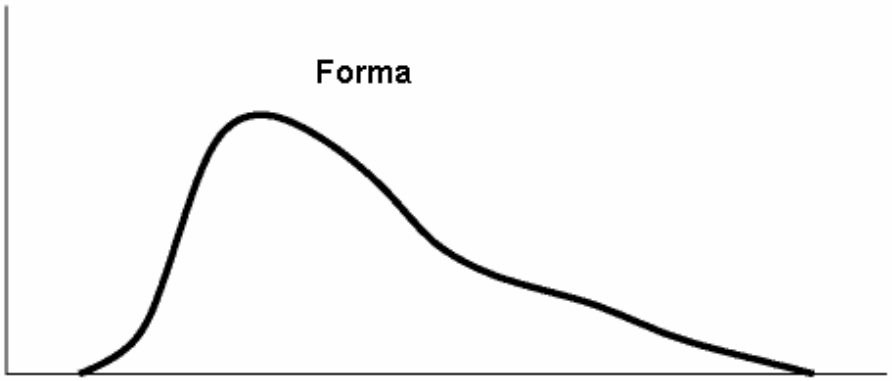
Diagramas de cajas




Centro


Dispersión


Posición
1/4
1/2
1/4


Forma

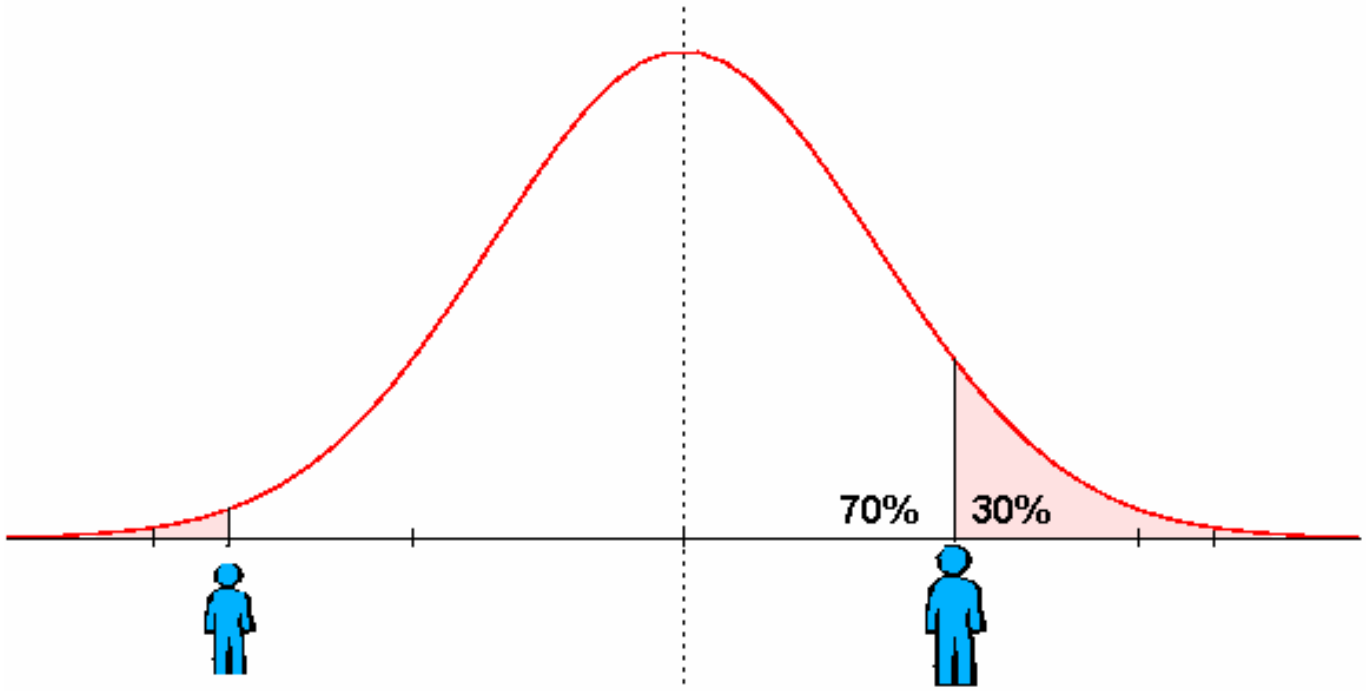
Un brevísimo resumen sobre estadísticos

- Posición
  - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
    - Cuantiles, percentiles, cuartiles, deciles,...
- Centralización
  - Indican valores con respecto a los que los datos parecen agruparse.
    - Media, mediana y moda
- Dispersión
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
    - Desviación típica, coeficiente de variación, rango, varianza
- Forma
  - Asimetría
  - Apuntamiento o curtosis

Estadísticos de posición

- Se define el **cuantil** de orden α como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada α.

- Casos particulares son los percentiles, cuartiles, deciles, quintiles,...



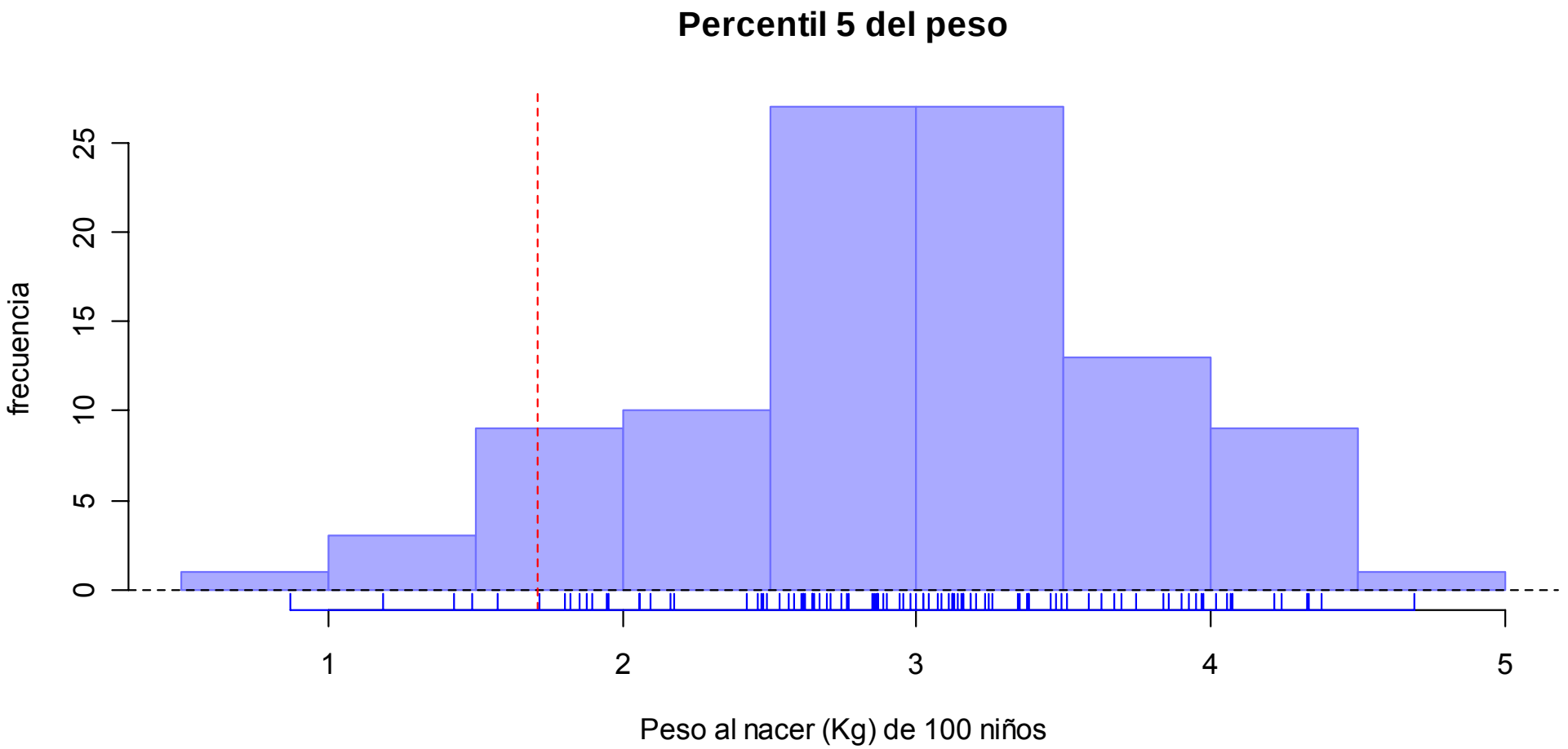
Estadísticos de posición

- **Percentil** de orden k = cuantil de orden k/100
  - La mediana es el percentil 50
  - El percentil de orden 15 deja por debajo al 15% de las observaciones. Por encima queda el 85%

- **Cuartiles**: Dividen a la muestra en 4 grupos con frecuencias similares.
  - Primer cuartil = Percentil 25 = Cuantil 0,25
  - Segundo cuartil = Percentil 50 = Cuantil 0,5 = mediana
  - Tercer cuartil = Percentil 75 = cuantil 0,75

Ejemplos

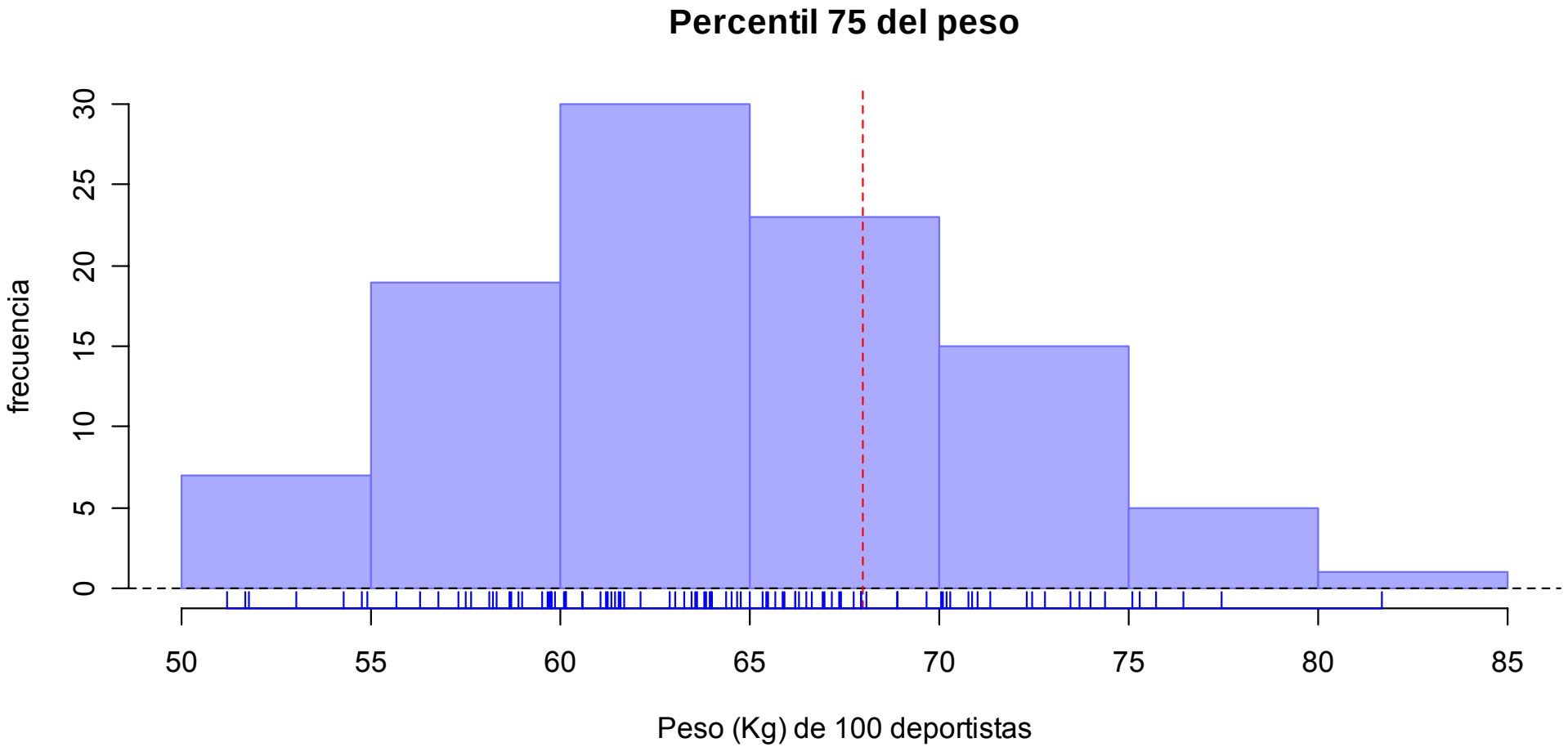
- El 5% de los recién nacidos tiene un peso demasiado bajo. ¿Qué peso se considera “demasiado bajo”?
  - Percentil 5 o cuantil 0,05



Ejemplos

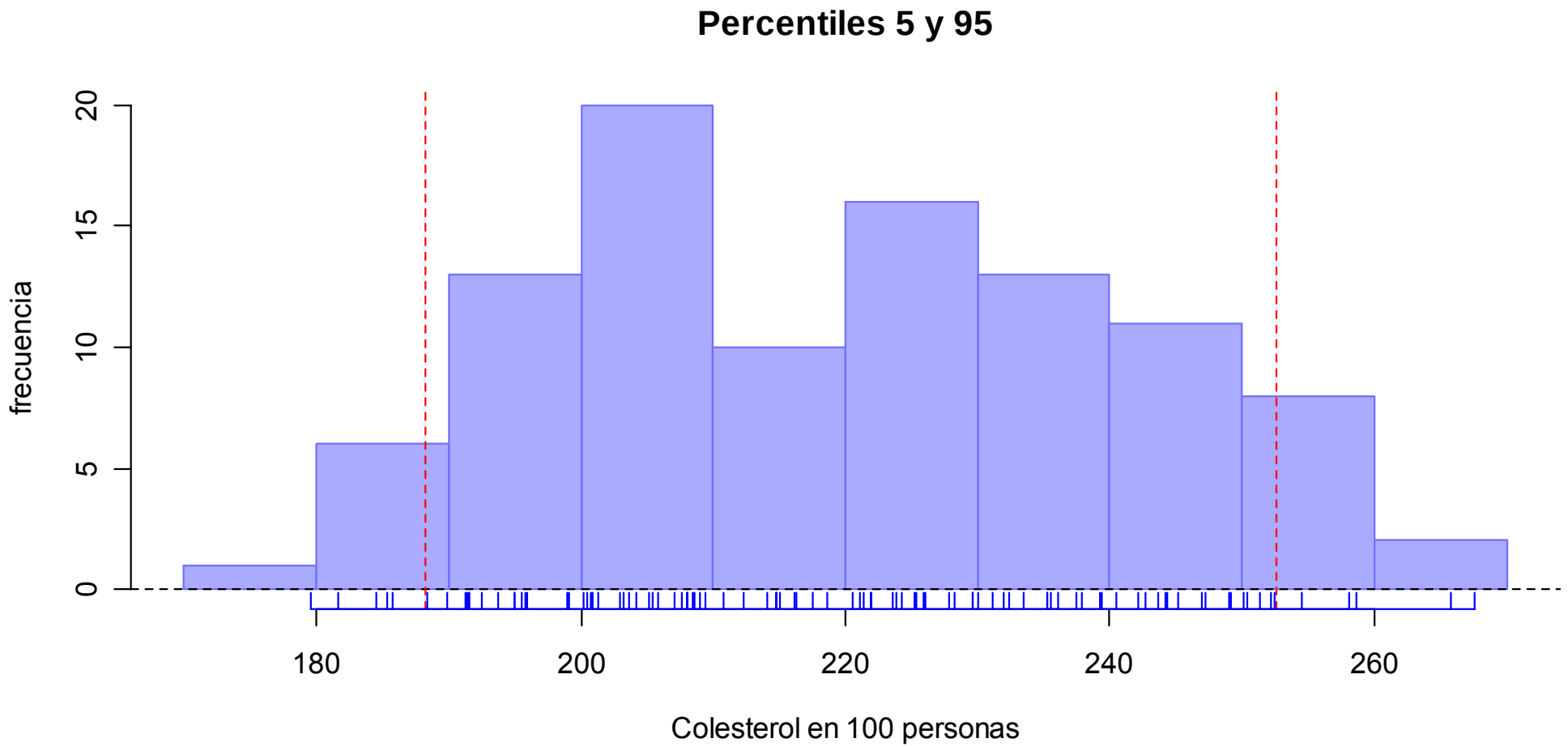
¿Qué peso es superado sólo por el 25% de los individuos?

- Percentil 75 o tercer cuartil



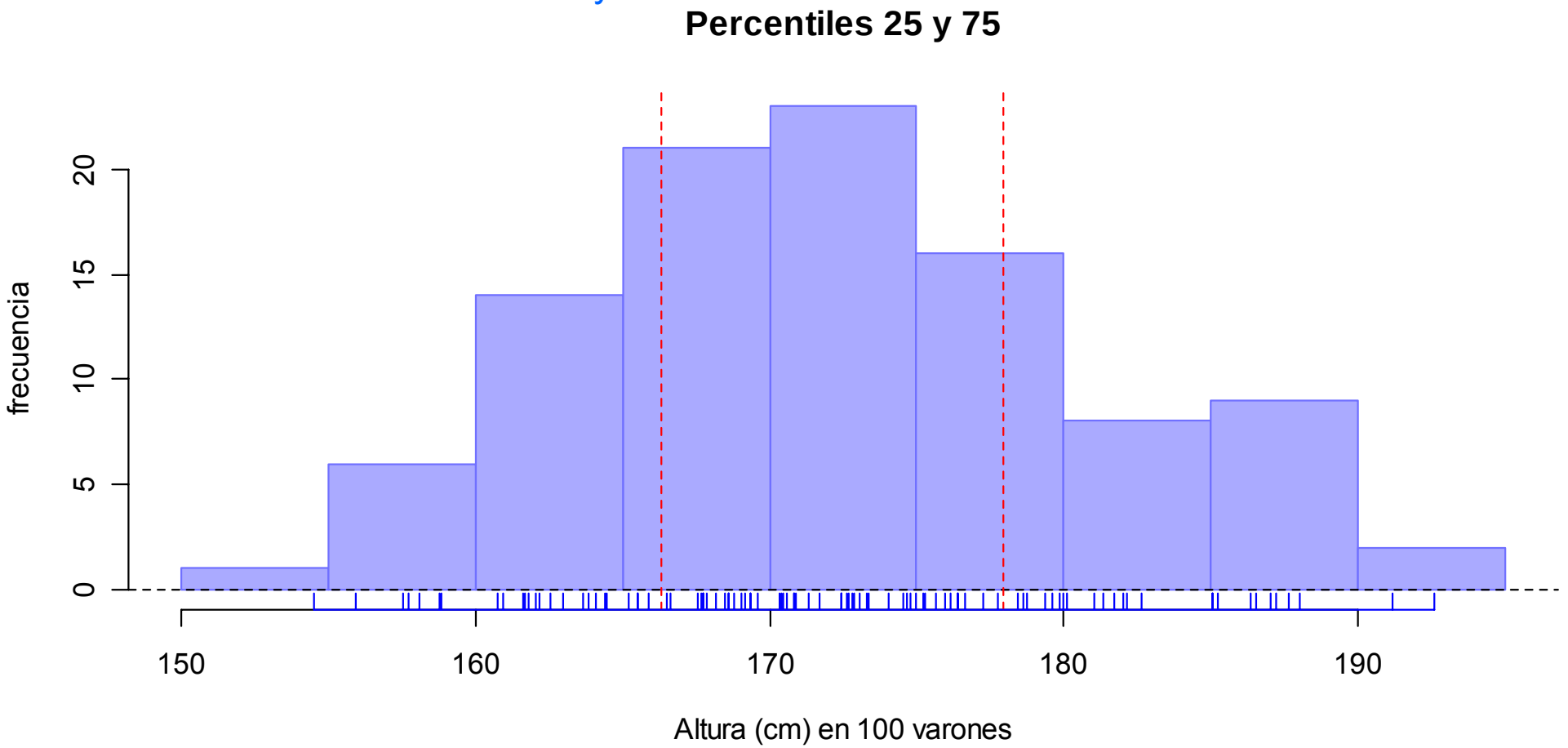
Ejemplos

- El colesterol se distribuye simétricamente en la población. Supongamos que se consideran patológicos los valores extremos. El 90% de los individuos son normales ¿Entre qué valores se encuentran los individuos normales?



Ejemplos

- ¿Entre qué valores se encuentran la mitad de los individuos “más normales” de una población?
  - Entre el cuartil 1º y 3º

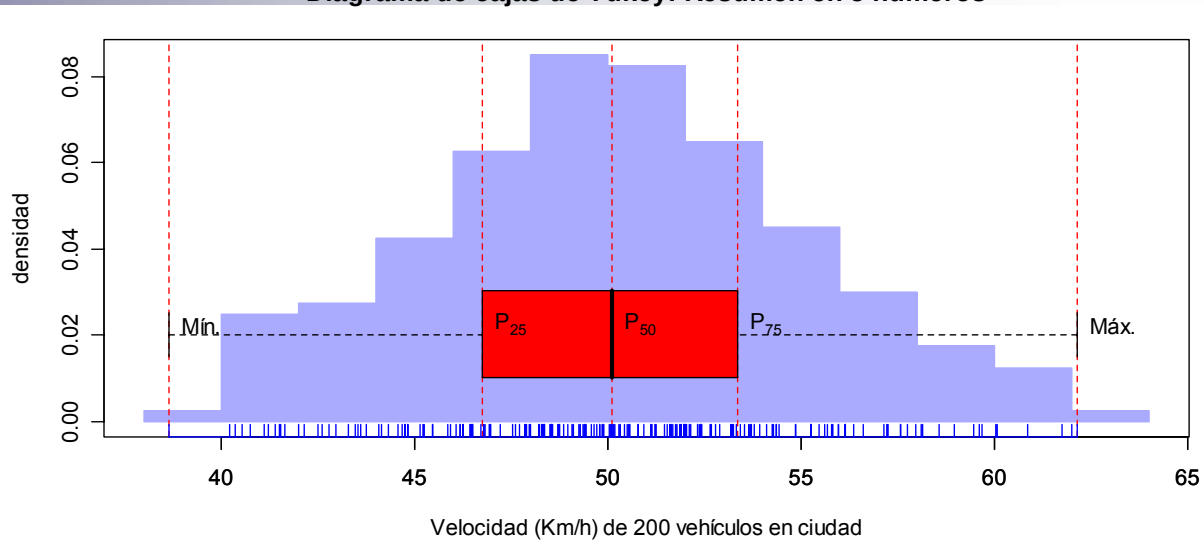


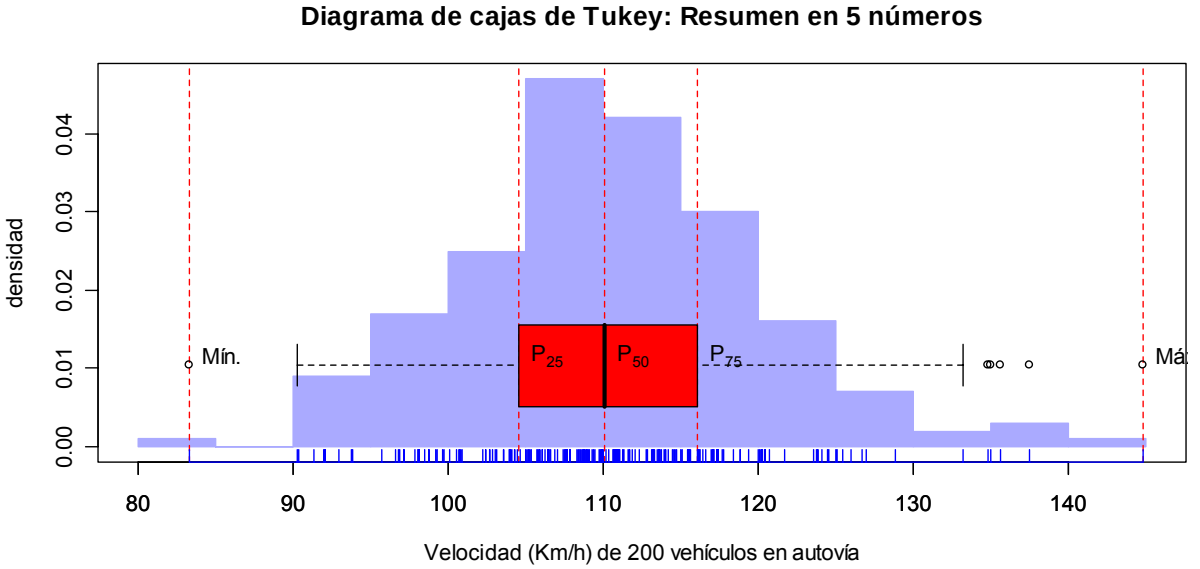
Diagramas de Tukey

- Resumen con 5 números:
  - Mínimo, cuartiles y máximo.
  - Suelen dar una buena idea de la distribución.

- La zona central, 'caja', contiene al 50% central de las observaciones.
  - Su tamaño se llama 'rango intercuartílico' (R.I.)

- Es costumbre que 'los bigotes', no lleguen hasta los extremos, sino hasta las observaciones que se separan de la caja en no más de 1,5 R.I.
  - Más allá de esa distancia se consideran anómalas, y así se marcan.





Ejemplo

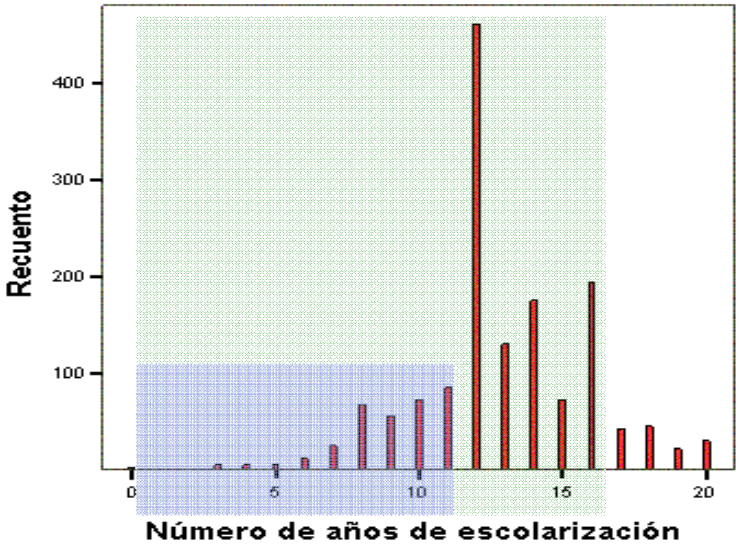
Número de años de escolarización

	Frecuencia	Porcentaje	Porcentaje acumulado	
3	5	,3	,3	
4	5	,3	,7	
5	6	,4	1,1	
6	12	,8	1,9	
7	25	1,7	3,5	
8	68	4,5	8,0	
9	56	3,7	11,7	
10	73	4,8	16,6	
11	85	5,6	22,2	≥20%?
12	461	30,6	52,8	
13	130	8,6	61,4	
14	175	11,6	73,0	
15	73	4,8	77,9	
16	194	12,9	90,7	≥ 90%?
17	43	2,9	93,6	
18	45	3,0	96,6	
19	22	1,5	98,0	
20	30	2,0	100,0	
Total	1508	100,0		

Estadísticos

Número de años de escolarización

N	Válidos	1508
	Perdidos	0
Media		12,90
Mediana		12,00
Moda		12
Percentiles	10	9,00
	20	11,00
	25	12,00
	30	12,00
	40	12,00
	50	12,00
	60	13,00
	70	14,00
	75	15,00
	80	16,00
	90	16,00



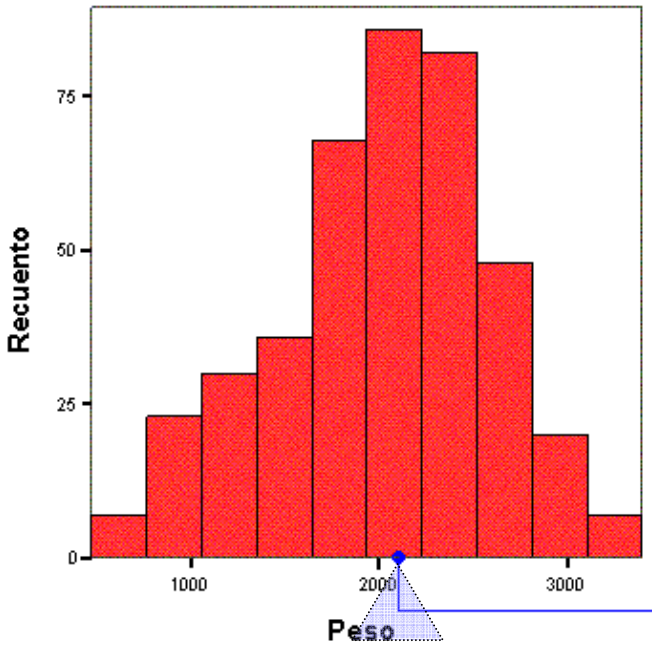
Estadísticos de centralización

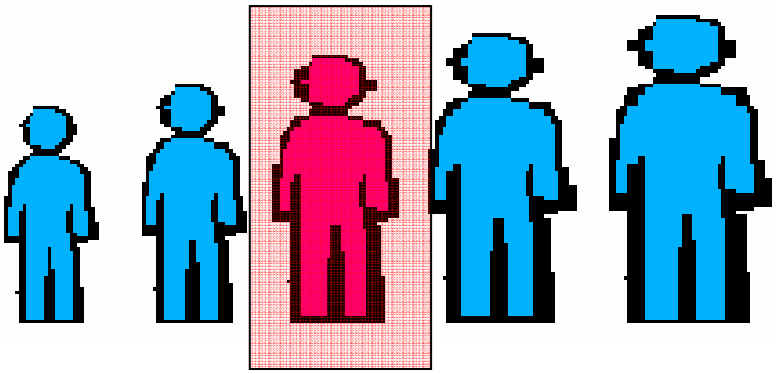
Añaden unos cuantos casos particulares a las medidas de posición. En este caso son medidas que buscan posiciones (valores) con respecto a los cuales los datos muestran tendencia a agruparse.

- Media Es la media aritmética (promedio) de los valores de una variable. Suma de los valores dividido por el tamaño muestral.
  - Media de 2,2,3,7 es (2+2+3+7)/4=3,5
  - Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.
  - Centro de gravedad de los datos

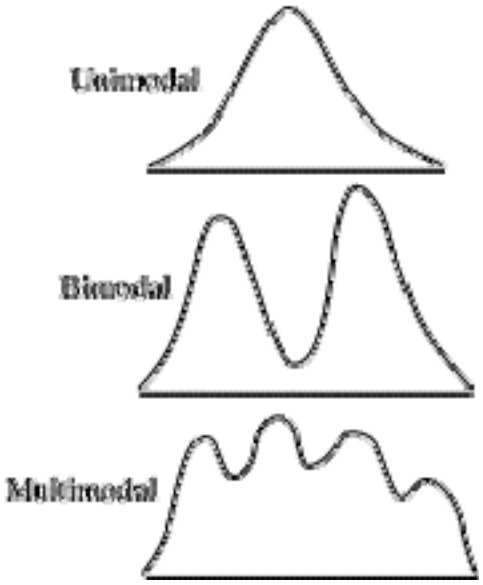
- Mediana Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
  - Mediana de 1,2,4,**5**,6,6,8 es 5
  - Mediana de 1,2,4,**5**,**6**,6,8,9 es (5+6)/2=5,5
  - Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.
    - Mediana de 1,2,4,**5**,6,6,800 es 5. ¡La media es 117,7!

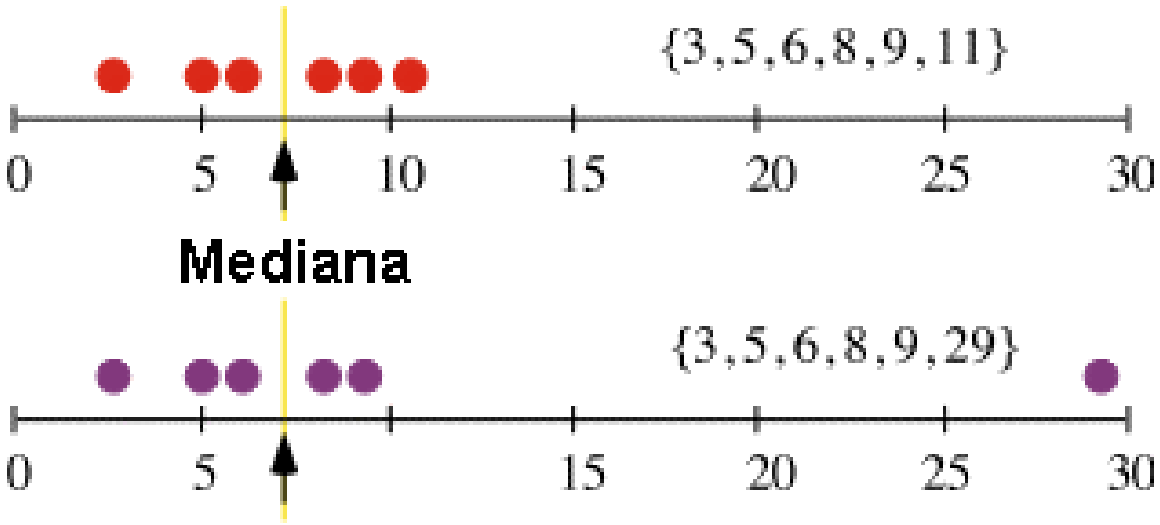
- Moda Es el/los valor/es donde la distribución de frecuencia alcanza un máximo.





Altura mediana





Algunas fórmulas

- Datos sin agrupar: $x_1, x_2, \ldots, x_n$
  - Media

$$\bar{x} = \frac{\sum_i x_i}{n}$$

- Datos organizados en tabla
  - si está en intervalos usar como x_i las marcas de clase. Si no ignorar la columna de intervalos.
  - Media

$$\bar{x} = \frac{\sum_i x_i n_i}{n}$$

Variable		fr.	fr. ac.
$L_0 - L_1$	x_1	n_1	N_1
$L_1 - L_2$	x_2	n_2	N_2
...			
$L_{k-1} - L_k$	x_k	n_k	N_k
n			

  - Cuantil de orden α
    - i es el menor intervalo que tiene frecuencia acumulada superior a α ·n
    - α=0,5 es mediana

$$C_\alpha = L_{i-1} + \frac{\alpha \cdot n - N_{i-1}}{n_i}(L_i - L_{i-1})$$

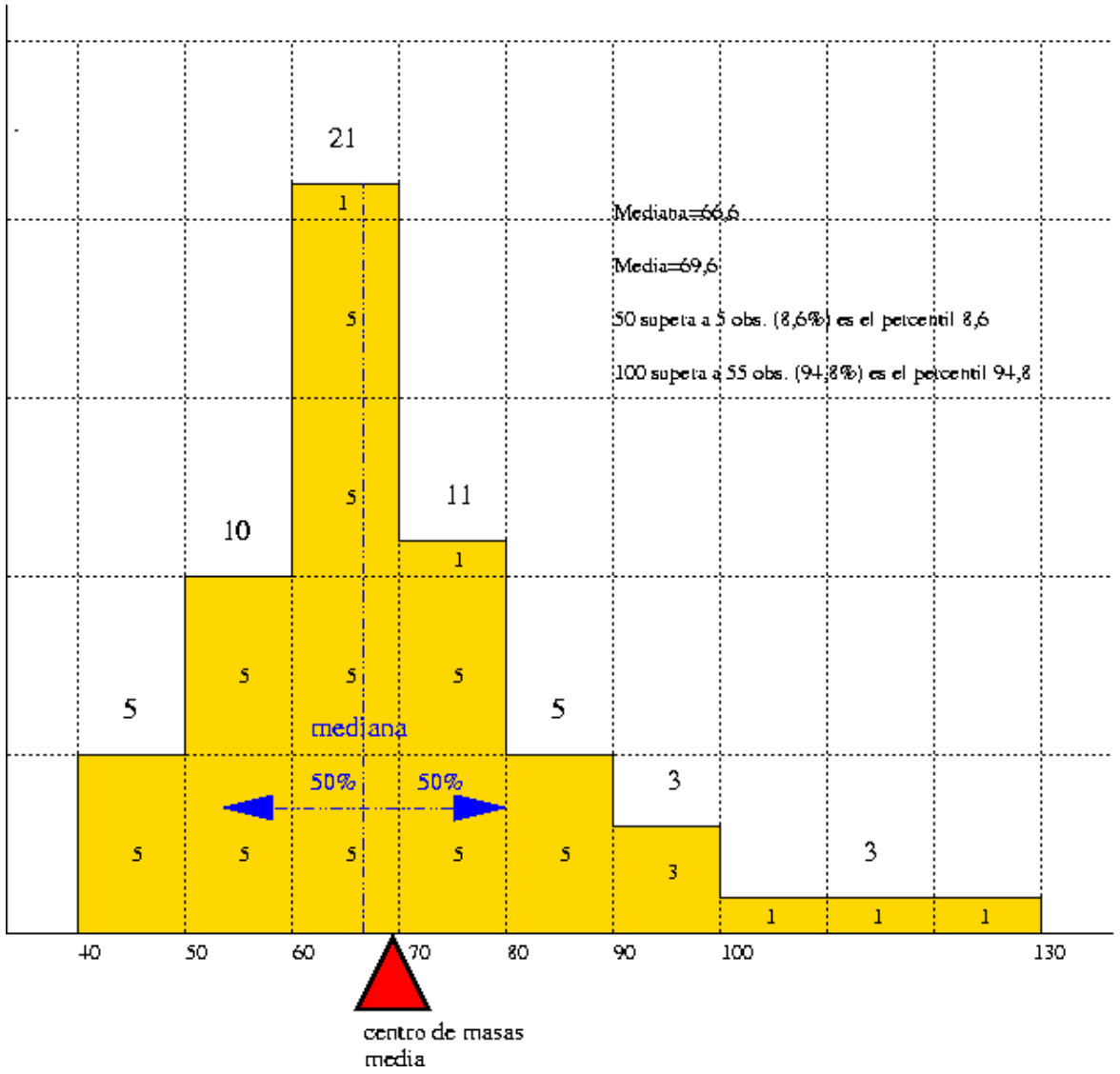
Ejemplo con variable en intervalos

Peso	M. Clase	frec	Fr. acum.
40 – 50	45	5	5
50 – 60	55	10	15
60 – 70	65	21	36
70 - 80	75	11	47
80 - 90	85	5	52
90 - 100	95	3	55
100 – 130	115	3	58

En el histograma se identifica “unidad de área” con “individuo”.

Para calcular la media es necesario elegir un punto representante del intervalo: La marca de clase.

La media se desplaza hacia los valores extremos. No coincide con la mediana. Es un punto donde el histograma “estaría en equilibrio” si tuviese masa.



Ejemplo (continuación)

Peso	M. Clase	Fr.	Fr. ac.
40 – 50	45	5	5
50 – 60	55	10	15
60 – 70	65	21	36
70 - 80	75	11	47
80 - 90	85	5	52
90 - 100	95	3	55
100 – 130	115	3	58
58			

$$\bar{x} = \frac{\sum_i x_i n_i}{n} = \frac{45 \cdot 5 + 55 \cdot 10 + \ldots + 115 \cdot 3}{58} = 69{,}3$$

$$Mediana = C_{0,5} = L_{i-1} + \frac{0{,}5 \cdot 58 - N_{i-1}}{n_i}(L_i - L_{i-1})$$

$$= 60 + \frac{0{,}5 \cdot 58 - 15}{21}(70 - 60) = 66{,}\widehat{6}$$

$$P_{75} = C_{0,75} = L_{i-1} + \frac{0{,}75 \cdot 58 - N_{i-1}}{n_i}(L_i - L_{i-1}) = 70 + \frac{43{,}5 - 36}{11}(80 - 70) = 76{,}8$$

- Moda = marca de clase de (60,70] = 65
  - Cada libro ofrece una fórmula diferente para la moda (difícil estar al día.)

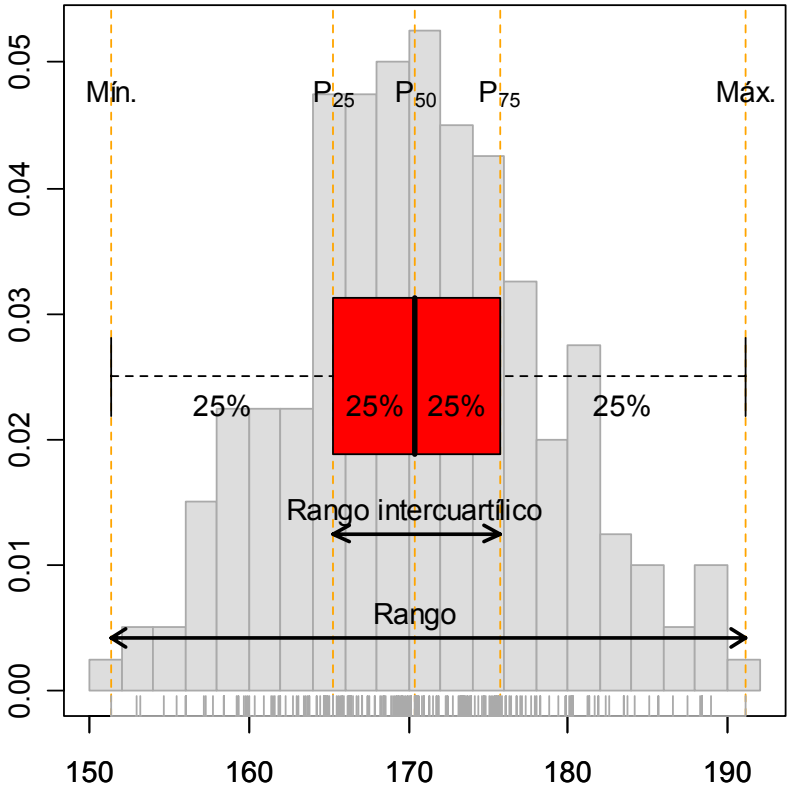
Variabilidad o dispersión

- Los estudiantes del propedeutico reciben diferentes calificaciones en la asignatura introducción a la estadística (variabilidad). ¿A qué puede deberse?
  - Diferencias individuales en el conocimiento de la materia.
- ¿Podría haber otras razones (fuentes de variabilidad)?
- Por ejemplo supongamos que todos los alumnos poseen el mismo nivel de conocimiento. ¿Las notas serían las mismas en todos? Seguramente No.
  - Dormir poco el día del examen, el croissant estaba envenenado...
    - Diferencias individuales en la habilidad para hacer un examen.
  - El examen no es una medida perfecta del conocimiento.
    - Variabilidad por error de medida.
  - En alguna pregunta difícil, se duda entre varias opciones, y al azar se elige la mala
    - Variabilidad por azar, aleatoriedad.

Medidas de dispersión

Miden el grado de dispersión (variabilidad) de los datos, independientemente de su causa.

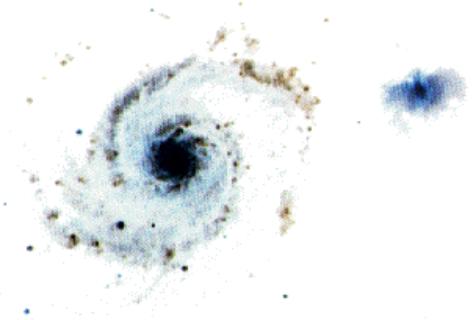
- Amplitud o Rango:
  Diferencia entre observaciónes extremas.
  - 2,1,4,3,8,4. El rango es 8-1=7
  - Es muy sensible a los valores extremos.



- Rango intercuartílico :
  - Es la distancia entre primer y tercer cuartil.
    - Rango intercuartílico = $P_{75} - P_{25}$
  - Parecida al rango, pero eliminando las observaciones más extremas inferiores y superiores.
  - No es tan sensible a valores extremos.

- Varianza S^2 : Mide el promedio de las desviaciones (al cuadrado) de las observaciones con respecto a la media.

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$



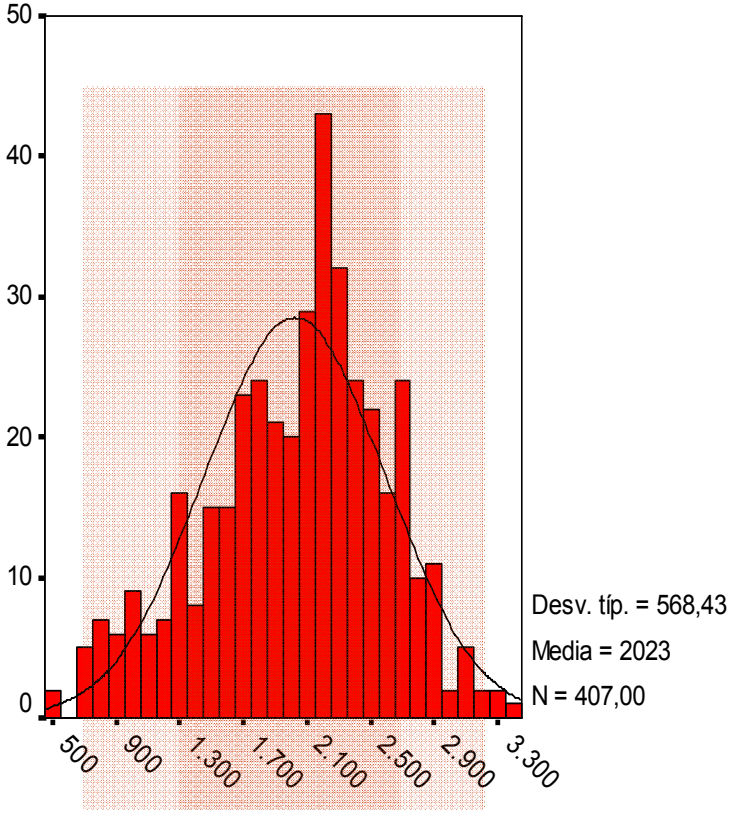
  - Es sensible a valores extremos (alejados de la media).
  - Sus unidades son el cuadrado de las de la variable.

Desviación típica

Es la raíz cuadrada de la varianza

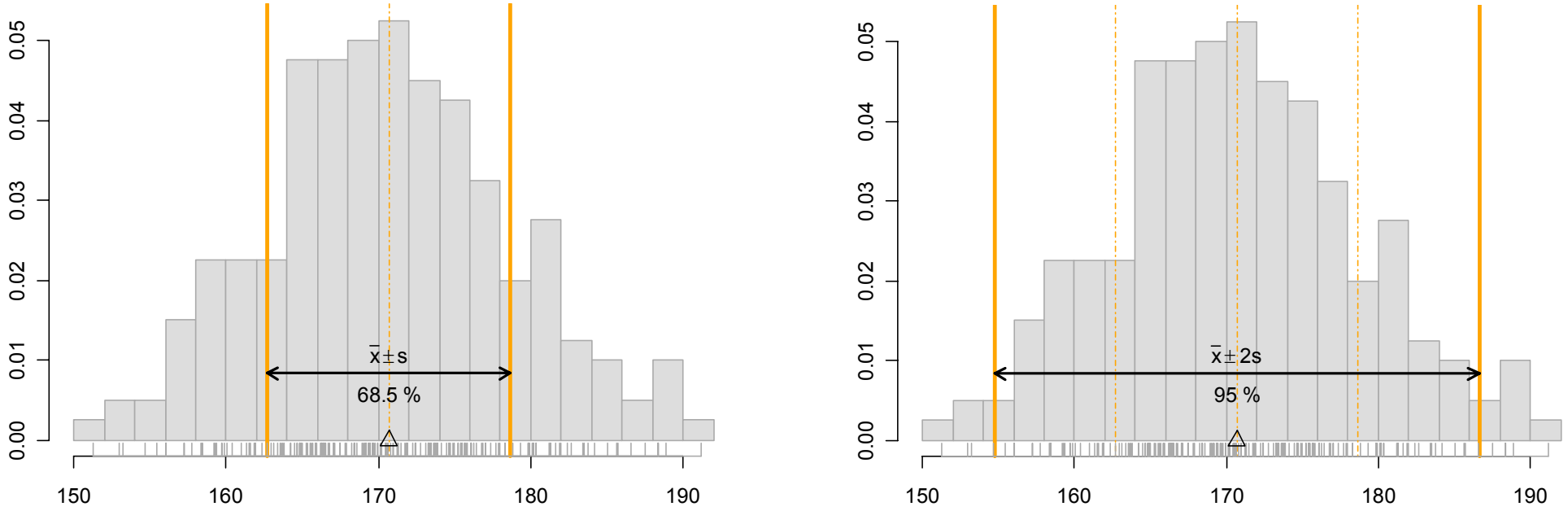
$$S = \sqrt{S^2}$$

- Tiene las misma dimensionalidad (unidades) que la variable. Versión ‘estética’ de la varianza.
- Cierta distribución que veremos más adelante (normal o gaussiana) quedará completamente determinada por la media y la desviación típica.
  - A una distancia de una desv. típica de la media hay más de la ‘más de la mitad’.
  - A una distancia de dos desv. típica de la media las tendremos casi todas.



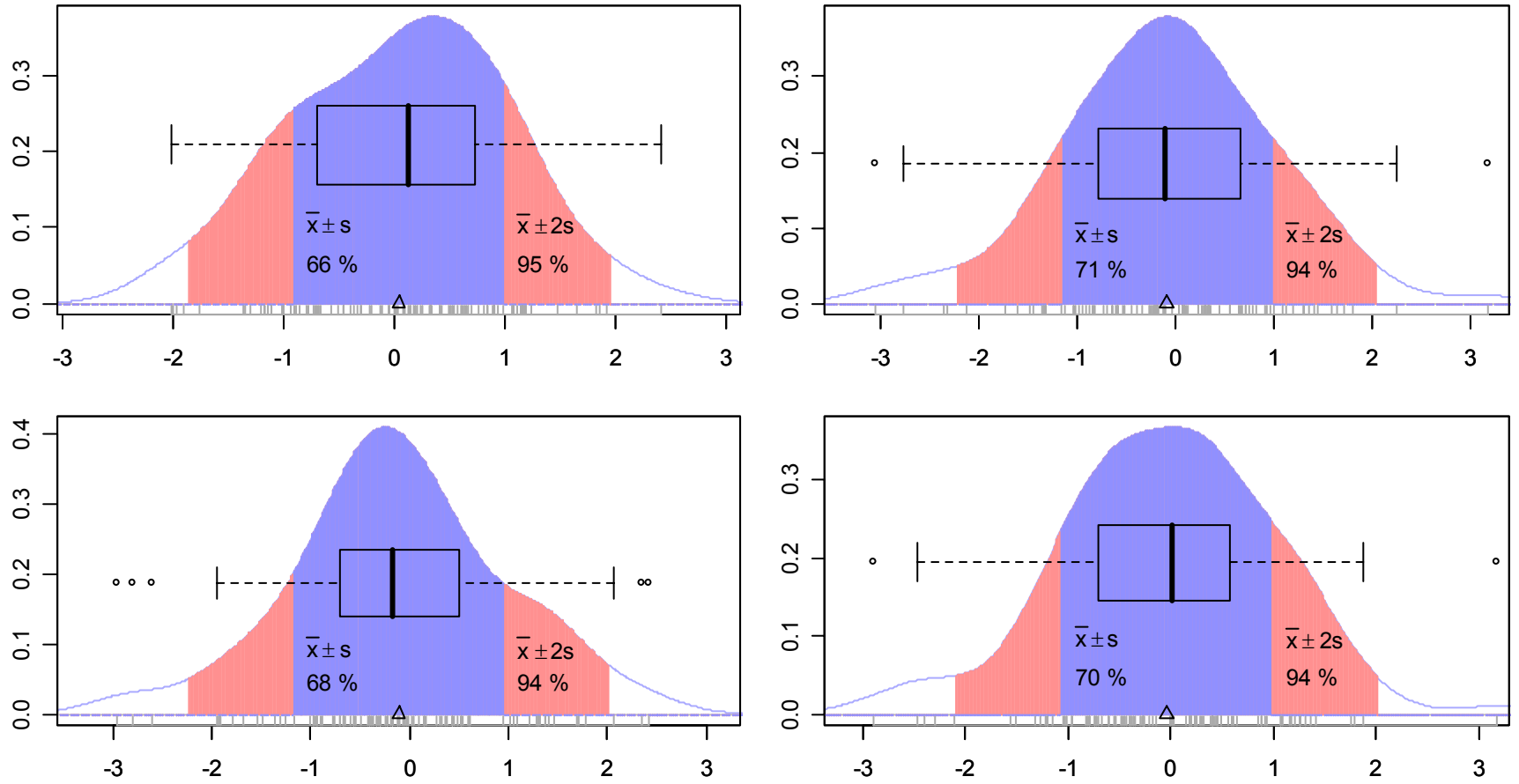
Peso recién nacidos en partos gemelares

Dispersión en distribuciones ‘normales’



- Centrado en la media y a una desv. típica de distancia hay aproximadamente el 68% de las observaciones.
- A dos desviaciones típicas tenemos el 95% (aprox.)

- Datos ‘casi normales’. Eje ‘x’ medido en desviaciones típicas…
  - ¿Encuentras relación entre rango intercuartílico y desviación típica?
  - ¿Y entre los ‘bigotes’ y dos desviaciones típicas? ¿Podrías caracterizar las observaciones anómalas?



Coeficiente de variación

$$CV = \frac{S}{\bar{x}}$$

Es la razón entre la desviación típica y la media.

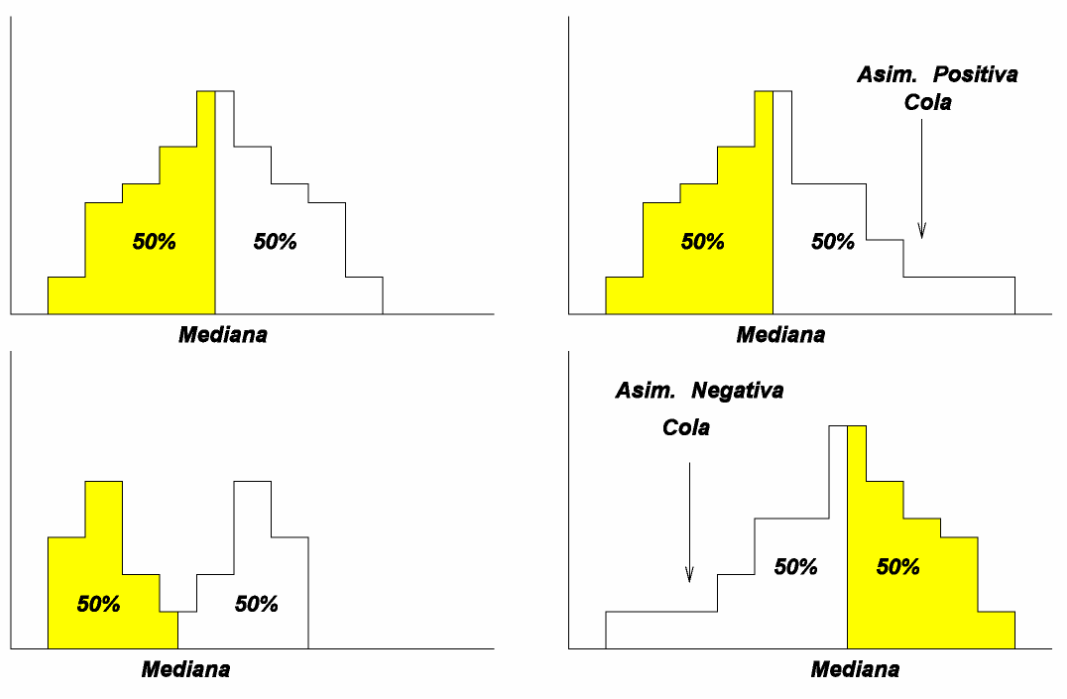
- Mide la desviación típica en forma de "qué tamaño tiene con respecto a la media"
- También se la denomina variabilidad relativa.
- Es frecuente mostrarla en porcentajes
  - Si la media es 80 y la desviación típica 20 entonces CV=20/80=0,25=25% (variabilidad relativa)

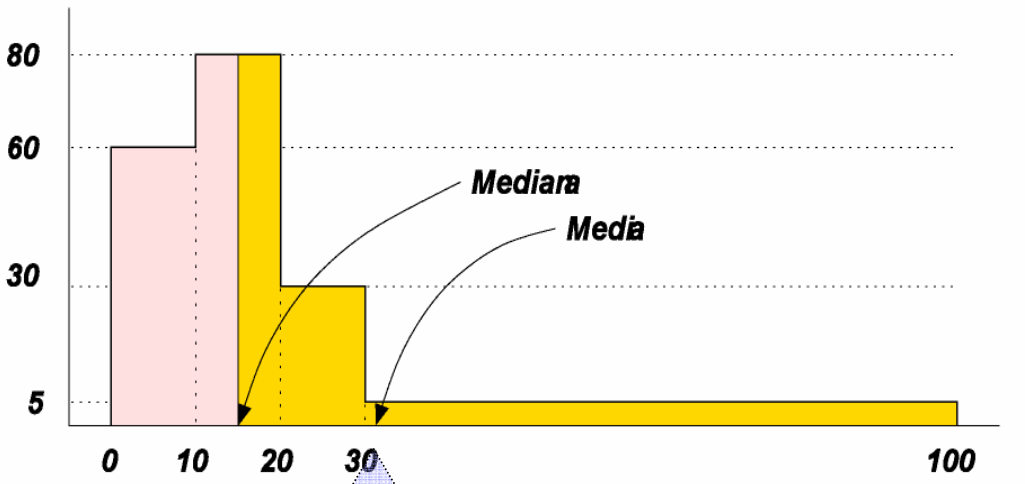
- Es una cantidad **adimensional**. Interesante para comparar la variabilidad de diferentes variables.
  - Si el peso tiene CV=30% y la altura tiene CV=10%, los individuos presentan más dispersión en peso que en altura.

- No debe usarse cuando la variable presenta valores negativos o donde el valor 0 sea una cantidad fijada arbitrariamente
  - Por ejemplo 0ºC ≠ 0ºF

Asimetría o Sesgo

- Una distribución es simétrica si la mitad izquierda de su distribución es la imagen especular de su mitad derecha.
- En las distribuciones simétricas media y mediana coinciden. Si sólo hay una moda también coincide
- La asimetría es positiva o negativa en función de a qué lado se encuentra la cola de la distribución.
- La media tiende a desplazarse hacia las valores extremos (colas).
- Las discrepancias entre las medidas de centralización son indicación de asimetría.

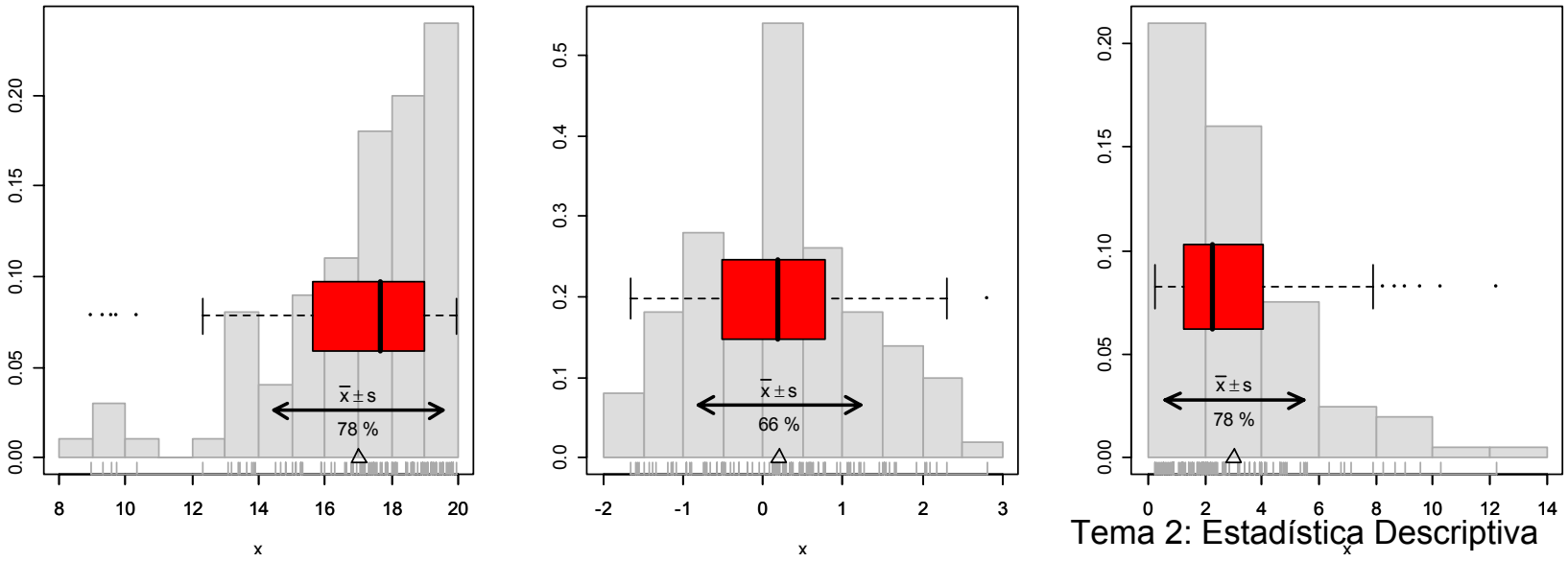




Estadísticos para detectar asimetría

- Hay diferentes estadísticos que sirven para detectar asimetría.
  - Basado en diferencia entre estadísticos de tendencia central.
  - Basado en la diferencia entre el 1º y 2º cuartiles y 2º y 3º.
  - Basados en *desviaciones con signo al cubo con respecto a la media*.
    - Los calculados con ordenador. Es pesado de hacer a mano.

- En función del signo del estadístico diremos que la asimetría es positiva o negativa.
  - Distribución simétrica ➔ asimetría nula.



Apuntamiento o curtosis

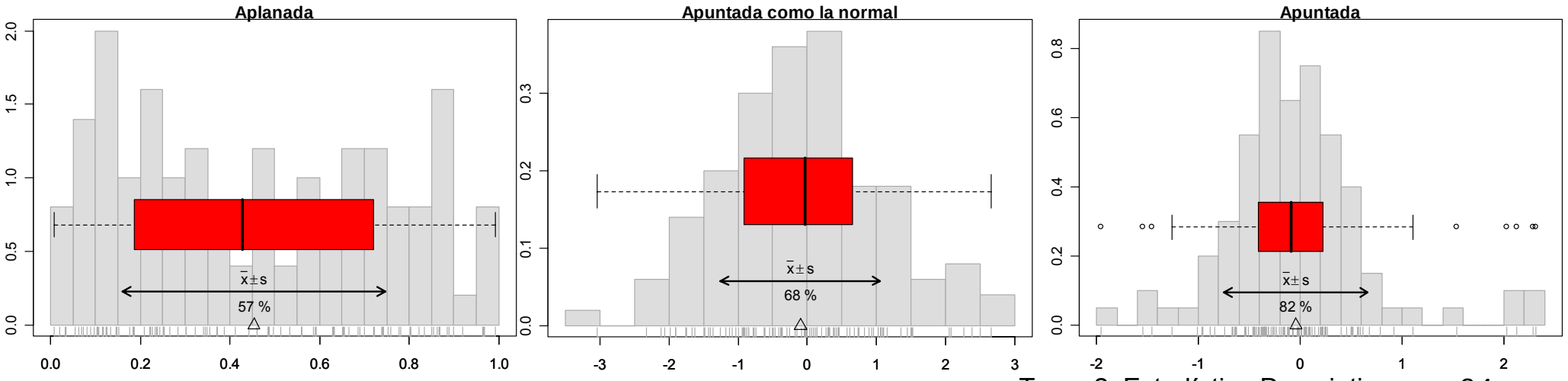
La curtosis nos indica el grado de apuntamiento (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana. Es adimensional.

Platicúrtica (aplanada): curtosis < 0

Mesocúrtica (como la normal): curtosis = 0

Leptocúrtica (apuntada): curtosis > 0

En el curso serán de especial interés las mesocúrticas y simétricas (parecidas a la normal).



PROPEDÉUTICO

Modulo: Introducción a la estadística

Guía de estudio para la Unidad 3: Modelos probabilísticos

UTILIZANDO LA INFORMACIÓN DE ESTA SECCIÓN ó DEL LIBRO BIOSTATISTICAL ANALYSIS, ZAR, J. PRENTICE-HALL 1984 Ó 1999 RESUELVE CADA UNO DE LOS INCISOS:

1. ¿Qué es una variable aleatoria?
2. ¿Qué es una función de probabilidad?
3. Define qué es una distribución de Bernoulli.
4. ¿Cuáles son los parámetros que describen a la distribución Normal?
5. Menciona las características de la distribución normal
6. ¿Qué es la estandarización y cómo se relaciona con el valor tipificado?
7. ¿Cuál es la fórmula para calcular el valor tipificado?
8. ¿Cuál es la distribución de Chi-cuadrado?

Introducción a la Estadística

Tema 3: Modelos probabilísticos

Variable aleatoria

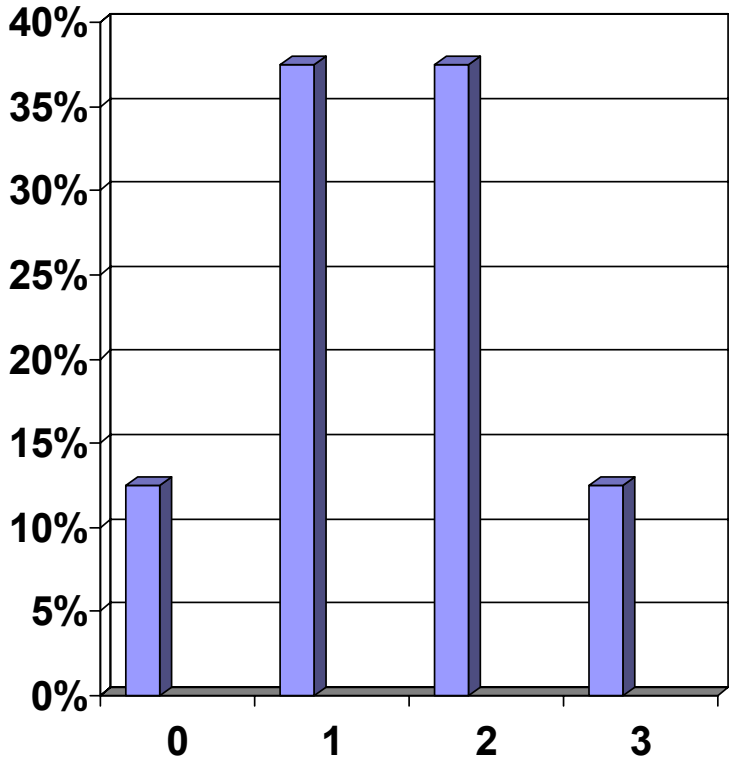
- El resultado de un experimento aleatorio puede ser descrito en ocasiones como una cantidad numérica.

- En estos casos aparece la noción de variable aleatoria
  - Función que asigna a cada suceso un número.

- Las variables aleatorias pueden ser discretas o continuas

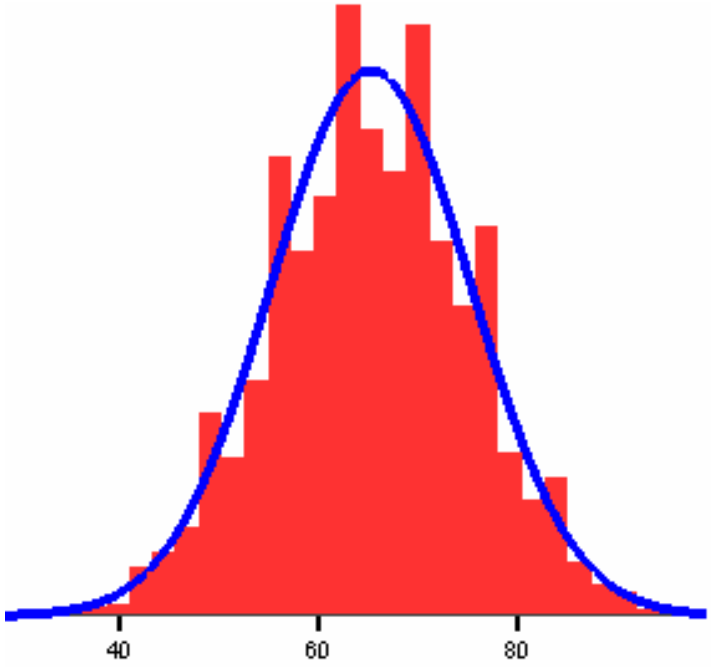
Función de probabilidad (V. Discretas)

- Asigna a cada posible valor de una variable discreta su probabilidad.
  - Recuerda los conceptos de frecuencia relativa y diagrama de barras.
- Ejemplo
  - Número de caras al lanzar 3 monedas.



Función de densidad (V. Continuas)

- Definición
  - Es una función no negativa de integral 1.
    - Piénsalo como la generalización del histograma con frecuencias relativas para variables continuas.



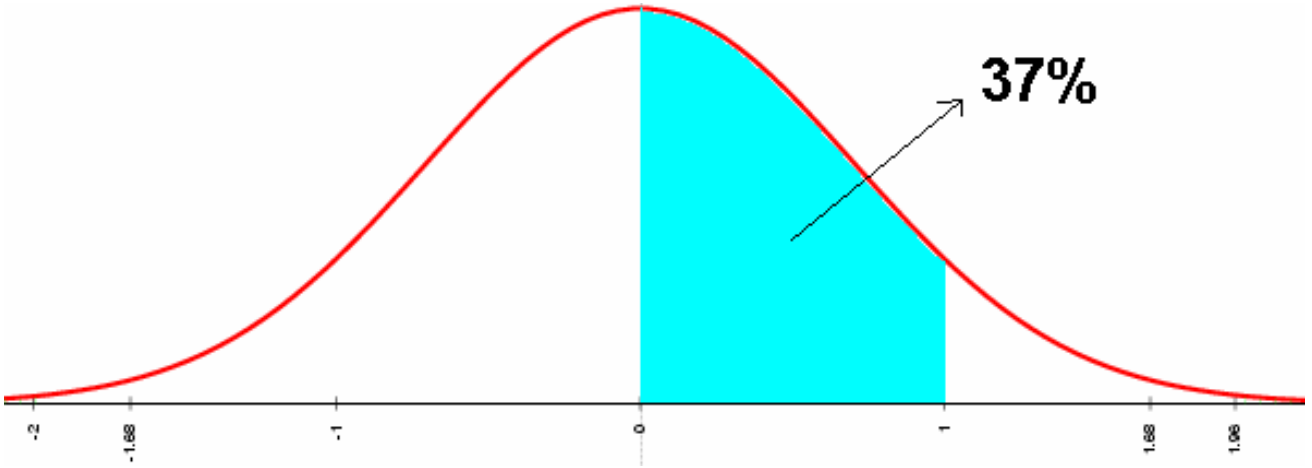
- ¿Para qué lo voy a usar?
  - Nunca lo vas a usar directamente.
  - Sus valores no representan probabilidades.

¿Para qué sirve la f. densidad?

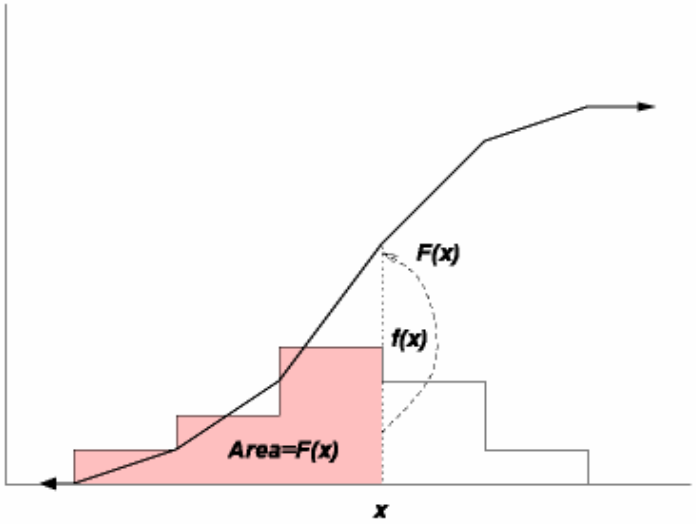
- Muchos procesos aleatorios vienen descritos por variables de forma que son conocidas las probabilidades en intervalos.

- La integral definida de la función de densidad en dichos intervalos coincide con la probabilidad de los mismos.

- Es decir, identificamos la probabilidad de un intervalo con el **área** bajo la función de densidad.



Función de distribución



- Es la función que asocia a cada valor de una variable, la probabilidad acumulada de los valores inferiores o iguales.
  - Piénsalo como la generalización de las frecuencias acumuladas. Diagrama integral.
    - A los valores extremadamente bajos les corresponden valores de la función de distribución cercanos a cero.
    - A los valores extremadamente altos les corresponden valores de la función de distribución cercanos a uno.
- Lo encontraremos en los artículos y aplicaciones en forma de "p-valor", valor observado de significancia,…

¿Para qué sirve la f. distribución?

- Contrastar lo anómalo de una observación concreta.
  - Sé que una persona de altura 210cm es "anómala" porque la función de distribución en 210 es muy alta.
  - Sé que una persona adulta que mida menos de 140cm es "anómala" porque la función de distribución es muy baja para 140cm.
  - Sé que una persona que mida 170cm no posee una altura nada extraña pues su función de distribución es aproximadamente 0,5.
- Relaciónalo con la idea de cuantil.
- En otro contexto (pruebas de hipótesis) podremos observar unos resultados experimentales y contrastar lo "anómalos" que son en conjunto con respecto a una hipótesis determinada.

Valor esperado y varianza de una v.a. X

- Valor esperado
  - Se representa mediante E[X] ó μ
  - Es el equivalente a la media
    - Más detalles: Ver libros de estadística.

- Varianza
  - Se representa mediante VAR[X] o σ^2
  - Es el equivalente a la varianza
  - Se llama desviación típica a σ
    - Más detalles: Ver libro de estadística

Algunos modelos de v.a.

- Hay v.a. que aparecen con frecuencia en las Ciencias de Biolióg**i**cas.
  - Experimentos dicotómicos.
    - Bernoulli
  - Contar éxitos en experimentos dicotómicos repetidos:
    - Binomial
  - Y en otras muchas ocasiones…
    - Distribución normal (gaussiana, campana,…)
- El resto del tema está dedicado a estudiar estas distribuciones especiales.

Distribución de Bernoulli

- Tenemos un experimento de Bernoulli si al realizar un experimento sólo son posibles dos resultados:
  - X=1 (**éxito**, con probabilidad p)
  - X=0 (**fracaso**, con probabilidad q=1-p)
    - Lanzar una moneda y que salga cara.
      - p=1/2
    - Elegir una persona de la población y que esté enfermo.
      - p=1/1000 = probabilidad de tener la enfermedad
    - Aplicar un tratamiento a un planta y que ésta se cure.
      - p=95%, probabilidad de que el individuo se cure

- Como se aprecia, en experimentos donde el resultado es dicotómico, la variable queda perfectamente determinada conociendo el parámetro **p**.

Ejemplo de distribución de Bernoulli.

- Se ha observado estudiando 2000 accidentes de tráfico con impacto frontal y cuyos conductores no tenían cinturón de seguridad, que 300 individuos quedaron con secuelas. Describa el experimento usando conceptos de v.a.

- Solución.
  - La noc. frecuentista de prob. nos permite aproximar la probabilidad de tener secuelas mediante 300/2000=0.15=15%
  - X="tener secuelas tras accidente sin cinturón" es variable de Bernoulli
    - X=1 tiene probabilidad $p \approx 0.15$
    - X=0 tiene probabilidad $q \approx 0.85$

Ejemplo de distribución de Bernoulli.

- Se ha observado estudiando 2000 accidentes de tráfico con impacto frontal y cuyos conductores sí tenían cinturón de seguridad, que 10 individuos quedaron con secuelas. Describa el experimento usando conceptos de v.a.

- Solución.
  - La noc. frecuentista de prob. nos permite aproximar la probabilidad de quedar con secuelas por 10/2000=0.005=0.5%
  - X="tener secuelas tras accidente usando cinturón" es variable de Bernoulli
    - X=1 tiene probabilidad p ≈ 0.005
    - X=0 tiene probabilidad q ≈ 0.995

Observación

- En los dos ejemplos anteriores hemos visto cómo enunciar los resultados de un experimento en forma de estimación de parámetros en distribuciones de Bernoulli.
  - Sin cinturón: $p \approx 15\%$
  - Con cinturón: $p \approx 0.5\%$

- En realidad no sabemos en este punto si ambas cantidades son muy diferentes o aproximadamente iguales, pues en otros estudios sobre accidentes, las cantidades de individuos con secuelas hubieran sido con seguridad diferentes.

- Para decidir si entre ambas cantidades existen diferencias estadísticamente significativas necesitamos introducir conceptos de estadística inferencial (extrapolar resultados de una muestra a toda la población).

- Es muy pronto para resolver esta cuestión ahora. Se utilizan las pruebas de X^2.

Distribución binomial

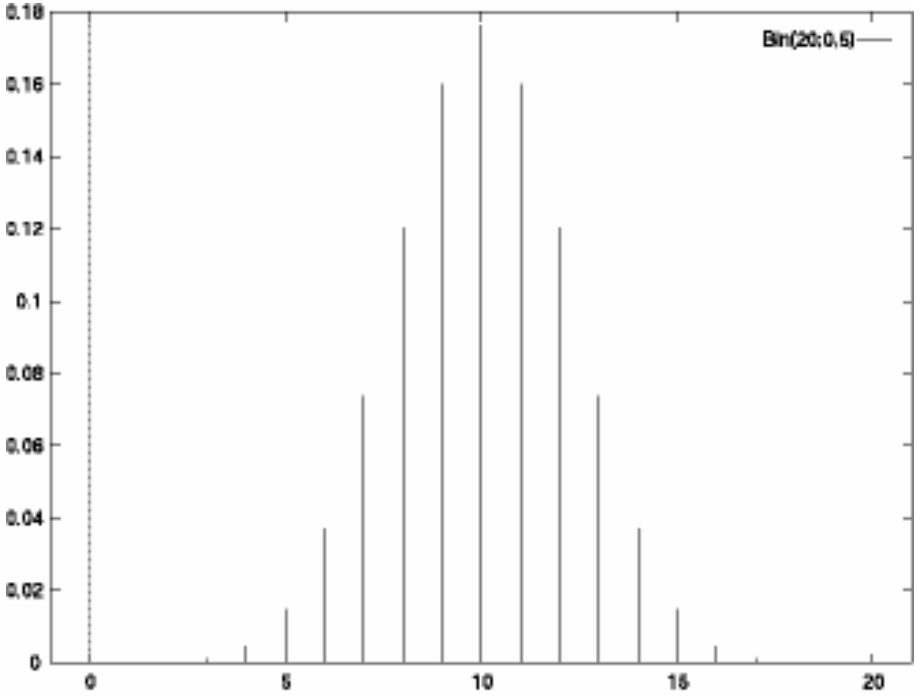
- Función de probabilidad

$$P[X=k]=\binom{n}{k}p^k q^{n-k},\ 0\le k\le n$$

  - Problemas de cálculo si n es grande y/o p cercano a 0 o 1.

- Media: $\mu = n\,p$

- Varianza: $\sigma^2 = n\,p\,q$



Distribución Binomial

- Si se repite un número fijo de veces, **n**, un experimento de Bernoulli con parámetro **p**, el número de éxitos sigue una distribución **binomial** de parámetros (n,p).

  - Lanzar una moneda 10 veces y contar las caras.
    - Bin(n=10, p=1/2)
  - Lanzar una moneda 100 veces y contar las caras.
    - Bin(n=100, p=1/2)
    - Difícil hacer cálculos con esas cantidades. El modelo normal será más adecuado.
  - El número de personas que enfermará (en una población de 500 000 personas) de una enfermedad que desarrolla una de cada 2000 personas.
    - Bin(n=500.000, p=1/2000)
      - Difícil hacer cálculos con esas cantidades.

Distribución normal o de Gauss

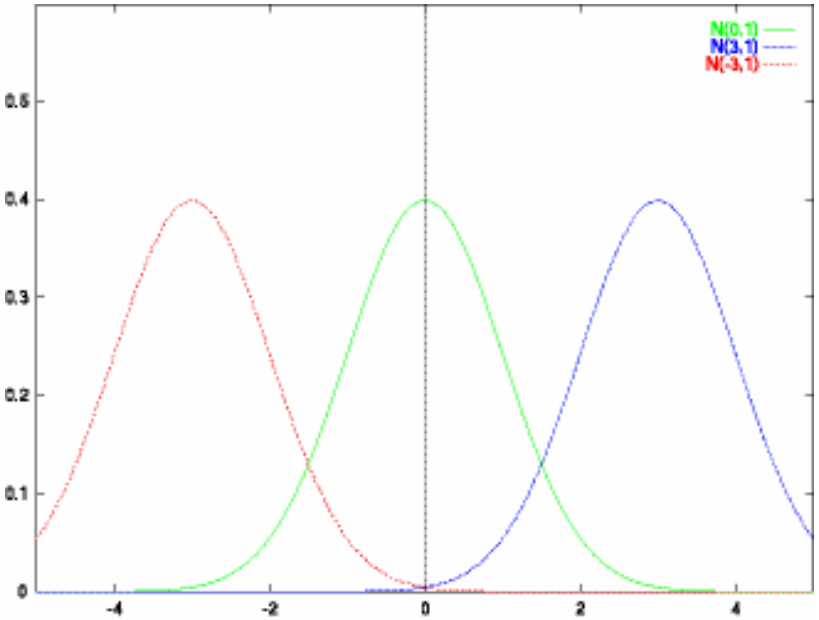
- Está caracterizada por dos parámetros: La media, μ, y la desviación típica, σ.

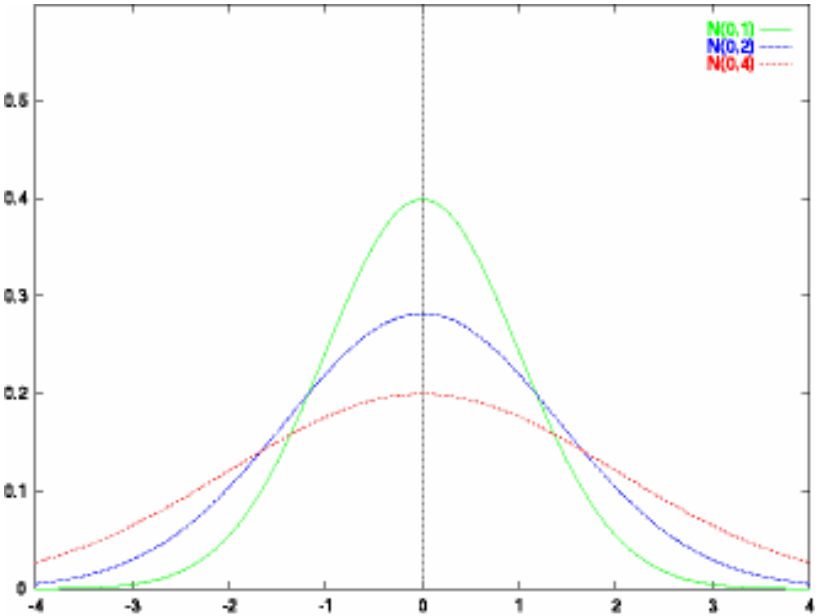
- Su función de densidad es:

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

N(μ, σ): Interpretación geométrica

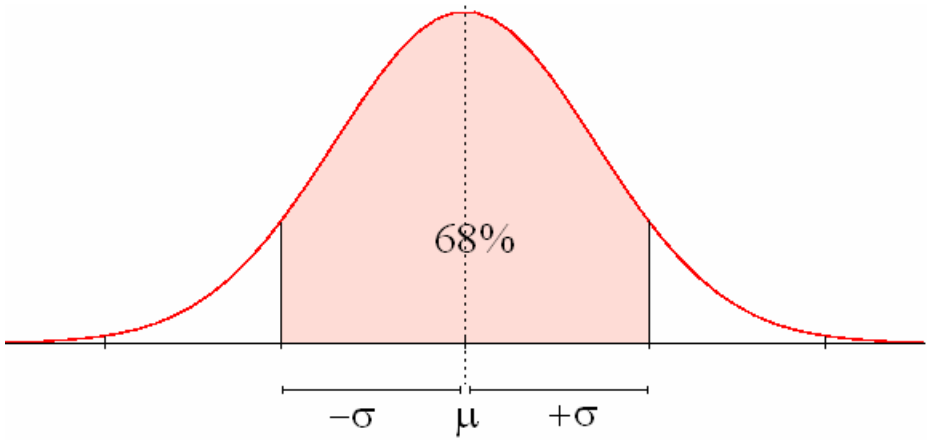
- Se puede interpretar la media como un factor de traslación.
- Y la desviación estándar como un factor de escala, grado de dispersión,…



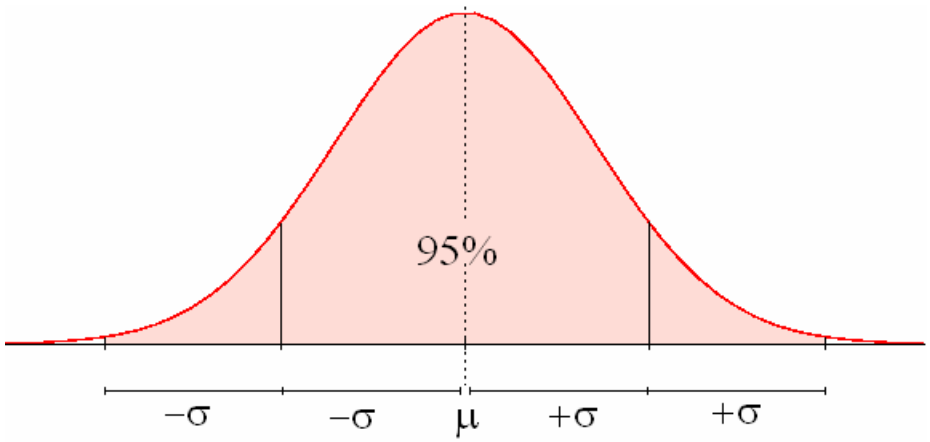


N(μ, σ): Interpretación probabilista

- Entre la media y una desviación típica tenemos siempre la misma probabilidad: aprox. 68%



- Entre la media y dos desviaciones típicas aprox. 95%



Algunas características

- La función de densidad es simétrica, mesocúrtica y unimodal.
  - Media, mediana y moda coinciden.

- Los puntos de inflexión de la fun. de densidad están a distancia **σ** de μ.

- Si tomamos intervalos centrados en μ, y cuyos extremos están…
  - a distancia **σ**, ➔ tenemos probabilidad **68%**
  - a distancia **2 σ**, ➔ tenemos probabilidad **95%**
  - a distancia **2'5 σ** ➔ tenemos probabilidad **99%**

- Todas las distribuciones normales N(μ, σ), pueden ponerse mediante una traslación μ, y un cambio de escala σ, como **N(0,1)**. Esta distribución especial se llama **normal estandarizada**.
  - Justifica la técnica de tipificación, cuando intentamos comparar individuos diferentes obtenidos de sendas poblaciones normales.

Estandarización

- Dada una variable de media μ y desviación típica σ, se denomina valor tipificado,z, de una observación x, a la distancia (con signo) con respecto a la media, medido en desviaciones típicas, es decir

$$z = \frac{x - \mu}{\sigma}$$

- En el caso de variable X normal, la interpretación es clara: Asigna a todo valor de N(μ, σ), un valor de N(0,1) que deja **exáctamente** la misma probabilidad por debajo.

- Nos permite así comparar entre dos valores de dos distribuciones normales diferentes, para saber cuál de los dos es más extremo.

Tabla N(0,1)

Z es normal estandarizada.

Calcular P[Z<1.85]

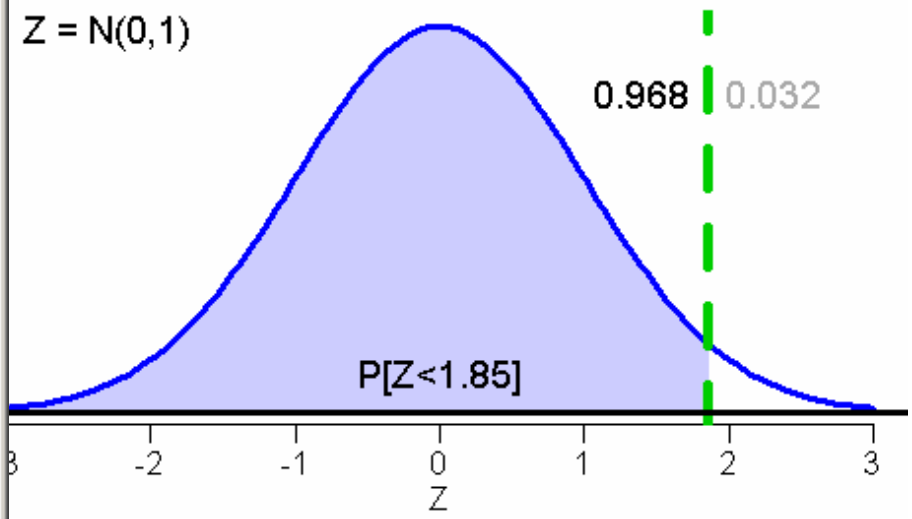
normal.ods - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

A1 | Función de distribución (acumulativa) de la distribución normal tipificada.

	A	B	C	D	E	F	G	H	I	J	K
3		**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
4	**0,0**	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	**0,1**	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	**0,2**	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	**0,3**	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	**0,4**	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	**0,5**	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	**0,6**	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	**0,7**	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	**0,8**	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	**0,9**	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	**1,0**	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	**1,1**	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
16	**1,2**	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
17	**1,3**	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
18	**1,4**	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
19	**1,5**	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
20	**1,6**	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
21	**1,7**	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
22	**1,8**	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
23	**1,9**	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
24	**2,0**	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
25	**2,1**	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
26	**2,2**	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
27	**2,3**	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
28	**2,4**	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
29	**2,5**	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995

Normal_0_1



Solución: 0.968 = 96.8%

Tabla N(0,1)

Z es normal estandarizada.

normal.ods - OpenOffice.org Calc

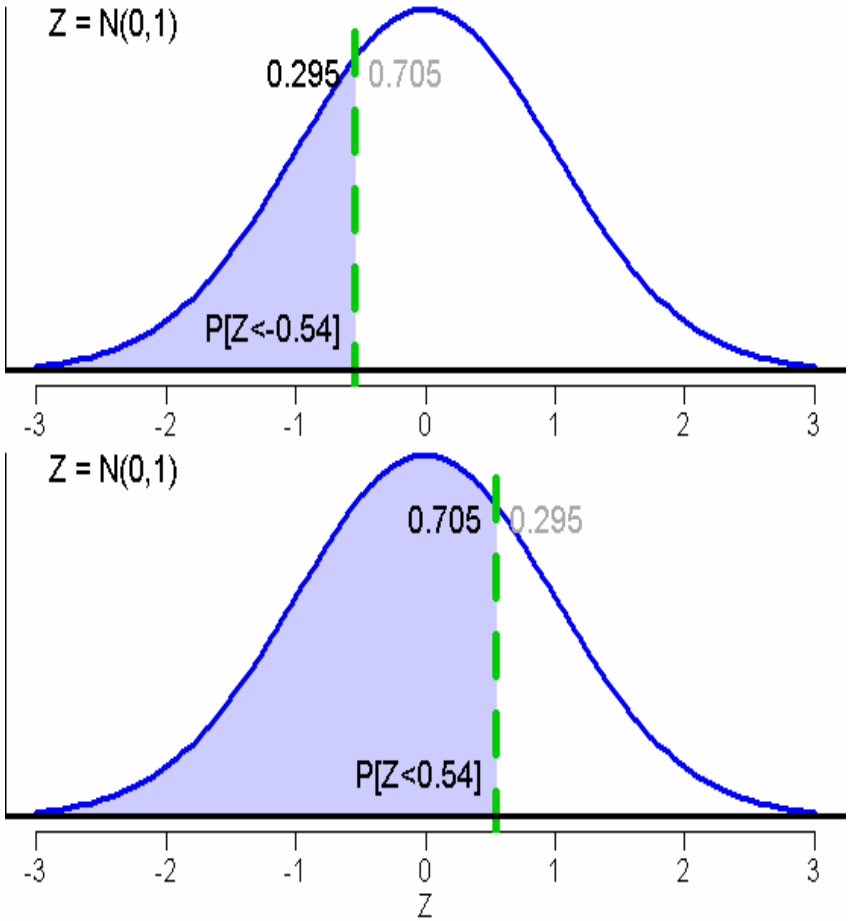
File Edit View Insert Format Tools Data Window Help

A1 | Función de distribución (acumulativa) de la distribución normal tipificada.

	A	B	C	D	E	F	G	H	I	J	K
3		**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
4	**0,0**	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	**0,1**	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	**0,2**	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	**0,3**	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	**0,4**	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	**0,5**	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	**0,6**	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	**0,7**	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	**0,8**	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	**0,9**	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	**1,0**	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	**1,1**	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
16	**1,2**	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
17	**1,3**	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
18	**1,4**	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
19	**1,5**	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
20	**1,6**	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
21	**1,7**	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
22	**1,8**	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
23	**1,9**	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
24	**2,0**	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
25	**2,1**	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
26	**2,2**	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
27	**2,3**	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
28	**2,4**	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
29	**2,5**	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995

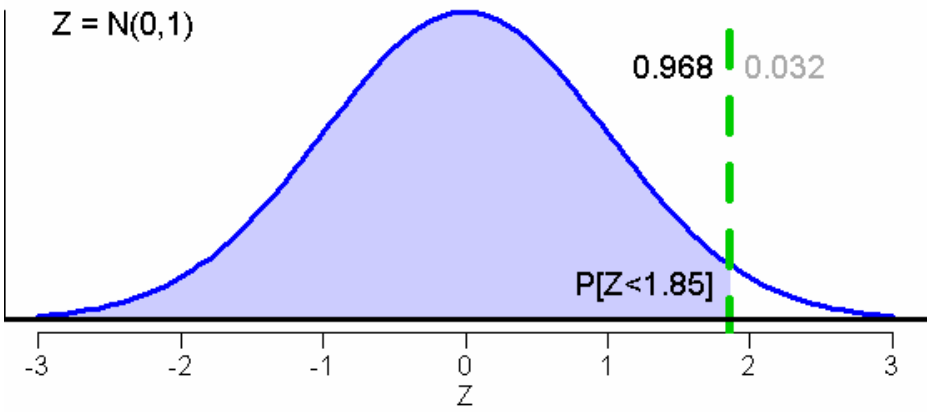
Normal_0_1

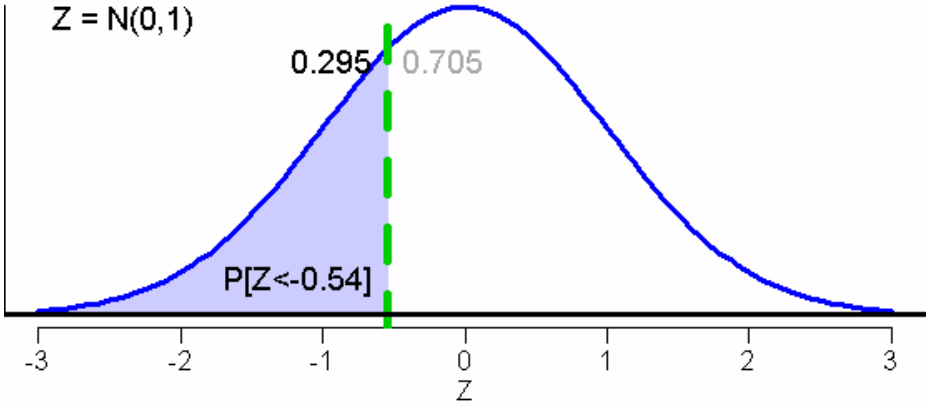
Calcular P[Z<-0.54]



Solución: 1-0.705 = 0.295

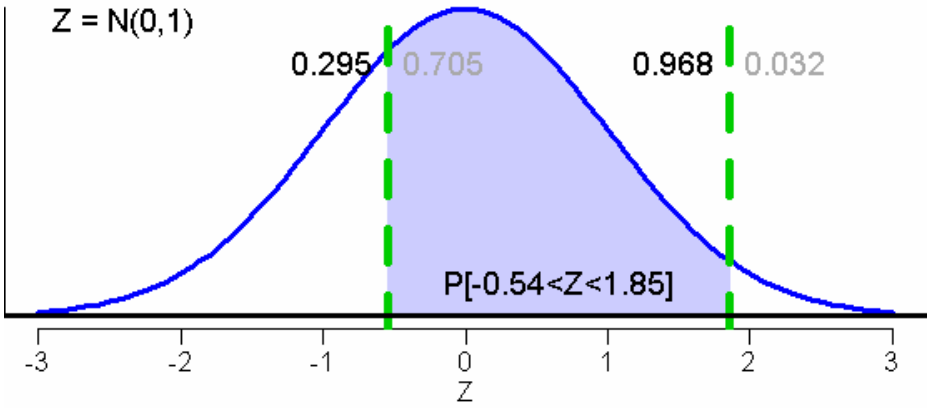
Tabla N(0,1)





Z es normal tipificada.

Calcular P[-0.54<Z<1.85]

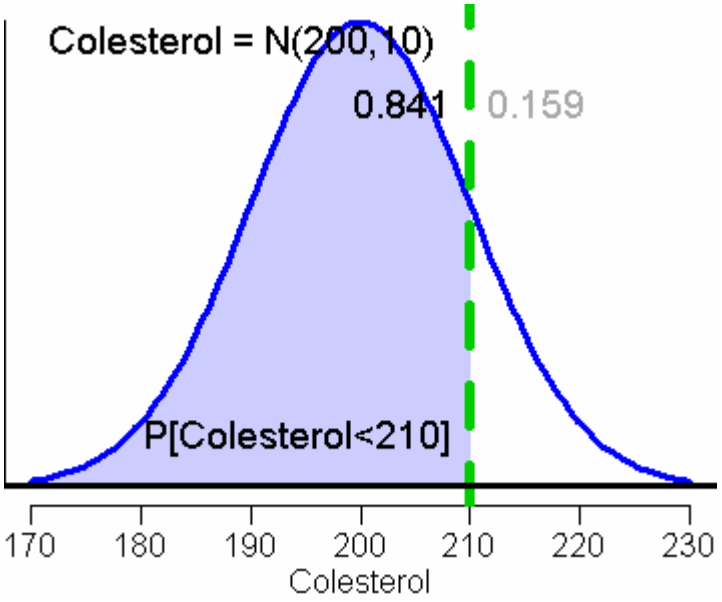


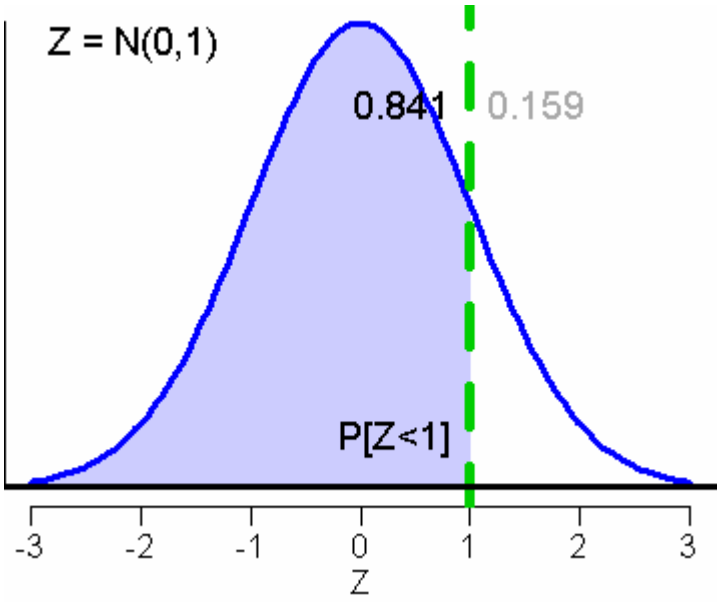
Solución: 0.968-0.295= 0.673

Ejemplo: Cálculo con probabilidades normales

- El colesterol en la población tiene distribución normal, con media 200 y desviación 10.
- ¿Qué porcentaje de individuos tiene colesterol inferior a 210?
- Qué valor del colesterol sólo es superado por el 10% de los individuos.

- Todas las distribuciones normales son similares salvo traslación y cambio de escala: estandaricemos.





$$z = \frac{x-\mu}{\sigma} = \frac{210-200}{10} = 1$$

normal.ods - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

A1 — Función de distribución (acumulativa) de la distribución normal tipificada.

	A	B	C	D	E	F	G	H	I	J	K
3		0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
4	0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	1,0	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883

Normal_0_1_

Sheet 1 / 1 | PageStyle_Normal_0_1_ | 100% | STD | Sum=0

$$P[Z < 1{,}00] = (\text{ver tabla}) = 0{,}841$$

- El valor del colesterol que sólo supera el 10% de los individuos es el percentil 90. Calculemos el percentil 90 de la N(0,1) y deshacemos la estandarización.

normal.ods - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

A1 | Función de distribución (acumulativa) de la distribución normal tipificada.

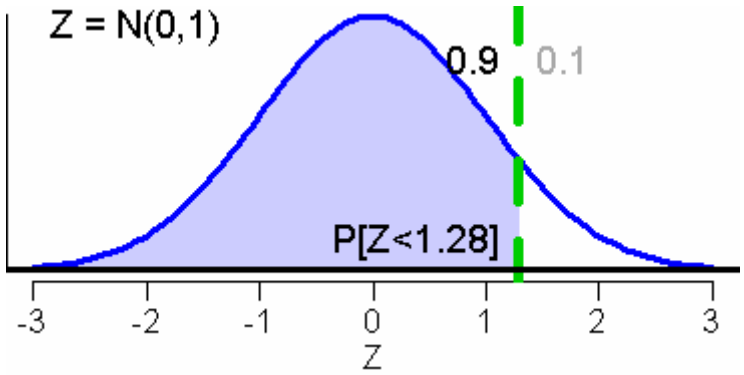
	A	B	C	D	E	F	G	H	I	J	K
3		**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
4	**0,0**	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	**0,1**	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	**0,2**	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	**0,3**	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	**0,4**	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	**0,5**	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	**0,6**	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	**0,7**	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	**0,8**	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	**0,9**	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	**1,0**	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	**1,1**	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
16	**1,2**	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
17	**1,3**	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
18	**1,4**	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
19	**1,5**	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
20	**1,6**	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954

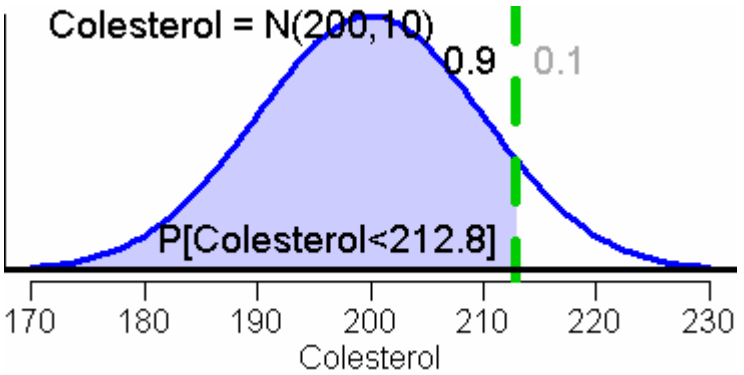
Normal_0_1

$$Z = \frac{x - \mu}{\sigma}$$

$$1{,}28 = \frac{x - 200}{10}$$

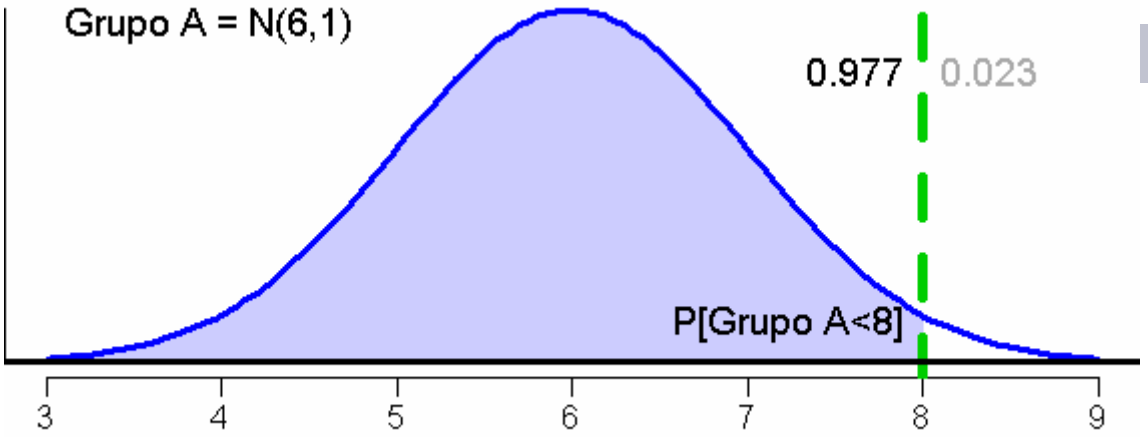
$$x = 200 + 10 \times 1{,}28 = 212{,}8$$

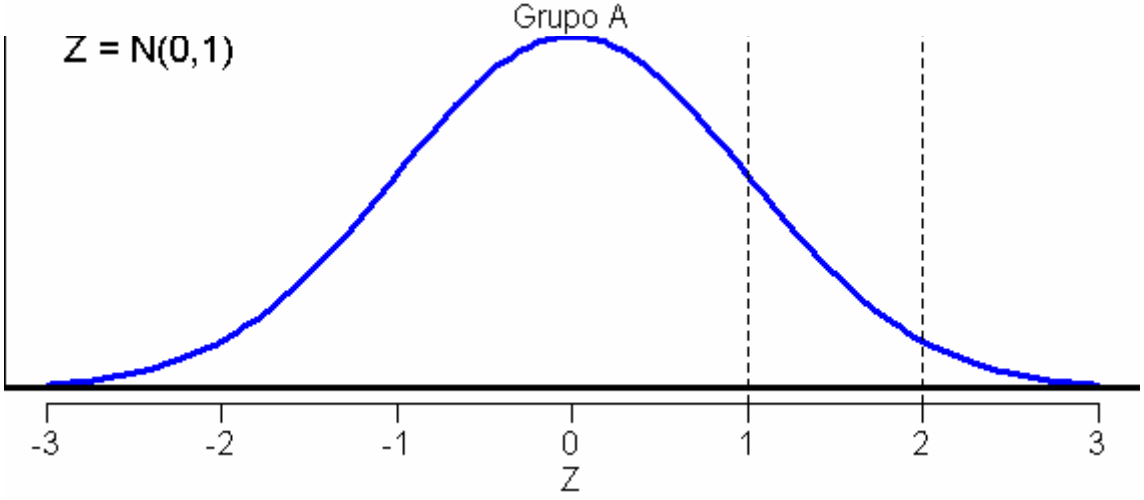


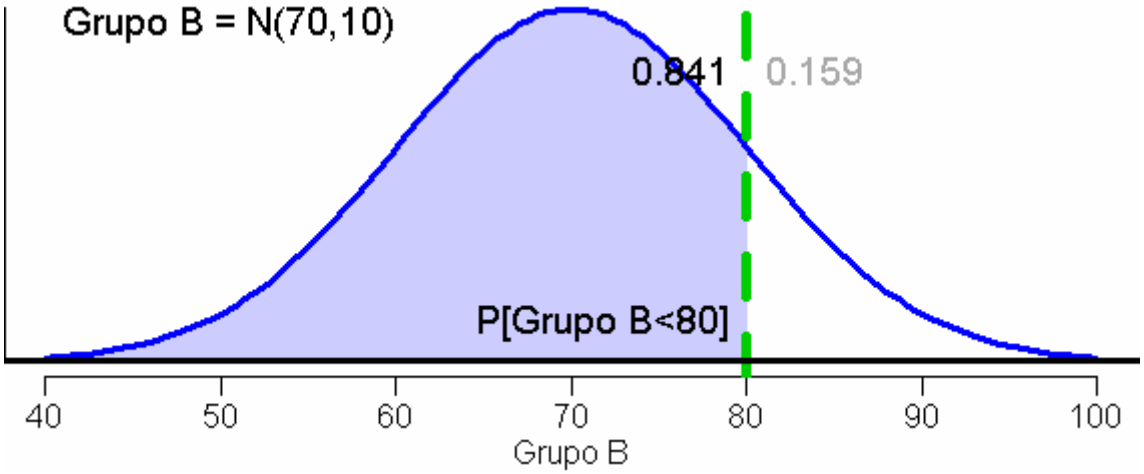


Ejemplo: Estandarización

- Se quiere dar una beca a uno de dos estudiantes de sistemas educativos diferentes. Se asignará al que tenga mejor expediente académico.
  - El estudiante A tiene una calificación de 8 en un sistema donde la calificación de los alumnos se comporta como N(6,1).
  - El estudiante B tiene una calificación de 80 en un sistema donde la calificación de los alumnos se comporta como N(70,10).
- **Solución**
  - No podemos comparar directamente 8 puntos de A frente a los 80 de B, pero como ambas poblaciones se comportan de modo normal, podemos tipificar y observar las puntuaciones sobre una distribución de referencia N(0,1)







$$z_A = \frac{x_A - \mu_A}{\sigma_A} = \frac{8-6}{1} = 2$$

$$z_B = \frac{x_B - \mu_B}{\sigma_B} = \frac{80-70}{10} = 1$$

Como $Z_A > Z_B$, podemos decir que el porcentaje de compañeros del mismo sistema de estudios que ha superado en calificación el estudiante A es mayor que el que ha superado B.

Podríamos pensar en principio que A es mejor candidato para la beca.

¿Por qué es importante la distribución normal?

- Las propiedades que tiene la distribución normal son interesantes, pero todavía no hemos hablado de por qué es una distribución especialmente importante.

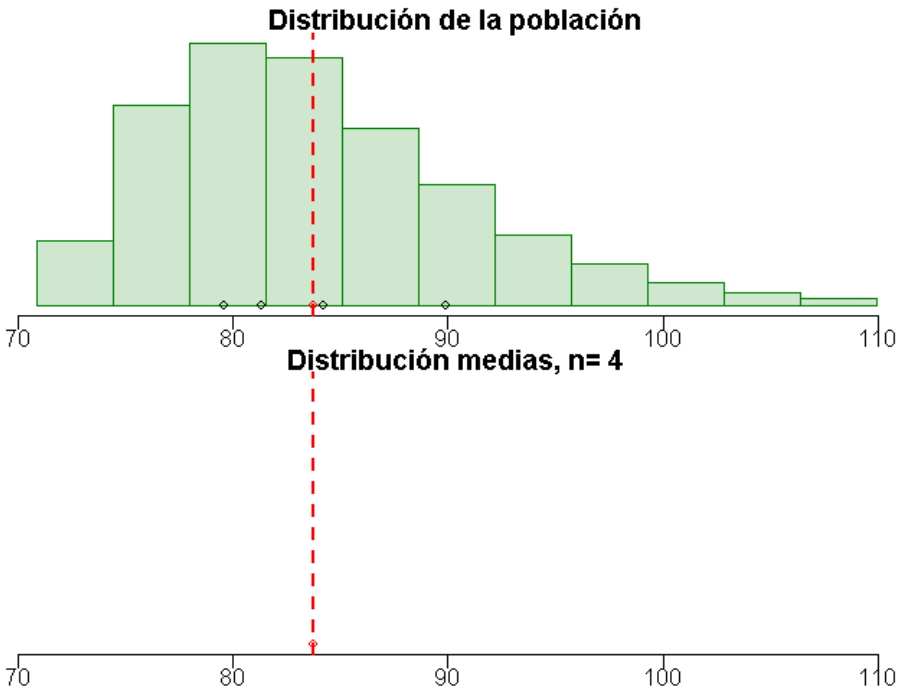
- La razón es que aunque una v.a. no posea distribución normal, ciertos estadísticos/estimadores calculados sobre muestras elegidas al azar sí que poseen una distribución normal.

- Es decir, tengan las distribución que tengan nuestros datos, los 'objetos' que resumen la información de una muestra, posiblemente tengan distribución normal (o asociada).

Aplic. de la normal: Estimación en muestras

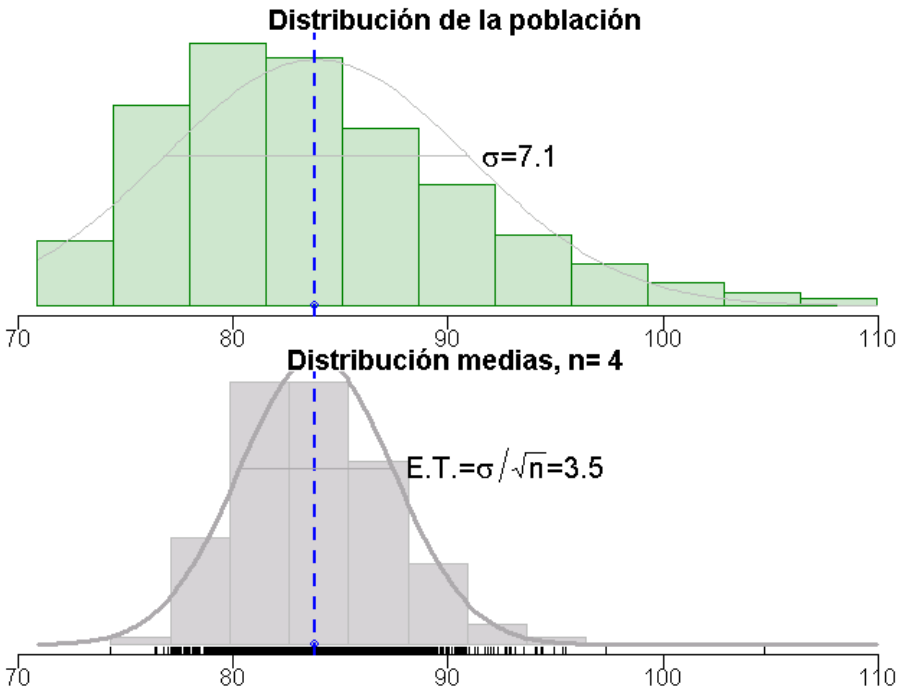
- Como ilustración mostramos una variable que presenta valores distribuidos de forma muy asimétrica. Claramente no normal.

- Saquemos muestras de diferentes tamaños, y usemos la media de cada muestra para estimar la media de la población.



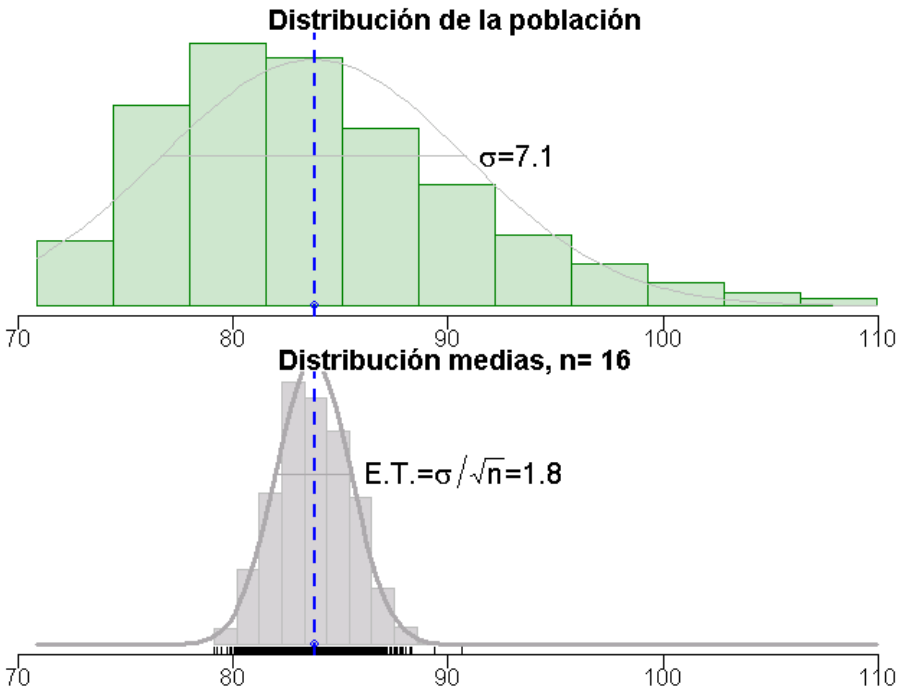
Aplic. de la normal: Estimación en muestras

- Cada muestra ofrece un resultado diferente: La media muestral es variable aleatoria.
- Su distribución es más parecida a la normal que la original.
- También está menos dispersa. A su dispersión ('desv. típica del estimador media muestral'… ¿les gusta el nombre largo?) se le suele denominar error típico.



Aplic. de la normal: Estimación en muestras

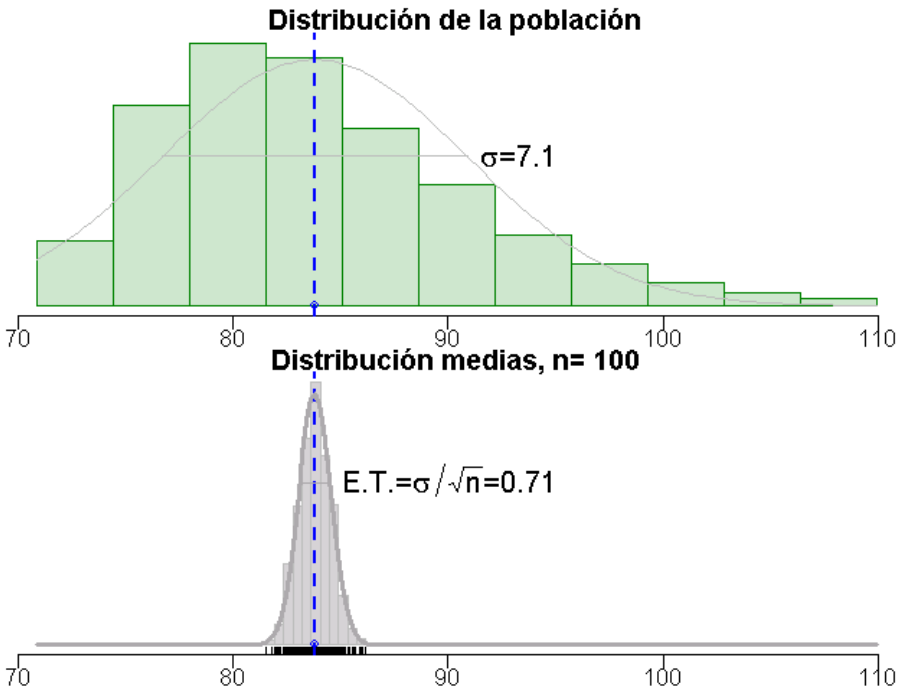
- Al aumentar el tamaño, n, de la muestra:
  - La normalidad de las estimaciones mejora
  - El error típico disminuye.



Aplic. de la normal: Estimación en muestras

- Puedo ‘garantizar’ medias muestrales tan cercanas como quiera a la verdadera media, sin más que tomar ‘n bastante grande’

- Se utiliza esta propiedad para dimensionar el tamaño de una muestra antes de empezar una investigación.



Resumen: Teorema del límite central

- Dada una v.a. **cualquiera**, si extraemos muestras de tamaño n, y calculamos los promedios muestrales, entonces:
  - dichos promedios tienen distribución aproximadamente normal;
  - La media de los promedios muestrales es la misma que la de la variable original.
  - La desviación típica de los promedios disminuye en un factor "raíz de n" (error estándar).
  - Las aproximaciones anteriores se hacen exactas cuando n tiende a infinito.
- Este teorema justifica la importancia de la distribución normal.
  - Sea lo que sea lo que midamos, cuando se promedie sobre una muestra grande (**n>30**) nos va a aparecer de manera natural la distribución normal.

Distribuciones asociadas a la normal

- Cuando queramos hacer inferencia estadística hemos visto que la distribución normal aparece de forma casi inevitable.

- Dependiendo del problema, podemos encontrar otras (asociadas):
  - X^2 (chi cuadrado)
  - t- student
  - F-Snedecor

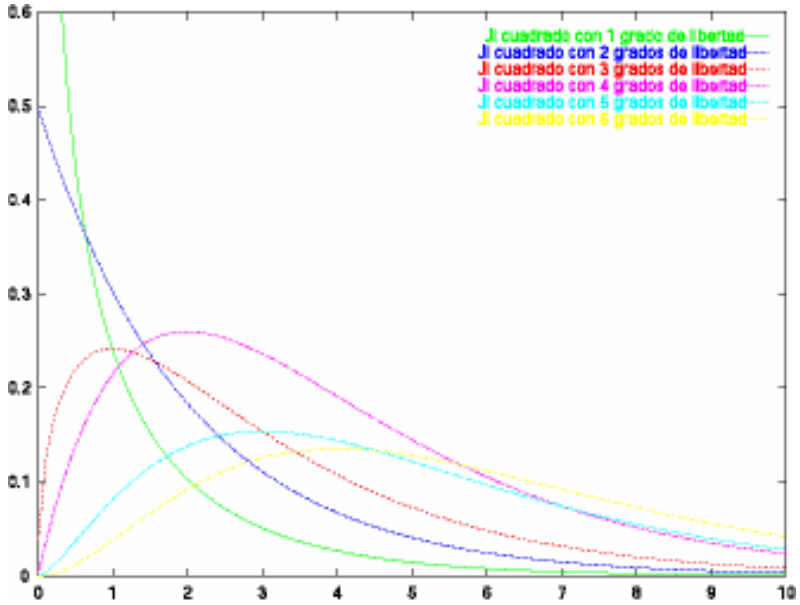
- Estas distribuciones resultan directamente de operar con distribuciones normales. Típicamente aparecen como distribuciones de ciertos estadísticos.

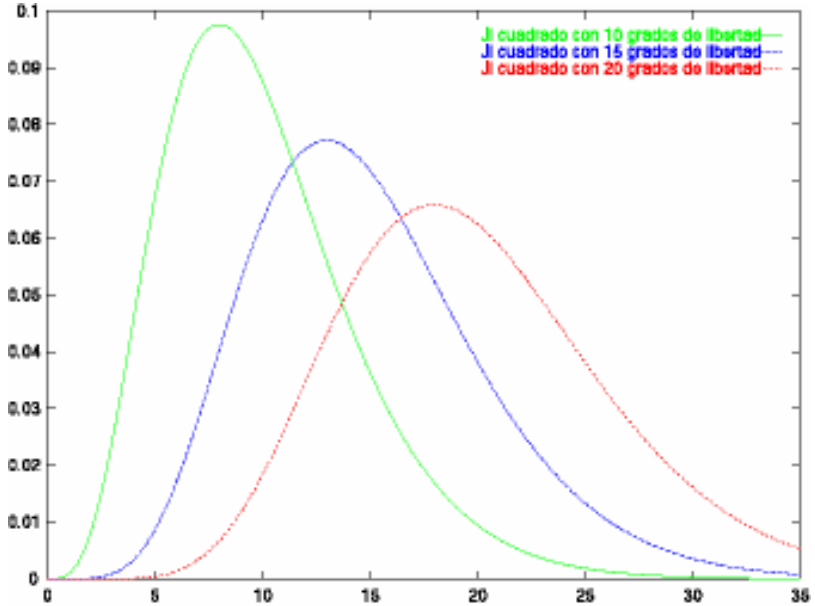
- Veamos algunas propiedades que tienen (superficialmente). Para más detalles consultar algún libro.

- Sobre todo nos interesa saber qué valores de dichas distribuciones son "atípicos".
  - Significación, p-valores,…

Chi cuadrado

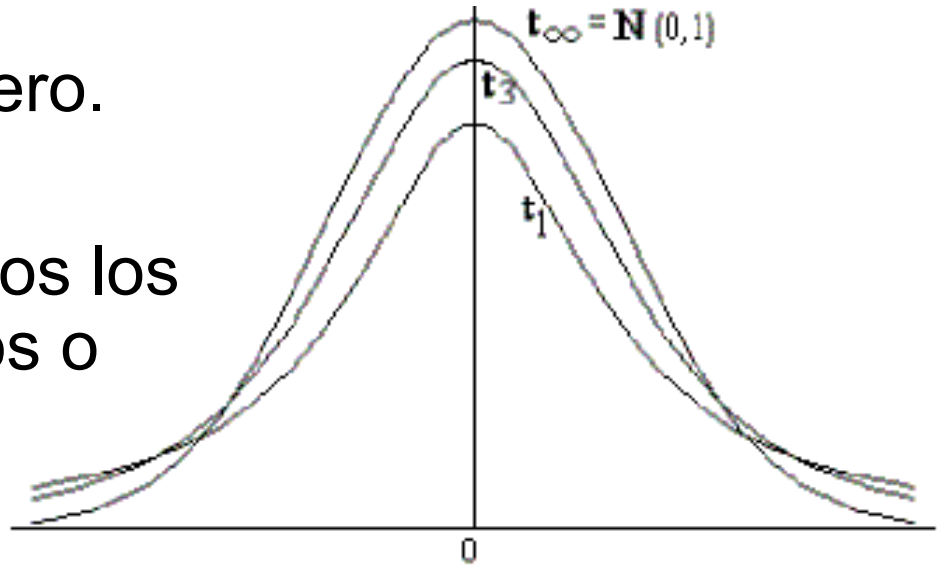
- Tiene un sólo parámetro denominado grados de libertad.
- La función de densidad es asimétrica positiva. Sólo tienen densidad los valores positivos.
- La función de densidad se hace más simétrica incluso casi gausiana cuando aumenta el número de grados de libertad.
- Normalmente consideraremos anómalos aquellos valores de la variable de la "cola de la derecha".



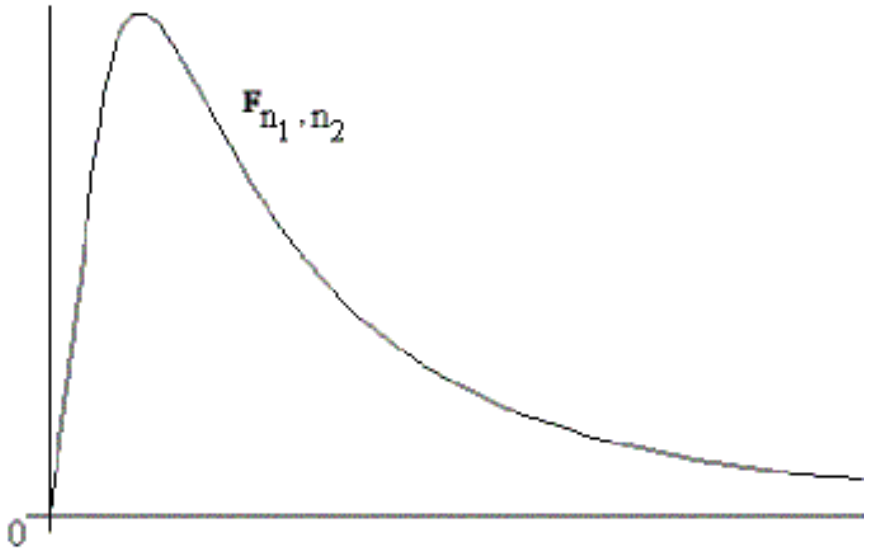


T de student

- Tiene un parámetro denominado grados de libertad.
- Cuando aumentan los grados de libertad, más se acerca a N(0,1).
- Es simétrica con respecto al cero.
- Se consideran valores anómalos los que se alejan de cero (positivos o negativos).



F de Snedecor



- Tiene dos parámetros denominados grados de libertad.
- Sólo toma valores positivos. Es asimétrica.
- Normalmente se consideran valores anómalos los de la cola de la derecha.

PROPEDÉUTICO

Modulo: Introducción a la estadística

Guía de estudio para la Unidad 4: Inferencia estadística

UTILIZANDO LA INFORMACIÓN DE ESTA SECCIÓN ó DEL LIBRO BIOSTATISTICAL ANALYSIS, ZAR, J. PRENTICE-HALL 1984 Ó 1999 RESUELVE CADA UNO DE LOS INCISOS:

1. ¿Qué es el Teorema de Límite Central y qué establece?
2. ¿Qué es un estimador de la población? Define.
3. ¿Qué es un intervalo de confianza y cuál es su relación con la estimación confidencial?
4. ¿Cuál es la fórmula generalizada para calcular el intervalo de confianza alrededor de la media y qué indica este intervalo?
5. ¿Cómo se relaciona el estadístico t con el intervalo de confianza de la μ en el caso general?
6. Fórmula de la estimación puntual insesgada.

Introducción a la Estadística

Tema 6: Inferencia estadística

¿Por qué es importante la distribución normal?

- Las propiedades que tiene la distribución normal son interesantes, pero todavía no hemos hablado de por qué es una distribución especialmente importante.

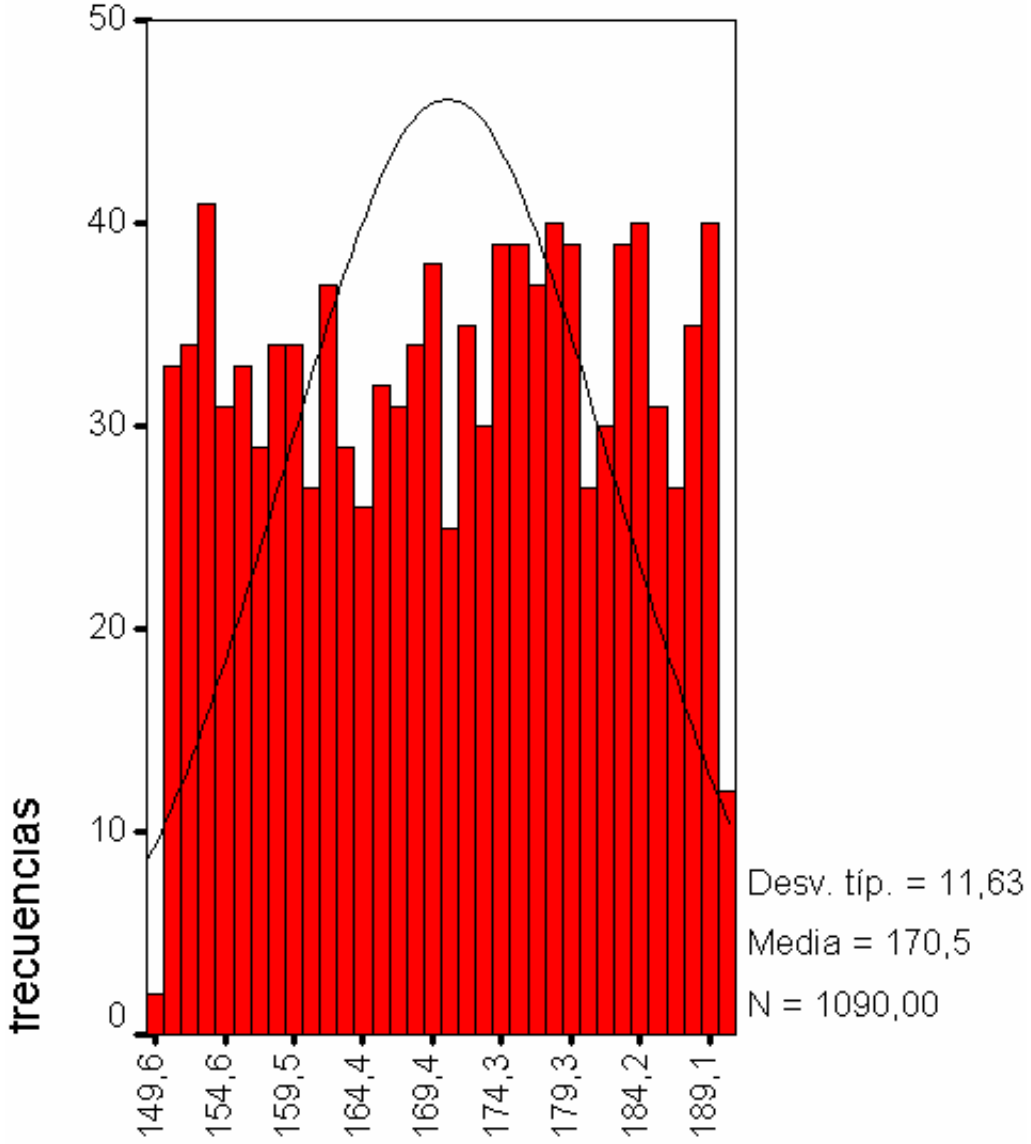
- La razón es que aunque una v.a. no posea distribución normal, ciertos estadísticos/estimadores calculados sobre muestras elegidas al azar sí que poseen una distribución normal.

- Es decir, tengan las distribución que tengan nuestros datos, los 'objetos' que resumen la información de una muestra, posiblemente tengan distribución normal (o asociada).

Veamos aparecer la distribución normal

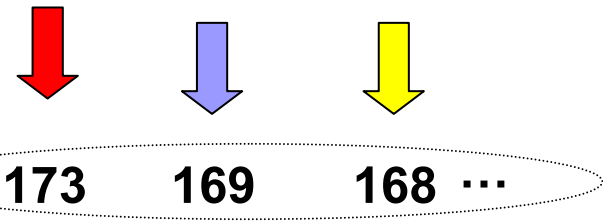
- Como ilustración mostramos una variable que presenta valores distribuidos más o menos uniformemente sobre el intervalo 150-190.

- Como es de esperar la media es cercana a 170. El histograma no se parece en nada a una distribución normal con la misma media y desviación típica.

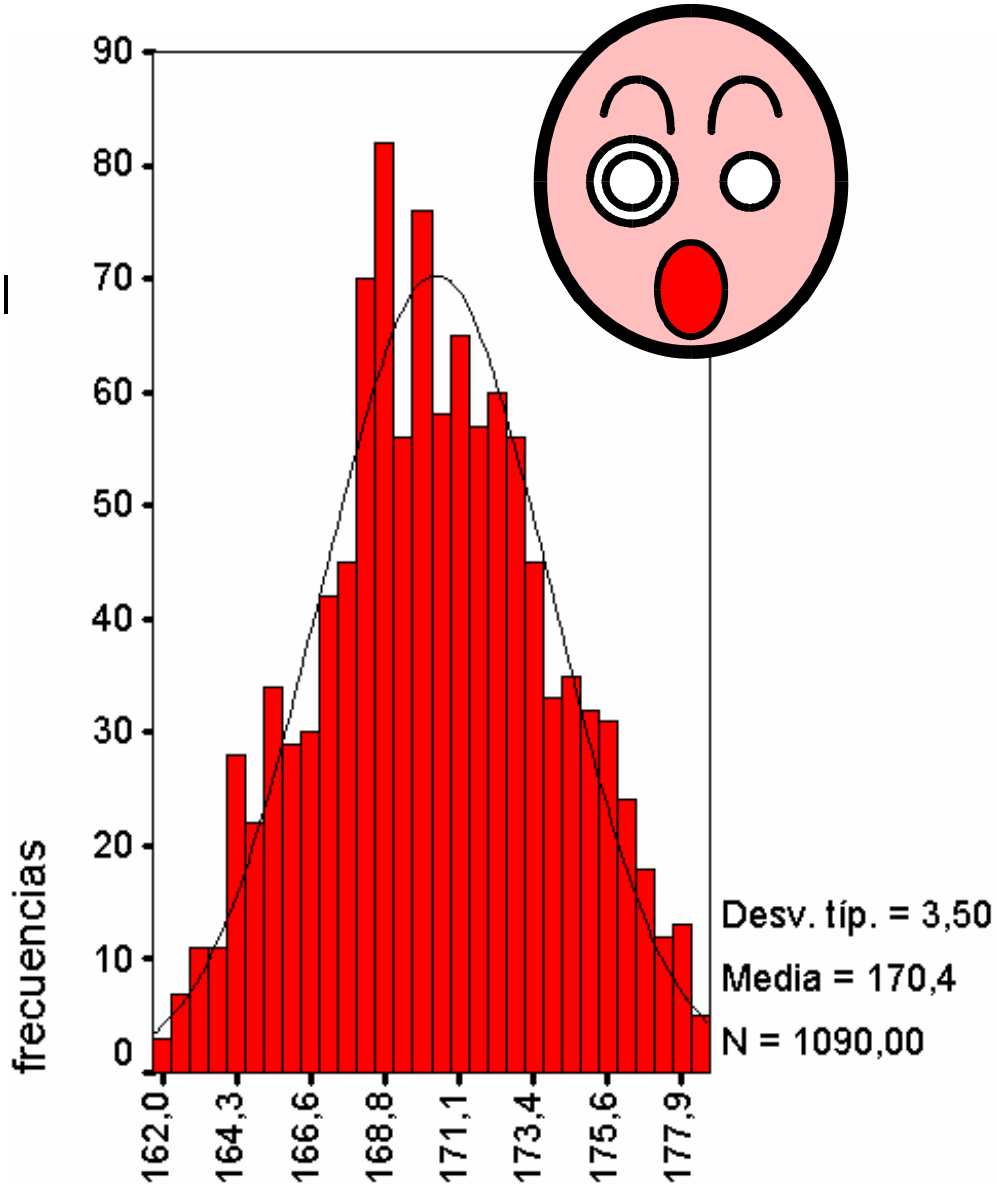


- A continuación elegimos aleatoriamente grupos de 10 observaciones de las anteriores y calculamos el promedio.
- Para cada grupo de 10 obtenemos entonces una nueva medición, que vamos a llamar promedio muestral.
- Observa que las nuevas cantidades están más o menos cerca de la media de la variable original.
- Repitamos el proceso un número elevado de veces. En la siguiente transparencia estudiamos la distribución de la nueva variable.

Muestra		
1ª	2ª	3ª
185	190	179
174	169	163
167	170	167
160	159	152
172	179	178
183	175	183
188	159	155
178	152	165
152	185	185
175	152	152



- La distribución de los promedios muestrales sí que tiene distribución aproximadamente normal.
- La media de esta nueva variable (promedio muestral) es muy parecida a l de la variable original.
- Las observaciones de la nueva variable están menos dispersas. Observa el rango. Pero no sólo eso. La desviación típica es aproximadamente 'raiz de 10' veces más pequeña. Llamamos error estándar a la desviación típica de esta nueva variable.
- **Nada** de lo anterior **es casualidad**.



Teorema del límite central

- Dada una v.a. **cualquiera**, si extraemos muestras de tamaño n, y calculamos los promedios muestrales, entonces:
- dichos promedios tienen distribución aproximadamente normal;
- La media de los promedios muestrales es la misma que la de la variable original.
- La desviación estándar de los promedios disminuye en un factor "raíz de n" (error estándar).
- Las aproximaciones anteriores se hacen exactas cuando n tiende a infinito.
  - Este teorema justifica la importancia de la distribución normal.
  - Sea lo que sea lo que midamos, cuando se promedie sobre una muestra grande (**n>30**) nos va a aparecer de manera natural la distribución normal.

Teorema del límite central

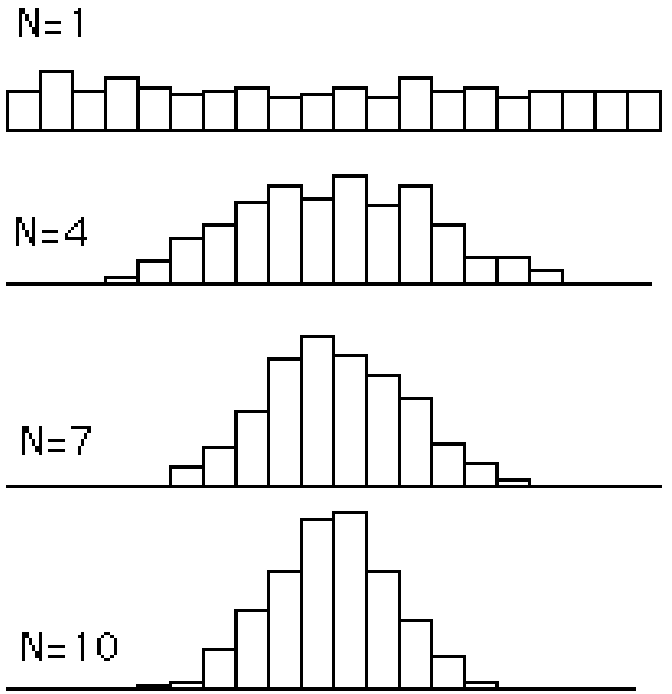
- Dada una distribución con media μ y varianza σ^2, La distribución de la media se aproxima a una distribución normal con media (μ) y una varianza σ^2/n cuando n, el tamaño de muestra, se incrementa

- Lo sorprendente acerca del teorema central del limite es que no importa la forma de la distribución original, la distribución de la media se aproxima a una distribución normal. Además, para la mayoría de las distribuciones, la distribución normal se aproxima tan rápido como n incrementa.

- Nótese que n es el tamaño de muestra para cada media y no el número de muestras

Teorema del límite central

- En el diagrama se muestra la distribución de frecuencias basada en 500 medias. Para n = 1, 4, 7 y 10, se obtuvieron 500 muestras de tamaño n a partir de una distribución uniforme.
- La distribución tiende a una normal cuando n incrementa
- La dispersión de la distribución tiende a decrecer, cuando n incrementa



Estimación

- Un estimador es una cantidad numérica calculada sobre una muestra y que esperamos que sea una buena aproximación de cierta cantidad con el mismo significado en la población (parámetro).

- En realidad ya hemos trabajado con estimadores cada vez que hacíamos una práctica con muestras extraídas de una población y suponíamos que las medias, etc… eran próximas de las de la población.

  - Para la media de una población:
    - "El mejor" es la media de la muestra.

  - Para la frecuencia relativa de una modalidad de una variable:
    - "El mejor" es la frecuencia relativa en la muestra.

¿Es útil conocer la distribución de un estimador?

- Es la clave para hacer inferencia. Ilustrémoslo con un ejemplo que ya tratamos en el tema anterior (teorema del límite central).

  - Si de una variable **conocemos μ y σ**, sabemos que para muestras "grandes", la **media muestral** es:
    - aproximadamente normal,
    - con la misma media y,
    - desviación típica mucho menor (error estándar)

$$EE = \frac{\sigma}{\sqrt{n}}$$

  - Es decir si por ejemplo **μ=60 y σ=5**, y obtenemos muestras de tamaño **n=100**,
    - La desv. típica de la media muestral (error estándar) es EE=5/raiz(100)=0,5
    - como la media muestral es aproximadamente normal, el 95% de los estudios con muestras ofrecerían estimaciones entre 60±1
    - Dicho de otra manera, al hacer un estudio tenemos una confianza del 95% de que la verdadera media esté a una distancia de ±1.

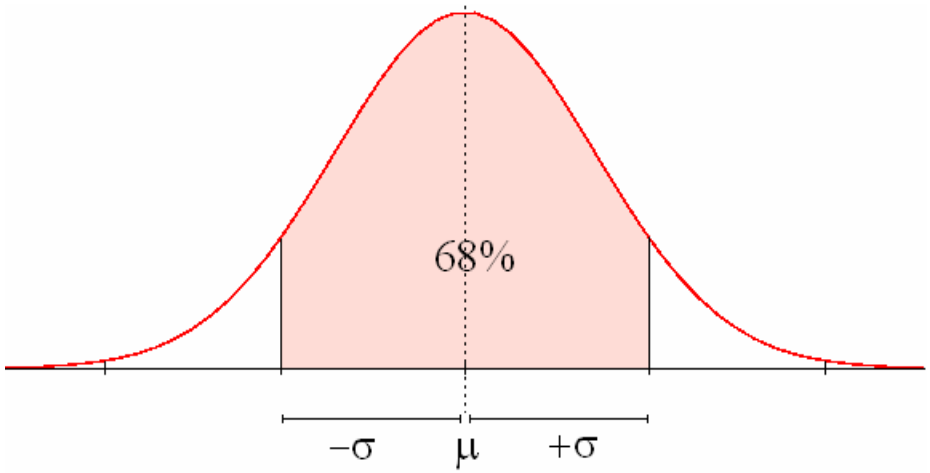
- En el ejemplo anterior la situación no era muy realista, pues como de todas maneras no conozco σ desconoceré el intervalo exacto para μ.

- Sin embargo también hay estimadores para σ y puedo usarlo como aproximación.

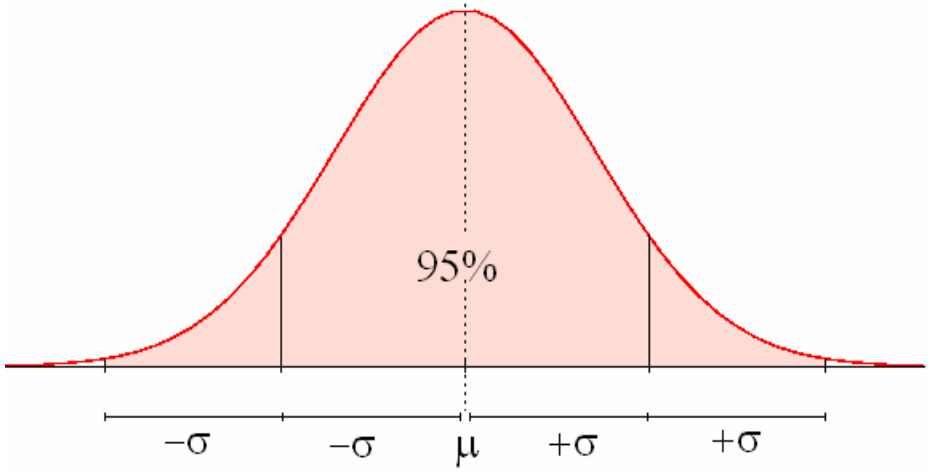
- Para tener una idea intuitiva, analicemos el siguiente ejemplo. Nos servirá como introducción a la estimación puntual y por intervalos de confianza.

- Ejemplo: Una muestra de n=100 individuos de una población tiene media de peso 60 kg y desviación 5kg.
  - Dichas cantidades pueden considerarse como aproximaciones (estimaciones puntuales)
    - 60 kg estima a μ
    - 5 kg estima a σ
    - 5/raiz(n)= 0,5 estima el error estándar (típico) EE
      - Estas son las llamadas estimaciones puntuales: un número concreto calculado sobre una muestra es aproximación de un parámetro.
  - Una estimación por intervalo de confianza es una que ofrece un intervalo como respuesta. Además podemos asignarle una probabilidad aproximada que mida nuestra confianza en la respuesta:
    - Hay una confianza del 68% de que μ esté en 60±0,5
    - Hay una confianza del 95% de que μ esté en 60±1.

- Entre la media y una desviación típica tenemos siempre la misma probabilidad: aprox. 68%



- Entre la media y dos desviaciones típicas aprox. 95%



- Se denomina estimación puntual de un parámetro al ofrecido por el estimador sobre una muestra.

- Se denomina estimación confidencial o intervalo de confianza para un nivel de confianza 1-α dado, a un intervalo que ha sido construido de tal manera que con frecuencia 1-α realmente contiene al parámetro.

  - Obsérvese que la probabilidad de error (no contener al parámetro) es α.
    - En el siguiente tema se llamará prob. de error de tipo I o nivel de significancia.
    - Valores típicos: α=0,10 ; **0,05** ; 0,01

  - En general el tamaño del intervalo disminuye con el tamaño muestral y aumenta con 1-α.

  - En todo intervalo de confianza hay una noticia buena y otra mala:
    - La buena: hemos usado una técnica que en % alto de casos acierta.
    - La mala: no sabemos si ha acertado en nuestro caso.

Intervalo para la media, si se conoce la varianza

Este caso que planteamos es más a nivel teórico que práctico: difícilmente vamos a poder conocer con exactitud σ^2 mientras que μ es desconocido. Sin embargo nos aproxima del modo más simple a la estimación confidencial de medias.

Para estimar μ, el estadístico que mejor nos va a ayudar es $\overline{X}$, del que conocemos su ley de distribución:

$$\overline{X} \rightsquigarrow \underbrace{\mathbf{N}\left(\mu, \frac{\sigma^2}{n}\right)}_{\text{un parámetro desconocido}}$$

Intervalo para la media, si se conoce la varianza

Esa ley de distribución depende de μ (desconocida). Lo más conveniente es hacer que la ley de distribución no dependa de ningún parámetro desconocido, para ello tipificamos:

$$Z = \underbrace{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}_{\begin{array}{c}\text{par. desconocido}\\ + \\ \text{estimador} \\ + \\ \text{cosas conocidas}\end{array}} \rightsquigarrow \underbrace{\mathbf{N}(0,1)}_{\text{tabulada}}$$

Intervalo para la media, si se conoce la varianza

Este es el modo en que haremos siempre la estimación puntual: *buscaremos una relación en la que intervengan el parámetro desconocido junto con su estimador y de modo que estos se distribuyan según una ley de probabilidad que es bien conocida y a ser posible tabulada.*

De este modo, fijado $\alpha \in (0,1)$, consideramos la v.a. $Z \rightsquigarrow \mathbf{N}(0,1)$ y tomamos un intervalo que contenga una masa de probabilidad de $1-\alpha$. Este intervalo lo queremos tan pequeño como sea posible. Por ello lo mejor es tomarlo simétrico con respecto a la media (0), ya que allí es donde se acumula más masa (véase la figura 8.1). Así las dos colas de la distribución (zonas más alejadas de la media) se repartirán a partes iguales el resto de la masa de probabilidad, α.

Intervalo para la media, si se conoce la varianza

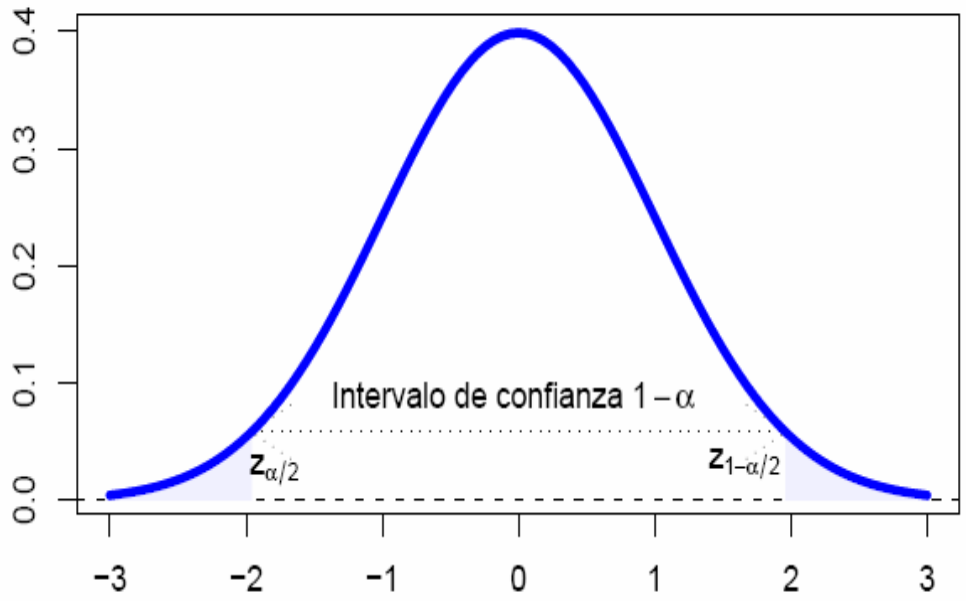


Figura 8.1: La distribución $\mathbf{N}(0,1)$ y el intervalo más pequeño posible cuya probabilidad es $1-\alpha$. Por simetría, los cuantiles $z_{\alpha/2}$ y $z_{1-\alpha/2}$ sólo difieren en el signo.

Intervalo para la media, si se conoce la varianza

Vamos a precisar cómo calcular el intervalo de confianza:

- Sea $z_{\alpha/2}$ el percentil $100 \cdot \frac{\alpha}{2}$ de Z, es decir, aquel valor de $\mathbb{R}$ que deja por debajo de si la cantidad $\frac{\alpha}{2}$ de la masa de probabilidad de Z, es decir:

$$\mathcal{P}[Z \leq z_{\alpha/2}] = \frac{\alpha}{2}$$

- Sea $z_{1-\alpha/2}$ el percentil $100 \cdot \frac{1-\alpha}{2}$, es decir,

$$\mathcal{P}[Z \leq z_{1-\alpha/2}] = 1 - \frac{\alpha}{2}$$

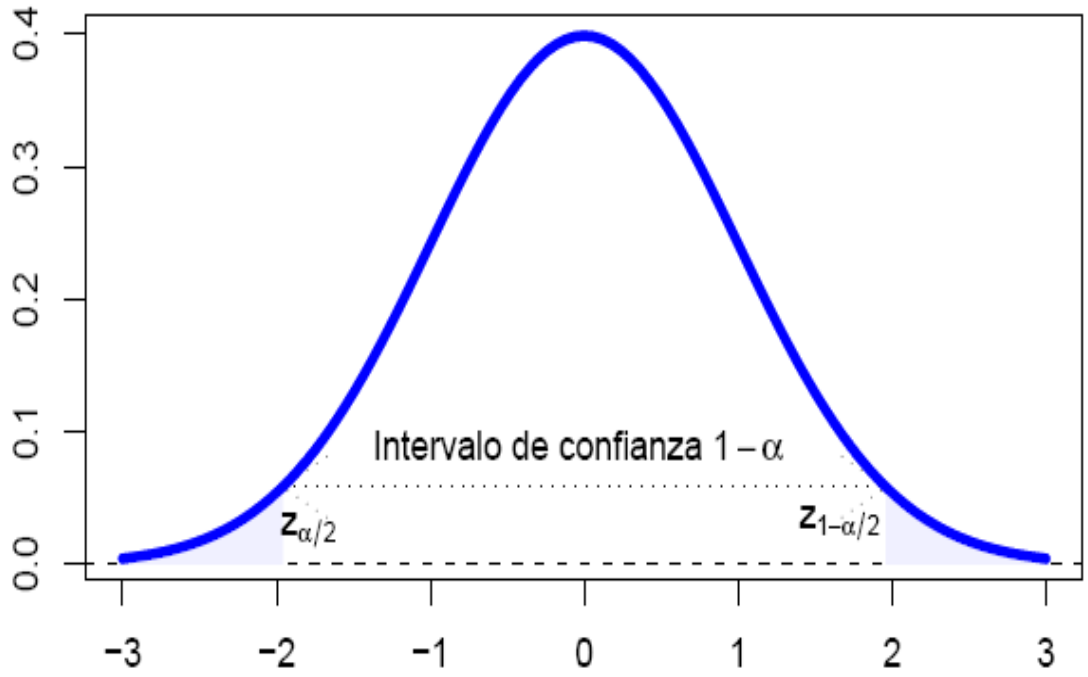
Intervalo para la media, si se conoce la varianza

Es útil considerar en este punto la simetría de la distribución normal, y observar que los percentiles anteriores son los mismos aunque con el signo cambiado:

$$z_{\alpha/2} = -z_{1-\alpha/2}$$

Intervalo para la media, si se conoce la varianza

- El intervalo alrededor del origen que contiene la mayor parte de la masa de probabilidad $(1-\alpha)$ es el intervalo siguiente (cf. Figura 8.1):



Intervalo para la media, si se conoce la varianza

De este modo un intervalo de confianza al nivel $1-\alpha$ para la esperanza de una normal de varianza conocida es el comprendido entre los valores

$$x_{\alpha/2} = \overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$x_{1-\alpha/2} = \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\boxed{\mu = \overline{X} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}$$

Ejemplo

Se sabe que el peso de los recién nacidos sigue una distribución normal con una desviación típica de 0,75 kg. Si en una muestra aleatoria simple de 100 de ellos se obtiene una media muestral de 3 kg, y una desviación típica de 0,5 kg, calcular un intervalo de confianza para la media poblacional que presente una confianza del 95 %.

Solución: En primer lugar hay que mencionar que la situación planteada no es habitual, ya que si somos capaces de obtener $\sigma = 0,75$, es natural que hayamos podido calcular tambien μ, y no necesitariamos una muestra aleatoria para estimar μ confidencialmente. Esto ocurre porque el ejemplo tiene utilidad puramente académica.

Ejemplo

Para calcular μ usamos el estadístico:

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

que como se observa no depende de la dispersión de la muestra, ya que tenemos la "fortuna" de disponer de la dispersión exacta de la población.

Esto no es lo habitual en una situación práctica, y como veremos más adelante, el papel del la dispersión exacta de la población (desconocido) será sustituido por el de la dispersión de la muestra.

Ejemplo

Un intervalo de confianza al 95 % se calcula teniendo en cuenta que $Z \rightsquigarrow \mathbf{N}(0,1)$, y dicha distribución presenta un 95 % de probabilidad de ocurrir entre sus cuantiles $z_{0,025} = -1,96$ y $z_{0,975} = 1,96$ (son de signo opuesto por simetría de la distribución normal). Luego con una confianza del 95 % ocurre:

$$-1,96 \leq Z \leq +1,96 \Leftrightarrow |Z| \leq +1,96 \Leftrightarrow |\overline{x}-\mu| \leq +1,96\frac{\sigma}{\sqrt{n}} \Leftrightarrow |\mu-3| \leq 0,147$$

Es decir con una confianza del 95 % tenemos que $\mu = 3 \pm 0,147kg$. Esto debe ser interpretado como que la técnica que se usa para el calcular el intervalo de confianza da una respuesta correcta en 95 de cada 100 estudios basados en una muestra aleatoria simple diferente sobre la misma población.

Intervalo para la media, caso general

El intervalo de confianza al nivel $1-\alpha$ para la esperanza de una distribución gaussiana cuando sus parámetros son desconocidos es:

$$\mu = \overline{X} \pm t_{n-1,1-\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Intervalo para la media, caso general

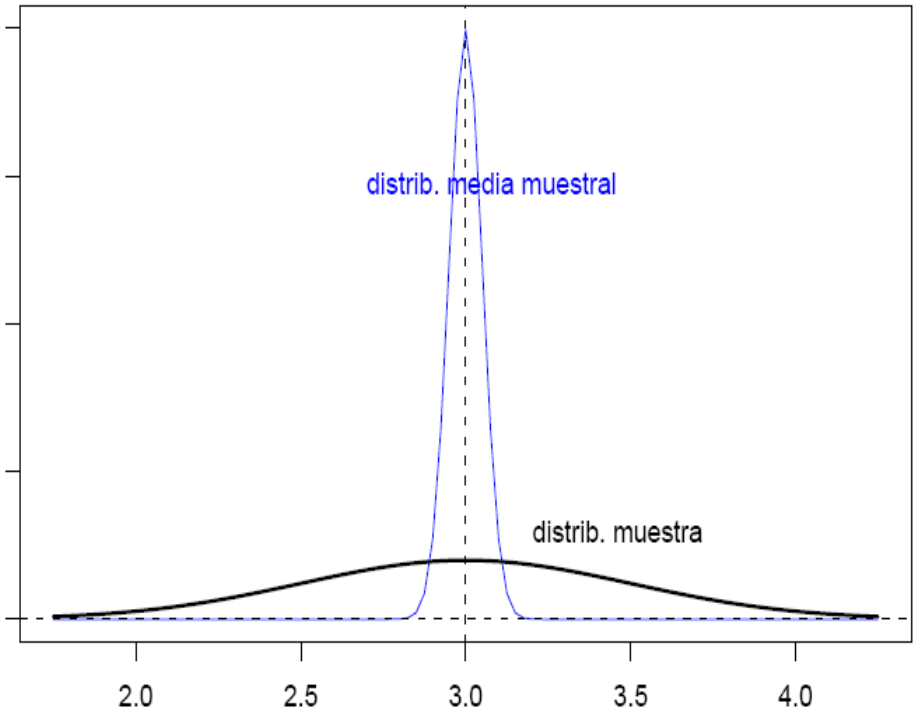


Figura 8.2: Un intervalo de confianza para la media podemos visualizarlo como el que correspondería a una distribución normal con el mismo centro que la de la población, pero cuya desviación está reducida en $\sqrt{n}$.

Intervalo para la media, caso general

Se sabe que el peso de los recién nacidos sigue una distribución normal. Si en una muestra aleatoria simple de 100 de ellos se obtiene una media muestral de 3 kg, y una desviación típica de 0,5 kg, calcular un intervalo de confianza para la media poblacional que presente una confianza del 95 %.

Intervalo para la media, caso general

Solución: Para calcular μ usamos el estadístico:

$$T = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}} \rightsquigarrow \mathbf{t}_{n-1}$$

que a diferencia del ejemplo mencionado anteriormente, no depende se σ (desconocido) si no de su estimación puntual insesgada:

$$\hat{S} = \sqrt{n/(n-1)}\, S = \sqrt{100/99}\, 0,5 = 0,503$$

Intervalo para la media, caso general

Un intervalo de confianza al 95 % se calcula teniendo en cuenta que $T \rightsquigarrow \mathbf{t}_{n-1}$, y dicha distribución presenta un 95 % de probabilidad de ocurrir entre sus cuantiles $T_{n-1;0,025} = -1,98$ y $T_{n-1;0,975} = 1,98$ (son de signo opuesto por simetría de la distribución de Student). Luego con una confianza del 95 % ocurre:

$$|\overline{x} - \mu| \leq +1,98 \frac{\hat{S}}{\sqrt{n}} \Leftrightarrow |\mu - 3| \leq 0,1$$

Es decir con una confianza del 95 % tenemos que $\mu = 3 \pm 0,1kg$.

Ejemplo

Se quiere estimar un intervalo de confianza al nivel de significación $\alpha = 0,05$ para la altura media μ de los individuos de una ciudad. En principio sólo sabemos que la distribución de las alturas es una v.a. X de distribución normal. Para ello se toma una muestra de $n = 25$ personas y se obtiene

$$\begin{aligned} \overline{x} &= 170 \text{ cm} \\ \mathcal{S} &= 10 \text{ cm} \end{aligned}$$

Ejemplo

Solución:

Este ejemplo es similar al anterior, pero vamos a resolverlo de una manera más detallada.

En primer lugar, en estadística inferencial, los estadísticos para medir la dispersión más convenientes son los insesgados. Por ello vamos a dejar de lado la desviación típica muestral, para utilizar la cuasidesviación típica:

$$\mathcal{S} = 10 \Longrightarrow \hat{\mathcal{S}} = \mathcal{S}\sqrt{\frac{n}{n-1}} = 10\sqrt{\frac{25}{24}} = 10'206$$

Ejemplo

$$\mu = 170 \pm 2,06 \cdot \frac{10,206}{5} = 170 \pm 4,204$$

o dicho de forma más precisa: Con un nivel de confianza del 95 % podemos decir que la media poblacional está en el intervalo siguiente:

$$\mu \in [165,796\,;\,174,204]$$

Ejemplo

Este ejemplo se puede considerar como una introducción a los contrastes de hipótesis. La variable IL se presenta en los niños recién nacidos con una distribución normal de media 2,5. En un grupo de 31 niños con sepsis neonatal se encuentra que el valor medio de IL es de $\overline{x} = 1,8$ y $\hat{S} = 0,2$. ¿Cree que presenta la presencia de sepsis neonatal afecta el valor de IL?

Ejemplo

Solución: Si no hubiese relación entre la sepsis neonatal y el valor de IL debería ocurrir que el valor de IL en niños nacidos con sepsis se comporte del mismo modo que en los niños normales. Por tanto debería seguir una distribución normal. Además un intervalo de confianza al 95 % para la media de la población de niños sépticos, calculado a partir de los datos de la muestra debería contener (con una confianza del 95 %) a la media de la población de niños normales. Si no fuese así habría que pensar que la variable IL está relacionada con la presencia de sepsis.

Ejemplo

Calculemos el intervalo de confianza para la media de los niños con sepsis. Para ello elegimos el estadistico más adecuado a los datos que poseemos:

$$T = \frac{\overline{x} - \mu}{\hat{S}/\sqrt{31}} \rightsquigarrow \mathbf{t}_{30}$$

Un intervalo de confianza al 95 % se calcula teniendo en cuenta que $T \rightsquigarrow \mathbf{t}_{30}$, y dicha distribución presenta un 95 % de probabilidad de ocurrir entre sus cuantiles $T_{30;0,025} = -2,04$ y $T_{30;0,975} = 2,04$ (son de signo opuesto por simetría de la distribución de Student). Luego con una confianza del 95 % ocurre:

Ejemplo

$$|1,8-\mu| \leq +2,04\frac{0,2}{\sqrt{31}} \Leftrightarrow |\mu-1,8| \leq 0,07$$

Por tanto podemos afirmar (con una confianza del 95 %) que la media poblacional de los niños con sepsis estaría comprendida entre los valores 1,73 y 1,87, que están muy alejados de 2,5 (media de los niños normales). Por tanto, podemos afirmar con una confianza del 95 % que están relacionados la IL y la sépsis en niños recien nacidos.

PROPEDÉUTICO

Modulo: Introducción a la estadística

Guía de estudio para la Unidad 5: Prueba de hipótesis

UTILIZANDO LA INFORMACIÓN DE ESTA SECCIÓN ó DEL LIBRO BIOSTATISTICAL ANALYSIS, ZAR, J. PRENTICE-HALL 1984 Ó 1999 RESUELVE CADA UNO DE LOS INCISOS:

1. Define qué es una hipótesis estadística.

2. Define qué es una hipótesis alternativa y cuál es su relación con la hipótesis nula.

3. Define que es la región crítica en una distribución de probabilidades y dónde se encuentra.

4. ¿Qué es el error tipo II?

5. Define que es α y qué relación guarda con el error tipo I.

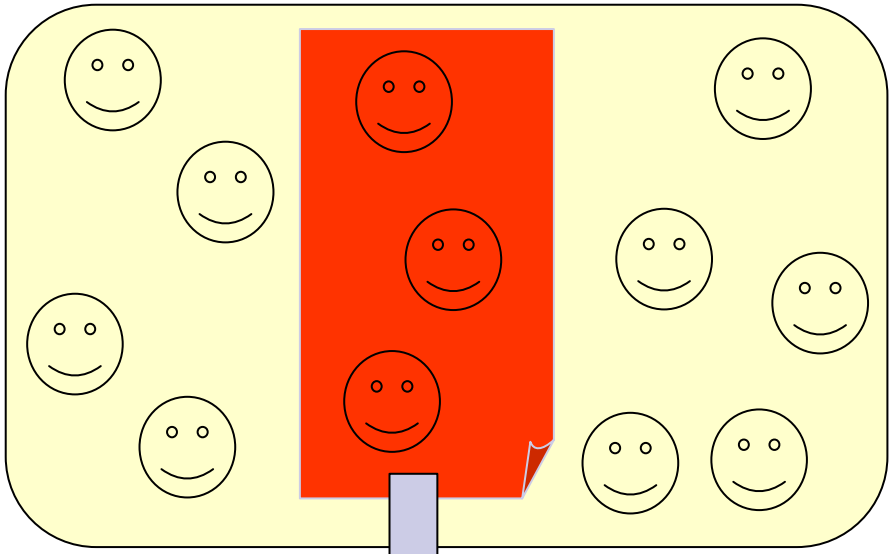
Introducción a estadística

Tema 5: Pruebas de hipótesis

Objetivos del tema

- Conocer el proceso para probar hipótesis y su relación con el método científico.
- Diferenciar entre hipótesis nula y alternativa
- Nivel de significancia
- Nivel observado de significancia
- Toma de decisiones, tipos de error y cuantificación del error.

Probando una hipótesis

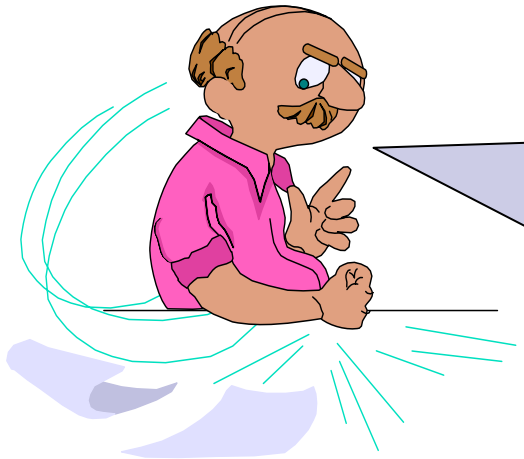


Muestra aleatoria

$\overline{X} = 20$ años

Son demasiados...

Creo que la edad media es **40** años...



¡Gran diferencia!

Rechazo la hipótesis

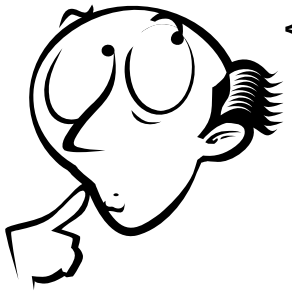
¿Qué es una hipótesis?

- Una creencia sobre la población, principalmente sus parámetros:
  - Media
  - Varianza
  - Proporción/Tasa

- OJO: Si queremos probarla, debe establecerse antes del análisis.

Identificación de hipótesis

- Hipótesis nula H_o
  - La que probaremos
  - Los datos pueden refutarla
  - No debería ser rechazada sin una buena razón.
- Hipótesis Alternativa H_1
  - Niega a H_0
  - Los datos pueden mostrar evidencia a favor
  - No debería ser aceptada sin una gran evidencia a favor.



$$\begin{cases} H_0: & p = 50\% \\ H_1: & p \neq 50\% \end{cases}$$

$=\ ,\ \leq\ ,\ \geq$

$\neq\ ,\ <\ ,\ >$

¿Quién es H_0?

- Problema: ¿La osteoporosis está relacionada con el género?
- Solución:
  - Traducir a lenguaje estadístico: $p = 50\%$
  - Establecer su opuesto: $p \neq 50\%$
  - Seleccionar la hipótesis nula

$$H_0 : p = 50\%$$

¿Quién es H_0?

- Problema: ¿El colesterol medio para la dieta mediterránea es 6 mmol/l?

- Solución:

  - Traducir a lenguaje estadístico: $\mu = 6$

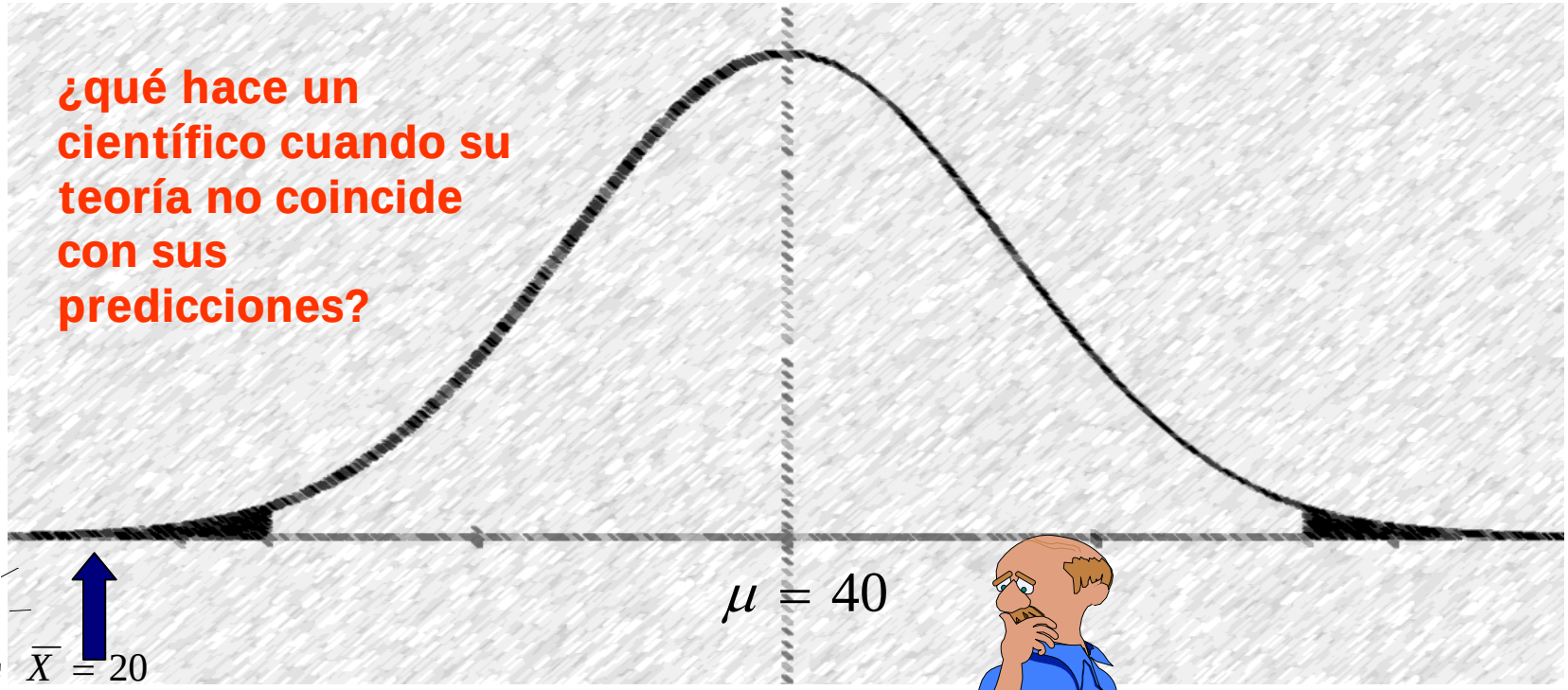
  - Establecer su opuesto: $\mu \neq 6$

  - Seleccionar la hipótesis nula

$$H_0 : \mu = 6$$

Razonamiento básico

Si supongo que H_0 es cierta...

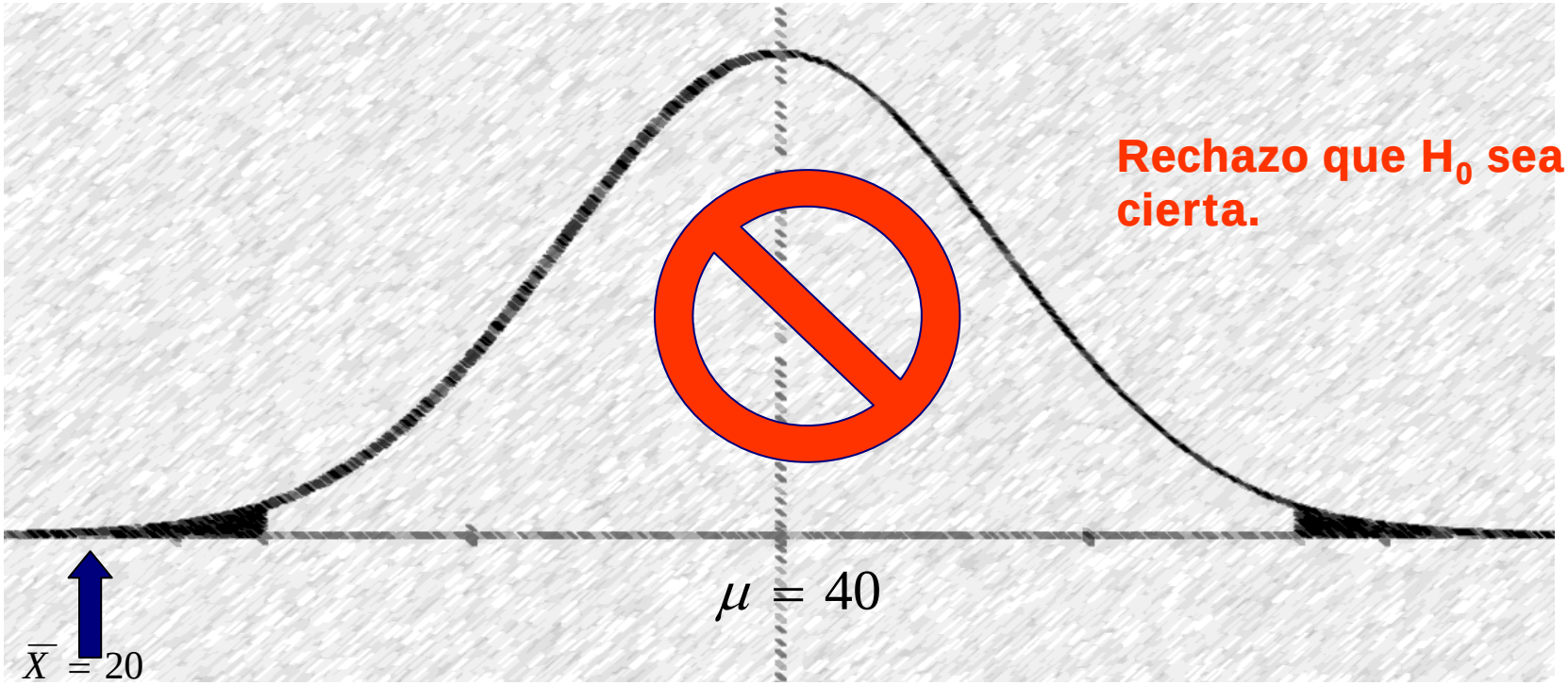


... el resultado del experimento sería improbable.

Sin embargo **ocurrió**.

Razonamiento básico

Si supongo que H_0 es cierta...



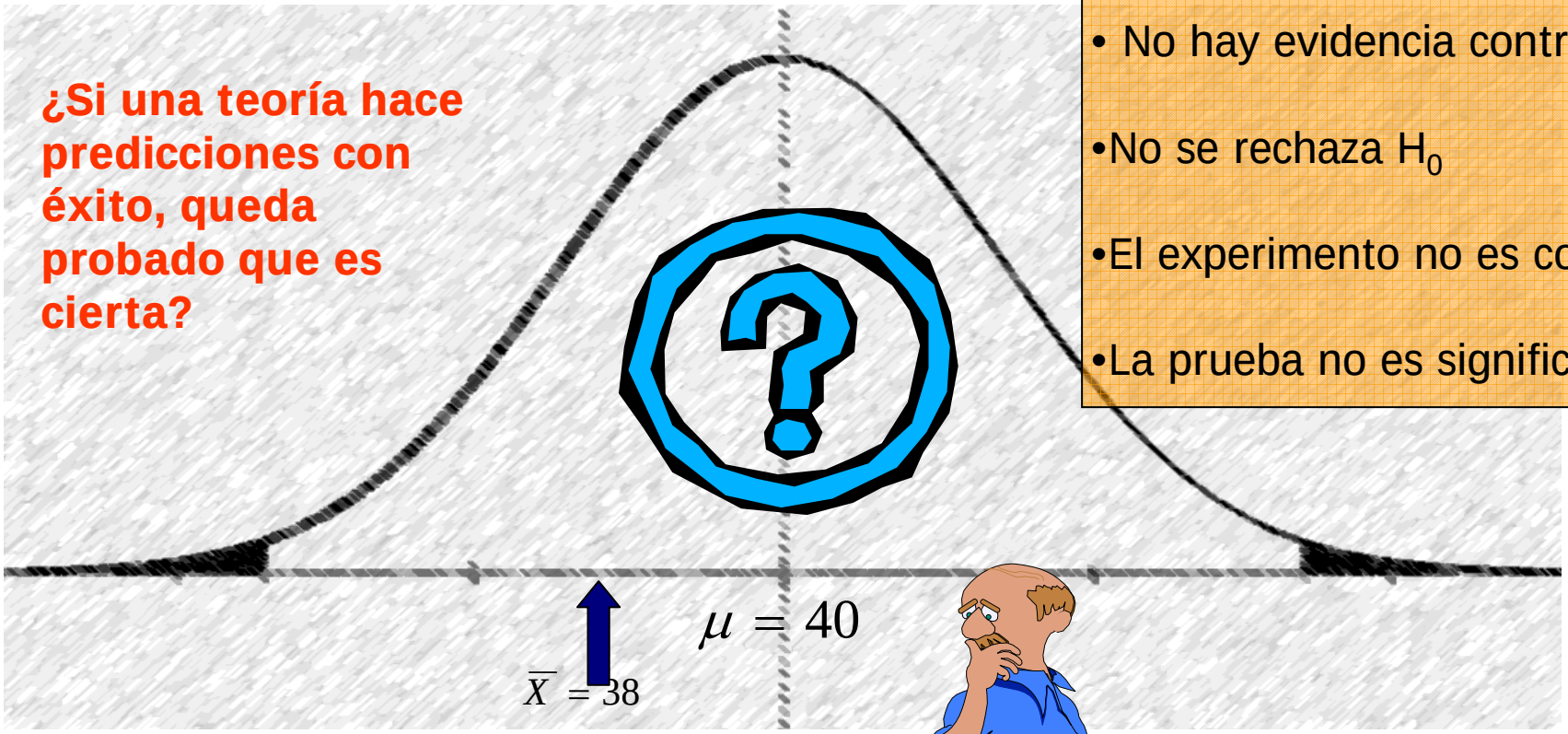
... el resultado del experimento sería improbable.

Sin embargo **ocurrió**.

Razonamiento básico

Si supongo que H_0 es cierta...

¿Si una teoría hace predicciones con éxito, queda probado que es cierta?



- No hay evidencia contra H_0
- No se rechaza H_0
- El experimento no es concluyente
- La prueba no es significativa

... el resultado del experimento es coherente.

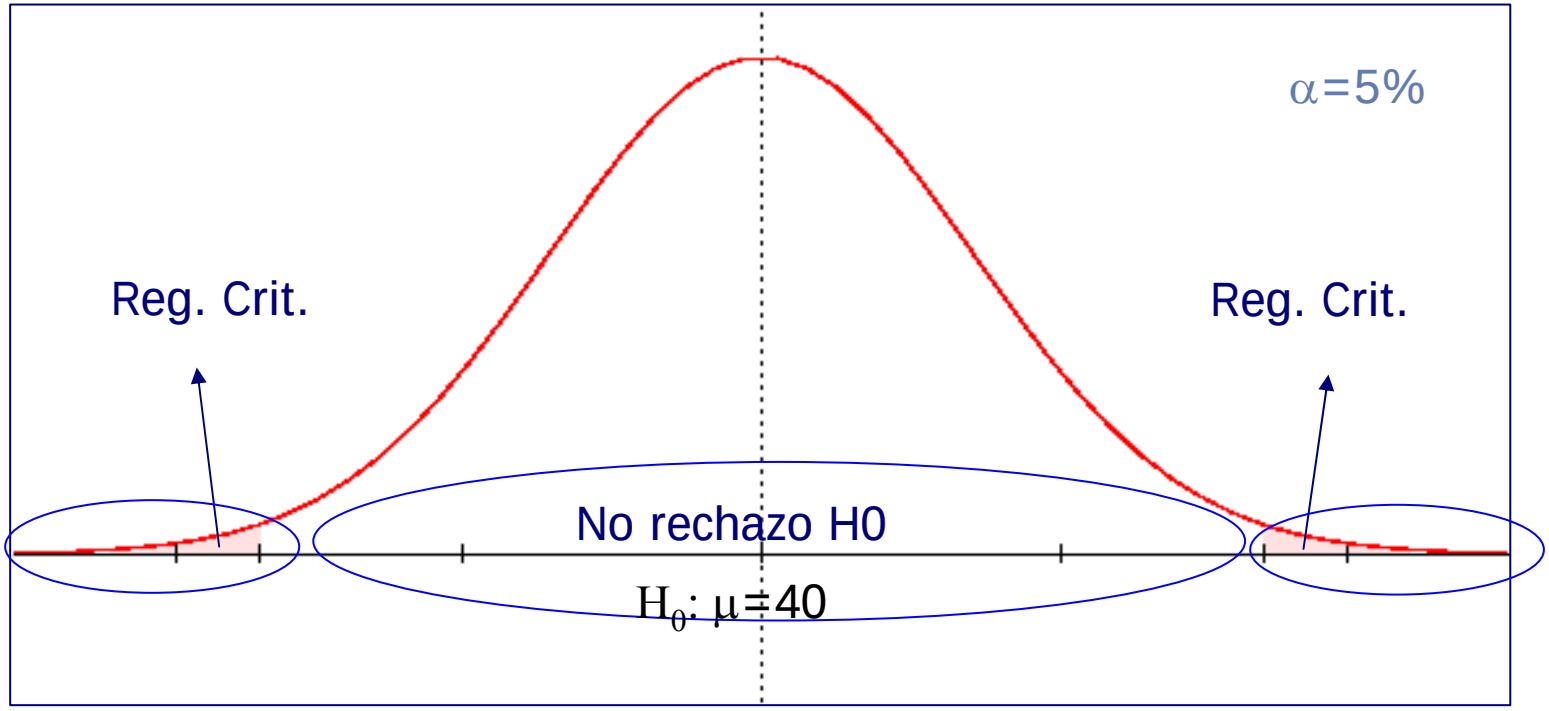
Región crítica y nivel de significancia

Región crítica

- Valores '*improbables*' si...
- Es conocida antes de realizar el experimento: resultados experimentales que refutarían H_0

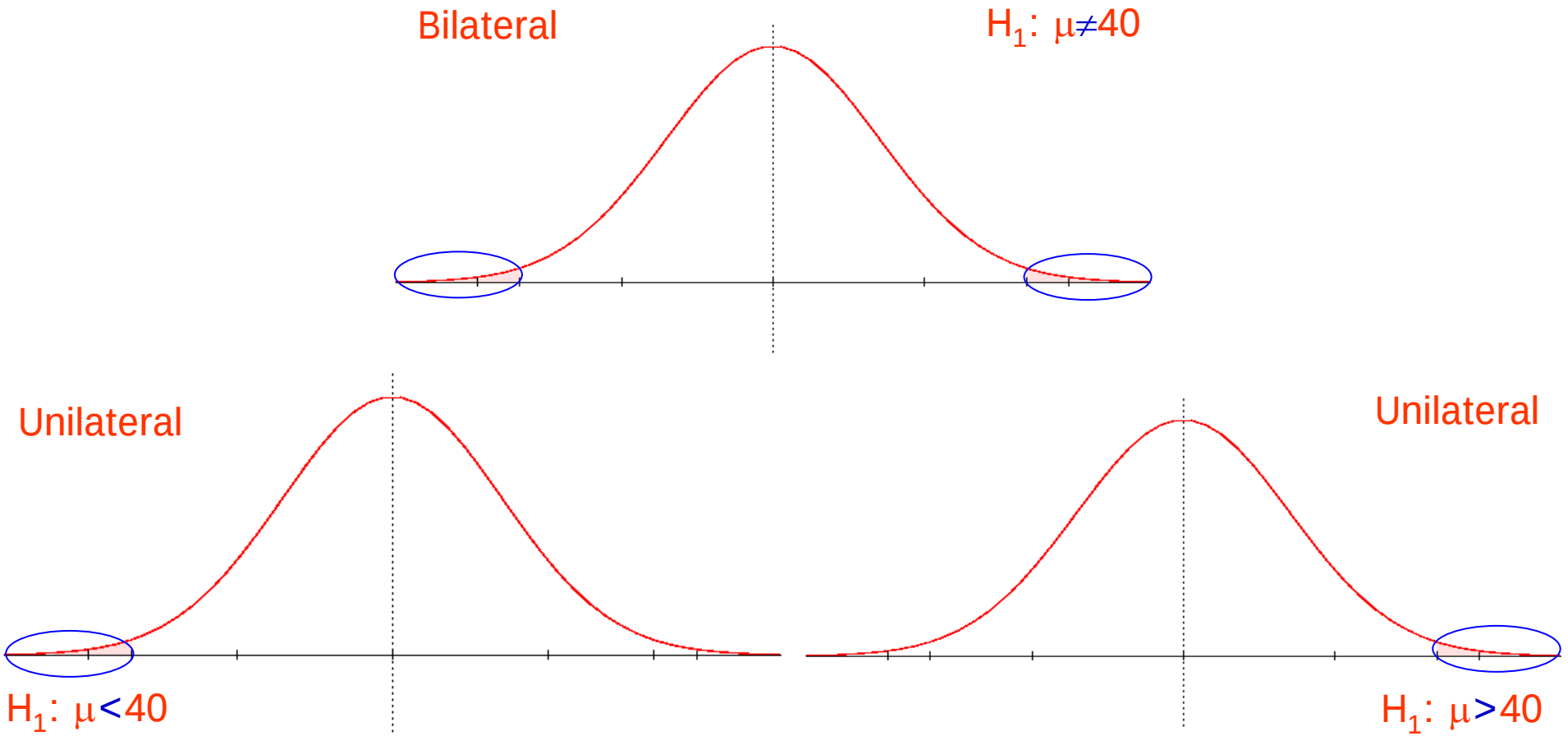
Nivel de significancia: α

- Número pequeño: 1% , 5%
- Fijado de antemano por el investigador
- Es la probabilidad de rechazar H_0 cuando es cierta

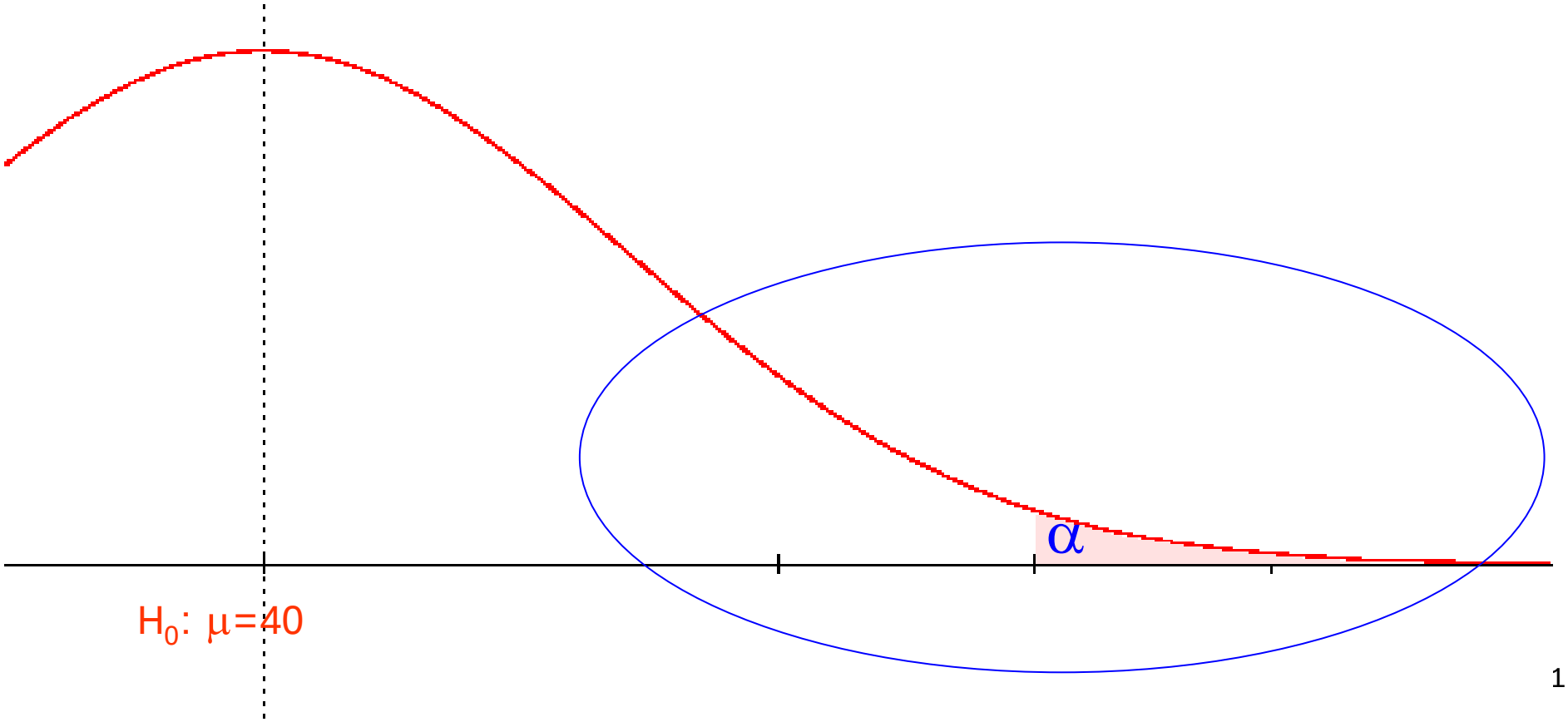


Hipótesis: unilateral y bilateral

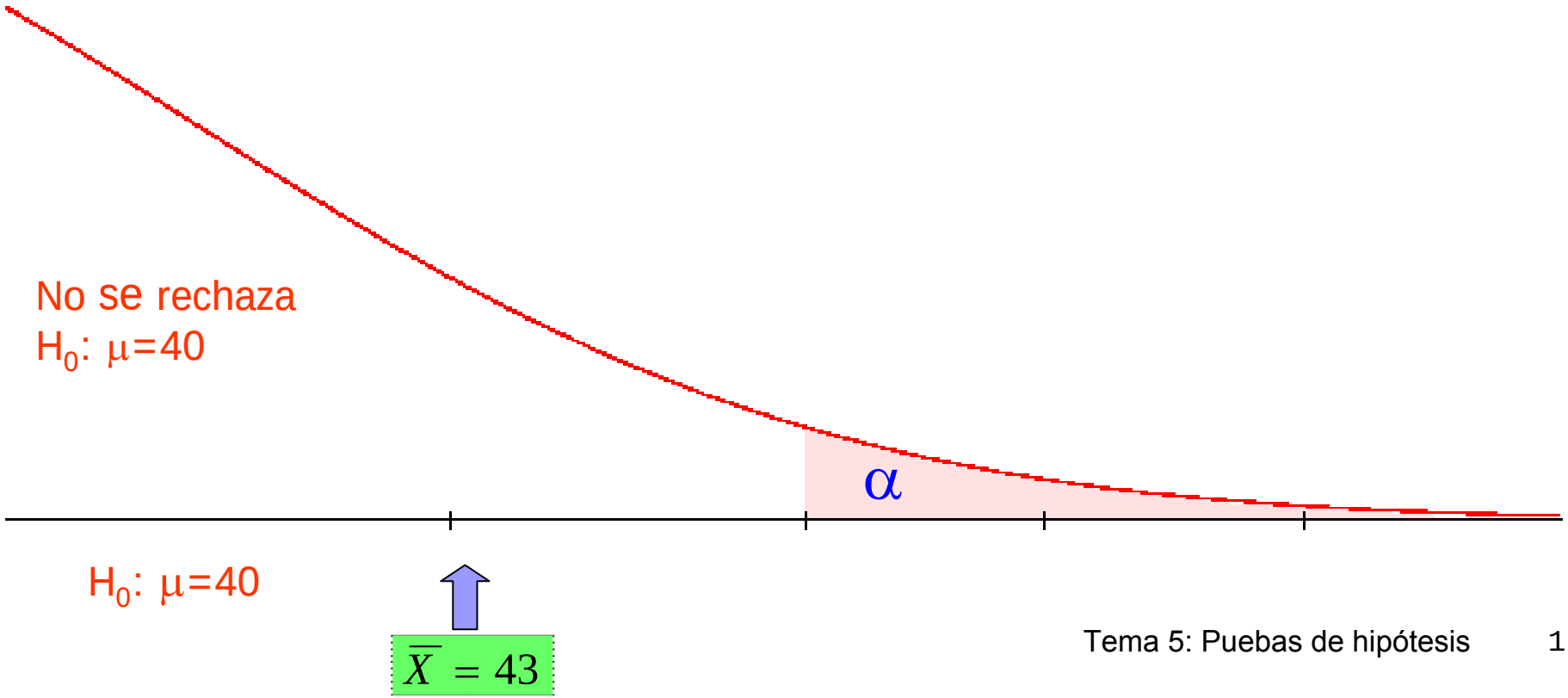
La posición de la región crítica depende de la hipótesis alternativa



Nivel observado de significancia: p

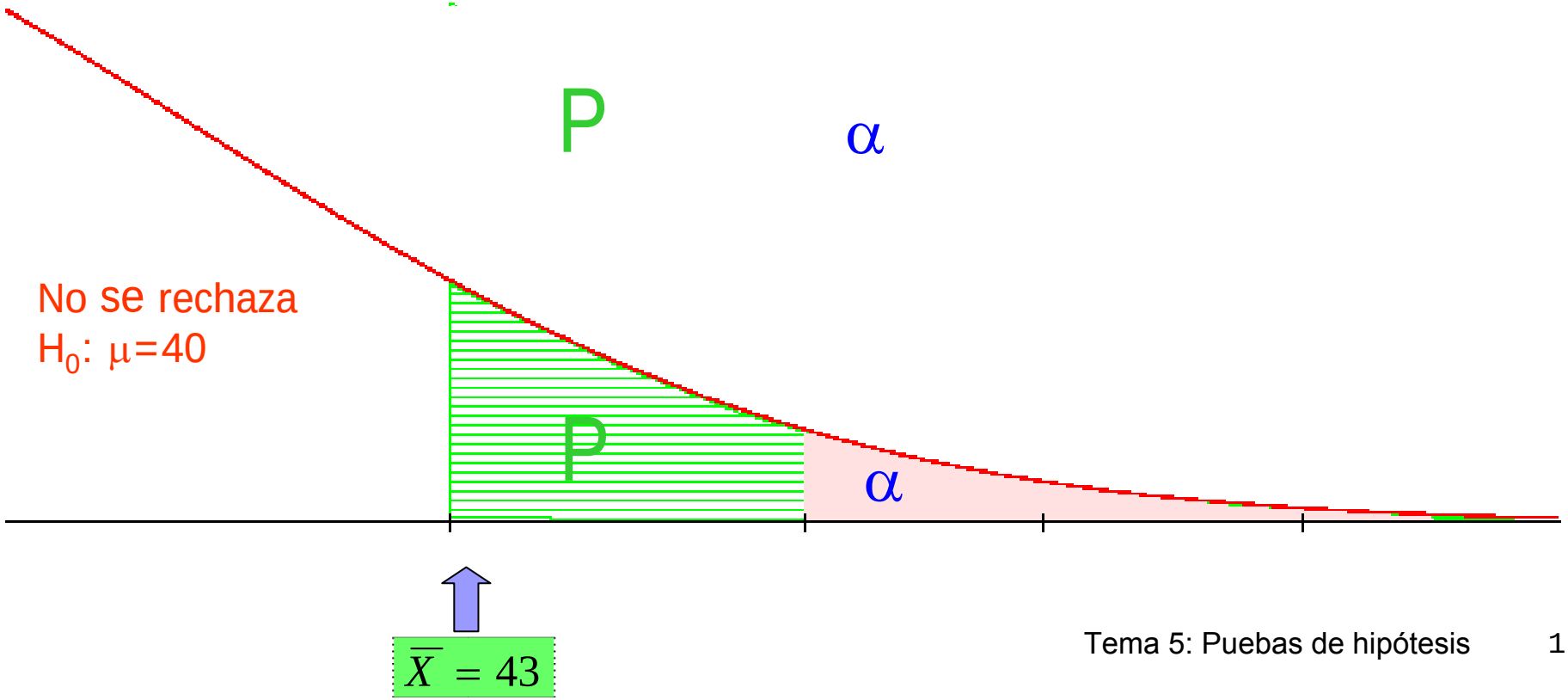


Nivel observado de significancia: p

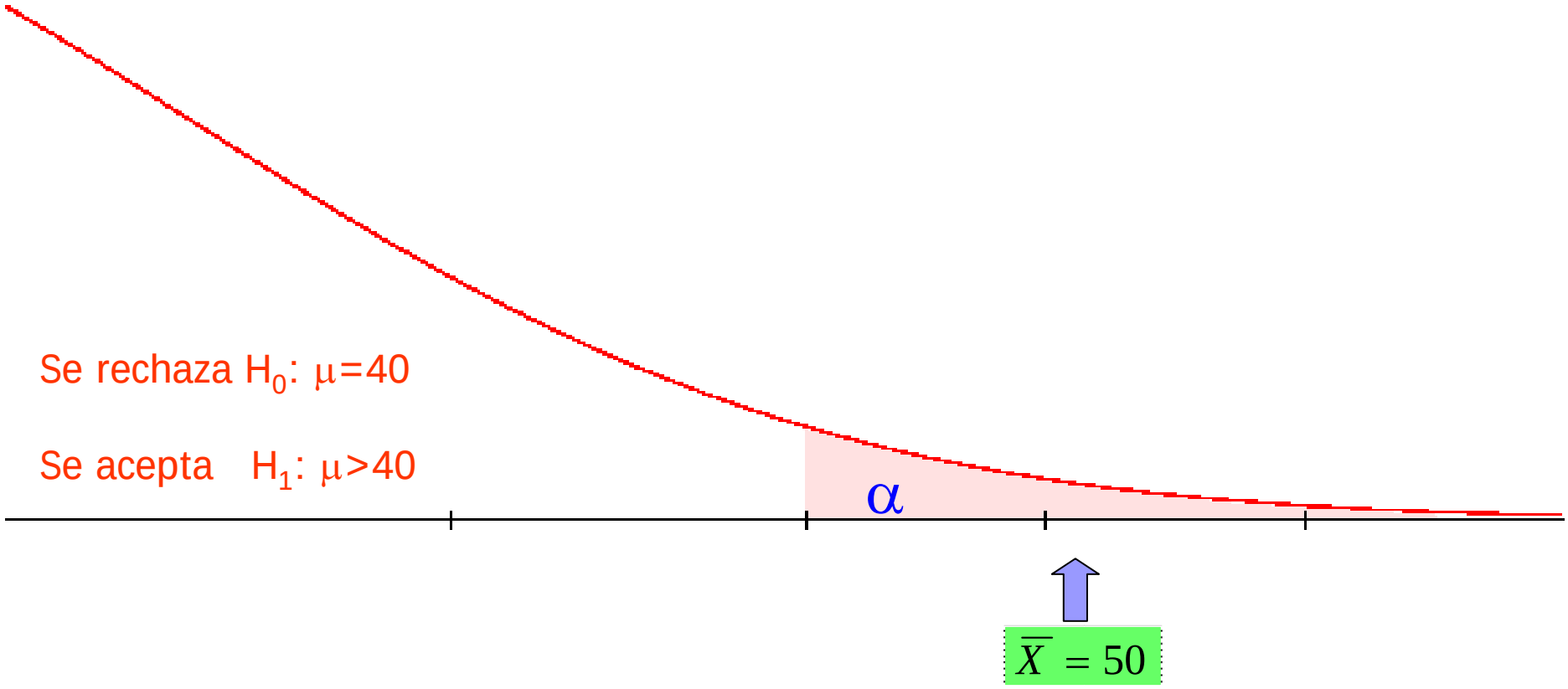


Nivel observado de significancia: p

Es la probabilidad que tendría una región crítica que comenzase exactamente en el valor del estadístico obtenido de la muestra.
p es conocido después de realizar el experimento aleatorio
La hipótesis es no significativa cuando $p>\alpha$



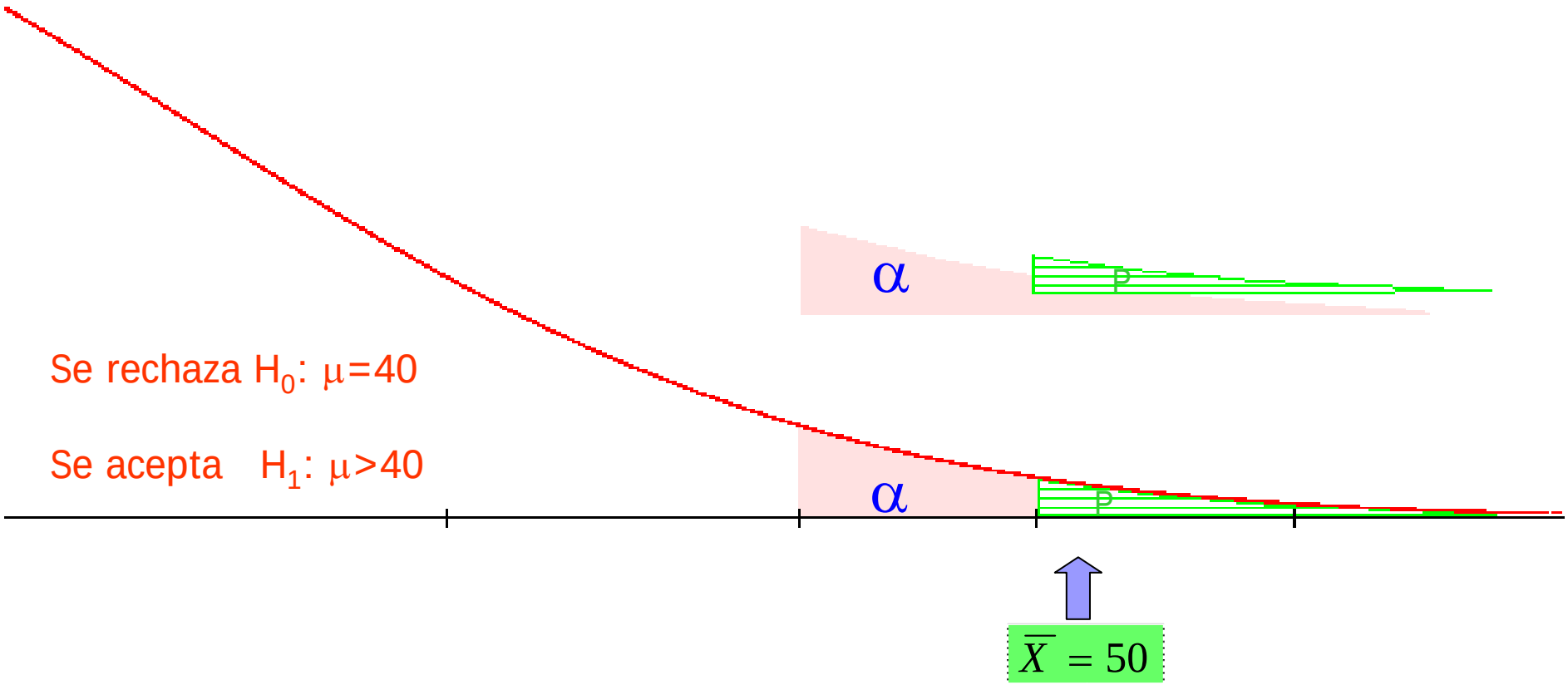
Nivel de observado de significancia : p



Nivel observado significancia : p

La hipótesis es estadísticamente significativa cuando $p<\alpha$

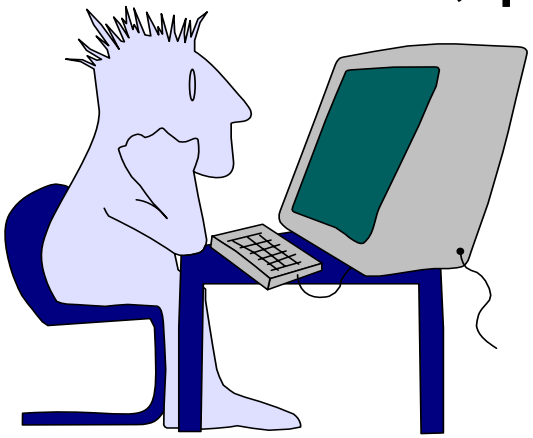
Es decir, si el resultado experimental discrepa más de "lo tolerado" *a priori*.



Resumen: α, p y criterio de rechazo

- Sobre α
  - Es número pequeño, preelegido al diseñar el experimento
  - Conocido α sabemos todo sobre la región crítica
- Sobre p
  - Es conocido tras realizar el experimento
  - Conocido p sabemos todo sobre el resultado del experimento
- Sobre el criterio de rechazo
  - Hipótesis significativa = p menor que α

Resumen: α, p y criterio de rechazo



Estadísticos de contraste [a]

	Edad del encuestado
U de Mann-Whitney	259753,500
W de Wilcoxon	462319,500
Z	-2,317
Sig. asintót. (bilateral)	,021

a. Variable de agrupación: Sexo del encuestado

A significación level of P < 0,05 was selected using t
for Windows V-5, 1992 in a IBM compatible DX 486

MATERIAL Y METODO

Se realizó un estudio de tipo prospectivo y longitudir
estuvo constituido por todos los recién nacidos de m
servicio de Neonatología del Hospital Gineco-Obstétr
Ciudad de La Habana en el período comprendido ent
1993, a los que se le realizó un seguimiento en la se
Hospital por un equipo multitudinario.

La muestra del presente estudio estuvo constituida p
que cumplieran como criterio de inclusión haber com
primeros dos años de edad corregida como mínimo.

Con los 86 pacientes se conformaron 2 grupos:

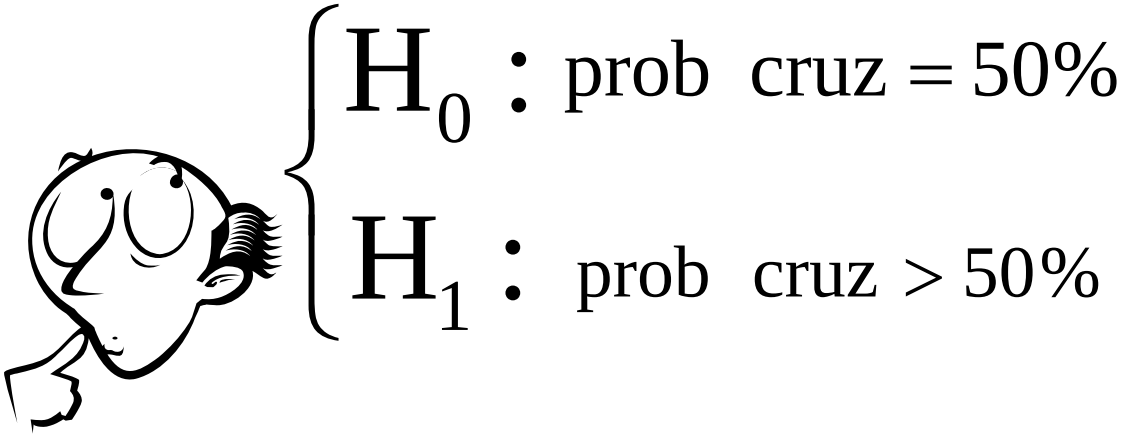
. Ventilados (26 pacientes): formado por aquell
mecánica a presión positiva intermitente durar
neonatal.

b. No Ventilados (60 pacientes): Integrado por lo
no recibieron ninguna modalidad de apoyo ver

- Sobre el criterio de rechazo
  - Hipótesis significativa = p menor que α

Ejemplo

- Problema: ¿Está sesgada la moneda?



$$\begin{cases} H_0 : \text{prob cruz} = 50\% \\ H_1 : \text{prob cruz} > 50\% \end{cases}$$

Experimento: Lanzar la moneda repetidamente:


P=50%


P=25%


P=12.5%


P=6.25%


P=3%


P=1.5%

Riesgos al tomar decisiones

Ejemplo 1: Se juzga a un individuo por la *presunta* comisión de un delito

- H_0: Hipótesis nula
  - Es inocente
- H_1: Hipótesis alternativa
  - Es culpable

Los datos pueden refutarla

La que se acepta si las pruebas no indican lo contrario

Rechazarla por error tiene graves consecuencias

No debería ser aceptada sin una gran evidencia a favor.

Rechazarla por error tiene consecuencias consideradas menos graves que la anterior

Riesgos al probar hipótesis

Ejemplo 2: Se cree que un nuevo tratamiento ofrece buenos resultados

Ejemplo 3: Parece que hay una incidencia de enfermedad más alta de lo normal

- H_0: Hipótesis nula
  - (Ej.1) Es inocente
  - (Ej.2) El nuevo tratamiento no tiene efecto
  - (Ej.3) No hay nada que destacar

No especulativa

- H_1: Hipótesis alternativa
  - (Ej.1) Es culpable
  - (Ej.2) El nuevo tratamiento es útil
  - (Ej. 3) Hay una situación anormal

Especulativa

Tipos de error al tomar una decisión

		Realidad	
		Inocente	Culpable
veredicto	Inocente	OK	Error Menos grave
	Culpable	Error Muy grave	OK

Tipos de error al probar hipótesis

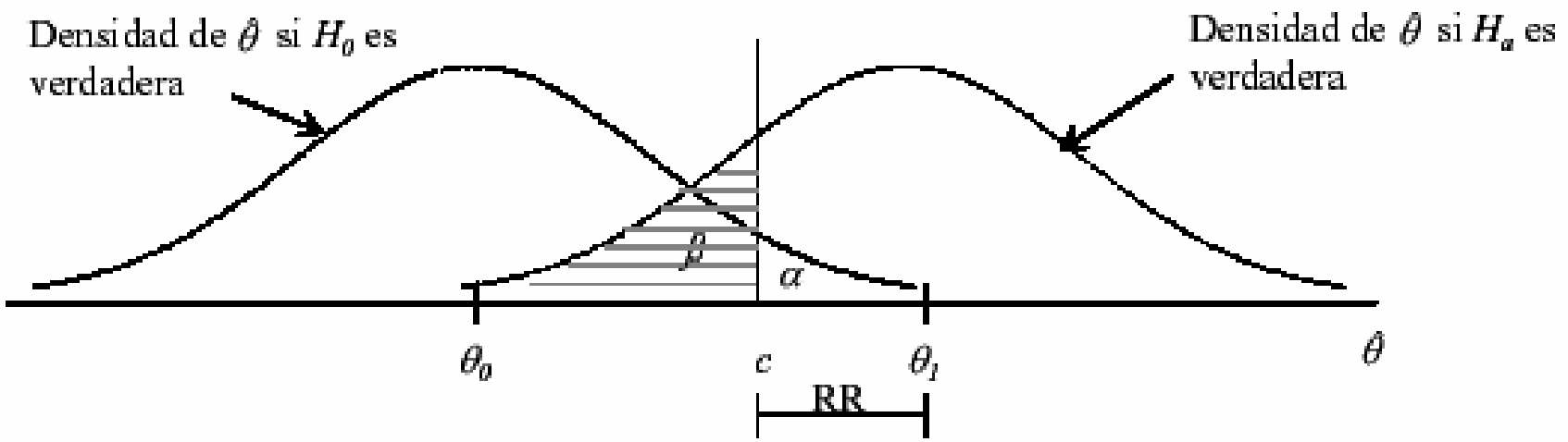
	Realidad	
	H_0 cierta	H_0 Falsa
No Rechazo H_0	**Correcto** El tratamiento no tiene efecto y así se decide.	**Error de tipo II** El tratamiento si tiene efecto pero no lo percibimos. Probabilidad β
Rechazo H0 Acepto H_1	**Error de tipo I** El tratamiento no tiene efecto pero se decide que sí. Probabilidad α	**Correcto** El tratamiento tiene efecto y el experimento lo confirma.

No se puede tener todo

- Para un tamaño muestral fijo, no se pueden reducir a la vez ambos tipos de error.

- Para reducir β, hay que aumentar el tamaño muestral.

Tipos de error al probar hipótesis



La figura muestra que para hacer que α decrezca debemos recorrer c a la derecha, en cuyo caso β crece. La bondad de una prueba estadística de una hipótesis se mide a través de las probabilidades de cometer errores de tipo I y II.

Para un tamaño de muestra n fijo, α y β están en relación inversa: a medida que una aumenta, la otra disminuye.

Conclusiones

- Las hipótesis no se plantean después de observar los datos.
- En ciencia, las hipótesis nula y alternativa no tienen el mismo papel:
  - H_0 : Hipótesis científicamente más simple.
  - H_1 : El peso de la prueba recae en ella.
- α debe ser pequeño
- Rechazar una hipótesis consiste en observar si $p<\alpha$
- Rechazar una hipótesis no prueba que sea falsa. Podemos cometer error de tipo I
- No rechazar una hipótesis no prueba que sea cierta. Podemos cometer error de tipo II
- Si decidimos rechazar una hipótesis debemos mostrar la probabilidad de equivocarnos.

Pruebas de hipótesis para muestras de la distribución normal

Prueba estadística para μ cuando la "muestra es grande" o σ conocida

Resumen de las pruebas de hipótesis con muestras grandes.

Hipótesis:

$$H_0 : \theta = \theta_0$$

$$H_a : \begin{cases} \theta > \theta_0 & \text{(alternativa de cola superior)} \\ \theta < \theta_0 & \text{(alternativa de cola inferior)} \\ \theta \neq \theta_0 & \text{(alternativa de dos colas)} \end{cases}$$

Estadístico de prueba: $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$

$$\text{Región de rechazo} : \begin{cases} z > z_\alpha & \text{(RR de cola superior)} \\ z < -z_\alpha & \text{(RR de cola inferior)} \\ |z| > z_{\alpha/2} & \text{(RR de dos colas)} \end{cases}$$

Ejemplo

Por investigaciones anteriores se sabe que el peso medio de los cerdos de 6 semanas es de 45 Kg, con una desviación típica de 8 Kg. Se ensaya otro tipo de alimentación con un grupo número de cerdos y se selecciona una muestra al azar de 36 cerdos con un peso medio de 48 Kg ¿Aumentó el peso de los cerdos? Utilice α=5%.

Solución

La variable de respuesta o de estudio es Y=peso de los cerdos de 6 semanas.

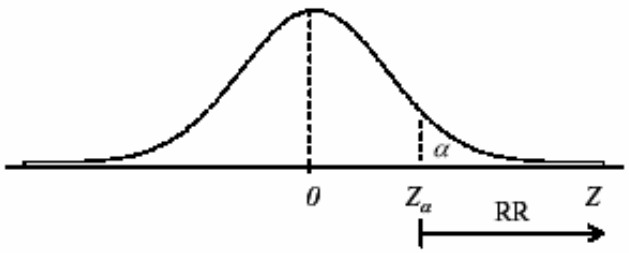
El parámetro en consideración es μ=peso promedio de los cerdos de 6 semanas.

1. Hipótesis: $H_0 : \mu = 45$ vs. $H_a : \mu > 45$
2. Estadístico de prueba: Si $H_0 : \mu = 45$ es cierta, entonces:

$$Z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} = \frac{\overline{Y} - \mu_0}{\sigma_{\overline{X}}} = \frac{\overline{Y} - \mu_0}{\sigma / \sqrt{n}}$$

Solución

3. Región de rechazo: $RR = \{z > z_\alpha\}$ donde



$$P(Z > z_\alpha) = \alpha$$
$$P(Z > z_{0.05}) = 0.05$$
$$\Rightarrow Z_{0.05} = 1.645 \Rightarrow RR = \{z > 1.645\}$$

4. Cálculo del estadístico de prueba: El valor observado de Z es

$$z_c = \frac{\overline{Y} - 45}{8/\sqrt{36}} = \frac{48 - 45}{8/6} = 2.25$$

5. Decisión: Como $z_c = 2.25 > 1.645$ se rechaza H_0 y decidimos por H_a, es decir, los datos presentan suficiente evidencia con α = 5% de que la nueva dieta aumenta el peso promedio de los cerdos de 6 semanas.

Pruebas de hipótesis para muestras de la distribución normal

Prueba de hipótesis respecto a la media de una población μ cuando σ se desconoce

Cuando se desconoce σ^2 en la prueba para μ no puede usarse Z como estadístico de prueba. En su lugar usamos (con $H_0 : \mu = \mu_0$ verdadera) la variable aleatoria

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} = \sqrt{n}\,\frac{\overline{Y} - \mu_0}{S} \sim t_{n-1} \text{ g.l.}$$

con variable de respuesta $Y \sim N(\mu = \mu_0, \sigma^2)$.

Pruebas de hipótesis para muestras de la distribución normal

Resumen de la prueba de hipótesis para μ de una distribución normal con varianza desconocida

Supuesto: La variable de respuesta se distribuye aproximadamente normal.

Hipótesis

$$H_0 : \mu = \mu_0$$

$$H_a : \begin{cases} \mu > \mu_0 & \text{(alternativa de cola superior)} \\ \mu < \mu_0 & \text{(alternativa de cola inferior)} \\ \mu \neq \mu_0 & \text{(alternativa de dos colas)} \end{cases}$$

Estadístico de prueba

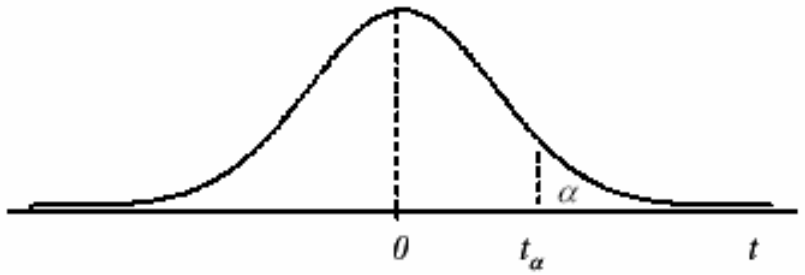
Si la $H_0 : \mu = \mu_0$ es cierta entonces $T = \dfrac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$ g.l.

Pruebas de hipótesis para muestras de la distribución normal

Resumen de la prueba de hipótesis para μ de una distribución normal con varianza desconocida

$$\text{Región de Rechazo}: \begin{cases} t > t_\alpha & \text{(Región de rechazo de cola superior)} \\ t < -t_\alpha & \text{(Región de rechazo de cola inferior)} \\ |t| > t_{\alpha/2} & \text{(Región de rechazo de dos colas)} \end{cases}$$

Los valores de t_α indican el valor de t tal que a su derecha se encuentra un área α, es decir, t_α corresponde a $P(T > t_\alpha) = \alpha$ para una distribución t con $n-1$ grados de libertad.



86

Ejemplo

Cada especie de luciérnaga tiene un modo peculiar de centelleo. Para una determinada especie consiste en un destello corto de luz seguido por un período de reposo que se piensa tenga una duración media de menos de cuatro segundos. Se obtuvieron los siguientes datos acerca del período de reposo entre centelleos para una muestra de 16 luciérnagas de esta especie.

3.9	3.2	3.8	4
3.8	3.5	3.7	4
3.5	3.6	4.2	3.6
4.1	3.7	3.4	4.3

¿El tiempo de reposo medio propuesto es de menos de cuatro segundo?

Solución

La variable de respuesta es Y=tiempo de reposo entre centelleos.
El parámetro en consideración es μ= tiempo promedio de reposo entre centelleos.

Solución

1. Hipótesis

$$H_0 : \mu = 4 \qquad \text{vs} \qquad H_1 : \mu < 4$$

2. Estadístico de prueba

Si la $H_0 : \mu = \mu_0$ es cierta entonces $T = \dfrac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$ g.l.

3. Cálculo del estadístico de prueba

De los datos muestrales obtenemos la media y desviación estándar:

$\overline{x} = 3.77 \qquad s = 0.30$

$$\Rightarrow t_c = \frac{\overline{x} - 4}{s/\sqrt{n}} = -3.06$$

Solución

4. Región de rechazo.

Para $\alpha = 1\% = 0.01$ la RR es $t_c < -t_\alpha = -t_{0.01,15} = -26025$

5. Decisión

Como $t_c = -3.06 < -2.6025$ rechazo H_0, con $\alpha = 1\%$, y decido por $H_1 : \mu < 4$ seg., es decir, el tiempo medio de reposo es menor de cuatro segundos.