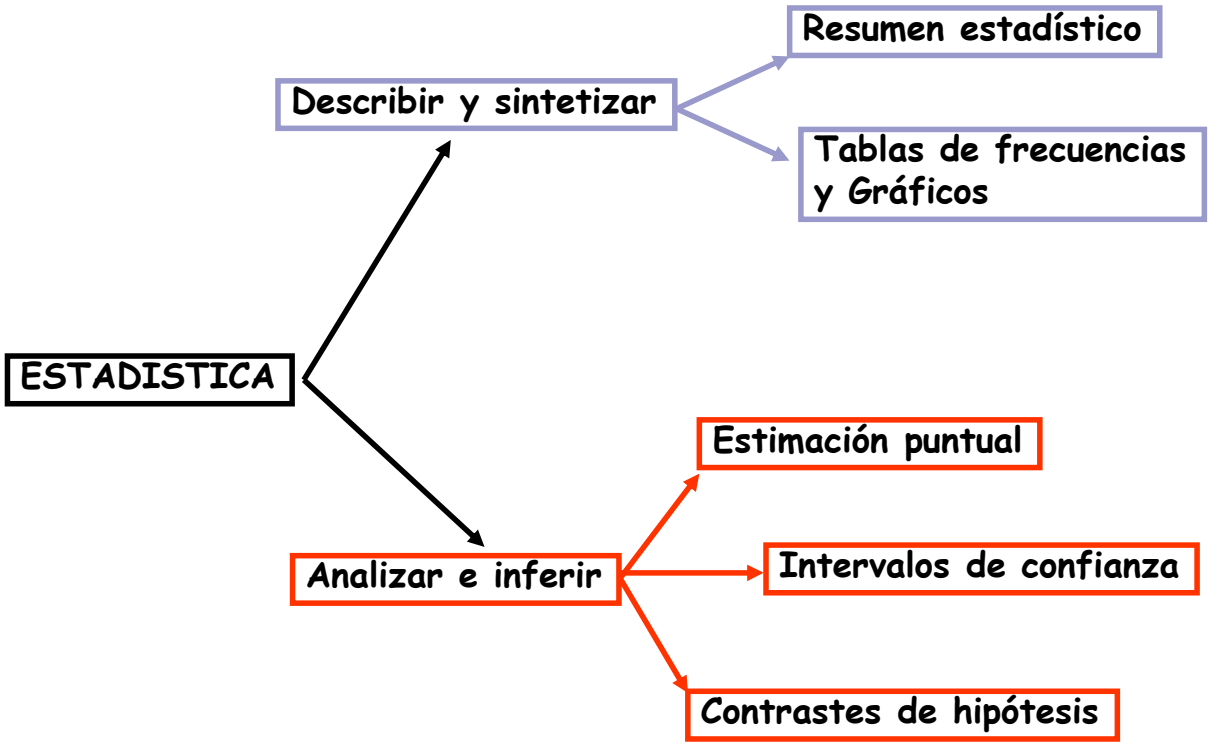
TÉCNICAS ESTADÍSTICAS APLICADAS EN NUTRICIÓN Y SALUD

Análisis descriptivo y exploratorio de datos

Francisco M. Ocaña Peinado @ocanapaco
http://www.ugr.es/local/fmocan
Departamento de Estadística e Investigación Operativa. UGR



Formas de representar los datos

- Resumen estadístico
  - Resumir las observaciones muestrales mediante algunos estadísticos que de una idea global del comportamiento del conjunto de datos
    - Centralización, Dispersión, Posición o Forma
- Tabla de Frecuencias
  - Ordenan y resumen la información de la muestra de forma que no se pierda nada de información (o poca)
    - Frecuencias absolutas, relativas y acumuladas (absolutas/relativas)
- Representación Gráfica
  - Se pretende que en una imagen se visualice la información de la muestra
    - Existen diferentes gráficos según sea la naturaleza de la variable: Barras, sectores, tallo y hojas, diagrama de caja, histograma…

Análisis Exploratorio de Datos

- **Análisis Exploratorio de Datos (AED)**: Conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas.

- Para conseguir este objetivo el **AED** proporciona métodos sistemáticos para **organizar y preparar los datos**, detectar fallos en el diseño y recogida de los mismos, **tratamiento y evaluación de datos ausentes** (*missing*), **identificación de casos atípicos** (*outliers*) y **comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes**: (*normalidad, linealidad y homocedasticidad*).

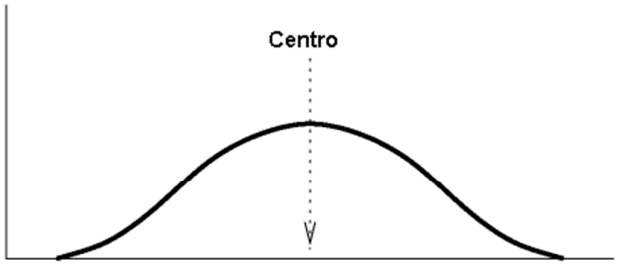
- El **AED es complementario** a la Estadística Descriptiva clásica: se pretende **maximizar la información** de la variable **previo** a establecer cualquier hipótesis de **Inferencia.**

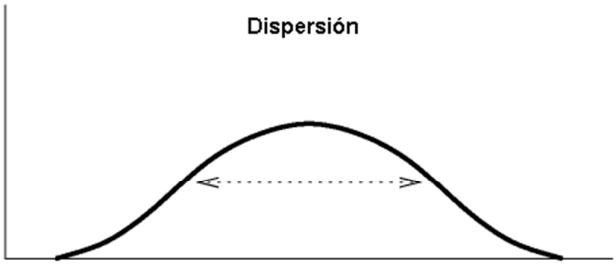
Un breve resumen sobre estadísticos

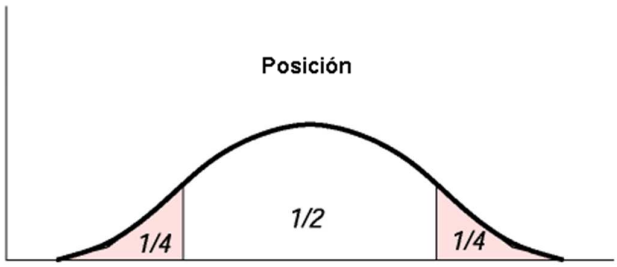
- Posición
  - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos
    - Percentiles, cuartiles, deciles,...
- Centralización
  - Indican valores con respecto a los que los datos parecen agruparse
    - Media, mediana y moda.
    - Estimadores robustos: Trimedia, media recortada, media de cuartiles, centrimedia...
- Dispersión
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización
    - Varianza, Desviación típica, Coeficiente de Variación, RIQ, CVI.
- Forma
  - Asimetría o Sesgo
  - Apuntamiento o Curtosis

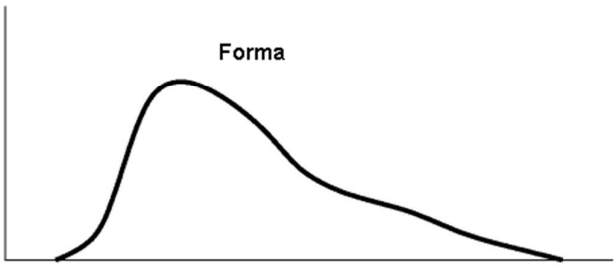
Un breve resumen sobre estadísticos

Analizar/ Estadísticos descriptivos/ Frecuencias
Analizar/ Estadísticos descriptivos/ Descriptivos
Analizar/ Estadísticos descriptivos/ Explorar



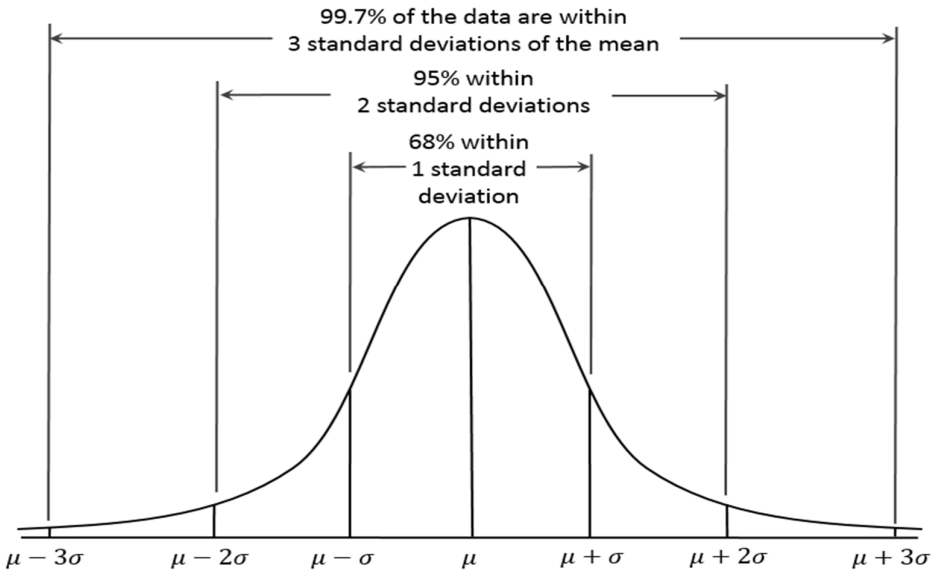






Detectar normalidad de la variable

- Objetivo
  - Determinar si una variable sigue una distribución Normal
  - **Esto es una cuestión decisiva, puesto que condiciona la elección de un método de Inferencia paramétrica o no paramétrica**



Estadísticos de centralización

Son medidas que buscan posiciones (valores) con respecto a los cuales los datos muestran tendencia a agruparse.

- Media ('mean') 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 ('median') 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 ('mode') Es el/los valor/es donde la distribución de frecuencia alcanza un máximo.

Estadísticos de centralización

- **Me** es una ***medida robusta***: menos sensible que la media a la variación de las observaciones muestrales.
- La media: muy **influenciada** por observaciones muy grandes o muy pequeñas, no así la Me.
- Me: más **recomendable** que la media cuando hay **asimetría** (existen una o muy pocas observaciones en uno de los extremos).
- **Moda** tiene el inconveniente de no ser necesariamente única.
- Las **unidades** en que vienen expresadas la media, mediana y moda corresponden a las de la variables en estudio.
- ***Estadísticos robustos***: Centrimedia, Media recortada, Promedio de Cuartiles o la Trimedia

Estadísticos de dispersión

- **Objetivo:** Medir el grado de proximidad de los datos con respecto a una medida de tendencia central (la media).

Ej: Se estudia la pérdida de peso con dos dietas A y B (en kg):

A: 9, 12, 15, 15, 55, 50 **B: 23, 24, 25, 26, 26, 27, 28, 29**

Ambas medias son 26 kg. ¿Es representativa la media?

Pretendemos dar una medida de la variabilidad en los datos para saber si la media es muy representativa o poco representativa del conjunto de observaciones. Primera idea: medir la *desviación* de cada dato a la media

Estadísticos de dispersión

Varianza, S^2 ('variance'): Media de los cuadrados de las *desviaciones*

La expresión de la varianza:

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

$$S_A^2 = 428.8\, kg^2$$

$$S_B^2 = 2.1667\ kg^2$$

Problema de la varianza: Viene expresada en las unidades al cuadrado de los datos, lo cual no tiene interpretación. En el ejemplo tendríamos que reducción media de peso es de 26 kg con $S^2 = 2.1667$ kg^2.

Este problema se resuelve tomando la raíz cuadrada positiva de la varianza, conocida como **desviación típica ('standard deviation')** y se representa como **S.**

$$S = \sqrt{S^2} = \sqrt{2.1667} = 1.4719\ kg$$

Estadísticos de dispersión

Coeficiente de variación, C.V. ('variation coefficient') Objetivo: comparar la dispersión de dos variables. Definición: cociente entre desviación típica y media. Por tanto, **no tiene unidades** y suele expresarse en %.

Pesos $\bar{x} = 3.09\ kg$ con S = 0.423 *kg*

Alturas $\bar{x} = 53.6\, cm$ con S = 3.9038 *cm*

¿Qué variable presenta mayor dispersión?

$$CV\ (Peso\ ) = \frac{S}{\bar{x}} = \frac{0.423}{3.09} = 0.1368 \quad (13.68\ \%)$$

$$CV\ (Altura\ ) = \frac{S}{\bar{x}} = \frac{3.9038}{53.6} = 0.072 \quad (7.2\ \%)$$

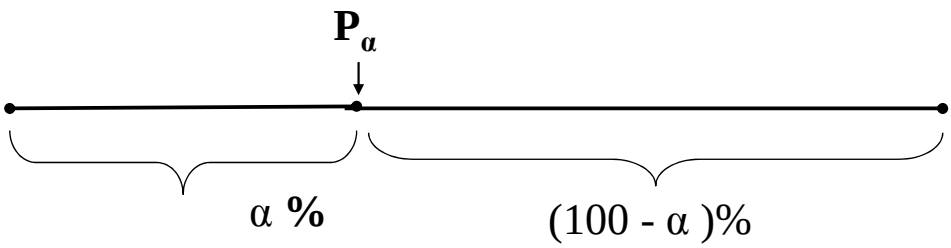
Estadísticos de dispersión

- Las medidas de dispersión son siempre no negativas. Son 0 si todas las observaciones muestrales son idénticas.
- S^2 tiene el inconveniente de tener como unidades las de la variable original al cuadrado. Por ello, se emplea S.
- S^2 y S son muy sensibles a la variación de cada una de las observaciones, ya que su valor depende de todos los datos muestrales.
- No se *debería* usar S^2 o S si la media no es la medida de tendencia central:
  - Media de las desviaciones absolutas respecto de Me
  - Mediana de las desviaciones absolutas respecto de Me.
  - RIQ = Rango entre cuartiles: mide la variabilidad de la mitad central de los datos; RIQ = $Q_3 - Q_1$
  - Coeficiente de Variación Intercuartílico: $(Q_3 - Q_1)/(Q_3 + Q_1)$

Estadísticos de posición: Percentiles

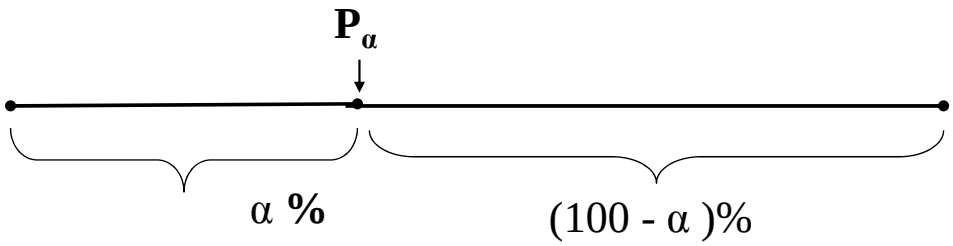
Percentil ('percentile') de orden α, se representa P_α :
valor que deja el α % de las observaciones a su izquierda, una vez ordenadas de menor a mayor.

$P_\alpha = x_i$ que ocupa el lugar $\alpha \cdot (n+1)/100$



El cálculo de un percentil es análogo al de la mediana.
La mediana es el P_{50}

Estadísticos de posición: Percentiles



- **Cuartiles (Q_1,Q_2,Q_3):** Dividen al conjunto en 4 partes.
- **Deciles (D_1,D_2,D_3,...,D_9):** Dividen al conjunto en 10 partes.
- **Percentiles (P_1,P_2,P_3,...,P_{99}):** Dividen al conjunto en 100 partes.

Estadísticos de posición: Percentiles

Estadísticos

edad

N	Válidos	52
	Perdidos	0
Percentiles	5	47,30
	25	55,00
	50	62,00
	75	66,75

Estadísticos

edad

N	Válidos	52
	Perdidos	0
Percentiles	25	55,00
	50	62,00
	75	66,75

Estadísticos

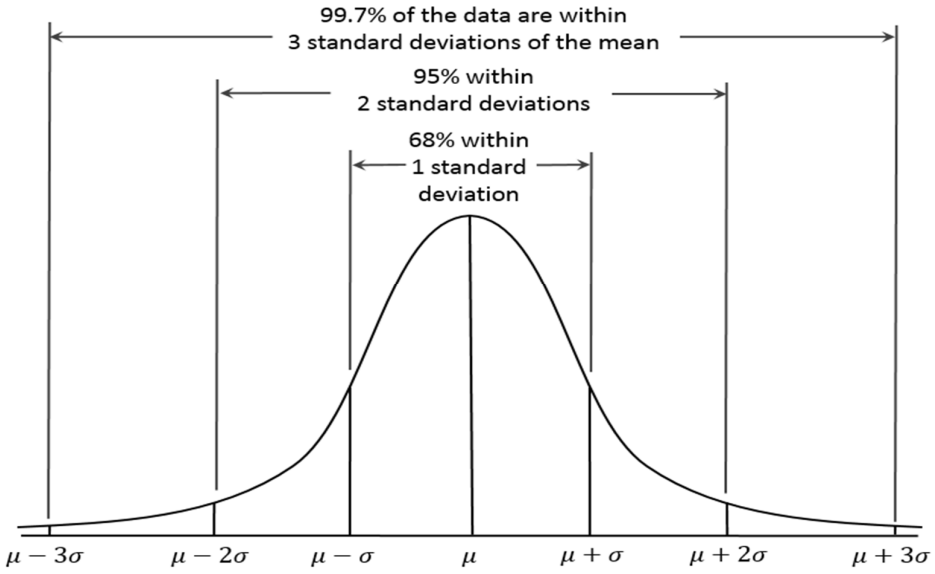
edad

N	Válidos	52
	Perdidos	0
Percentiles	10	50,30
	20	54,60
	25	55,00
	30	55,00
	40	57,20
	50	62,00
	60	64,00
	70	65,10
	75	66,75
	80	68,00
	90	70,00

- Analizar/ Estadísticos descriptivos/ Frecuencias
- Analizar/ Estadísticos descriptivos/ Explorar

Medidas de Forma: Asimetría/Sesgo y Curtosis

Medidas que comparan la forma que tiene la variable, en relación con el modelo Normal

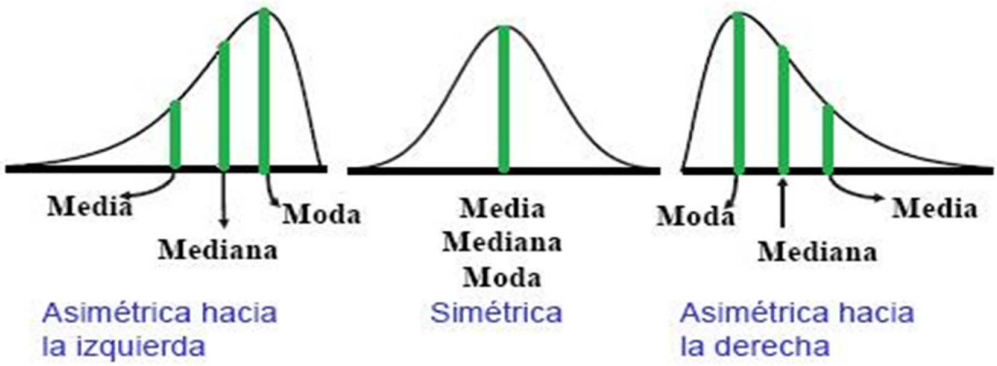


Asimetría o Sesgo

- Una distribución es simétrica si la mitad izquierda de su distribución es la imagen exacta de su mitad derecha.
- En las distribuciones simétricas media y mediana coinciden. Si sólo hay una moda también coincide. **Las discrepancias entre las medidas de centralización son indicación de asimetría.**
- 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).

- Analizar/ Estadísticos descriptivos/ Descriptivos
- Analizar/ Estadísticos descriptivos/ Explorar

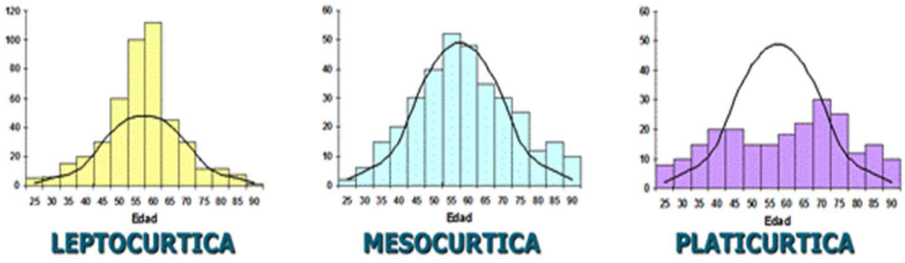
Asimetría o Sesgo



En función del signo del coeficiente diremos que la asimetría es positiva o negativa:

- Distribución simétrica ➔ coeficiente asimetría nulo.
- Distribución asimétrica a la derecha ➔ coeficiente de asimetría positivo.
- Distribución asimétrica a la izquierda ➔ coeficiente de asimetría negativo.

Apuntamiento o curtosis



La curtosis nos indica el grado de apuntamiento/aplastamiento de una distribución con respecto a la distribución Normal. Es adimensional.

- Leptocúrtica (más apuntada que la normal): coef. de curtosis > 0
- Mesocúrtica (como la normal): coef. de curtosis = 0
- Platicúrtica (más aplanada o aplastada que la normal): coef. de curtosis < 0

Tablas de frecuencias

- 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): Contabiliza la proporción de individuos
  - Frecuencias acumuladas: Sólo tienen sentido para variables ordinales y numéricas
    - Útiles para detectar percentiles
      - ¿Qué porcentaje de individuos tiene al menos 3 hijos? Sol: 83,8
      - ¿Entre 4 y 6 hijos? Soluc 1ª: 8,4%+3,6%+1,6%= **13,6%**.

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

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		

Tabla de frecuencias

➢ Analizar/ Estadísticos descriptivos/ Frecuencias

Sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Masculino	87	36,3	36,3	36,3
	Femenino	153	63,8	63,8	100,0
	Total	240	100,0	100,0	

Hábito tabáquico

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No fumador	130	54,2	54,2	54,2
	Ex-fumador	28	11,7	11,7	65,8
	Fumador	82	34,2	34,2	100,0
	Total	240	100,0	100,0	

Colesterol HDL

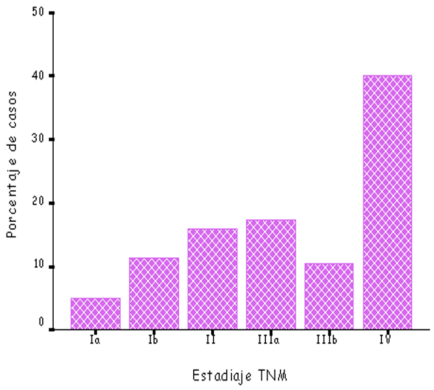
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	36,00	1	,4	3,3	3,3
	37,00	1	,4	3,3	6,7
	41,00	1	,4	3,3	10,0
	42,00	1	,4	3,3	13,3
	43,00	2	,8	6,7	20,0
	47,00	2	,8	6,7	26,7
	50,00	2	,8	6,7	33,3
	51,00	1	,4	3,3	36,7
	52,00	1	,4	3,3	40,0
	54,00	1	,4	3,3	43,3
	56,00	2	,8	6,7	50,0
	58,00	2	,8	6,7	56,7
	59,00	1	,4	3,3	60,0
	61,00	2	,8	6,7	66,7
	62,00	1	,4	3,3	70,0
	65,00	1	,4	3,3	73,3
	72,00	1	,4	3,3	76,7
	73,00	1	,4	3,3	80,0
	80,00	1	,4	3,3	83,3
	83,00	1	,4	3,3	86,7
	85,00	1	,4	3,3	90,0
	90,00	1	,4	3,3	93,3
	96,00	1	,4	3,3	96,7
	104,00	1	,4	3,3	100,0
	Total	30	12,5	100,0	
Perdidos	Sistema	210	87,5		
Total		240	100,0		

Representaciones gráficas

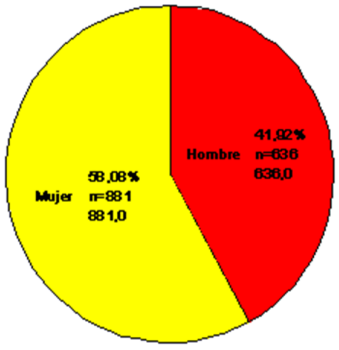
- Objetivo
  - Se pretende que en una imagen se visualice la información de la muestra
  - **El objetivo de los gráficos es que la información impacte con poco esfuerzo, de manera que pueda interpretarse de forma rápida.**
- Requisitos indispensables
  - Deben indicarse claramente las escalas y unidades de medida.
  - El área de cada modalidad debe ser proporcional a la frecuencia.
- Tipos de Representaciones Gráficas
  - Diferentes gráficos según sea la naturaleza de la variable
    - Variables cualitativas: Diagrama de Barras, de Sectores
    - Variables cuantitativas discretas: Diagrama de Barras o de Sectores.
    - Variables cuantitativas continuas: Histograma, Diagrama de Caja (Box & Whisker Plot) y de Tallo y Hojas (Stem & Leaf Plot)

Gráficos para v. cualitativas

- Diagramas de barras
  - Alturas proporcionales a las frecuencias (absolutas o relativas)
  - Se pueden aplicar también a variables discretas



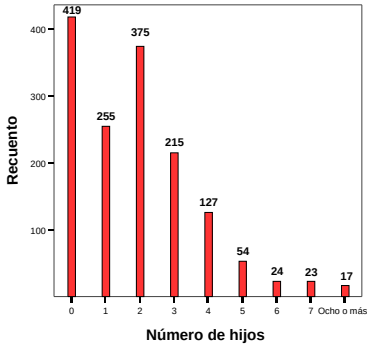
- Diagramas de sectores
  - El área de cada sector es proporcional a su frecuencia (absolutas o relativas)
  - No conveniente usarlo para variables ordinales

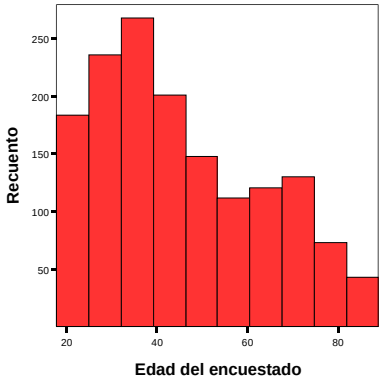


➢ Analizar/ Estadísticos descriptivos/Frecuencias/Gráficos

Gráficos para variables cuantitativas

- **Gráficos de barras**: Diferentes en función de que las variables sean discretas o continuas. Pueden ser de frecuencias absolutas o relativas.
  - **Diagramas barras para v. discretas**
    - Se deja un hueco entre barras para indicar los valores que no son posibles
  - **Histograma 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.



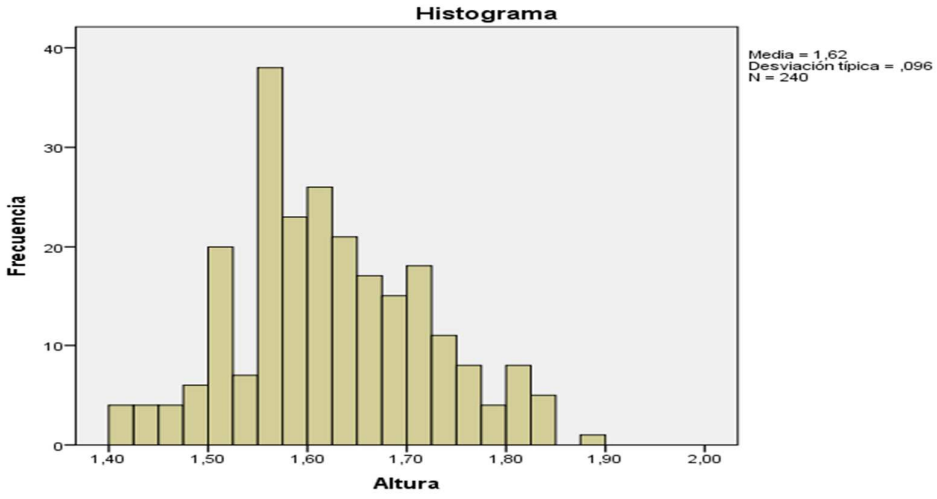


Gráficos para variables cuantitativas: Histograma

Representación para una variable **cuantitativa continua** que muestra la frecuencia (abs/rel) de la variable a lo largo de diferentes intervalos (deben ser iguales)

La altura de cada barra es la frecuencia absoluta o relativa de ese intervalo

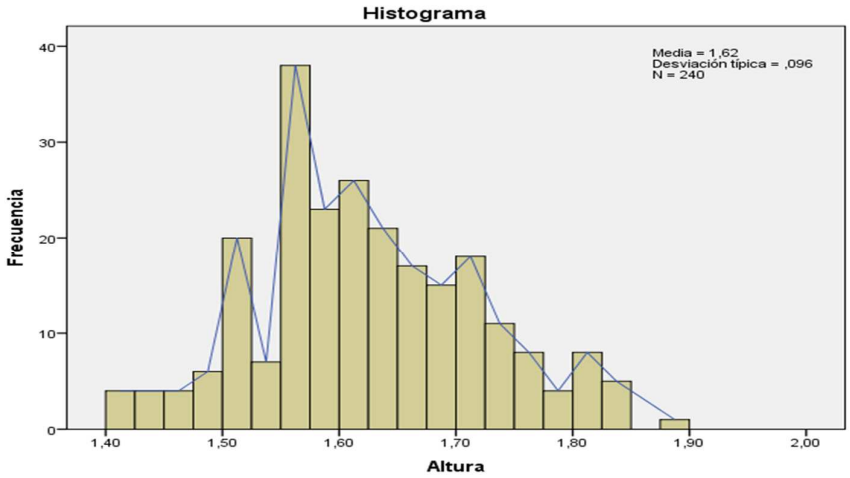
Muestra la forma, el centro y dispersión de la variable



➢ Analizar/ Estadísticos descriptivos/Frecuencias/Gráficos

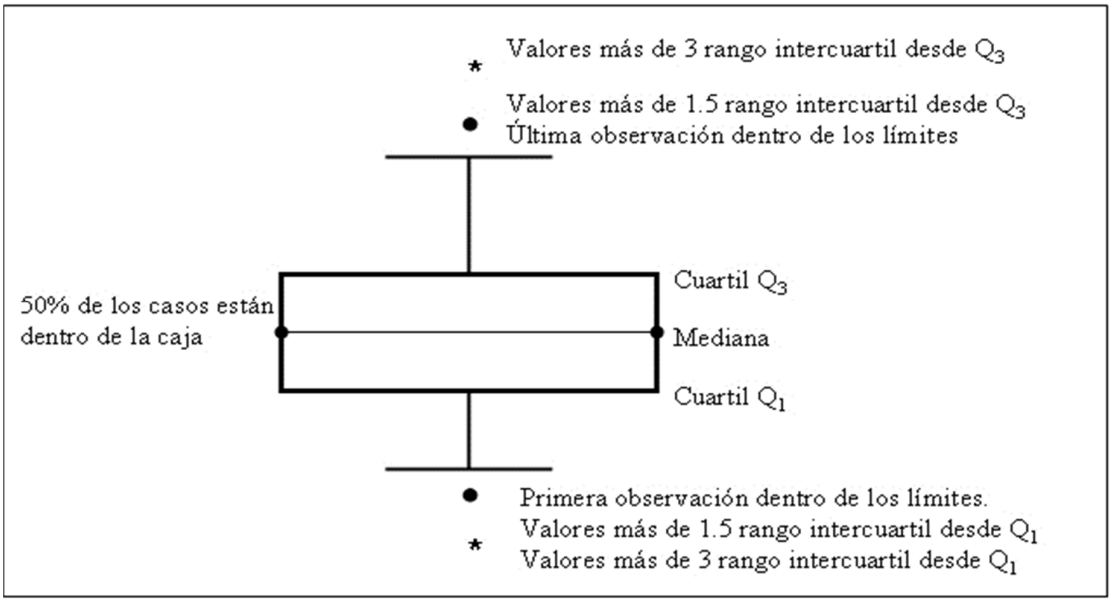
Gráficos para var.cuantitativas: Polígono de Frecuencias

El polígono de frecuencias resulta de la unión mediante una línea quebrada de los puntos medios de las bases superiores de los rectángulos del histograma.



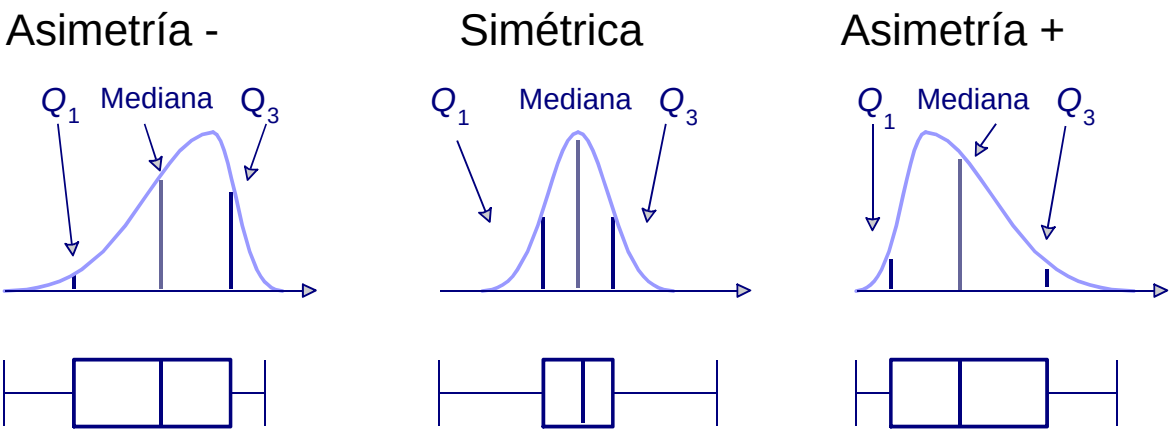
- Editar Histograma/Elementos/Línea de interpolación
- Gráficos/Líneas/Simple/Resumen para grupos de casos

Gráficos para var.cuantitativas: Gráfico de Caja



Se representan: Me, P_α, dispersión y permite detectar valores extremos y asimetría.

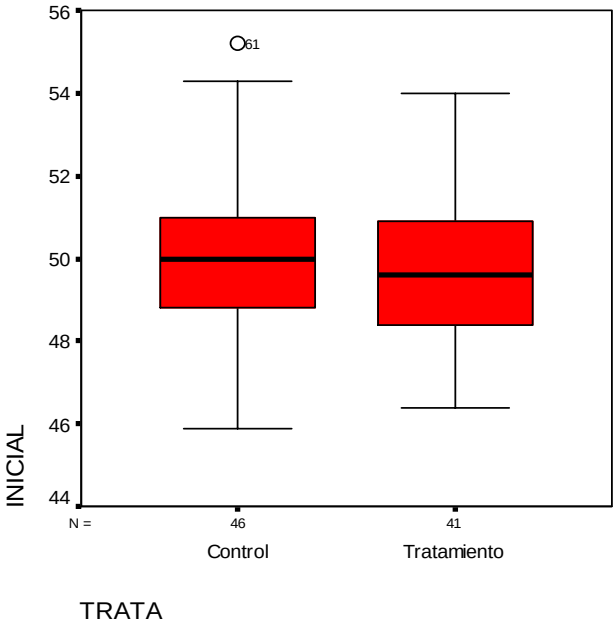
Analizar/ Estadísticos descriptivos/ Explorar/ Gráficos/ Gráfico de caja



- Me en el centro de la caja: indicio de variable simétrica
- Me tiende al límite superior de la caja: indicio de asimetría a la izquierda
- Me tiende al límite inferior de la caja: indicio de asimetría a la derecha
- Si el «bigote» que sale desde el límite inferior de la caja es más largo: indicio de asimetría a la izquierda
- Si el «bigote» que sale desde el límite superior de la caja es más largo: indicio de asimetría a la derecha
- Los 2 «bigotes» tienen aproximadamente igual longitud: indicio de variable simétrica

Gráficos para var.cuantitativas: Gráfico de Caja

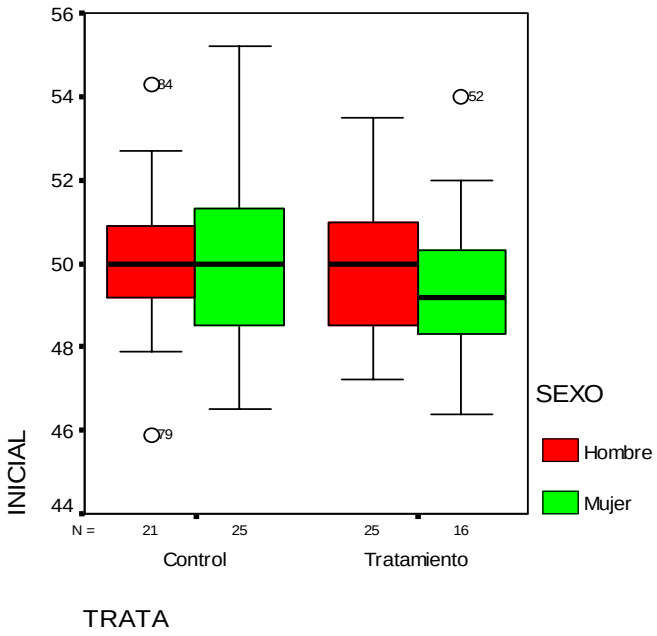
Permite además comparar gráficamente el comportamiento de una variable en distintos grupos



- Las dos distribuciones son similares
- El grupo control presenta algo más de dispersión
- En el grupo control se observa un valor extremo (caso 61)
- Se deben verificar los casos extremos

Gráficos para var.cuantitativas: Gráfico de Caja

Permite comparar gráficamente el comportamiento de una variable en distintos grupos



- *En el grupo control, las mujeres presentan mayor dispersión, así como asimetría a la derecha*
- *En el grupo tratamiento, los hombres presentan mayor dispersión y cierta asimetría a la derecha*
- *Se identifican 3 casos extremos*

Gráficos para var.cuantitativas: Diagrama tallo y hojas

Anchura cintura Stem-and-Leaf Plot

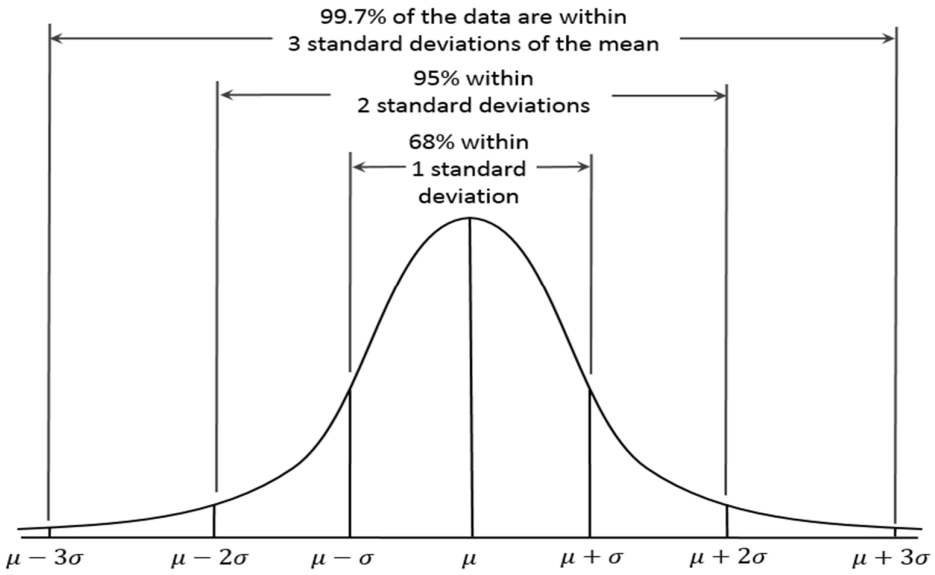
Frequency	Stem &	Leaf
8,00	6 .	01133444
15,00	6 .	555666777788889
24,00	7 .	000000011112233333344444
15,00	7 .	556666667777999
22,00	8 .	0000001111222233444444
17,00	8 .	55566666778888889
35,00	9 .	00000011111222233333333334444444444
36,00	9 .	55555555666677777778888888899999999
23,00	10 .	00000111111222222333344
24,00	10 .	55666778888888889999999
12,00	11 .	011122333344
3,00	11 .	557
6,00	12 .	001122

Stem width: 10,00
Each leaf: 1 case(s)

Analizar/ Estadísticos descriptivos/ Explorar/ Gráficos/ Gráfico de tallo y hojas

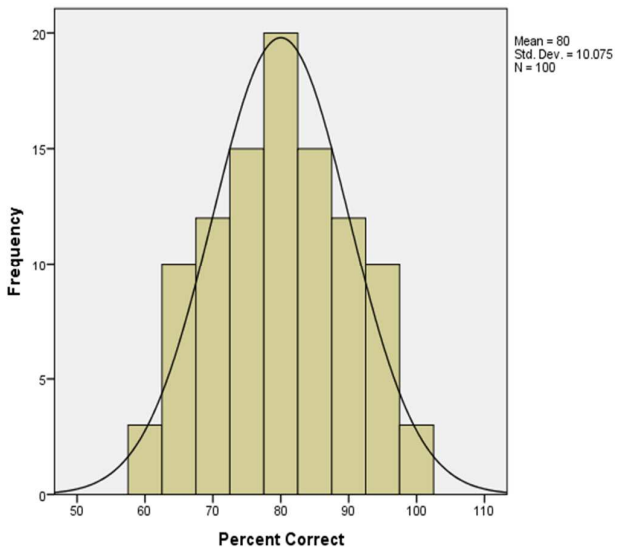
Gráficos específicos para detectar la normalidad

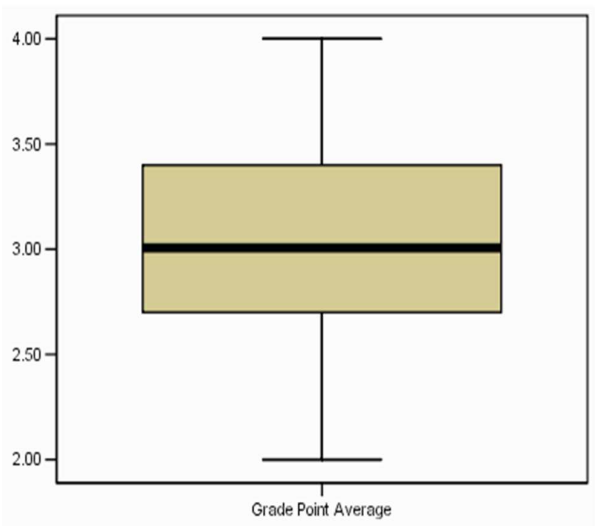
- Objetivo
  - Determinar si una variable sigue una distribución Normal
  - **Esto es una cuestión decisiva, puesto que condiciona la elección de un método de Inferencia paramétrica o no paramétrica**



Gráficos para detectar la normalidad

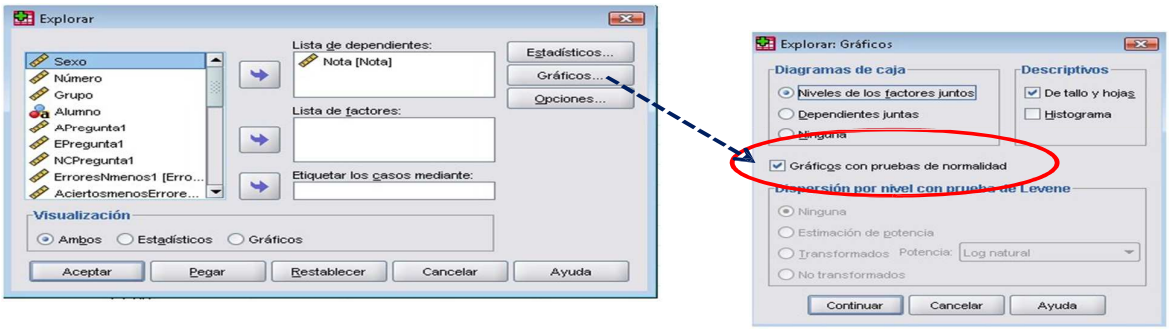
- Histograma y Gráfico de Caja
  - Con pocos datos no es fácil obtener conclusiones consistentes, de ahí que se hayan ideado gráficos concretos para observar la normalidad





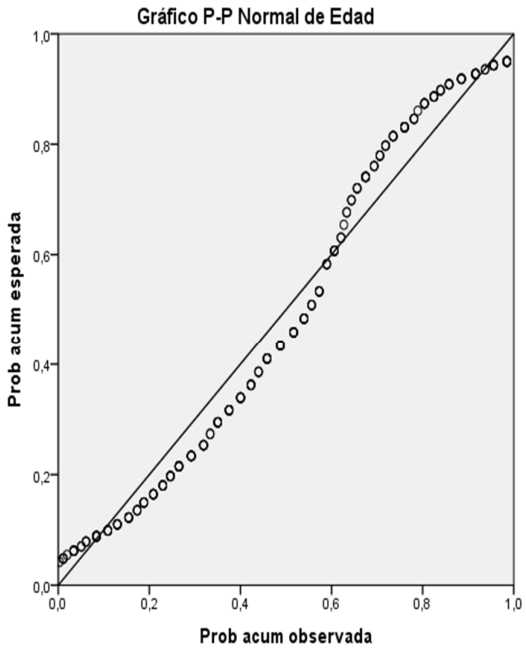
Gráficos para detectar la normalidad

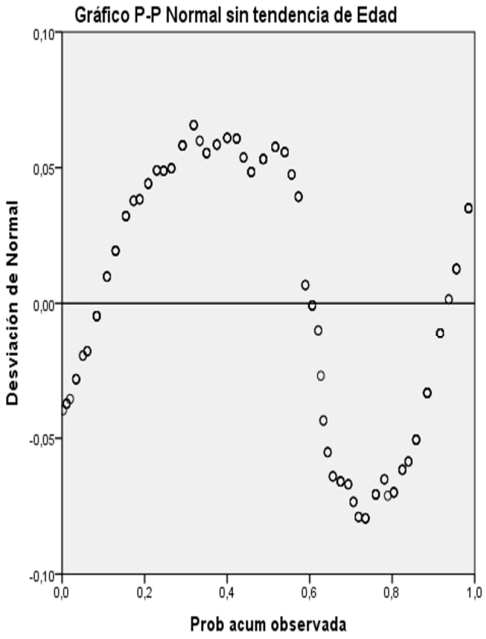
- Se basan en ordenar y estandarizar los datos
- Se representan dichos datos frente a los datos esperados de una distribución normal tipificada, N(0,1)
- SPSS ofrece 2 tipos de gráficos
  - Gráficos de probabilidad normal «P-P plots»: basados en probabilidades acumuladas
  - Gráfico de cuantiles normales «Q-Q plots»: basados en cuantiles



Analizar/ Estadísticos descriptivos/ Explorar/ Gráficos/ Gráficos con prueba de normalidad

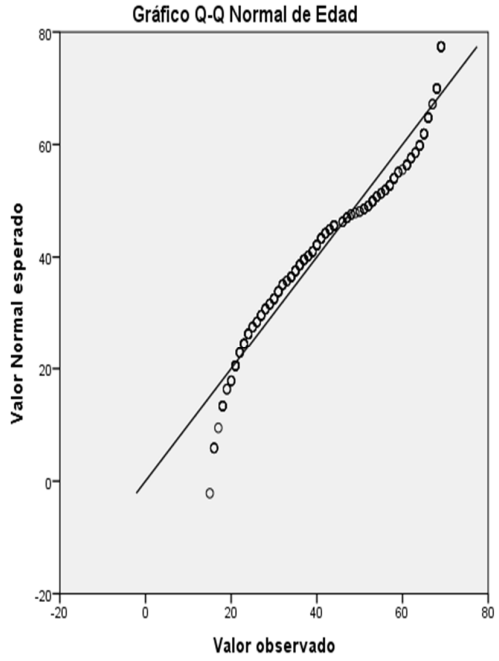
Gráficos para detectar la normalidad: P-P plots

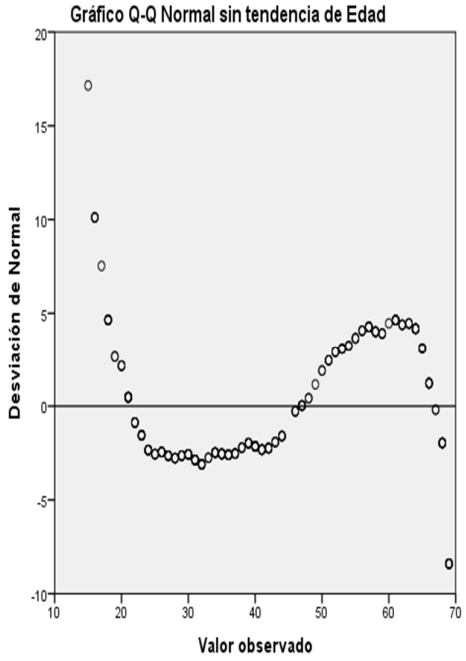




Analizar/ Estadísticos descriptivos/ Gráficos P-P

Gráficos para detectar la normalidad: Q-Q plots





Analizar/ Estadísticos descriptivos/ Gráficos Q-Q

Gráficos para detectar la normalidad

- Ventajas:
  - Sencillez de interpretación
  - Fáciles de obtener (cualquier software estadístico los implementa)
  - No requieren muestras tan numerosas como algunos tests de normalidad
- Inconvenientes:
  - **Subjetividad** de la interpretación visual
  - Un gráfico no ofrece una medida objetiva que nos permita tomar una decisión acerca de la normalidad o no de la variable
- Conclusión:
  - Los gráficos orientan sobre la procedencia o no de la muestra de una población normal. Por ello, es necesario trabajar con un test/contraste estadístico que certifique la normalidad o no de una variable.

Contraste/Test de hipótesis de normalidad

- Determina si una variable se ajusta a una Normal o no
- Coeficientes (asimetría y curtosis) y gráficos (tallo y hojas, caja, histograma, P-P, Q-Q plots), son mero apoyo a la decisión
- SPSS: 3 pruebas estadísticas que nos permiten determinar si una distribución de datos se ajusta a una normal o no:
  - **Test de Kolmogorov-Smirnov**: excesivamente conservador con la hipótesis de normalidad
  - **Test de Kolmogorov-Smirnov con la corrección de Lilliefors**: test más potente para n ≥ 50
  - **Test de Shapiro-Wilks**: es el test más recomendable para n < 50

Contraste/Test de hipótesis de normalidad

Objetivo: Determinar si una variable se ajusta a una normal

H_0: Normalidad de la variable

H_1: No Normalidad de la variable

α = nivel de significación Usualmente α = 0.05

α = P(Rechazar H_0 / H_0 cierta)

SPSS, a partir de los datos de la muestra, nos ofrece el p-valor, clave para tomar la decisión:

- **Si p-valor ≤ α se rechaza H_0**
- **Si p-valor > α no se rechaza H_0**

Contraste/Test de hipótesis de normalidad

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Anchura cintura	,082	240	,000	,977	240	,001

a. Corrección de la significación de Lilliefors

Analizar/ Estadísticos descriptivos/ Explorar/ Gráficos/ Gráficos con prueba de normalidad

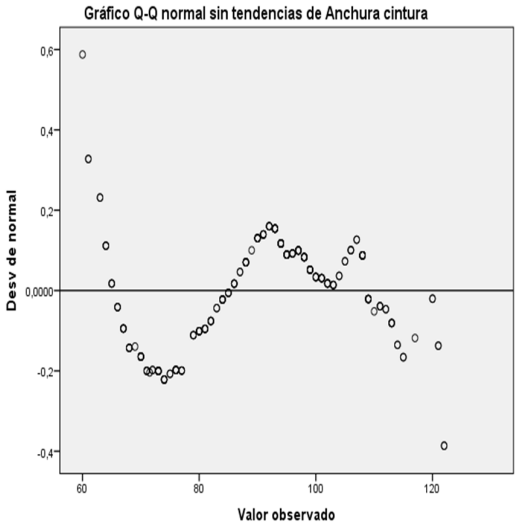
Prueba de Kolmogorov-Smirnov para una muestra

		Anchura cintura
N		240
Parámetros normales[a,b]	Media	90,5479
	Desviación típica	14,88649
Diferencias más extremas	Absoluta	,082
	Positiva	,065
	Negativa	-,082
Z de Kolmogorov-Smirnov		1,272
Sig. asintót. (bilateral)		,079

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Analizar/ Pruebas no Paramétricas/ K-S de una muestra



Inferencia paramétrica y no paramétrica

- **Inferencia Paramétrica**
  - Se conoce el modelo de distribución de la población objeto de estudio y se desconoce un número finito de parámetros de dicha distribución que hay que estimar con los datos de la muestra
  - Comparan los grupos a través de un *parámetro* de la distribución
  - Requieren conocer la distribución de la muestra para poder realizar inferencias sobre la población (por ejemplo la Normal)

- **Inferencia No Paramétrica**
  - Son métodos de distribución libre. No requieren conocer la distribución de la cual proviene la muestra
  - Se realizan con procedimientos de ordenación, rangos y recuentos
  - Se utilizan estadísticos cuya distribución se determina con independencia de cuál sea la distribución de la población

Inferencia paramétrica

- **Supuestos**
  - **Normalidad** de la variable
  - **Homocedasticidad**: La variabilidad de los diferentes grupos tienen que ser iguales
  - En análisis multivariante, hay otros supuestos adicionales, como **la normalidad e independencia** de los residuos de la regresión, o el hecho de que la ecuación refleja una **composición aditiva** de las fuentes de variación de la variable objetivo

- **Observación (basada en el TLC):**
  - En caso de *indicios de normalidad* en los datos, el supuesto de normalidad se puede flexibilizar siempre y cuando el tamaño de muestra sea ***grande*** (n ≥ 30 para algunos autores, n ≥ 60 para otros)

Transformaciones de los datos

Escalera de Transformaciones de Tukey

Técnica de AED que pretende conseguir la normalidad de la variable o al menos corregir la asimetría.

Importancia:

- Las distribuciones que muestran una clara asimetría son difíciles de estudiar.
- Los valores originales aparentemente atípicos se encontrarán más cercanos al grueso de los datos.
- Los métodos estadísticos de Inferencia paramétrica suelen emplear la **media aritmética**; pero la media de una distribución asimétrica no es un buen índice del conjunto de datos.

Transformaciones de los datos

Escalera de Transformaciones de Tukey

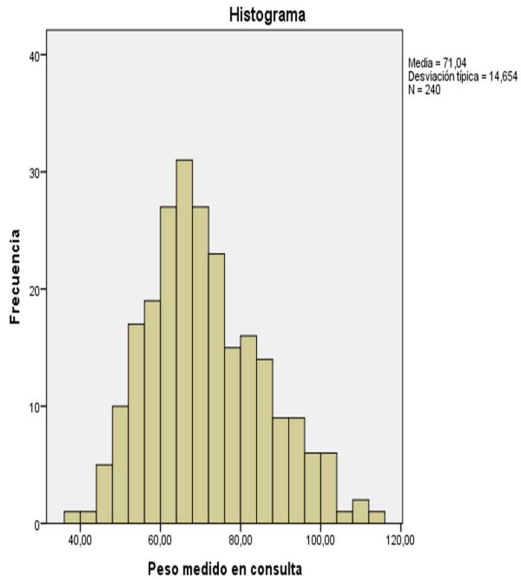
Según sea la intensidad de la asimetría o la dirección en la que van los casos extremos, se tiene:

- Corregir asimetría a la derecha ($p < 1$):
  - Suave: $X^{1/2}$, $X^{1/3}$, logaritmo
  - Fuerte: Transformaciones 1/X ó $1/X^2$
- Corregir asimetría a la izquierda ($p > 1$):
  - Suave: X^2 , X^3
  - Fuerte: Exponencial
- Observaciones:
  - Orientativa: Pueden hacer falta varias transformaciones
  - Si hay poca asimetría, puede que al transformar se tenga mayor asimetría que la original (trabajar entonces con datos originales)

Transformaciones de Tukey

Descriptivos

			Estadístico	Error típ.
Peso medido en consulta	Media		71,0396	,94593
	Intervalo de confianza para la media al 95%	Límite inferior	69,1762	
		Límite superior	72,9030	
	Media recortada al 5%		70,5731	
	Mediana		69,0000	
	Varianza		214,746	
	Desv. típ.		14,65421	
	Mínimo		39,00	
	Máximo		115,00	
	Rango		76,00	
	Amplitud intercuartil		18,38	
	Asimetría		,507	,157
	Curtosis		-,098	,313



Pruebas de normalidad

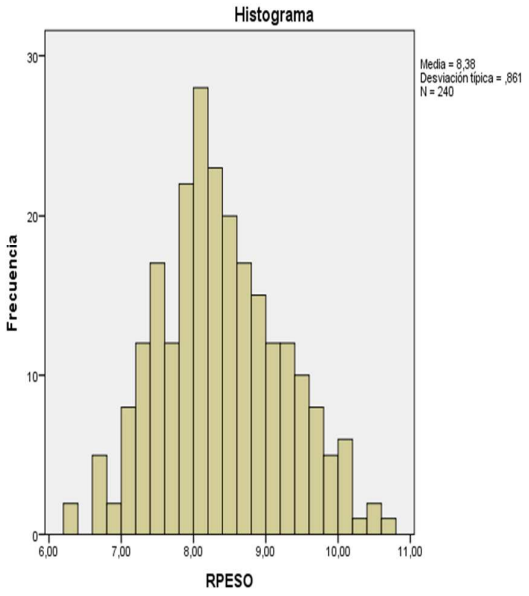
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Peso medido en consulta	,074	240	,003	,979	240	,001

a. Corrección de la significación de Lilliefors

Transformaciones de Tukey

Descriptivos

			Estadístico	Error típ.
RPESO	Media		8,3846	,05556
	Intervalo de confianza para la media al 95%	Límite inferior	8,2752	
		Límite superior	8,4941	
	Media recortada al 5%		8,3719	
	Mediana		8,3066	
	Varianza		,741	
	Desv. típ.		,86072	
	Mínimo		6,24	
	Máximo		10,72	
	Rango		4,48	
	Amplitud intercuartil		1,15	
	Asimetría		,259	,157
	Curtosis		-,274	,313



Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
RPESO	,055	240	,077	,991	240	,148

a. Corrección de la significación de Lilliefors

Transformaciones de Tukey

Propiedades:

- **Preservan el orden** de los valores: los valores mayores/menores originales seguirán siendo los valores mayores/menores en la escala transformada.

- Modifican la distancia entre los valores. Para potencias ***p*<1 se comprimen** los datos en la parte superior de la distribución en relación a los valores menores. Para ***p*>1** se tiene el **efecto contrario.**

- **El efecto sobre la forma de la distribución cambia sistemáticamente con *p***. Si raíz x hace menos pronunciada la asimetría positiva de una variable, el log x provocará que la distribución resultante sea aún menos asimétrica positiva (en relación a raíz x).

TÉCNICAS ESTADÍSTICAS APLICADAS EN NUTRICIÓN Y SALUD

Análisis descriptivo y exploratorio de datos

Francisco M. Ocaña Peinado @ocanapaco
http://www.ugr.es/local/fmocan
Departamento de Estadística e Investigación Operativa. UGR