BIOESTADISTICA

TEMA 1: NECESIDAD DEL CONOCIMIENTO DE LA ESTADISTICA POR EL PROFESIONAL DE LAS CIENCIAS DE LA SALUD.

Las ciencias de la salud es una ciencia experimental, es decir, aquella que utiliza el método inductivo (de lo particular a lo general). Este es el único método que asegura que lo que estamos realizando posee valor. El método deductivo obtiene deducciones de lo general a lo particular. Con el método deductivo puedes cometer más errores.

La estadística se hace necesaria en esta ciencia debido a:

- Variabilidad de los sujetos estudiados (ejemplo cada uno puede reaccionar de una forma a un medicamento).
- La naturaleza pide más cuantificación.

El método científico está compuesto por tres fases: diseño, recopilación de datos y análisis de los resultados. La estadística se encarga de que las tres fases sean correctas. Tenemos que tener en cuenta que el trabajo clínico requiere de un rigor científico. Y la estadística es necesaria por tanto también a nivel comunitario (ejemplo para gestionar el dinero del ministerio de la salud).

La estadística proporciona métodos para:

a) La recopilación, organización y análisis (de datos de un grupo de individuos).>>>>Estadística descriptiva: la cual se realiza sobre la muestra. Es la ciencia que se dedica a describir las regularidades o características existentes en un conjunto de datos.
b) Decidir: aceptación o rechazo de la hipótesis.>>>Estadística inferencia: trata de obtener conclusiones de la población a través del conocimiento de las características de la muestra.

La estadística va a estudiar los fenómenos que ocurren en torno a un conjunto de sujetos, denominados población (conjunto de todos los individuos que poseen alguna característica observable y en los que se desea estudiar un determinado fenómeno). Mientras que la muestra es la parte de la población sobra la cual se efectúa el estudio del fenómeno detectado en una población determinada.

El índice que mide la media, incorporada a la población, de una característica es el parámetro (μ). Por su parte la media de una muestra es la denominada media estadística(X).

LA ESTADISTICA DESCRIPTIVA:

- Recopilación >>>muestra:
  - Tipo de muestreo.
  - Tamaño de la muestra.
- Organización de los datos:

    - Cuadros de distribución de frecuencias.
    - Representación grafica.
- Análisis de los datos:
    - Medidas de tendencia central.
    - Medidas de dispersión.
    - Medidas de posición.

Modalidades: son las distintas categorías que tiene una variable (toda característica empírica que es objeto del proceso de estudio). Las variables son:

- Cualitativas: aquellas cuyos valores se presentan como cualidades o atributos. (por ejemplo: las profesiones).
- Cuasi cuantitativas: aquellas cuyos valores se presentan como cualidades que son susceptibles de ser ordenadas de menor a mayor. (por ejemplo: una película).
- Cuantitativas: aquellas cuyos valores pueden ser contados y medidos numéricamente.
    - Discreta: no existe valores intermedios entre otros dos valores consecutivos de la variable. (ejemplo el número de hijos).
    - Continua: existen infinitos valores intermedios entre dos valores de la variable.

Las modalidades tienen que ser:

- Exhaustivas: todas las posibilidades que hay en la población de una variable deben de estar introducidas en dicha modalidad.
- Excluyentes: un determinado individuo, únicamente puede ser introducido en una única modalidad.

La cantidad de sujetos que dependen en una determinada modalidad es lo denominado frecuencia absoluta. El tamaño de la muestra o cantidad de sujetos de estudio es el N. y la proporción o frecuencia relativa es pi.

Para la **<u>Representación de variables cualitativas</u>**:

- Diagramas de barras: es una representación sobre un eje de coordenadas. En el eje de las abscisas (horizontal) colocamos las variables y en el eje de las ordenadas, se representan las frecuencias. Tenemos que tener en cuenta la escala (1/1 o 1/2, el dos a favor de la abscisa).
- Diagrama de sectores o quesitos: 360º_____100% X_____Ejempl%

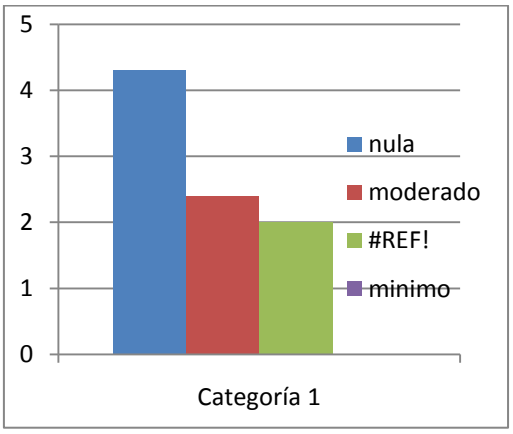
- Pictogramas.

Representación de variables cauasicuantitativas.

X Grado de escara	n_i	p_i	P_i	n_a	p_a	P_a
Mínimo	134	0’216	21’6 %	622	1	100%
Moderado	212	0’341	34’1%	488	0’784	78%
Leve	129	0’207	20’7%	276	0’443	44’3%
Nulo	147	0’236	23’6%	147	0’236	23’6%
Total	622	1	100%			

El n_a es la frecuencia acumulada. Se acumula desde la cantidad inferior a la superior. (ej.: 129+147=276)

Diagrama de barras:



A diferencia del diagrama de las variables cualitativas, las modalidades en este se encuentran unidas (pegadas), porque no hay valores intermedios.

Representación de variable cuantitativa

Un grupo de enfermeros de un CS determinado, desean conocer la distribución de las edades en que se distribuye la vacunación de la poliomielitis que ellos/as administran:

a) Para lo cual recogen una muestra de los niños y niñas que acuden al mismo, durante un mes determinado (empíricamente).
b) Para….muestra (representativa de su población y mediante muestreo aleatorio) de 36 niños y niñas de edades comprendidas entre 2 y 13 años.

Problemas: ¿entre que edades tenemos más demandas? ¿A qué grupo van dirigidos nuestros programas de vacunación sanitaria?

La puntuación que se recoge de las variables de cada individuo que se está estudiando, se denominan puntuación directa. Ordenamos las puntuaciones de mayor a menor. Cuando las muestras son grandes, los ordenamos por agrupaciones o intervalos.

Limites aparentes Xi	Limites exactos Xi	n_i	Xj	p_i	P_i	n_a	p_a	P_a
12-13	11.5-13.5	1	12.5	0.027	2.7%	36	0.997	99.7%
10-11	9.5-11.5	5	10.5	0.138	13.8%	35	0.97	97%
8-9	7.5-9.5	9	8.5	0.25	25%	30	0.832	83.2%
6-7	5.5-7.5	11	6.5	0.305	30.5%	21	0.582	58.2%
4-5	3.5-5.5	7	4.5	0.194	19.4%	10	0.277	27.7%
2-3	1.5-3.5	3	2.5	0.083	8.3%	3	0.083	8.3%
		36						

Los límites exactos les conseguimos al restar media unidad al límite aparente inferior y sumando media unidad al límite aparente superior de ese intervalo. La i es igual a la amplitud del intervalo, es decir la diferencia de unidades que existe entre el límite superior y el inferior de un intervalo.

(EXAM) en cualquier distribución todos tienen las mismas unidades de intervalos.

Nunca se trabaja con límites aparentes, porque en una variable cuantitativa continua jamás puede darse un salto. La amplitud del intervalo real es la de los límites exactos. En el ejemplo i=2. El Xj es el punto medio del intervalo que será igual al límite exacto superior más el límite exacto inferior, todo ello partido entre dos. Obtenemos la puntuación representativa de ese intervalo.

ESTADISTICA DESCRIPTIVA:

El análisis de los datos se realiza mediante:

1. MTC: medidas de tendencia central
2. MD: medidas de dispersión
3. MP: medidas de posición

1: es aquel valor de la variable n torno al cual se agrupan los datos. (Media, mediana y moda).

2: es el valor de la variable que indica la dispersión o variabilidad de la distribución. (Varianza, amplitud o rango, desviación típica y amplitud semintercuartil).

3: valor de la variable que indica una determinada posición. (Cuartiles y deciles).

MEDIDAS DE TENDENCIA CENTRAL EN DATOS NO AGRUPADOS:

a. Media:

M = $\frac{\sum x_i}{N}$ ejemplo: 2-5-7-7-8 M = $\frac{2+5+7+7+8}{5} = \frac{29}{5} = 5'8$

Se utilizará cuando la muestra cumpla unos requisitos. No la elegiremos como MTC cuando en la muestra existan puntuaciones extremas (los picos de la grafica son muy altos). Ej.: 6-6-7-7-8 X=6'8

La media es muy sensible a la variación de las puntuaciones. En el momento en el que varíe una puntuación, me varía la media.

b. Mediana: es aquel valor de la variable que deja por encima y por debajo se sí el 50% de los casos o de las observaciones. En todos los casos se ordenan las puntuaciones de menor a mayor (importante¡¡).

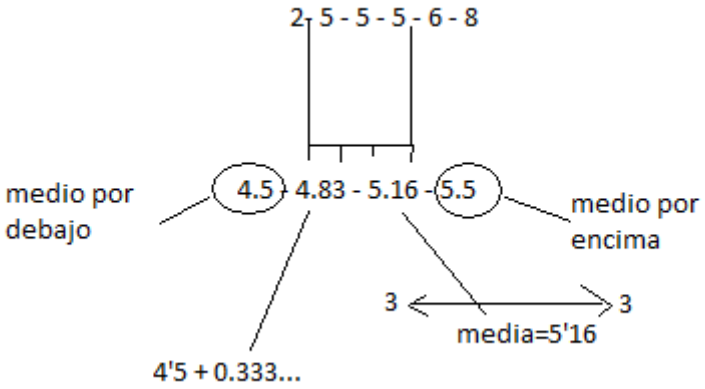
1. 1-3-7-12-15-21-29 (número de datos impares)

mediana

2. número de datos pares: 2-5-7-8-10-12

$$md = \frac{7+8}{2} = 7'5$$

1. Repetición de la puntación donde cae la mediana: si se repiten las puntuaciones donde cae la mediana no se hace esto.



1 dividido entre el numero de veces que se repite el nuemro(en este caso el numero 5 se repite 3 veces).

Lo calculamos en todos los casos donde la media no sea recomendable su calculo o donde haya puntuaciones extremas. Ante la duda entre la media y mediana, se calcula la mediana porque es menos sensible, a no ser que tenga que realizar alguna medida de dispersion y que haga falta la media. No calculamos la media cuando medimos constantes (ej. Presión).

c. Moda: es el valor de una variable que con mayor frecuencia se repite.
2-3-3-3-4-5-6-8-10-10-10 en este caso encontramos dos modas: el 3 y el 10, por lo que decimos que es bimodal. Existen también trimodales, etc. un multimodal es en aquel en el cual se repiten todos los números las mismas veces (ej. 2-3-5-6).

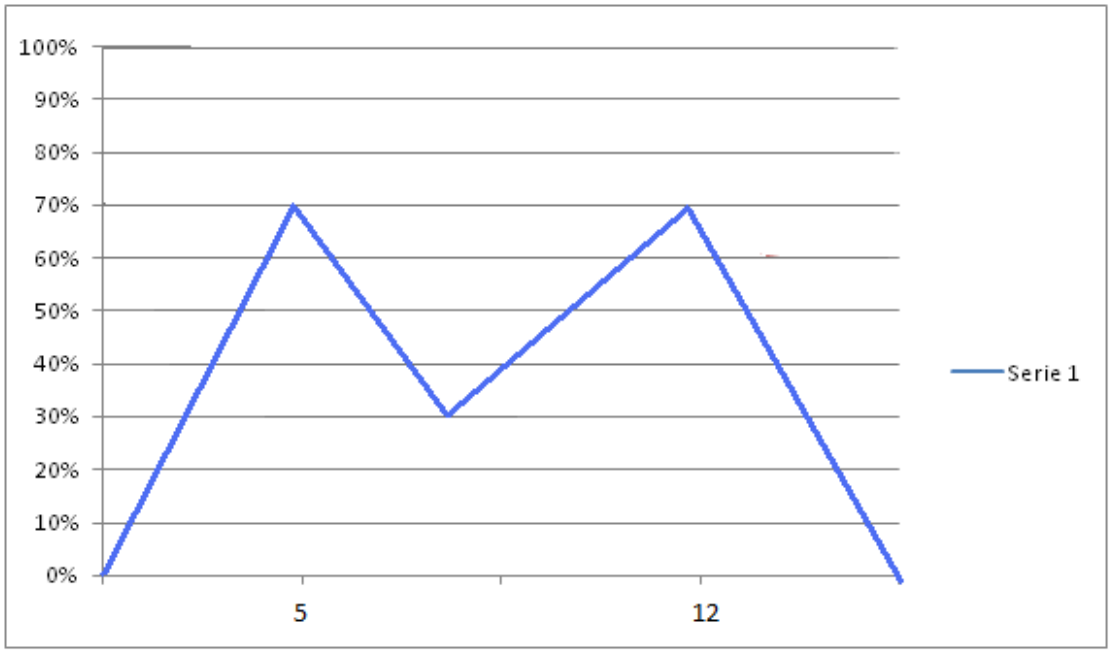
La moda es poco utilizada. Lo único que nos da es una referencia momentánea, es poco fiable.

MEDIDAS DE TENDENCIA CENTRAL EN DATOS AGRUPADOS:

1. La moda es el punto medio de mayor frecuencia.

X	n_i	X_j	$n_i X_j$
11’5-13’5	1	12’5	12’5
9’5-11’5	5	10’5	52’5
7’5-9’5	9	8’5	76’5
5’5-7’5	11	6’5	71’5
3’5-5’5	7	4’5	31’5
1’5-3’5	3	2’5	7’5
	36		252

En este caso la moda es igual a 6’5, ya que la mayor frecuencia es 11.



En este otro caso la moda será 5 y 12 ya que en ambas se repite la frecuencia. La moda se emplea cuando se quiere calcular de manera rápida la medida de tendencia central.

2. Media aritmética:

$$X = \frac{\sum n_i x_j}{N} = \frac{\sum n_{i1} x_{j1} + n_{i2} x_{j2} + \cdots + n_{in} x_{jn}}{N};$$

X	n_i	X_j	$n_i X_j$
11’5-13’5	1	12’5	12’5
9’5-11’5	5	10’5	52’5
7’5-9’5	9	8’5	76’5
5’5-7’5	11	6’5	71’5
3’5-5’5	7	4’5	31’5
1’5-3’5	3	2’5	7’5
	36		252

En este caso: X = $\frac{252}{36} = 7\ años$

3. Mediana: aquella puntuación de la variable que deja el 50% por arriba y el otro 50% por debajo.

X	n_i	X_j	$n_i X_j$
11'5-13'5	1	12'5	12'5
9'5-11'5	5	10'5	52'5
7'5-9'5	9	8'5	76'5
5'5-7'5	11	6'5	71'5
3'5-5'5	7	4'5	31'5
1'5-3'5	3	2'5	7'5
	36		252

$$Md = Li + \left(\frac{N/_2 - n_d}{n_c}\right) \cdot i$$

Li = limite exacto inferior del intervalo crítico.

n_d = Número de frecuencias por debajo del intervalo crítico.

n_c = Número de frecuencias en el intervalo crítico.

1º. Calcular $\frac{N}{2}$ es la media de los sujetos.

2º. Buscar donde está el sujeto de la media en la n_a.

El intervalo crítico es aquel que contiene una determinada puntuación.

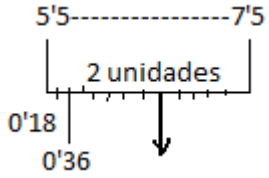
3º. Aplicar la formula:

Li = 5'5 N/2 = 36/2 = 18 (50% de los casos)

n_d = 10 $i = 7'5 - 5'5 = 2$ (amplitud)

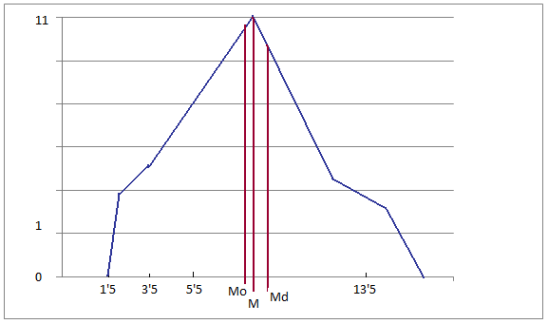
n_c = 11

$$Md = 5'5 + \left(\frac{18-10}{11}\right) \cdot 2; \quad Md = 6'95 \text{ años}$$



Dividimos las dos unidades entre el número de intervalos, en este caso 11. Desde el 10 hasta el sujeto 18 (que es el que buscamos), faltan 8 individuos, por eso de las 11 divisiones cogemos la octava. Dicha puntuación la sumamos al 5'5 y obtenemos la media. Se puede hacer sumando 0'18 + 0'18… hasta llegar al octavo. O como nos hace falta el octavo hacemos $0'18 \cdot 8$

tiende a coincidir media, mediana y moda.
Mo= 6'5
M=7
Md=6'95



Por lo que concluimos que es una distribucion simétrica. El estadistico mas representativo en estos casos es la media. Cuando tenemos una distribucion en la que M, Mo y Md no tienden a coincidir, le medida mas representativa será la mediana.

MEDIDAS DE DISPERSION: nos indican la variabilidad de los datos de la variables que estamos estudiando.

Medidas de tendencia central	Medidas de dispersion
Media: X	Varianza: S^2_X; desviacion tipica: S_x
Mediana: Md	Amplitud semiintercuartil: Q
Moda: Mo	Aplitud o rango: A

A. Desviacion tipica/ varianza: estudia cual es la cuantia de cada puntuacion con relacion a la media. Y en el caso de datos agrupados estudia cual es la cuantia de cada punto medio con relación a la distribución.

$$s_x^2 \rightarrow \sqrt{s_x^2} = s_x\,; \qquad s_x^2 = \frac{\sum(x_j - X)}{N}\,;$$

x_i	n_i	x_j	$n_i \cdot x_j$	$(x_j - X)$	$(x_j - X)^2$	$n_i \cdot (x_j - X)^2$
1'5-3'5	3	2'5	7'5	-4'5	20'25	60'75
3'5-5'5	7	4'5	31'5	-2'5	6'25	43'75
5'5-7'5	11	6'5	71'5	-0'5	0'25	2'75
7'5-9'5	9	8'5	76'5	1'5	2'25	20'25
9'5-11'5	5	10'5	52'5	3'5	12'25	61'25
11'5-13'5	1	12'5	12'5	5'5	30'25	30'25
	36					219

1º. Calcular la media:

$$X = \frac{\sum n_i x_j}{N};\ X = \frac{252}{36};$$ X = 7 años.

2º. Aplicar la formula:

$$s_x^2 = \frac{219}{36}\,; s_x^2 = 6'08;\ s_x^2 = 6'08\ años;\ \sqrt{s_x^2} = s_x\,;$$ S_x= 2'47 años.

B. Amplitud semiintercuartil:
Los caurtiles son una medida de posicion e indican un valor de la variable que deja un determinado porcentaje por debajo de la distribucion. El Q_1 es aquella puntuacion de la variable que deja por debajo el 25% y por arriba el 75%. El Q2 deja tanto por arriba como por abajo el 50%, tendra el mismo valor que la mediana (EXAM). Y el Q3 deja por debajo el 75% y por arriba el 25% de los casos.

$$Q_1 = L_i + \left(\frac{\frac{25 \cdot N}{100} - n_d}{nc}\right) \cdot i \; ; \qquad Q = \frac{Q_3 - Q_1}{2};$$

1º. Aplicar la regla de tres:

N_____1000 36_____100 $X = \frac{36 \cdot 25}{100} = 9$

X_____25 X_____25

2º. Tenemos que buscar donde esta el sujeto 9, para poder definir el intervalo critico de Q_1.

X	n_i	n_a
11'5-13'5	1	36
9'5-11'5	5	35
7'5-9'5	9	30
5'5-7'5	11	21
3'5-5'5	7	10
1'5-3'5	3	3
	36	

$$Q_1 = 3.5 + \left(\frac{9 - 3}{7}\right) \cdot 2; \quad Q_1 = 5'21 a\tilde{n}os.$$

$$Q_1 = 7'5 + \left(\frac{\frac{75 \cdot 36}{100} - 21}{9}\right) \cdot 2; \; Q_1 = 8'83 a\tilde{n}os.$$

En una distribucion simetrica la distancia que hay entre Q1 y Q2 va a ser casi lo mismo que la de Q2 y Q3. En una asimetrica no ocurre lo mismo.

*Coeficiente de variacion (CD):

$$cv = \frac{s_x}{X}$$

Se utiliza para comparar la despersion de dos muestras:

a) Cuando se desea compara la dispersion de una misma muestra pero medida en unidades distintas (dos variables distintas).
b) Cuando queremos comparar la variabilidad de una misma variable en grupos de distinta poblacion.

REPASO:

- X = variable
- Xi = puntuaciones directas de la variables estudiada, que tomamos directamente de los sujetos.

- Ni = frecuencias absolutas de una puntuacion o puntuaciones (intervalos) determinadas.
- N = tamaño de la muestra = $\sum n_i$
- Xj = punto medio del lintervalo.

$(x_i - X)\ o\ (x_j - X)$ O puntuaciones diferenciales , que podrá ser positiva o negativa según este por encima o per debajo de la media.

TEMA 3. CURVA NORMAL

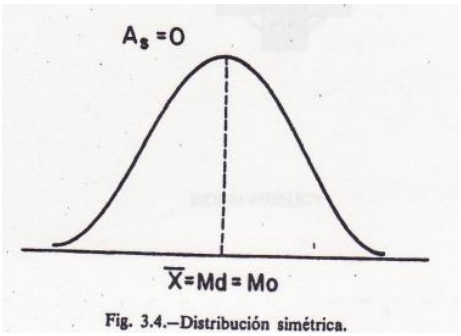


Fig. 3.4.—Distribución simétrica.

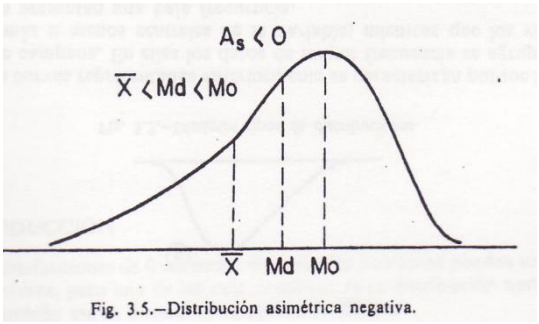


Fig. 3.5.—Distribución asimétrica negativa.

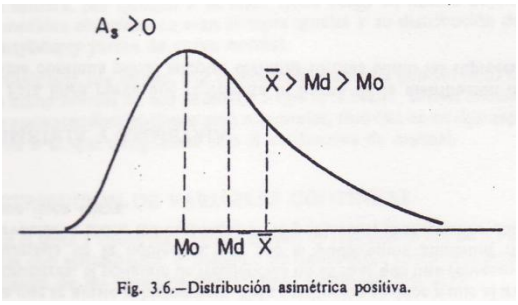


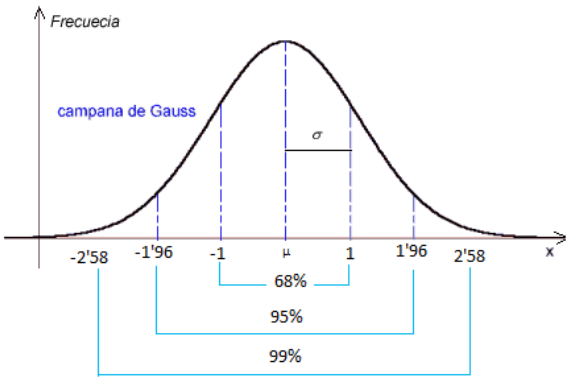
Fig. 3.6.—Distribución asimétrica positiva.

TIPOS DE DISTRIBUCION

En las distribuciones asimetricas:
A. positivas: X>Md>Mo
B. negativas: X<Md<Mo

En las distribuciones simetricas:
x = Md = Mo

Si son asimetricas el indice de tendencia central mas significativo, es la mediana.



Curva normal, campana de Gaus:

(libro pag 101) X= μ: media parámetro. Cuando se habla de una poblacion. Deja el 50% por encima y el otro 50% por debajo.

Desviacion tpica= σ

La curva normal es asíntota (que nunca se junta con el eje de abcisas) porque no es posible medidir el total de la poblacion.

Entre ± 1 desviacion se encuentra el 68% de los sujetos.

Entre $\pm 1'96$ desviacion se encuentra el 95% de los sujetos.

Entre $\pm 2'58$ desviacion se encuentra el 99% de los sujetos.

TIPIFICACION: es el calculo de las puntuaciones tipicas, Z, y su area en la curva normal. Z=puntuacion de desviacion. Se halla aplicando la siguiente fórmula:

$$datos\ agrupados: z = \frac{X_j - X}{S_x}$$

$$datos\ \ no\ agrupados: z = \frac{X_i - X}{S_x}$$

Cada puntuacion directa se puede tipificar.

X	n_i	x_j	$(x_j - X)$	Z	p_i	P_i
1'5-3'5	3	2'5	-4'5	-1'82	0'0344	3'44%
3'5-5'5	7	4'5	-2'5	-1'01	0'1562	15'62%
5'5-7'5	11	6'5	-0'5	-0'2	0'4207	42'07%
7'5-9'5	9	8'5	1'5		0'7257	72'57%
9'5-11'5	5	10'5	3'5		0'9222	92'22%
11'5-13'5	1	12'5	5'5		0'9827	98'27%
	36					

$$z_{-1'9} \rightarrow' 0287 = 0'0287 = 2'87\%$$

$$z_{-1'96} \rightarrow' 0250 = 0'0250 = 2'5\%$$

*para calcular la probabilidad por debajo de 4'5 años:

1. Tipificar primero 4'5 para calcular z y mirar en la tabla la probabilidad.

*y por encima?¿ 1 menos la probabilidad, ejemplo: 1-0'1562=

*¿Qué probabilidad hay entre 4'5 y 8'5?

0'7257-0'1562=0'5615

Nunca se realiza la probabilidad del primer intervalo, en este caso se tipificaria solo 4'5 y esa es la probabilidad. Tampoco se tipifica el limite exacto superior (al igual que el limite exacto inferior) , se le resta a 1 la probabilidad del limite del intervalo.

Ejercicio: datos: media= 7 años; desviacion=2'47 años; n=36

1. ¿Qué tanto por ciento de niños representan la mayor frecuencia de la variable, es decir se encuentra comprendido en el intervalo 7’5-9’5?
2. ¿Qué tanto por ciento de niños estan por debajo de los 6 años?
3. ¿Qué tanto por ciento de los niños estan por encima de los 9 años?
4. ¿Qué tanto por ciento de sujetos estan entre 6 y 9 años?

*en la pregunta 2 se tipificaria la puntuacion 6 y comprobariamos en la tabla la probabilidad.

$$z = \frac{6-7}{2'47} = -0'406 = -0'41$$

Z →0’3409 = 34’09%

Conocida una determinada proporcion, ¿a qué puntuacion pertenece? ((despejarlo de la formula))

*Problemas:

entre que puntuaciones se encuentra un determinado porcentaje?¡ (siempre va a ser el porcentaje central).

1: ¿entre que puntuaciones se encuentra el 95% de la poblacion?

2’5%-------0’025 Z=-1’96

97’5%------0’975 Z=+1’96

$Z = \frac{X_i - X}{S_x}$; $S_x = 2'47 años; Z = \pm 1'96\ años; N = 36; X = 7\ años.$

$X_i = X \pm (Z \cdot S_x);\ \ X_i = 7 \pm (4'84);\ \ X_i = 11'84\ años\ y\ 2'16\ años.$

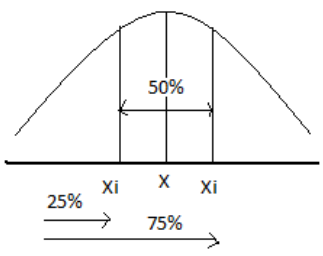
*conclusión: el 95% de la población de la distribucion se encuentra entre los 11’84 años y los 2’16.

2: entre que edades se encuentra el 50% de la población?

25%-------0’25 Z=-0’67

75%------0’75 Z=+0’67 se coge el valor más próximo (0’2514) a 0’25, ya que este ultimo no existe.

$Z = \frac{X_i - X}{S_x}; X_i = X \pm (Z \cdot S_x);\ X_i = 7 \pm (0'67 \cdot 2'47);\ X_i$ =8’65 años y 5’35 años.



Ejemplo: N=36, X=7 años, Sx=2’47 años, %= ¿ de 7 y 9 años.

$Z = \frac{X_i - X}{S_x};\ Z = \frac{7-7}{2'47}; z = 0\ \rightarrow 0'5 = 50\%.$

$Z = \frac{X_i - X}{S_x},\ Z = \frac{9-7}{2'47}$; z= 0’81→0’7910=79’1%

79’1-50=29’1%

Ejempl: entre el ± 1 se encuentra el 68%, entre ± 2 =?

$Z_{-2} = 0'0228$

$Z_2 = 0'9772$

%= 95’44.

ESTADISTICA INFERENCIAL:

1. Estudio de parametros.
2. Verificacion de hipostesis.

1.ESTIMACION DE PARAMETROS:

Se estiman dos valores , uno por arriba y otro por debajo, a los cuales vamos a llamar intervalos de confianza.

El error muestral es la diferencia que pueda existir entre el valor de la media de la muestra y el valor de la media de la poblacion.

El error tipico de la media es la desviacion tipica de la distribucion muestral de ese estadistico.

Nivel de confianza: es aquello que utilizamos para afirmar una determinada hipótesis, verificacion, etc. si la muestra es de mas de 30 individuos usamos un nivel de confianza del 95%, si la muestra es de menos de 30 individuos usamos un nivel de confianza del 99%. El nivel de significacion es lo que le resta al nivel de confianza, es decir, si el nivel de confianza es del 99% el nivel de significacion es del 1%.

A.Procedimiento:

1. Fijar el nivel de confianza: normalmente NC=95% o NC 99%.
2. Calcular la Z (tablas) correspondiente a NC: NC=95% ___ Z=$\pm 1'96$; NC=99%_____z=$\pm 2'58$
3. Calcular el error tipico de la media: Sx=$\frac{S_x}{\sqrt{N-1}}$;
4. Calcular el error muestral: E=Z· S_x
5. Construir el intervalo de confianza: IC=X$\pm E = X \pm Z \cdot S_x$

Ejemplo: N=36, X=7 años.,

$Sx = \frac{S_x}{\sqrt{N-1}}$; $Sx = \frac{2'47}{\sqrt{35}}$=0’417.

E=1’96X0’417; E= 0’817.

IC=X$\pm E$; IC= 7$\pm$0’817; IC= 7’82_____6’18

E= 2’58X0’417; E= 1’075; IC= 7$\pm 1'075$; IC=8’075 y 5’92

B.muestras pequeñas (N menor de 30):

1. Fijar el nivel de confianza: normalmente NC=95% o NC 99%.
2. Calcular los grados de libertad: gl=N-1 (aquellos elementos que pueden variar libremente; 5+7+3=15, escoges dos numeros libremente pero uno de ellos depende de los dos anteriores. N=3 por lo que el grado de libertad seria igual a N-1=2).
3. Calcular el valor de t (tablas) dependiendo del gl y NC.
4. Calcular el error tipico de la muestra: Sx=$\frac{S_x}{\sqrt{N-1}}$;
5. Determinar el error muestral: E=t· S_x
6. Construir el intervalo de confianza: IC=X$\pm E = X \pm t \cdot S_x$

*al ser la muestra pequeña la curva se estrecha: leptokurtica

Ejemplo: N=22, X=7 años; Sx=2’47 años.

Gl=N-1. gl= 21.

t= 99%=2’831. Y 95%= 2’080.

$Sx = \frac{S_x}{\sqrt{N-1}}$; $Sx = \frac{2'47}{\sqrt{21}}$=0’598.

E=2’080X0’59; E= 1’22. E= 2’831X0’59; E= 1’67;

IC=X$\pm E$; IC= 7$\pm$1’12; IC=8’22_____5’78 y IC= 8’67_____5’33

TEMA 8: CORRELACION

- Variable independiente: se investiga en el ambito biopsicosocial para determinar su influencia sobre las mediaciones individuales obtenidas, en otra variable, en el mismo grupo de individuos.
- Variable dependiente: varaibles que se suponen influidas por la independiente, en medida que va cambiando. Este estudio corresponde a la estadistica y el establecer si existe o no dependencia entre unas y otras variables.

Los índices de correlacion son índices estadisticos que estudian la posible variacion conjunta de dos variables.

Relacion entre dos variables cuantitativas:

- La correlacion trata de estudiar los problemas referentes a la variacion conjunta de dos variables cuantitativas.
- Coeficiente de correlacion de Pearson: es un índice estadistico que nos permite describir la intensidad y el sentido de la correlacion entre dos variables cuantitativas (X é Y).
    - Se representa por $r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} \cdot \sqrt{N\sum X^2 - (\sum Y)^2}}$;
    - Los valores de r_{xy} se encuentran entre -1<r_{xy}<+1.
    - Interpretacion:
        a. r_{xy}=0 (ausencia de correlacion entre variables X é Y)
        b. r_{xy}=+1 (correlacion perfecta positiva) cuando una variable sube la otra hace lo mismo.
        c. r_{xy}=-1 (correlacion perfecta negativa), cuando una sube la otra hace lo contrario.
- El coeficiente CP indica covariacion, NO RELACION CASUAL entre X é Y. (EXAM)

X	Y	XY	X^2	Y^2
1	278	278	1	77284
2	260	520	4	67600
3	198	594	9	39204
4	160	640	16	25600
5	154	770	25	23716
$\sum = 15$	1050	2802	55	233404

X= numero de pastillas

Y= TA.

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} \cdot \sqrt{N\sum X^2 - (\sum Y)^2}}$$

$N = 5$ $\sum Y = 1050$

$\sum X = 15$ $(\sum Y)^2 = 1050^2 = 1102500$

$\sum X^2 = 55$ $\sum Y^2 = 233404$

$(\sum X)^2 = 225$

- Aplicando la formula: $r_{xy} = -0'968$
- Interpretacion: $r_{xy} = -0'968$ indica una relacion muy elevada entre las variables X é Y, pero al ser de signo negativo quiere indicar que cuando aumeneta una, disminuye la otra (cuando aumenta X, disminuye Y).

Relacion entre dos variables cualitativas:

- Coeficiente de correlacion Ø (phi): se usa para estudiar la relacion entre dos variables cualitativas.

- V. dicotómicas: aquellas que por su propia naturaleza solo pueden manifestarse según dos modalidades (ejemplo el sexo).
    - Si no se presentan las variables dicotomizadas, las dicotomizamos. Ej: altura: bajo y alto. Expresando siempre a que se refiere cada cosa.
    - Se entiende por variables dicotomizadas aquellas que por su propia naturaleza pueden manifestarse según muchas modalidades pero que, para un determinado estudio, se les permite manifestarse según dos modalidades.
    - Los datos los ordenamos en las tablas de contigencia:

		X		
		0	1	
	1	a	b	(a+b)
Y	0	c	d	(c+d)
.		(a+c)	(b+d)	N

Para hallar la relacion de dos variables docotomizadas se emplea el "coeficiente de correlacion Ø".

$\emptyset = \frac{cb-ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$;

$ejemplo$:

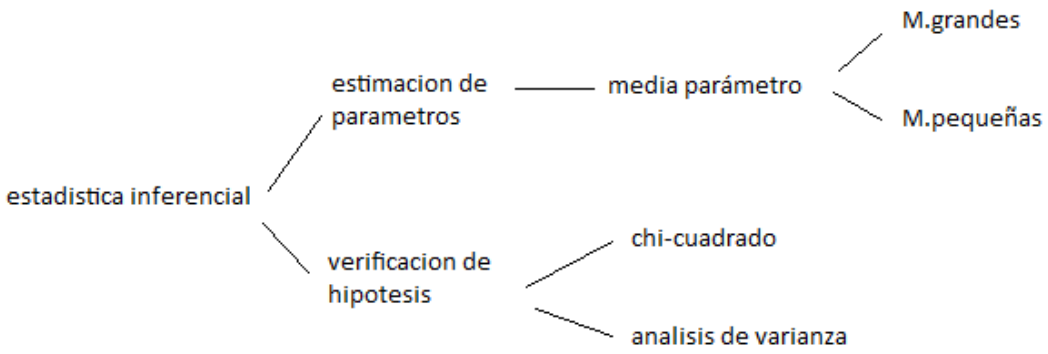
		X		
		No hepat.	Si hepatitis.	
	Heroína	30	60	(90)
Y	No heroi.	70	40	(110)
.		(100)	(100)	200

Aplicamos la fórmula:

$\emptyset = \frac{60 \cdot 70 - 30 \cdot 40}{\sqrt{90 \cdot 100 \cdot 110 \cdot 100}} = 0'3$; Ø en este caso es positivo, esto quiere decir que la relacion entre la heroína y la hepatitis es positiva, es decir, 1-1 y 0-0. Como r_{xy} los valores de Ø van de -1<Ø<+1.

1. Coeficientes de correlacion biseral puntual: r_{bp}
    - Mide la covariacion entre una variable cuantitativa (X) y una variable dicotómica, Y (1-0).
    - $r_{bp} = \frac{X_1 - X_0}{S_x} \cdot \sqrt{p \cdot q}$:
        - $X_1 =$ media de las puntuacion X que corresponden a 1 en la variable Y.
        - X_o: media de las puntuaciones X que corresponden a 0 en la variable Y.
        - S_x: desviacion típica de la variable X.
        - p: proporción de sujetos con 1 en Y.
        - q: proporción de sujetos con O en Y.
    - r_{bp}: los valores van desde -1<Ø<+1
    - $ejemplo$:
      X: variable cuantitativa= nivel de ansiedad (0 a 10)

Y: Variable dicotomizada: sexo= hombre y mujer (1 y 0)



$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$; fe: es la que obtenemos ft : lo que debería ser.

Estudia si existe diferencia entre lo empírico (lo que estudias) y lo teórico (lo que debería ser). Se utiliza en la verificación de hipótesis.

a. Variable cuantitativa: BONDAD DE AJUSTE (si la muestra se ajusta a la población).
b. Variable cualitativa: PRUEBA DE INDEPENDENCIA (posible dependencia o independencia entre dos variables cualitativas).

A.VEIFICACION DE HIPOTESIS: BONDAD DE AJUSTE

1. Establecer la Ho (hipótesis nula): "no hay diferencias entre la distribución de la muestra y la distribución de la población".
2. Fijar el NC.
3. Calcular $\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$
4. Calcular los grados de libertad: gl=h-1 (h: número de veces que se realiza el sumatorio).
5. Obtener $\chi 2$ en tablas con grado de libertad hallados y NC fijados.
6. Comparar $\chi 2$ obtenida con $\chi 2$ de las tablas:
   a. Si $\chi 2$ obtenida es < que $\chi 2$ de las tablas, aceptamos Ho (no hay diferencia entre la muestra y la población).
   b. Si $\chi 2$ obtenida es > que $\chi 2$ de las tablas, rechazamos Ho (hay diferencia entre la muestra y la población).

Ejemplo:
Unos DE desean estudiar la posible relación entre el numero de ingresos en planta y los días de la semana. Datos:

Dia	Lunes	Martes	Miercoles	Jueves	Viernes
Ingresos	7	5	1	0	2

1. Establecer la Ho (hipótesis nula): "no hay relación entre numero de ingresos que se producen con un determinado día de la semana".
2. Fijar el NC en 95%.
3. Calcular $\chi 2 = \frac{(7-3)^2}{3} + \frac{(5-3)^2}{3} + \frac{(1-3)^2}{3} + \frac{(0-3)^2}{3} + \frac{(2-3)^2}{3}$; $\chi 2 = 11'3$

4. Calcular los grados de libertad: gl=h-1 (h: número de veces que se realiza el sumatorio). **gl=5-1=4.**
5. Obtener $\chi 2$ en tablas con grado de libertad halados y NC fijados. **95%=9'488 y al 99%=13'277.**
6. Comparar $\chi 2$ obtenida con $\chi 2$ de las tablas:
    a. Al 95% la $\chi 2$ obtenida es igual a 11'3>9'488 por lo que rechazamos la Ho
    b. Al 99%: la $\chi 2$ obtenida es igual a 11'3<13'277 por lo que aceptamos la Ho
7. Conclusión: como $\chi 2$ calculada 11'3 es mayor que $\chi 2$ obtenida en las tablas 9'49, rechazamos la Ho a un NC del 95% y 4 grados de libertad, y concluimos que si existe relación entre que haya un determinado número de ingresos de pacientes en planta y el día de la semana en el que estamos.

B.PRUEBA DE INDEPENDENCIA (EXAM)

1. Establecer la Ho (hipótesis nula): "no existen diferencias entre los resultados obtenidos y los resultados teóricos, es decir, las dos variables estudiadas son independientes". "no existen diferencias significativas entre las fe y las ft", "no existe relación entre las dos variables estudiadas, luego ambas son independientes".
2. Fijar el NC.
3. Calcular $\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$
4. Calcular los grados de libertad: gl= (f-1)(c-1) entablas de 2x2 siempre será 1.
5. Obtener $\chi 2$ en tablas con grado de libertad hallados y NC fijados.
6. Comparar $\chi 2$ obtenida con $\chi 2$ de las tablas:
    c. Si χ^2 obtenida es $\leq$ que χ^2 de las tablas, aceptamos Ho (no hay diferencia entre la muestra y la población).
    d. Si χ^2 obtenida es > que χ^2 de las tablas, rechazamos Ho (hay diferencia entre la muestra y la población).

Ejemplo:

1. Queremos estudiar la posible relación entre el hábito de fumar y la aparición del infarto de miocardio. Para ello tenemos una muestra y un seguimiento de 1000 sujetos, obtenida en la siguiente tabla:
2.

		X		
		infartos	No infartos	((FILAS))
	fumardores	240	360	600
Y	No fumad.	100	300	400
((COLUMN))		340	660	1000

3. Calculamos f_t:

$casilla\ a$: $\frac{600x340}{1000} = 204$

$casilla\ b$: $\frac{600x600}{1000}$ =396

$casilla\ c$: $\frac{400x346}{1000} = 136$

$casilla\ d$: $\frac{400x660}{1000} = 264$

$f_e = \frac{total\ filas\ x\ total\ columnas}{total\ global}$

4. Fijar el NC: 95%
5. Calcular $\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$;
   $\chi^2 = \frac{(240-204)^2}{204} + \frac{(360-396)^2}{396} + \frac{(100-136)^2}{136} + \frac{(300-264)^2}{264}$=24'06
6. Calcular los grados de libertad: gl= (f-1)·(c-1).
   Gl= (2-1)(2-1)=1
7. Obtener χ^2 en tablas con grado de libertad hallados y NC fijados. $\chi^2 = 3'841$
8. Comparar χ^2 obtenida con χ2 de las tablas:
   Como χ^2 obtenida es mayor que χ^2 de las tablas, rechazamos la hipótesis nula y decimos que si que existe relación entre fumar y el infarto de miocardio.
   En este caso como la diferencia entre la obtenida y la de las tablas es tanta, se puede rechazar la hipótesis nula en todos los NC.

Prueba χ^2: bondad de ajuste en datos agrupados.

X_i	$f_e = n$	f_t
13'5-16'5	**9**	
10'5-13'5	**9**	
7'5-10'5	**14**	
4'7-7'5	**10**	
1'5-4'5	**8**	

$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$;

Para hallar las ft: (EXAM)

1. $Z = \frac{X_i - X}{S_x}$
2. Encontrar la z en las tabas: hallar las probabilidades.
3. $p_i = \frac{n_i}{N}$; $n = p_i \cdot N$;

X_i	$f_e = n$	f_t
13'5-16'5	**9**	**6'57**
10'5-13'5	**9**	**11'4**
7'5-10'5	**14**	**14'62**
4'7-7'5	**10**	**11'16**
1'5-4'5	**8**	**6'26**

$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$; $\chi^2 = \frac{(8-6'26)^2}{6'2} + \frac{(10-11'16)^2}{11'16} + \frac{(14-14'62)^2}{14'62} + \frac{(9-11'4)^2}{11'4} + \frac{(9-6'57)^2}{6'57}$=2’035.

4. Calculo de gl: gl=h·m-1, siendo h el número de veces que se realiza el sumatorio; y m el numero de estadísticos calculados, en este caso dos (la media y la desviación típica). Gl=5-2-1=2.

   *la Z no es un estadístico, es una puntuación típica o de desviación (EXAM)