BIOESTADÍSTICA

CLASIFICACIÓN DE VARIABLES

TIPOS DE VARIABLES

- Cualitativas
- Cuantitativas
  - Discretas
  - Continuas

ESCALAS DE MEDIDA

- Nominal
- Ordinal
- Intervalo
- Razón

GRÁFICOS PARA VARIABLES CUALITATIVAS

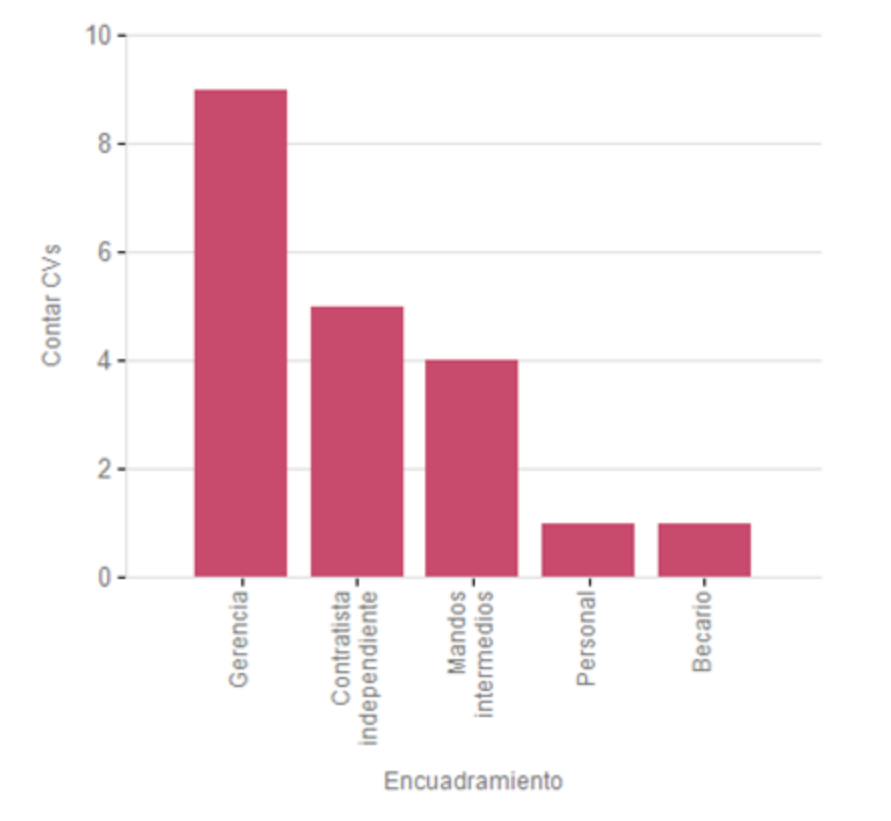
GRÁFICOS DE BARRAS SIMPLES Y AGRUPADAS

Los gráficos más frecuentes para representar **variables cualitativas** son los diagramas de barras.

Se representa en el eje de ordenadas las modalidades y en abscisas las frecuencias absolutas o las frecuencias relativas.

Si se intentan comparar varias poblaciones entre sí, usando el diagrama, existen otras modalidades como las **barras agrupadas**

GRÁFICOS DE BARRAS SIMPLES Y AGRUPADAS



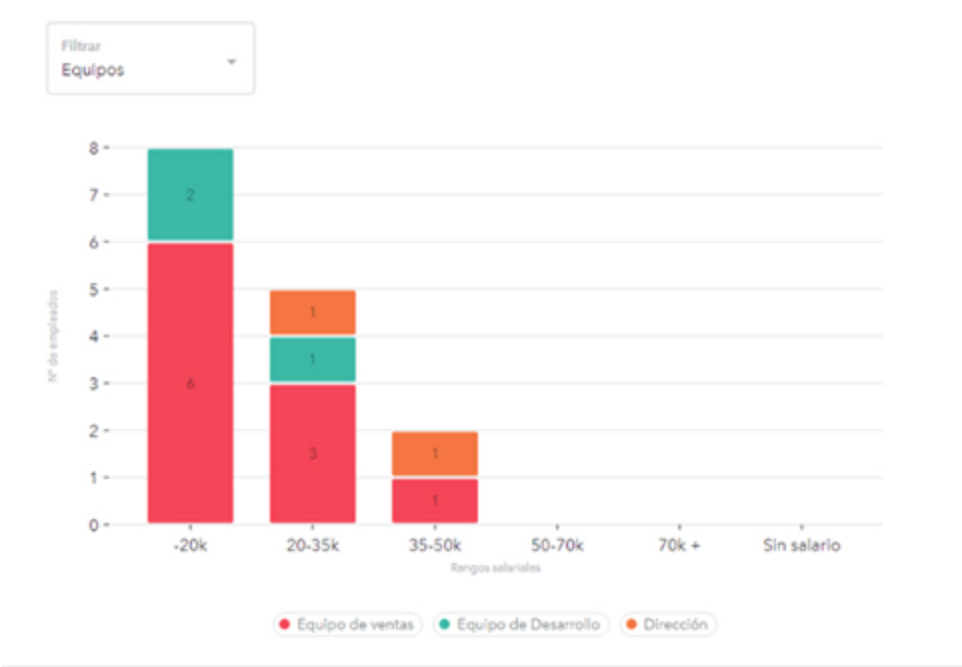


GRÁFICO DE SECTORES

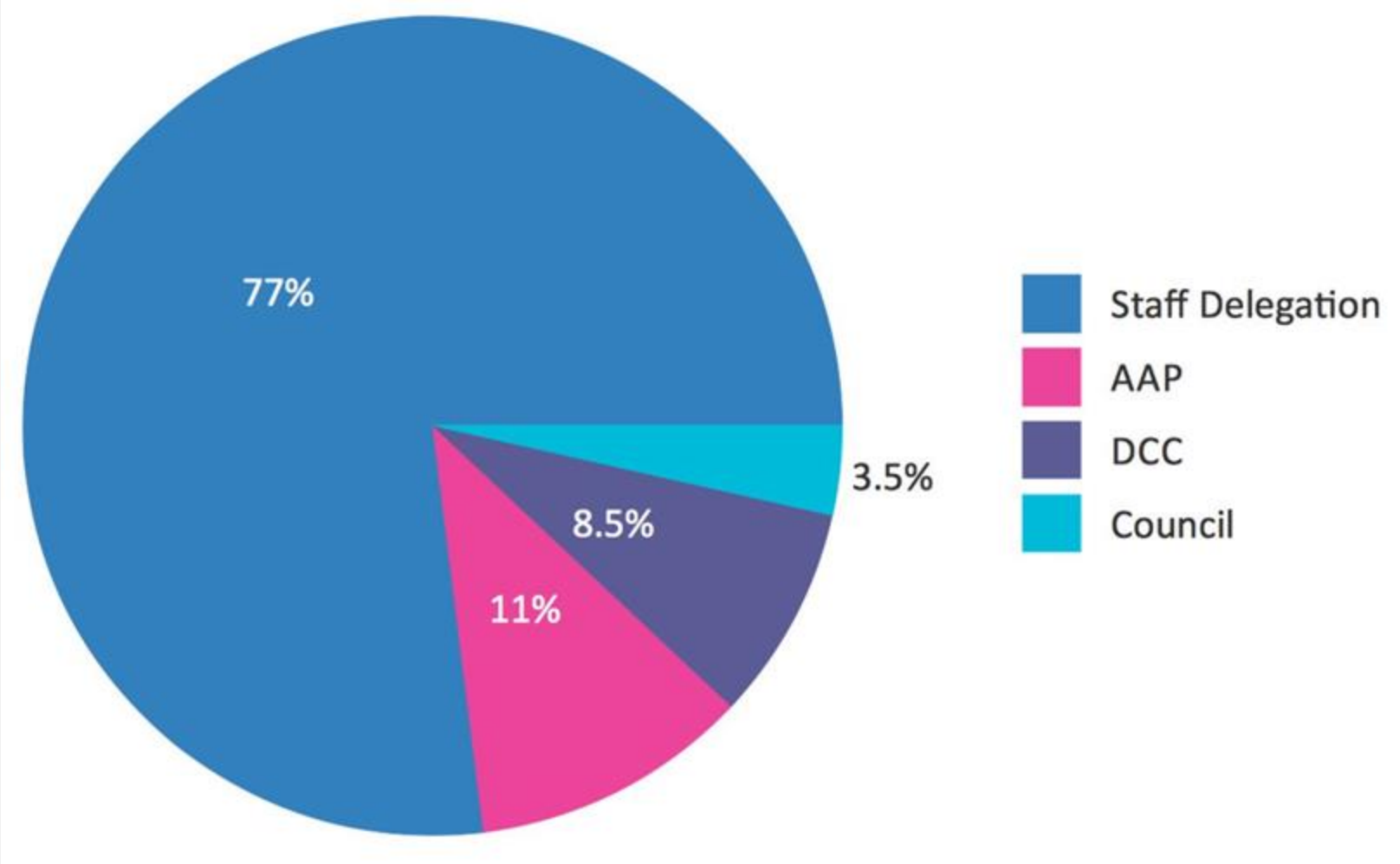
Es el otro tipo de gráfico para representar **variables cualitativas.**

Se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa.

El arco de cada porción se calcula con una regla de tres entre 360 grados, tamaño de la muestra y frecuencia de la clase.

Se pueden comparar dos poblaciones con círculos concéntricos.

GRÁFICO DE SECTORES



GRÁFICOS PARA VARIABLES CUANTITATIVAS

CONSIDERACIONES GENERALES

Para las variables cuantitativas, consideraremos dos tipos de gráficos, en función de que para realizarlos se usen las frecuencias (absolutas o relativas) o las frecuencias acumuladas.

Se le llaman **diagramas diferenciales** a los primeros (F y f) y **diagramas integrales** a los últimos (FA)

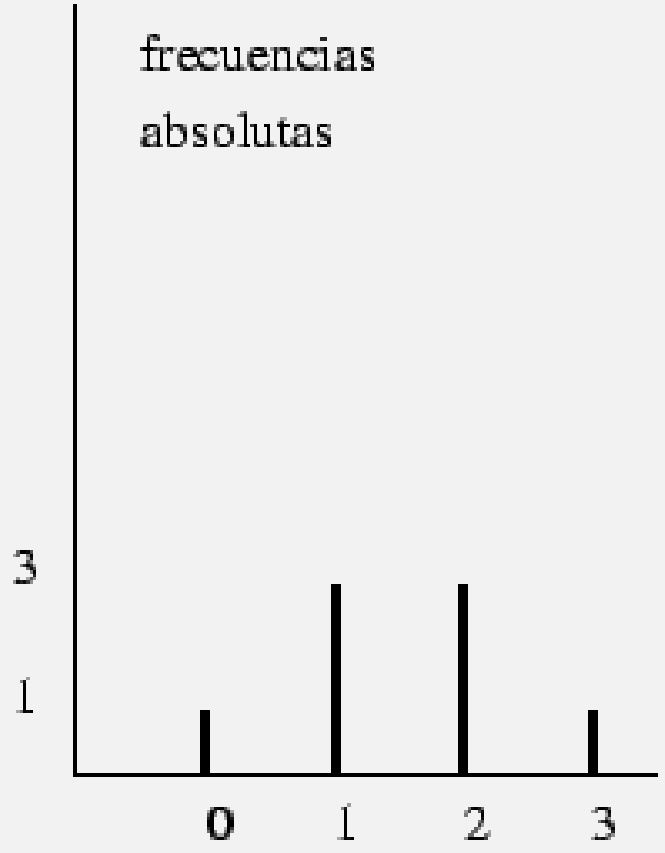
VARIABLES DISCRETAS

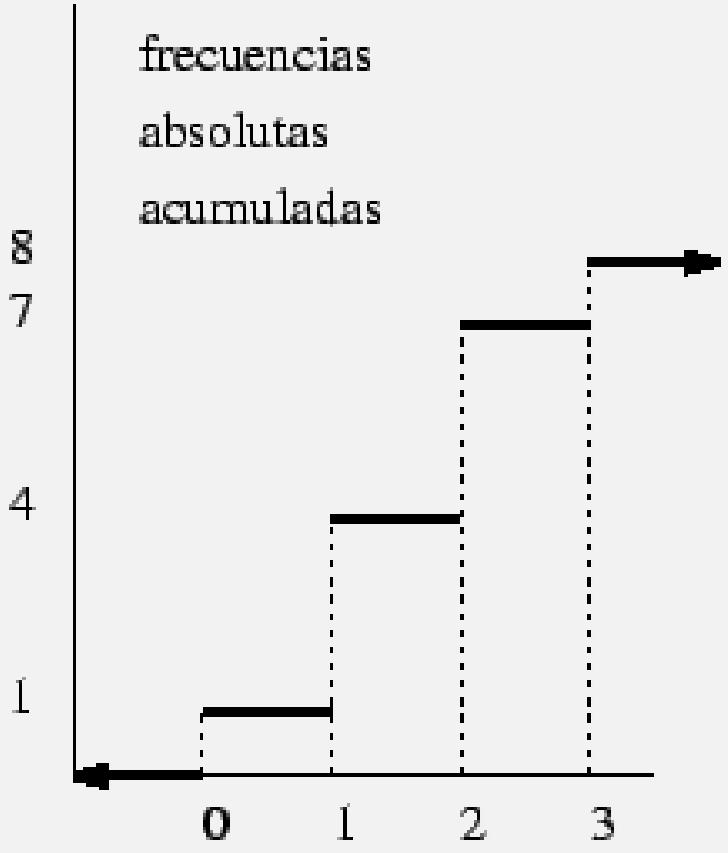
Usamos el *diagrama de barras* cuando pretendemos hacer una **gráfica diferencial**.

Las barras deben ser estrechas para representar que los valores que toma la variable son discretos.

El **diagrama integral** o *acumulado* tiene, por la naturaleza de la variable, forma de escalera.

GRÁFICOS DE BARRAS Y ESCALERAS





VARIABLES CONTINUAS

Cuando las variables son **continuas**, utilizamos como diagramas diferenciales los *histogramas* y los *polígonos de frecuencias*.

HISTOGRAMA

Un histograma se construye a partir de la tabla estadística, representando sobre cada intervalo, un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de *mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de cada intervalo y el área de los mismos*

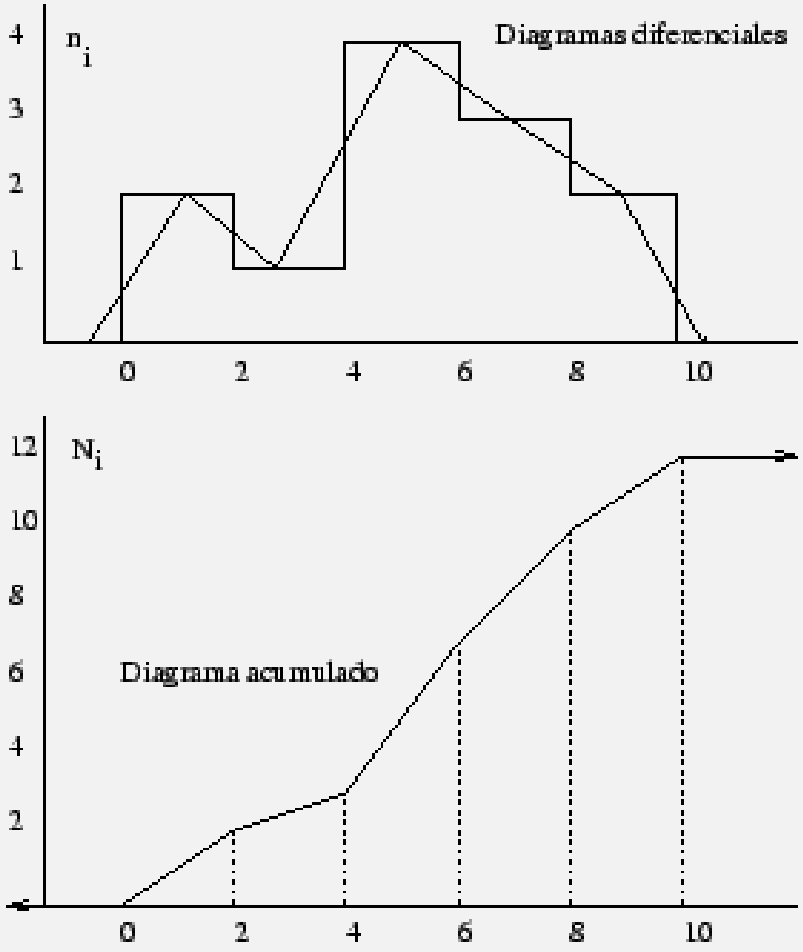
POLÍGONO DE FRECUENCIA

El polígono de frecuencias se construye fácilmente si tenemos representado previamente el histograma, ya que consiste en unir mediante líneas rectas los puntos del histograma que corresponden a las marcas de clase. Para representar el polígono de frecuencias en el primer y último intervalo, suponemos que adyacentes a ellos existen otros intervalos de la misma amplitud y frecuencia nula, y se unen por una línea recta los puntos del histograma que corresponden a sus marcas de clase.

POLÍGONO DE FRECUENCIAS ACUMULADAS (OJIVA)

El diagrama integral para una variable **continua** se denomina también *polígono de frecuencias acumulado*, y se obtiene como la poligonal definida en abscisas a partir de los extremos de los intervalos en los que hemos organizado la tabla de la variable, y en ordenadas por alturas que son proporcionales a las frecuencias acumuladas.

DIAGRAMAS DE VARIABLES CONTINUAS



BOX AND WHISKER PLOT

AKA diagrama de caja y bigote o *boxplot*.

Son una forma de representar de manera gráfica una serie de datos utilizando los cuartiles (caja) y otra serie de datos (bigotes).

CONSTRUCCIÓN DE UN *BOX PLOT*

Los límites de la caja son el primer y tercer cuartil, la banda central es la mediana y se puede representar la media con un punto dentro de la caja.

Los bigotes pueden representar una serie de posibles valores alternativos, como por ejemplo:

- Mínimo y máximo (dando idea del rango de la distribución)
- Media ± Desviación estándar
- Percentil 9 y 91
- Percentil 1 y 99
- Cuartil 1 y 3 ± 1,5 recorridos intercuartílicos

Todos los datos que no entren en los bigotes se tienen que representar con un punto o estrella.

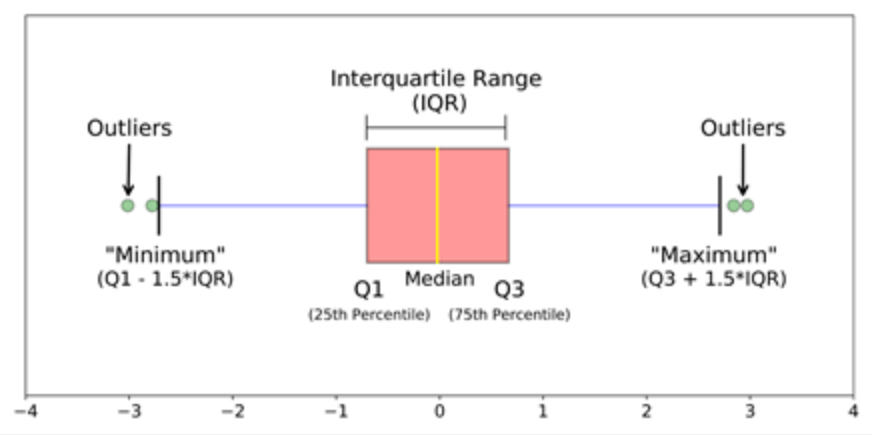
USOS

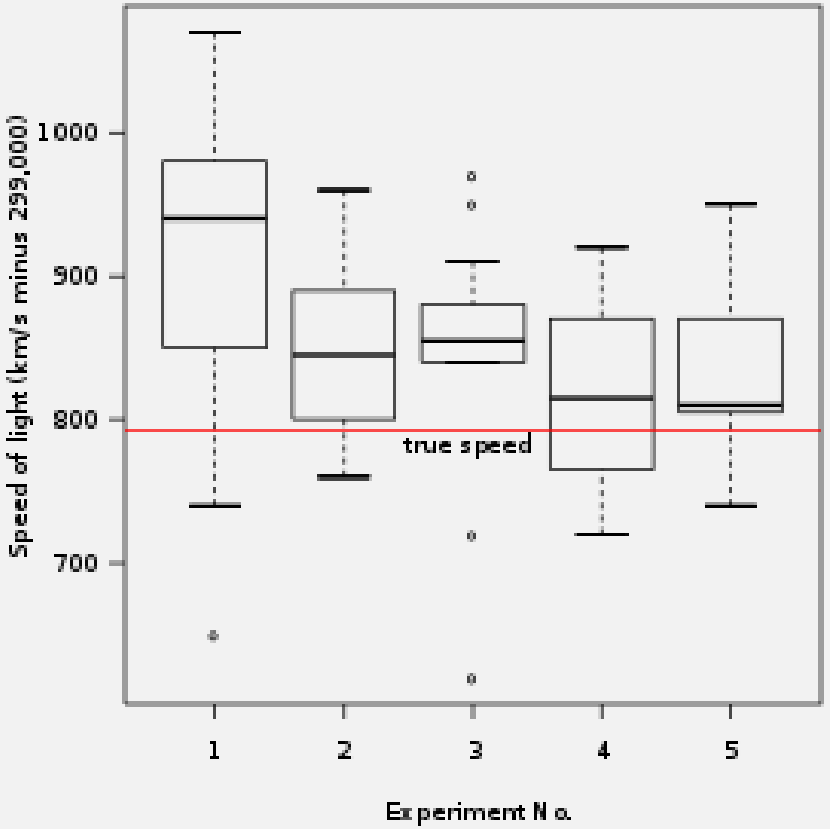
Los box plot permiten mostrar rápidamente un gran grupo de medidas de resumen (Cuartiles, mediana, media, desviación y varianza o mínimo y máximo) y muchas otras que se desprenden del análisis rápido del gráfico (CV, Q, rango, asimetría, curtosis).

Como ocupan poco espacio permiten comparar muchos grupos juntos.

La desventaja que presentan es que omiten valores específicos de la distribución, mostrando solo medidas de resumen.

EJEMPLOS


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4


Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.

ESTADÍSTICA DESCRIPTIVA

En un estudio taxonómico para identificar ecotipos de una especie de pez de río se midió la longitud furcal obteniéndose los siguientes datos:

longitud	78-82	82-86	86-90	90-94	94-98	98-102	102-106	106-110	110-114	114-118	118-122	122-126
marca de clase	80	84	88	92	96	100	104	108	112	116	120	124
f.a.	3	6	10	25	35	50	42	38	22	14	4	1

En qué escala de medida está la variable?

Calcular: Media, Mediana, Moda, Varianza Centrada, Q, Amplitud Total.

MEDIDAS DE RESUMEN

Media: 101,792

Desviación: 8,399

Varianza: 70.543

Mediana:

Valor de la observación ordenada numero 125:

Esta observación está en la clase 98 - 102: 101,68

Moda: La clase modal es 98 - 102: 100,6087

Cuartil 3: 107.7368

Cuartil 1: 96,1143

Q: 5,811

CV = 8,25 %

REGLAS DE PROBABILIDAD

Probabilidad total S = Espacio muestral **P(S)=1**

Acontecimientos complementarios $P(A) + P(A^c) = S = 1 \quad P(A^c) = 1 - P(A)$

Acontecimientos mutuamente excluyentes $P(A \cup B) = P(A) + P(B)$

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

Probabilidad condicional $P(A \mid B) = P(A \cap B) / P(B)$

Acontecimientos independientes $P(A \cap B) = P(A).\ P(B)$

Teorema de Bayes $$P(B_i \mid A) = \frac{P(A \mid B_i).P(B_i)}{\sum_{i}^{j} P(A \mid B_j).P(B_j)}$$

PROBABILIDAD I

Los mosquitos machos no pican. El 20% de los mosquitos hembras pican al ser humano (y a los gorilas). Sólo el 30% de la población de mosquitos local son hembras. Si se nos posa en el brazo un mosquito, cual es la probabilidad de que no nos pique?

PROBABILIDAD 2

Los exámenes para una enfermedad rara da resultados positivos en el 90% de los casos (es decir que al 90% de los animales con la enfermedad el examen le da positivo). Además, al 1% de la población el examen le da un falso positivo (es decir que el examen le dice que tiene la enfermedad cuando en realidad no la tiene). La prevalencia de esta enfermedad es de 0.0005

A. ¿Cuál es la probabilidad de que el examen de positivo si el animal examinado está enfermo?
B. ¿Cuál es la probabilidad de que el examen de negativo si el animal esta sano?
C. ¿Cuál es la probabilidad de que un examen positivo detecte efectivamente la enfermedad? (Pista: es decir que la persona tuviera efectivamente TB si el examen le dio positivo).
D. ¿Es cierta la siguiente frase? (Justifique): "En promedio, de un grupo de 23 animales a los que el examen les da positivo solo uno de ellos tiene la enfermedad."

PROBABILIDAD 3

Los cocodrilos son menos agresivos de lo que la fama dice de ellos. Así, sólo el 5% demuestran ataques no provocados con los humanos. De las tres especies (llamémosle A, B y C). Los A, que forman el 30% del total de cocodrilos, son los menos agresivos con los humanos, de hecho sólo el 1% de ellos tienen estas conductas. De la especie B que supone el 40%, el 3% de ellos atacan. El resto de cocodrilos forman la especie C. Si una persona es atacada por un cocodrilo sin haberlo provocado, cual es la probabilidad de que el agresor fuera de la especie C?

FUNCIONES DE PROBABILIDAD

El exceso de lluvias en ciertos campos bajos produce parasitosis en el ganado. Precipitaciones anuales mayores a 2000 mm son consideradas, en algunos casos, perjudiciales y provocan infestaciones que involucran al 20% del ganado bovino destetado. Si se toma una muestra de 12 terneros de un lote de 1000, Cuál es la probabilidad de que:

A) ninguno esté parasitado?
B) 5 o 6 no estén parasitados?
C) mas de 1 estén parasitados?

FUNCIONES DE PROBABILIDAD

La probabilidad que se produzcan alteraciones fetales en embriones cuyas madres han recibido penicilina en el tratamiento de infecciones durante el período de gravidez, es de 0,0003. Si 2000 hembras recibieron este antibiótico, calcular la probabilidad de que (considerando que cada una tuvo un solo hijo):

A) 2 fetos tengan alteraciones
B) Más de 2 fetos cuenten con alteraciones
C) Cuanto vale la desviación estándar de las alteraciones fetales?

FUNCIÓN DE DENSIDAD I

La calificación de un examen de bioestadística se distribuye normalmente con $\mu = 5.7$ y $\sigma = 1$.

Calcular la probabilidad de que un estudiante tomado al azar obtenga una calificación que no difiera de la media en más de 0.5 puntos.

FUNCIÓN DE DENSIDAD 2

El tiempo medio de respuesta obtenido en la aplicación de prostaglandinas a hembras caninas preñadas entre 35 y 55 días es de 60 horas con un desvío de 15 horas. ¿Cual es la probabilidad de obtener la reacción en los siguientes escenarios?

A) Luego de 50 horas.
B) Antes de 30 horas.
C) Entre 30 y 60 horas.
D) Antes de 80 horas.
E) Luego de 90 horas.

ESTIMACIÓN

Al examinar 140 vacas Holando Argentino de un tambo, 35 dieron reacción positiva para brucelosis.

A) Estimar puntualmente y a través de un intervalo de confianza del 95% la proporción de vacas no brucelosas.
B) ¿Cuántas vacas habría que examinar como mínimo para que el error no supere el 5% a un 95% de confianza?

ESTIMACIÓN

Un investigador necesita conocer el nivel medio de una enzima en una población. Los datos disponibles corresponden a determinaciones hechas en un conjunto de 400 individuos y se obtuvo una media de 91 unidades y se sabe que el CV poblacional es de 15%.

A) ¿Que cantidad de determinaciones deberá realizar para estimar la media poblacional con una confianza del 90% y que no difiera del verdadero valor en más de 5 unidades?
B) Estimar el intervalo de la media poblacional con un nivel de confianza del 99%

ESTIMACIÓN Y PRUEBA DE HIPÓTESIS

En un estudio de población salvaje se tomó una muestra de 1000 patos barcinos (Anas flavirostris) y se encontraron 400 machos y 600 hembras.

A) ¿Cuántos ejemplares habría que contar para estimar la proporción de machos a nivel de confianza del 95% con error menor a 0,01?
B) Estimar la proporción de machos con 95% de confianza.
C) ¿En la población hay tantos machos como hembras? (Hipótesis con alfa=0,05)

MUESTRAS INDEPENDIENTES Y DEPENDIENTES

Las muestras independientes son aquellas que se seleccionan de forma aleatoria para que sus observaciones no dependan de los valores de otras observaciones.

Muchos análisis estadísticos se basan en el supuesto de que las muestras son independientes. Otros se diseñan para evaluar muestras que no son independientes.

MUESTRAS INDEPENDIENTES Y DEPENDIENTES

Las muestras dependientes se les conoce también como pareadas porque generalmente los datos de estas se pueden representar como pares formados por un dato de una muestra, comparado con un dato de la otra.

EJEMPLO

Consideremos un laboratorio farmacéutico que desea probar la efectividad de un nuevo fármaco para reducir la presión arterial. El laboratorio podría recolectar los datos de dos maneras:

OPCIÓN 1

Tomando muestras de la presión arterial de las mismas personas antes y después de administrarles una dosis.

OPCIÓN 1

Tomando muestras de la presión arterial de las mismas personas antes y después de administrarles una dosis.

Las dos muestras son dependientes, porque se toman de las mismas personas. Es probable que las personas con la presión arterial más alta en la primera muestra también tengan la presión arterial más alta en la segunda muestra.

OPCIÓN 2

Dando a un grupo de personas un medicamento activo y dando a otro grupo de personas un placebo inactivo, para luego comparar los valores de presión arterial entre los grupos.

OPCIÓN 2

Dando a un grupo de personas un medicamento activo y dando a otro grupo de personas un placebo inactivo, para luego comparar los valores de presión arterial entre los grupos.

Es muy probable que estas dos muestras sean independientes, porque las mediciones corresponden a personas diferentes. Saber algo sobre la distribución de los valores de la primera muestra no le indica nada con respecto a la distribución de los valores de la segunda.

COMPARACIÓN DE MEDIAS

Inspectores de calidad desean comparar dos laboratorios para determinar si sus exámenes de sangre proporcionan resultados similares. Envían a ambos laboratorios muestras de sangre extraídas de un mismo grupo de 10 individuos, para ser analizadas.

¿Existe diferencia entre los resultados de un laboratorio y otro?

1	0.8	0.7
2	4.8	5
3	7.9	7.8
4	15.7	16.3
5	21.2	20.2
6	9.7	9.4
7	38.7	44
8	5.1	5.1
9	29	26.9
10	75.2	74.6

COMPARACIÓN DE MEDIAS (PAREADOS)

1	0.8	0.7	0.1
2	4.8	5	0.2
3	7.9	7.8	0.1
4	15.7	16.3	0.6
5	21.2	20.2	1
6	9.7	9.4	0.3
7	38.7	44	5.3
8	5.1	5.1	0
9	29	26.9	2.1
10	75.2	74.6	0.6

Los grupos no son independientes

$$t_c = \frac{\bar{d}}{\hat{s}_d / \sqrt{n}} \qquad gl = n - 1$$

đ = 1.03
s_d = 1.625
n = 10

t_c = 2.004
t_t = 2.262

Acepto hipótesis nula de igualdad de medias (α=0.05)

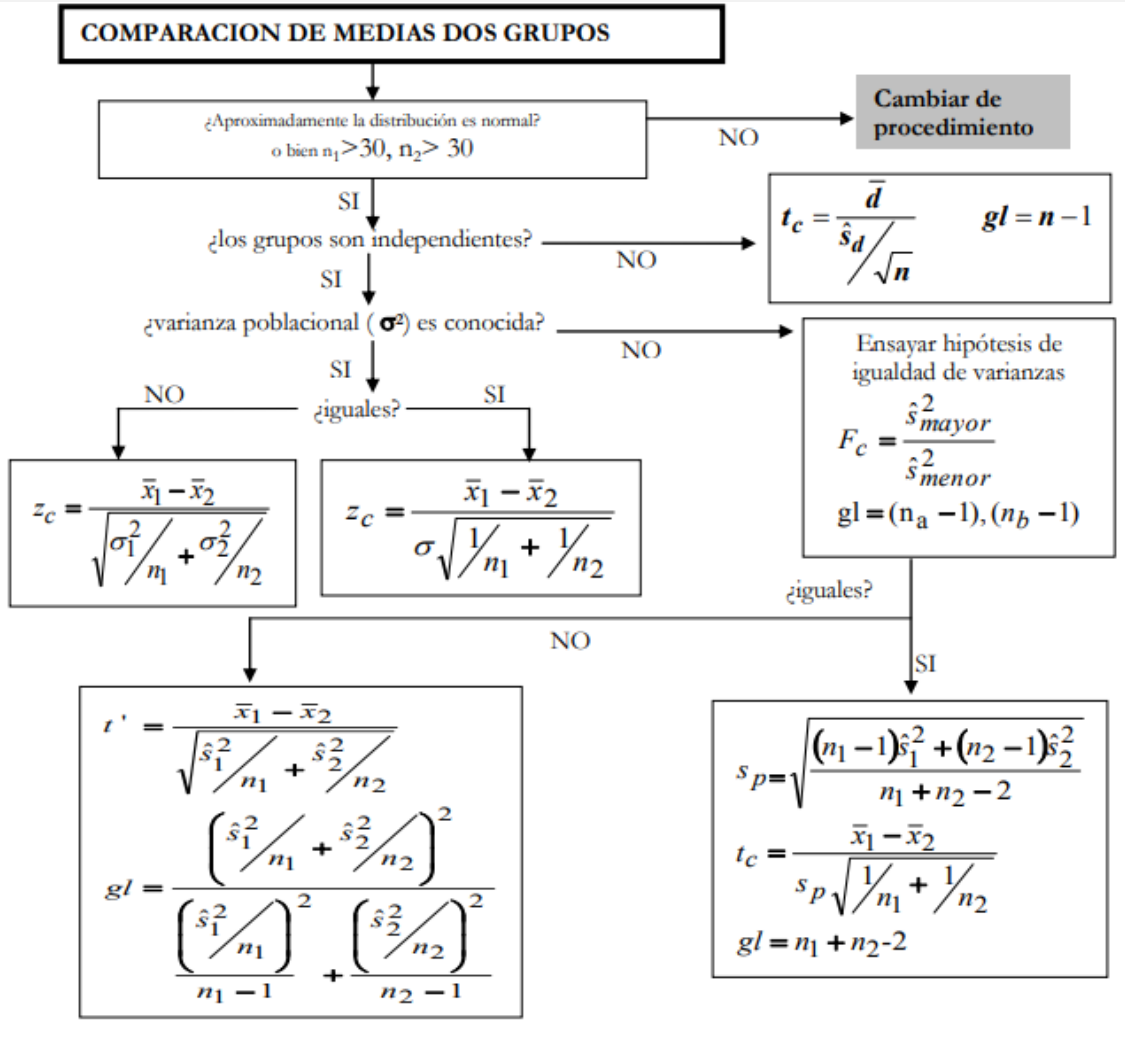
COMPARACIÓN DE MEDIAS

La siguiente tabla muestra la longitud de dos muestras de la misma especie de pez de río de dos criaderos distintos.

Criadero A	30	25	30	33	27	26	27	20	26	23	25	21	26	23
Criadero B	23	25	18	23,6	21,5	23	27							

¿Hay diferencia de talla entre las locaciones?

COMPARACIÓN DE MEDIAS (INDEPENDIENTES)

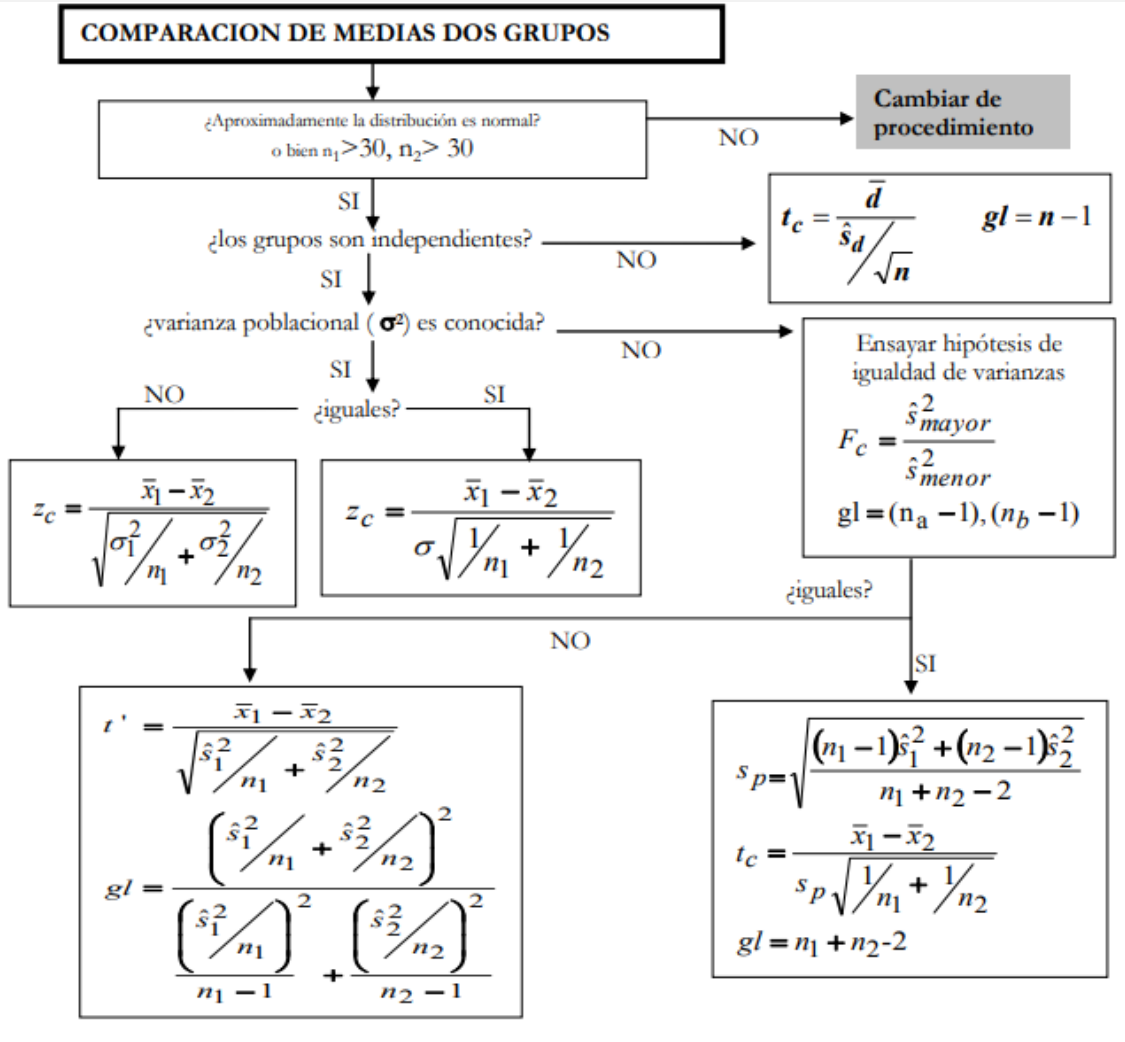


Como la varianza poblacional es desconocida primero se ensaya hipótesis de igualdad de varianzas con estadístico F

Fc = 1,587
g.l. = 13 ; 6
Ft = 3,98

Se acepta hipótesis nula de igualdad de varianzas (son muestras homocedasticas)

COMPARACIÓN DE MEDIAS (INDEPENDIENTES)



Sp = 3,335
Tc = 1,846
g.l. = 19
Tt = 2,093

Acepto hipótesis nula de igualdad de medias

COMPARACIÓN DE MEDIAS

Los niveles medios de sulfato en la capa superior de suelo pueden ser indicador de su calidad. Se sabe que al norte del río negro este valor tiene un desvío de 3 y al sur presenta desvío de 4.
Para saber si hay una diferencia de medias se toman medidas en 35 campos del norte y 30 del sur del río negro.
La media muestral del norte es de 15, la del sur de 10.

¿Se puede afirmar que la media varía significativamente entre ambas zonas?

COMPARACIÓN DE MEDIAS

¿Se puede afirmar que la media varía significativamente entre ambas zonas?

Tenemos un caso de muestras independientes, con varianza poblacional conocida, y estas son diferentes (heterocedásticas), por lo que se utiliza estadístico z

$$z_c = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Zc = 5,624

Rechazo la hipótesis de igualdad de medias

ANOVA

Cuando se nos presentan situaciones que requieren comparar una variable de respuesta entre mas de 2 grupos divididos por otra variable llamada factor estudiado debemos realizar un Análisis de Varianza (ANalysis Of VAriance)

Este parte de una hipótesis nula de igualdad de medias y se resuelve con el estadístico F.

Si el Fc es menor que el Ft aceptamos la hipótesis nula de igualdad de medias de los grupos

ANOVA

El exceso de ozono es un indicador temprano de contaminación, como primer etapa de un estudio de impacto ambiental, se tomaron muestras en 2 establecimientos cercanos a posibles fuentes de contaminación y 1 en uno que previamente se conocía que no estaba afectado por contaminantes.

Zona 1	15,5	12,5	17,2	14,6
Zona 2	14,7	18,5	14,3	16,1
Zona 3	18,5	15,5	20,3	17,0

¿Cuáles son las variables en estudio?

¿Qué supuestos asumo para realizar una prueba de hipótesis?

Probar si hay diferencia significativa en los niveles de ozono en las 3 zonas

ANOVA

$SC\ total = \sum x^2 - (T^2/n)$

$SC\ entre = \sum T_i^2/n_i - (T^2/n)$

$SC\ dentro = SC\ total - SC\ entre$

$MC\ entre = SC\ entre\ /\ (a - 1)$

$MC\ dentro = SC\ dentro\ /\ (n - a)$

$Fc = MC\ entre/MC\ dentro;\ gl = (a - 1);\ (n - a)$

a = número de grupos

$\mathbf{n_i}$ = número de observaciones del i-ésimo grupo

n= total de observaciones

$\mathbf{T_i}$ = Suma de observaciones del i-ésimo grupo

T= Suma total de observaciones

Fuente de Variación	SC	gl	MC	F
Entre grupos (tratamientos)	SC entre	a-1	MC entre	$F_{calculado}$
Dentro de grupos (residual)	SC dentro	n-a	MC dentro	
total	SC total	n-1		

FV	SC	gl	MC	Fc
Entre	17,165	2	8,5825	2,2096
Dentro	34,958	9	3,8842	
Total	52,123	11		

Ft = 4,26

ANOVA

- La variable de respuesta es el nivel de ozono.
- El factor estudiado es la fuente de contaminación.
- Los supuestos para realizar cualquier ANOVA son:
  - Variable normal
  - Homocedasticidad
  - Observaciones independientes
- Se acepta la hipótesis nula de igualdad de los grupos.

ANOVA

La otra forma de presentar ejercicios de ANOVA es con una tabla parcialmente completa.

Esta forma es mas liviana en operaciones pero requiere recordar la relación entre las celdas de la tabla de ANOVA.

FV	SC	gl	MC	Fc
Entre			302	
Dentro				
Total	3213	22		

*Suponer datos divididos en 4 grupos

ANOVA

La otra forma de presentar ejercicios de ANOVA es con una tabla parcialmente completa.

Esta forma es mas liviana en operaciones pero requiere recordar la relación entre las celdas de la tabla de ANOVA.

FV	SC	gl	MC	Fc
Entre	906	3	**302**	*2,4872*
Dentro	2307	19	121,421	
Total	**3213**	**22**		

Ft = 3,13

ANÁLISIS DE FRECUENCIAS

Pueden ser pruebas de homogeneidad e independencia, o de bondad de ajuste.

Son pruebas que se realizan en grupos separados por un factor estudiado para ver si las frecuencias obtenidas de la variable de respuesta (nominal) son iguales a las esperadas asumiendo una característica particular.

Estos ejercicios se resuelven a partir de tablas de contingencia y el estadístico Chi-cuadrado

ANÁLISIS DE FRECUENCIAS

El cáncer mas común que afecta la cavidad oral de rumiantes es de tipo carcinoma. Se quiere determinar si la frecuencia en que se presenta es igual en paladar, encía y lengua, con 95% de confianza.
Se tomó una muestra de 90 animales a los que se les diagnosticó cáncer bucal y se los agrupó según si eran carcinomas u otros tipos de cáncer y a su vez según su localización, obteniéndose los siguientes datos:

Recuento

		Localización del Cáncer			Total
		Paladar	Encía	Lengua	
Estirpe Celular del Cáncer	Carcinoma	6	11	40	57
	Otros Ca	10	17	6	33
Total		16	28	46	90

Determinar si hay diferencia de frecuencia según localización y tipo.

ANÁLISIS DE FRECUENCIAS

Determinar si hay diferencia de frecuencia según localización y tipo.

Recuento

		Localización del Cáncer			Total
		Paladar	Encía	Lengua	
Estirpe Celular del Cáncer	Carcinoma	6	11	40	57
	Otros Ca	10	17	6	33
Total		16	28	46	90

Se debe calcular las frecuencias esperadas asumiendo homogeneidad según la fórmula

$Eij = Oi \cdot Oj / n$

La tabla de frecuencias esperadas sería la siguiente

ANÁLISIS DE FRECUENCIAS

	Paladar	Encía	Lengua	Total
Carcinoma	10,133	17,733	29,133	**57**
Otros	5,867	10,267	17,967	**33**
Total	**16**	**28**	**46**	**90**

Con los valores de frecuencia esperada calculados se computa el estadístico chi cuadrado:

$$\chi_c^2 = \sum O_{ij}^2 / E_{ij} - n$$

$$gl = (filas - 1)(columnas - 1)$$

$$\chi_c^2 = \sum O_{ij}^2 / E_{ij} - n$$

$$gl = (filas - 1)(columnas - 1)$$

$$\sum \frac{O_{ij}^2}{E_{ij}} = 112.61 \qquad \chi_c^2 = \sum \frac{O_{ij}^2}{E_{ij}} - n = 22.61$$

	Paladar	Encía	Lengua	Total
Carcinoma	6 (10.13) [3.55]	11 (17.73) [6.82]	40 (29.13) [54.93]	57
Otros	10 (5.87) [17.04]	17 (10.27) [28.14]	6 (16.87) [2.13]	33
Total	16	28	46	90

ANÁLISIS DE FRECUENCIAS

ANÁLISIS DE FRECUENCIAS

En una empresa de acuicultura se quiere hacer un estudio sobre el nivel de parásitos en la producción de doradas. Para ello, se tomó una muestra de 5 individuos cada día, repitiendo el experimento durante 550 días. De cada muestra se analizaron los peces determinando cuantos de ellos contenían parásitos. ¿Se ajusta a un modelo de distribución Binomial?

X	0	1	2	3	4	5
O	17	81	152	180	104	16

$X \equiv B(n, p)$

ANÁLISIS DE FRECUENCIAS

X	0	1	2	3	4	5
O	17	81	152	180	104	16
p	0.026	0.141	0.301	0.322	0.173	0.037
E = Np	14.30	77.55	165.6	177.1	95.15	20.35

$X \equiv B(n, p)$

X = N° de individuos con parásitos
n = 5 individuos
P = ?

$X \equiv B\ (5; 0{,}517)$

$$P(x, n, p) = C_x^n \cdot p^x \cdot q^{n-x}$$

media = np = 2.584
p = media / n = 0.517

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i F_i}{N}$$

¡¡¡Ojo, esta "N" hace referencia al tamaño muestral!!! (n=550)

$$\chi_c^2 = \sum \frac{O_{ij}^2}{E_{ij}} - n = 3.187$$

REGRESIÓN Y CORRELACIÓN LINEAL

El análisis de correlación y el de regresión son dos herramientas de gran importancia y poder en la estadística, pueden verse como las dos caras de una moneda para analizar un problema.

REGRESIÓN Y CORRELACIÓN LINEAL

La **correlación** indica la fuerza y la dirección de una relación lineal entre dos variables estadísticas. Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si tenemos dos variables (A y B) existe correlación si al aumentar los valores de A lo hacen también los de B y viceversa. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad (*Cum hoc ergo propter hoc*)

REGRESIÓN Y CORRELACIÓN LINEAL

Este es expresado por un único valor llamado coeficiente de correlación (r), el cual puede tener valores que oscilan entre -1 y +1. Cuando "r" es negativo, significa que una variable (ya sea "x" o "y") tiende a decrecer cuando la otra aumenta (se trata entonces de una "correlación negativa", correspondiente a un valor negativo de "b" en el análisis de regresión). Cuando "r" es positivo, en cambio, esto significa que una variable se incrementa al hacerse mayor la otra (lo cual corresponde a un valor positivo de "b" en el análisis de regresión).

REGRESIÓN Y CORRELACIÓN LINEAL

La **regresión lineal** es un método matemático de ajuste, que permite obtener una fórmula matemática que se ajusta al cambio de una variable dependiente en relación al cambio de una variable independiente.

El método utilizado en el curso se basa en calcular los parámetros de una formula de recta (o de primer grado), de esto que se llame correlación *lineal*.

Los parámetros de una recta son llamados *a* y *b*, el primero representa las ordenadas en origen y el último a la pendiente de la recta.

REGRESIÓN Y CORRELACIÓN LINEAL

La **correlación lineal** nos permite comprobar la asociación lineal entre dos variables cuantitativas.

La **regresión lineal** nos permite obtener un modelo predictivo sobre dos variables correlacionadas.

REGRESIÓN Y CORRELACIÓN LINEAL

Aspectos "teóricos" a recordar:

- La relación entre x e y es lineal en el intervalo considerado por lo que las estimaciones a partir de un modelo de regresión lineal solo se pueden considerar válidas dentro del rango de los datos.
- La variable controlada por el experimentador es la independiente (x) y la otra es la dependiente (y).
- Se le llama centroide al punto en la coordenada correspondiente a la media de ambas variables.
- Las variables serán mejor estimadas cuanto mas cercanas al centroide se encuentren.

REGRESIÓN Y CORRELACIÓN LINEAL

Aspectos "teóricos" a recordar:

- En los ejercicios de correlación se puede mencionar la suma de cuadrados de la regresión y la suma de cuadrados total como datos para la resolución del ejercicio. Generalmente en este tipo de planteo se nos pide calcular R^2 (coeficiente de determinación)
- $R^2 = \frac{SCR}{SCT}$
- R^2 toma valores entre 0 y 1, y su raíz cuadrada es el valor absoluto de r del modelo lineal.