# Problemas resueltos

## 1. Estadística Descriptiva

**Problema 1.1.** En una encuesta sobre equipamiento y uso de las tecnologías de la información y comunicación en los hogares (viviendas familiares habitadas por al menos una persona de 16 a 74 años), realizada en el año 2014 en Valencia, 408200 personas indican que han utilizado Internet en los últimos tres meses. En la pregunta a este grupo de personas, sobre el uso de Internet para comprar, se obtuvieron los siguientes resultados:

| **Ha utilizado Internet para comprar** | **Nº personas** |
|---|---|
| En el último mes | 109900 |
| Hace más de un mes pero menos de 3 meses | |
| Hace más de tres meses pero menos de un año | 52406 |
| Hace más de un año | 17045 |
| No | 153800 |

a) ¿Qué población y qué variable aleatoria se están estudiando?

b) Completa la tabla.

c) ¿Qué tipo de frecuencias recoge la segunda columna? Calcula a partir de ellas otro tipo de frecuencias que permitan estudiar la distribución de la variable.

d) Representa el diagrama de sectores de la muestra.

**RESPUESTAS:**

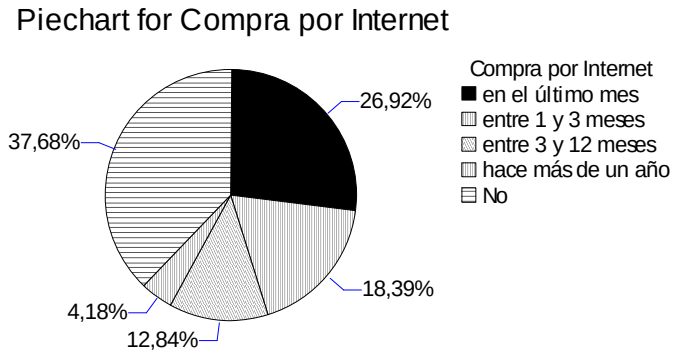
a) Población: personas de 16 a 74 años, en viviendas familiares de Valencia en el año 2014, que han utilizado Internet en los últimos 3 meses. Variable aleatoria: meses desde la última vez que se utilizó Internet para comprar.

b) La tabla se completa añadiendo en la tercera fila el nº de personas que han utilizado Internet para comprar por última vez, hace más de un mes pero menos de tres meses. Se calcula como 408200 – 109900 – 52406 – 17045 – 153800 = 75049

c) La segunda columna tiene las frecuencias absolutas. Las frecuencias relativas se calculan dividiendo los valores de esta columna por 408200.

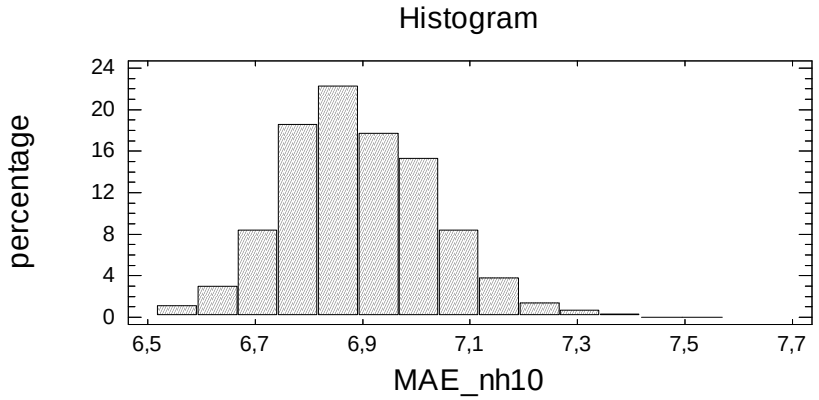
| **Ha utilizado Internet para comprar** | **Nº personas** | **%** |
|---|---|---|
| En el último mes | 109900 | 26,92 |
| Hace más de un mes pero menos de 3 meses | 75049 | 18,39 |
| Hace más de tres meses pero menos de un año | 52406 | 12,84 |
| Hace más de un año | 17045 | 4,18 |
| No | 153800 | 37,68 |

d) El gráfico de sectores es:

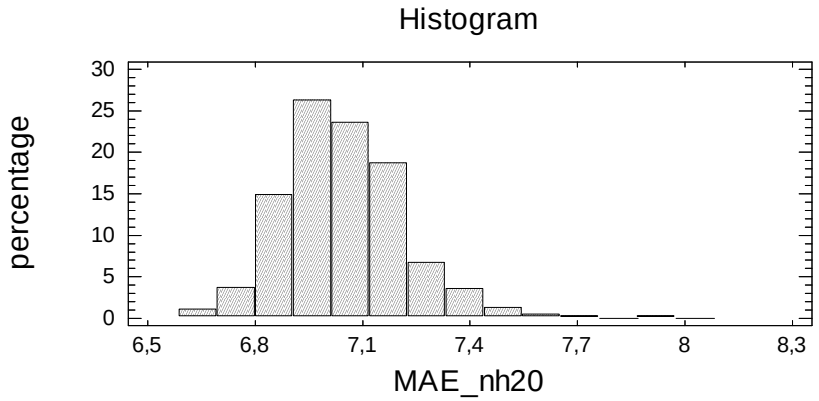



**Problema 1.2.** Se han evaluado las predicciones de la concentración media horaria de ozono, calculadas con la base de datos de la estación automática de medición de la calidad del aire, ubicada en un campus universitario. El modelo ha sido una red neuronal.

a) El histograma adjunto representa los promedios de los errores de predicción en valor absoluto (MAE_nh10), cuando en el modelo se aplicaron 10 neuronas en la capa oculta. ¿Cómo es la distribución de frecuencias? Estima a partir del gráfico, de forma aproximada, los parámetros de posición que sean adecuados en este caso. ¿Qué medida de dispersión es representativa de la distribución?




b) Cuando se utilizaron 20 neuronas en la capa oculta, el histograma para la misma medida de incertidumbre de las predicciones es:




Calcula en este caso los parámetros adecuados para posición. La dispersión en estas observaciones, ¿es mayor o menor? ¿Con qué número de neuronas es menor la incertidumbre de las predicciones?

**RESPUESTAS:**

a) Es una distribución asimétrica positiva. Por tanto, los parámetros adecuados de posición son la mediana y los cuartiles C1 y C3. Teniendo en cuenta los porcentajes del eje de ordenadas, la mediana se encuentra en el intervalo [6,8, 6,9], C1 entre [6,75,

6,82], y C3 entre [6,98, 7,06]. Para medir la dispersión, es representativo el recorrido intercuartílico, que se estimaría como C3-C1.

b) La distribución de frecuencias es también asimétrica positiva. Con lo que se utilizan como parámetros de posición la mediana y los cuartiles C1 y C3. A partir de los porcentajes del eje de ordenadas, aproximadamente son: mediana un valor entre [7,01, 7,13], C1 un valor en [6,91, 7,01] y C3 en [7,13, 7,24]. La dispersión es mayor en este histograma, teniendo en cuenta el rango de valores (recorrido (máximo-mínimo)) y los porcentajes de los intervalos.Con 10 neuronas en la capa oculta, la posición y dispersión de la medida de incertidumbre es menor. Por tanto las predicciones en ese caso tienen en un número mayor de casos, menor error.

**Problema 1.3**. La velocidad de carga de una página web (décimas de segundo), observada en 15 momentos diferentes del día, ha sido:

2,7 0,5 2,9 10,8 1,0 3,8 6,3 1,1 3,8 2,0 5,6 0,8 1,8 1,7 1,8

a) Representa el diagrama Box-Whisker de los datos.
b) La distribución de los datos es ¿simétrica o asimétrica?
c) ¿Qué parámetros de posición y dispersión son más adecuados para estos datos?

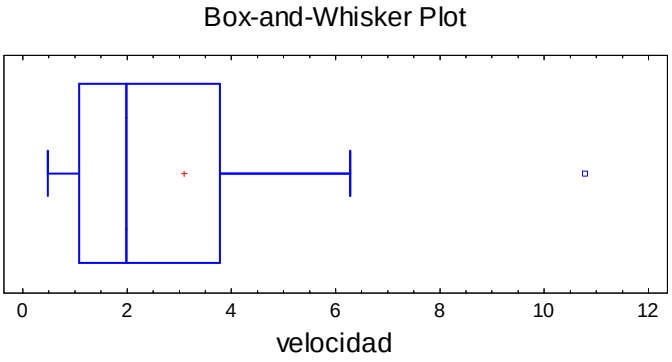
**RESPUESTAS:**

a) Datos ordenados: 0,5 0,8 1,0 1,1 1,7 1,8 1,8 2,0 2,7 2,9 3,8 3,8 5,6 6,3 10,8

Mediana =2 C1=1,1 C3=3,8

1,5(C3-C1)=4,05

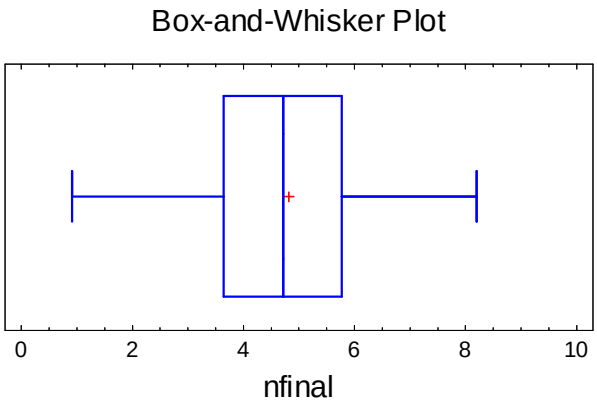
10,8 > 3,8 + 4,05 $\Rightarrow$ 10 ,8 es una observación anómala




b) La distribución es asimétrica positiva, ya que el bigote derecho es de mayor longitud que el izquierdo, la media es mayor que la mediana y hay un punto anómalo por la derecha.

c) Con datos asimétricos positivos y con anomalías, los parámetros de posición adecuados son la mediana=2 y los cuartiles C1=1,1 y C3=3,8, y de dispersión el recorrido intercuartílico RI = C3 - C1 = 2,7

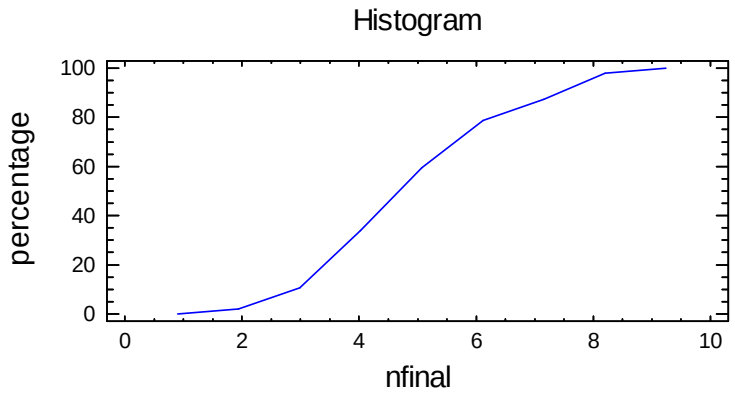
**Problema 1.4.** El diagrama Box-Whisker representa las notas finales obtenidas por un grupo de 47 alumnos en la asignatura Estadística, del grado en Ingeniería Informática de una determinada universidad:




a) Indica la población y la muestra de este estudio. ¿Qué variable aleatoria se está observando? ¿Qué tipo de variable es?

b) Es en este caso la media ¿es un buen parámetro de posición?

c) ¿Qué porcentaje de alumnos tiene una nota final por encima de 5? Confirma la respuesta a partir del polígono de frecuencias relativas acumuladas.




d) ¿Qué parámetro de dispersión es adecuado para describir esta muestra?

**RESPUESTAS:**

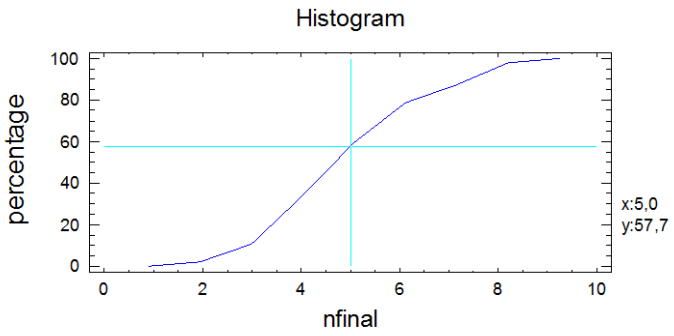
a) Población: Todos los estudiantes de la asignatura
Muestra: Los 47 estudiantes

Variable aleatoria: La nota final, es cuantitativa continua

b) La media si que es un parámetro representativo de posición, ya que la distribución de los datos es aproximadamente simétrica. Los bigotes tienen una longitud muy similar, media y mediana están cercanas, y centradas en la caja.

c) La mediana vale 4,7 (aproximadamente, en el porcentaje 50%). Como la media es superior a ella, pero muy cercana, el porcentaje de alumnos con nota superior a 5 es menor que el 50%. En el polígono de frecuencias acumuladas se obtiene que por debajo de 5, está el 57,7% de los alumnos. Por tanto, por encima de 5 estará el 100-57,7=42,3% restante.

d) La desviación típica es en este caso un parámetro de dispersión adecuado para describir la muestra, ya que hay simetría en los datos y no hay ninguna observación anómala.




**Problema 1.5.** En un estudio nacional sobre la situación laboral de los profesionales del sector de tecnologías de la información, se ha obtenido la tabla de frecuencias relativas de edad x tipo de contrato:

| | Fijo | Temporal | Funcionario | Becario/Prácticas |
|---|---|---|---|---|
| ≤ 30 | 61 | | 0,8 | 14,4 |
| [31, 45] | 74,2 | 16,1 | | 0,8 |
| ≥ 46 | 70 | 7,3 | 22 | 0,7 |

a) ¿Qué porcentaje de contratos son temporales en el grupo de los que tienen 30 años o menos?

b) ¿En qué franja de edad hay más funcionarios?

c) ¿Qué cuantifica el valor 0,8% de la segunda fila de la tabla?

**RESPUESTAS:**

a) Ese porcentaje es la frecuencia relativa de contrato temporal condicionada a edad ≤ 30. Resulta 100 - 61 - 0,8 - 14,4 = 23,8 %

b) En el grupo de edad ≤ 30, el porcentaje de funcionarios es 0,8%. En la siguiente franja de edad ([31, 45]), esa frecuencia relativa condicionada vale 100 – 74,2 – 16,1 – 0,8 = 8,9%. Por tanto la mayor proporción de funcionarios (22%) se presenta en el grupo de edades ≥ 46.

c) El valor 0,8% es el porcentaje que hay de funcionarios en el grupo con edad ≤ 30 años. Frecuencia relativa de funcionarios condicionada a edad menor o igual a 30 años.

**Problema 1.6.** En el mismo estudio del problema 1.5, se analizó la relación entre titulación y salario, con las siguientes frecuencias relativas:

Tabla de frecuencias titulación x salario (euros)

| | <12000 | [12001, 30000] | [30001, 60000] | >60000 |
|---|---|---|---|---|
| Doctor en Informática | 2,7 | | 68,3 | 4,9 |
| Ingeniero, Licenciado o Máster en Informática | 2,3 | 38,6 | | 9,9 |
| Ing.Técnico, Diplomado o Graduado en Informática | 6,5 | | 37,9 | 4,5 |
| FP TIC | | 60,5 | 20,3 | 1,9 |
| DEA | 3,1 | 34,6 | 53,1 | 9,2 |
| Sin formación reglada | 4,7 | 38,2 | 45,7 | 11,4 |

a) Completa la tabla. ¿Qué tipo frecuencias relativas incluye?

b) ¿Qué cuantifican los valores 1,9% y 68,3%?

c) ¿Hay relación entre la formación alcanzada y el salario percibido?

**RESPUESTAS:**

a) Las frecuencias relativas son las de salario condicionado a titulación. La tabla completa está en la página siguiente.

b) El 1,9% de los que tienen un ciclo formativo FP TIC, cobra más de 60000 euros.

El 68,3% de los Doctores en Informática tiene un salario entre 30001 y 60000.

c) Si que se observa relación entre titulación y salario. Los Doctores en Informática presentan un mayor porcentaje de salario en el intervalo [30001, 60000], pero el mayor porcentaje de salarios > 60000 se da en los profesionales sin formación reglada (11,4%). El mayor porcentaje de titulados con salarios < 12000, se observa en el grupo con un ciclo formativo FP TIC (17,3%).

Tabla de frecuencias titulación x salario (euros)

| | <12000 | [12001, 30000] | [30001, 60000] | >60000 |
|---|---|---|---|---|
| Doctor en Informática | 2,7 | **24,1** | 68,3 | 4,9 |
| Ingeniero, Licenciado o Máster en Informática | 2,3 | 38,6 | **49,2** | 9,9 |
| Ing.Técnico, Diplomado o Graduado en Informática | 6,5 | **51,1** | 37,9 | 4,5 |
| FP TIC | **17,3** | 60,5 | 20,3 | 1,9 |
| DEA | 3,1 | 34,6 | 53,1 | 9,2 |
| Sin formación reglada | 4,7 | 38,2 | 45,7 | 11,4 |

## 2. Conceptos Básicos del Cálculo de Probabilidades

**Problema 2.1.** En la población de ingenieros informáticos, un 82,7% son hombres. El 48,7% de los hombres tiene un salario superior a 30000 euros, y en las mujeres este porcentaje es 38,6%

a) Calcula el porcentaje que tiene un salario mayor que 30000 euros.

b) En el grupo que tiene un salario superior a 30000, ¿qué porcentaje hay de mujeres?

**RESPUESTAS:**

a)

P(hombre)=0,827

P(mujer)=1-0,827=0,173

P(salario>30000/hombre)=0,487

P(salario>30000/mujer)=0,386

Aplicando el Teorema de la Probabilidad Total

P(salario>30000)=
=P(hombre)xP(salario>30000/hombre)+P(mujer)xP(salario>30000/mujer)=

= 0,827 x 0,487 + 0,173 x 0,386 = 0,4027+ 0,0668= 0,4695

El porcentaje con salario superior a 30000 es 46,95%

b) Aplicando el Teorema de Bayes

P(mujer/salario>30000)=(P(mujer)xP(salario>30000/mujer))/P(salario>30000)=

=(0,173x0,386)/0,4695 = 0,1422

En el grupo que tiene un salario superior a 30000, el porcentaje de mujeres es 14,22%

**Problema 2.2.** En una ciudad determinada, se ha analizado la frecuencia de uso de ordenadores y el género. El porcentaje que utiliza el ordenador diariamente (al menos 5 días por semana) es 72,5%, entre 1 y 5 días por semana 21,3%, y menos de un día a la semana 6,2%.

En el grupo que utiliza el ordenador al menos 5 días por semana, hay un 49,8% de hombres. En el de frecuencia de uso de 1 a 4 días por semana, hay un 34,7% son hombres. Hay un 16,5% de hombres en el grupo que lo utiliza menos de un día por semana.

a) Calcula el porcentaje total de hombres.

b) En el grupo de las mujeres, ¿qué porcentaje utiliza el ordenador al menos cinco días por semana?

**RESPUESTAS:**

a) Definición de sucesos y datos

Frecuencia de uso del ordenador

A= al menos cinco días por semana P(A) = 0,725

B= entre cuatro y un día por semana P(B) = 0,213

C= menos de un día por semana P(C) = 0,062

E= género masculino

P(E/A) = 0,498 P(E/B) = 0,347 P(E/C) = 0,165

Aplicación del Teorema de la Probabilidad Total

P(E) = 0,725 x 0,498 + 0,213 x 0,347 + 0,062 x 0,165 = 0,4452

Porcentaje de hombres 44,52%

b) $\bar{E}$= género femenino $P(\bar{E}) = 1 - 0,4452 = 0,5548$

Aplicando el Teorema de Bayes

$P(A/\bar{E}) = (0,725 \text{ x } (1-0,498))/0,5548 = 0,656$

Porcentaje que utiliza el ordenador con frecuencia diaria en el grupo de mujeres = 65,6%

**Problema 2.3.** La fiabilidad de un dispositivo es del 80% a las 5000 horas. Se quiere incrementar dicha fiabilidad hasta el 95%, añadiendo una componente en paralelo. Para obtener ese resultado, ¿cuánto debe valer la fiabilidad de esa nueva componente a las 5000 horas?

**RESPUESTAS:**

Variable aleatoria: duración sin fallos (horas)

Fiabilidad a las t horas: P(duración > t)

Datos del problema:

Suceso: $D_1$ = "duración del dispositivo > 5000 horas"

Fiabilidad del dispositivo a las 5000 horas =

= P(duración del dispositivo > 5000) = 0,8

Objetivo:

Incrementar dicha fiabilidad al 95%, añadiendo una componente en paralelo.

¿Cuánto debe valor P(duración de nueva componente > 5000) para alcanzar este objetivo?

Sucesos:

C = "Duración de la nueva componente > 5000"

$D_2$ = "Duración del dispositivo tras añadir componente nueva > 5000"

$D_2 = D_1 \cup C$

$P(D_2) = P(D_1 \cup C) = P(D_1) + P(C) - P(D_1 \cap C) = 0,8 + P(C) – P(D_1) \times P(C) =$

$= 0,8 + P(C) – 0,8\ P(C) = 0,95$

$\Rightarrow 0,8 + 0,2\ P(C) = 0,95 \Rightarrow P(C) = (0,95-0,8)/0,2 = 0,75$

Por tanto la fiabilidad de la nueva componente en paralelo a las 5000 horas, debe ser del 75%.

**Problema 2.4.** En un dado, el resultado 3 sale con el doble de frecuencia y el número 5 con la mitad de frecuencia, que el resto de caras. ¿Cuánto vale la probabilidad de obtener una puntuación menor o igual que 3?

**RESPUESTAS:**

El conjunto de posibles valores de la variable

X= "puntuación obtenida al lanzar el dado"

Es:

E={1, 2, 3, 4, 5, 6}

Según indica el enunciado P(X=1) = p  P(X=2) = p  P(X=3) = 2p

P(X=4) = p  P(X=5) = p/2  P(X=6) = p

Como el suceso seguro E es la unión de los sucesos excluyentes:

$E=(X=1)\cup(X=2)\cup(X=3)\cup(X=4)\cup(X=5)\cup(X=6) \Rightarrow$

$P(E) = 1 = P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)+P(X=6) = 6{,}5p \Rightarrow$

p=1/6,5

Probabilidad de obtener una puntuación menor o igual a 3:

$P(X \leq 3) = P(X=1) + P(X=2) + P(X=3) = (4/6{,}5) = 4/6{,}5 = 0{,}6153 \Rightarrow$

$\Rightarrow$ **61,53% de probabilidad**

**Problema 2.5.** Se ha montado un circuito con 4 componentes en serie, siendo dos de tipo A y dos de tipo B (-A-A-B-B-). El 55% de las componentes tipo A falla antes de un año, y en el tipo B, ésto ocurre en un 36% de los casos.

a) Calcula la probabilidad de que el circuito falle antes de un año.

b) En un nuevo circuito, ¿cuántas componentes de tipo B hay que conectar en paralelo como mínimo, para obtener una probabilidad superior al 95% de que el dispositivo resultante funcione más de un año sin fallo?

**RESPUESTAS:**

a) Definición de sucesos
A= componente tipo A falla antes de un año
B= componente tipo B falla antes de un año
D= circuito -A-A-B-B- falla antes de un año

$D = A \cup A \cup B \cup B$
$P(D) = P(A \cup A \cup B \cup B) = 1- P(\overline{A \cup A \cup B \cup B}) =$
$= \text{Aplicando la ley de Morgan} = 1 - P(\bar{A} \cap \bar{A} \cap \overline{B} \cap \overline{B}) = 1 - (1-0{,}55)^2 \times (1-0{,}36)^2 =$
$= 0{,}917 \Rightarrow 91{,}7\%$

b) Definición de sucesos
$C_i$ = componente i de tipo B funciona más de un año

E = circuito con n componentes tipo B en paralelo funciona más de un año
Dato: P(E) > 0,95

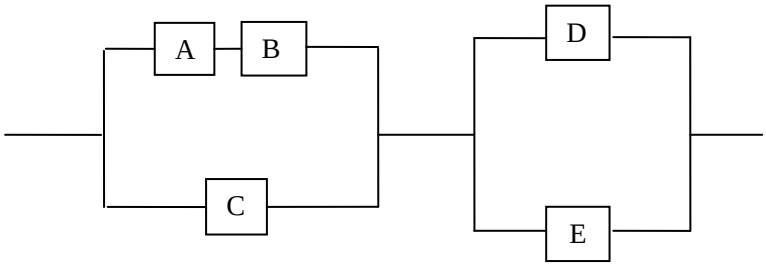
$E = C_1 \cup C_2 \cup \ldots \cup C_n$

$P(E) = P(C_1 \cup C_2 \cup \ldots \cup C_n) = 1 - P(\overline{C1 \cup C2 \cup \ldots \cup Cn}) =$
= Aplicando la ley de Morgan = $1 - P(\overline{C_1} \cap \overline{C_2} \cap \ldots \cap \overline{C_n}) = 1 - P(\overline{C_1})^n > 0,95$

$P(\overline{C_1}) = 0,36$

$1 - 0,36^n > 0,95 \Rightarrow 0,05 > 0,36^n \log 0,05 > n \log 0,36 \Rightarrow n > (\log 0,05)/(\log 0,36)$
$n > 2,93 \Rightarrow$ **n mínimo = 3**

**Problema 2.6.** El esquema muestra un dispositivo, en el que la fiabilidad a los tres años de A, B y C es 90%. Para el mismo periodo de tiempo, la de D y E es 98%. Calcula la fiabilidad del dispositivo a los tres años.




**RESPUESTAS:**

Definición de sucesos:

A= componente A dura más de tres años

B= componente B dura más de tres años

C= componente C dura más de tres años

D= componente D dura más de tres años

E= componente E dura más de tres años

DI= dispositivo dura más de tres años

$DI = [(A \cap B) \cup C] \cap (D \cup E)$

Fiabilidad del dispositivo a los tres años P(DI)

$P(DI) = P([(A \cap B) \cup C] \cap (D \cup E)) = P([(A \cap B) \cup C]) \times P(D \cup E) =$

$= 0{,}981 \times 0{,}9996 = 0{,}9806$

**Respuesta: 98,06%**

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

$= P(A) \times P(B) + P(C) - P(A) \times P(B) \times P(C) = 0{,}9^2 + 0{,}9 - 0{,}9^3 = 0{,}981$

$P(D \cup E) = 1 - P(\overline{D \cup E}) =$ Ley de Morgan $= 1 - P(\bar{D} \cap \bar{E}) =$

$= 1 - P(\bar{D}) \times P(\bar{E}) = 1 - 0{,}02^2 = 0{,}9996$

## 3. Distribuciones de Probabilidad

**Problema 3.1.** En una tienda de electrónica e informática, un 12% de los auriculares en venta supera los 60 euros.

a) Calcula la probabilidad de que en 5 seleccionados al azar, ninguno supere los 60 euros.

b) ¿Cuántos auriculares hay que seleccionar al azar como mínimo para que la probabilidad de que alguno supere los 60 euros, sea superior al 96%?

**RESPUESTAS:**

a) Definición de la variable

X= nº auriculares que supera los 60 euros en una muestra de cinco

Dato: P(un auricular tenga un precio > 60) = 0,12

Aplicando el modelo Binomial, X ~ B(n=5, p=0,12)

P(en cinco ninguno supere los 60 euros) = P(X = 0) = P(B(5, 0,12) = 0) =

$= \binom{5}{0} 0{,}12^0 0{,}88^5 = 0{,}5277 \Rightarrow 52{,}77\%$

b) n?

X= nº auriculares que supera los 60 euros en una muestra de n

X ~ B(n, p=0,12)

Datos: $P(X \geq 1) > 0{,}96 \Rightarrow P(X=0) < 0{,}04$

$\binom{n}{0} 0{,}12^0 0{,}88^n < 0{,}04 \Rightarrow 0{,}88^n < 0{,}04 \Rightarrow n \log 0{,}88 < \log 0{,}04 \Rightarrow$

$n > (\log 0{,}04)/(\log 0{,}88) = 25{,}18 \Rightarrow$ Habría que seleccionar al menos 26 auriculares.

**Problema 3.2.** En la asignatura de Estadística de ingeniería informática de una determinada universidad, en promedio suspende el 16% de los estudiantes.

a) Si hay un grupo en dicha asignatura de 40 alumnos, ¿cuánto vale la probabilidad de que aprueben todos?

b) ¿Cuánto vale la probabilidad de que suspendan entre 15 y 17, si el tamaño de grupo es 35?

**RESPUESTAS:**

a) Definición de la variable

X= nº alumnos que suspende en un grupo de 40

Datos: P(alumno suspenda) = 0,16$\Rightarrow$ X~B(n=40, p=0,16)

P(aprueben los 40 alumnos) = P(X= 0) = $\binom{40}{0} 0{,}16^0 0{,}84^{40} = 0{,}000936$

b) X= nº alumnos que suspende en un grupo de 35

Datos: P(alumno suspenda) = 0,16$\Rightarrow$ X~B(n=35, p=0,16)

P(15 $\leq$ X $\leq$ 17)= P(X=15) + P(X=16) + P(X=17) =

$= \binom{35}{15} 0{,}16^{15} 0{,}84^{20} + \binom{35}{16} 0{,}16^{16} 0{,}84^{19} + \binom{35}{17} 0{,}16^{17} 0{,}84^{18} =$

= 0,00011455 + 0,0000272737 + 0,00000580617 = 0,00014762987

**Problema 3.3.** El canal youtube de una institución recibe en promedio 8 visitas diarias, de usuarios de un determinado país.

a) Calcula el porcentaje de días que recibe 12 visitas.

b) Calcula la probabilidad de que reciba al menos 3 visitas diarias.

c) El video "La estadística y el reconocimiento de voz en los smartphones" es visionado en promedio 2 veces por día. Calcula la probabilidad de que en una semana sea visto más de 20 veces.

**RESPUESTAS:**

a) Definición de la variable

X= nº visitas diarias en el canal youtube

Dato: media = 8

Utilizando la distribución de Poisson

$$P(X=12) = e^{-8}\frac{8^{12}}{12!} = 0{,}048$$

porcentaje de días que recibe 12 visitas = 4,8%

b) $P(X \geq 3) = 1 - P(X \leq 2)$ = con el ábaco $\approx$ 1- 0,018 = 0,982

c) Definición de las variables:

Y= nº visionados diarios del vídeo

Z= nº visionados semanales del vídeo

Dato: media de Y=2 $\Rightarrow$ media de Z = 7 x 2 = 14

$P(Z > 20) = 1- P(Z \leq 20)$= con el ábaco $\approx$ 1- 0,95 = 0,05

**Problema 3.4.** Se controla la calidad de las partidas recibidas, de unidades de almacenamiento de datos USB, aplicando muestreo. Si en las unidades que se muestrean hay más de tres defectuosas, se rechaza la partida. ¿Cuántas unidades hay que muestrear con este control, para que la probabilidad de rechazar una partida, sea mayor del 98%, si hay al menos un 5% de defectuosas?

**RESPUESTAS:**

Definición de la variable:

X=nº de unidades defectuosas en la muestra de tamaño N

Datos:

Se rechaza la partida si X > 3

P(rechazar) > 0,98 si la P(defectuosa) $\geq$ 0,05

X sigue la distribución Binomial con parámetros N y p $\geq$ 0,05. Se resuelve el problema para el valor más difícil de detectar p=0,05.

Suponiendo que N será grande (>30), el problema se puede resolver también con la distribución de Poisson X~Poisson($\lambda$=N0,05)

P(rechazar) = P(X > 3) = P(Poisson($\lambda$) > 3) > 0,98 $\Rightarrow$ P(Poisson($\lambda$) $\leq$ 3) < 0,02 $\Rightarrow$

Con el ábaco, $\lambda$ > 9,2 $\Rightarrow$ N0,05 > 9,2 $\Rightarrow$ N > 9,2/0,05 = 184.

Por tanto, hay que muestrear al menos 185 unidades.

**Problema 3.5**. Una calculadora dispone de una función para generar valores aleatorios entre 0 y 2. Se ha programado de forma que la frecuencia de valores generados con la función en todo el rango, sea constante.

a) Calcula la función de densidad teórica. ¿Cuánto vale la varianza de esta variable?

b) ¿Qué porcentaje de valores está entre 0,35 y 0,72?

**RESPUESTAS:**

a)

Definición de la variable

X= valor resultante de ejecutar la función en la calculadora

Datos:

La frecuencia de valores generados por la función, es constante en el intervalo (0, 2) $\Rightarrow$

En la población de ejecuciones de dicha función, la función de densidad teórica, o densidad de probabilidad (frecuencia relativa) por unidad de longitud, es una constante c:

f(x) = c  Como en el intervalo (0, 2) está el total de valores:

$P(0 < X < 2) = 1 = \int_0^2 f(x)dx = \int_0^2 cdx = xc]_0^2 = c(2-0) \Rightarrow c = ½ = 0,5$

Está función de densidad coincide con la del modelo de probabilidad Uniforme en el intervalo (0,2).

Varianza de X en la población:

$\sigma^2 = E(X-m)^2$

La media $m = E(X) = \int_0^2 X0,5dx = 0,5\frac{X^2}{2}\Big]_0^2 = 0,5\ (2-0)^2/2 = 0,5\ 4/2 = 1$

$\sigma^2 = E(X-m)^2 = E(X-1)^2 = \int_0^2 (X-1)^2\ 0,5dx = 0,5\frac{(X-1)^3}{3}\Big]_0^2 =$

$= 0,5\ [(1/3)-(-1/3)] = 0,5\ 2/3 = 1/3$

b) Porcentaje de valores en el intervalo (0,35, 0,72)= longitud del intervalo multiplicada por la densidad de probabilidad por unidad de longitud f(x).

$P(0,35 < X < 0,72) = f(x) \times (0,72- 0,35) = 0,5 \times (0,72-0,35) = 0,185 \Rightarrow 18,5\%$

**Problema 3.6.** El tiempo de respuesta a un mensaje sigue distribución exponencial. El 95% de los mensajes tarda un tiempo menor o igual a 7 (u. t. unidades de tiempo).

a) Calcula el tiempo medio de respuesta.

b) Calcula la probabilidad de que un mensaje tenga un tiempo de respuesta mayor que 10 u.t.

c) Si en un mensaje enviado no se ha obtenido respuesta en 8 u.t., calcula la probabilidad de obtenerla en las 3 u.t. siguientes.

d) Calcula la probabilidad de que la suma de los tiempos de respuesta de 15 mensajes, sea superior a 50 u.t.

**RESPUESTAS:**

a) Definición de la variable

t: tiempo de respuesta de un mensaje

Datos: $t \sim Exp\ (\alpha)$

Porcentaje de mensajes con $t \leq 7 = P(t \leq 7) = 0,95$

$P(t \leq 7) = 1- e^{-\alpha 7} = 0,95 \Rightarrow e^{-\alpha 7} = 0,05 \Rightarrow -\alpha 7 = \ln 0,05 \Rightarrow \alpha = (-\ln 0,05)/7$

$\alpha = 0,4279$

media $= 1/\alpha = 1/0,4279 = 2,34$ u.t.

b) Probabilidad de que el tiempo de respuesta de un mensaje sea > 10 u.t. =

$= P(t > 10) = e^{-0,4279 \times 10} = 0,0138 \Rightarrow 1,38\ \%$

c) $P(t < 11/ t > 8)$

Se aplica la propiedad de que en el modelo exponencial, la distribución de $X/ X > x_0$ no depende de $x_0$ ("la dist. exponencial no tiene memoria").

$P(t < 11/ t > 8) = P(X < 3) = 1- e^{-0,4279 \times 3} = 0,723$

d) Definición de la variable

$T_{15}$= suma de los tiempos de respuesta de 15 mensajes

Aplicando las propiedades de la media y varianza

media = 15 x 2,34 = 35,1 u.t.

Suponiendo que los 15 tiempos de respuesta son independientes:

Varianza = 15 x $2{,}34^2$ = 82,134 $u.t.^2$ $\Rightarrow$ desviación típica = 9,06 u.t.

$T_{15}$ es la suma de 15 variables. Aplicando el Teorema Central del Límite,

$T_{15} \approx$ Normal (m=35,1, $\sigma$=9,06)

$P(T_{15} > 50) \approx P(N(35{,}1, 9{,}06) > 50) = P(N(0,1) > (50-35{,}1)/9{,}06)=$

$= P(N(0,1) > 1{,}64)$ = con la tabla de la Normal tipificada = 0,0505

**Problema 3.7.** El tiempo (minutos) de acceso a un fichero texto en red, en un aula informática, oscila con la distribución Normal de media m=3,87 y desviación típica $\sigma$=0,57.

a) Calcula la probabilidad de que el tiempo de acceso esté en el intervalo (1,25, 4).

b) Calcula la probabilidad de que se superen los 5 minutos al acceder al fichero.

c) Si se quiere optimizar el sistema para que la probabilidad anterior sea inferior a 0,1%, manteniendo la misma dispersión, ¿cuánto se tendría que tardar en promedio en acceder al fichero?

**RESPUESTAS:**

a) Definición de la variable:

t= tiempo de acceso al fichero

Datos t~N(m=3,87, $\sigma$=0,57)

$P(1{,}25 < t < 4) = P(1{,}25 < N(3{,}87, 0{,}57) < 4) =$

$=P((1{,}25-3{,}87)/0{,}57<N(0,1)<(4-3{,}87)/0{,}57) = P( -4{,}59 < N(0,1) < 0{,}23 ) =$

$= 1- P(N(0,1) > 0{,}23) – P(N(0,1) > 4{,}59)$ = utilizando la tabla de la $N(0,1)\approx$

$\approx 1- 0{,}4090 = 0{,}591 \Rightarrow 59{,}1\%$

b) $P(t > 5) = P(N(3,87, 0,57) > 5 ) = P(N(0,1) > (5-3,87)/0,57) =$

$= P(N(0,1) > 1,98) = 0,0238 \Rightarrow 2,38\%$

c) Datos :

$t \sim N(m, 0,57)$

$P(t > 5) < 0,001$

$P(N(m, 0,57) > 5) = P(N(0,1) > (5-m)/0,57) < 0,001$

Utilizando la tabla de la N(0,1):

$((5-m)/0,57) > 3,09 \Rightarrow 5-m > 0,57 \times 3,09 \Rightarrow m < 5 - 0,57 \times 3,09 \Rightarrow$

$m < 3,24$ minutos

**Problema 3.8.** Los vertidos mensuales de tóner (kg) en un ecoparque, siguen distribución Normal de media 69,7 y desviación típica 11,79.

a) Calcula la probabilidad de que los vertidos de tóner mensuales sean inferiores a 30 kg.

b) Calcula para esta variable un intervalo centrado en la media, de probabilidad 80%.

**RESPUESTAS:**

a) Definición de la variable

X= kg mensuales de vertidos tóner en el ecoparque

Datos: $X \sim N(m=60,7, \sigma=11,79)$

$P(X < 30) = P(N(0,1) < (30-60,7)/11,79) = P(N(0,1) < )$=con la tabla de la N(0,1)=

$= P(N(0,1) < -2,60) = 0,0047 \Rightarrow 0,47\%$

b) Datos:

$P( a < X < b) = 0,8$

Centro del intervalo [a, b]= 60,7

P(a < N(60,7, 11,79) < b) = 0,8

a = 60,7 – Zx11,79

b= 60,7 + Zx11,79

P(N(60,7, 11,79) >b) = P(N(60,7, 11,79) < a) =0,1

P(N(0,1) > (b-60,7)/11,79) = 0,1 $\Rightarrow$ (b-60,7)/11,79 $\approx$ 1,28 = Z

Intervalo:

a = 60,7 – Zx11,79 = 60,7 – 1,28x11,79= 45,61 Kg

b= 60,7 + Zx11,79 = 60,7 + 1,28x11,79= 75,79 kg

**Problema 3.9.** En una oficina municipal de información y defensa del consumidor, se ha registrado un promedio mensual de 194 reclamaciones y denuncias, relacionadas con telefonía e internet.

a) Calcula la probabilidad de que en un mes se registren más de 200 reclamaciones y denuncias.

b) Para reducir la probabilidad anterior al 1%, ¿cuánto debe valer el promedio mensual de esta variable?

**RESPUESTAS:**

a) Definición de la variable

X= nº de reclamaciones y denuncias mensuales relacionadas con telefonía e internet

Dato:

media de X = 194

Suponiendo distribución Poisson para esta variable discreta, X ~ Poisson ($\lambda$=194)

Como $\lambda$ > 9, se puede aproximar el cálculo con la distribución Normal de media m= 194 y desviación típica $\lambda=\sqrt{194}$ = 13,93

P(X > 200) $\approx$con corrección de continuidad $\approx$ P(N(194, 13,93) > 200,5) =

= P(N(0,1) > (200,5-194)/13,93) =

= P(N(0,1) > 0,47) = 0,3192 $\Rightarrow$ 31,92%

b) Datos:

$X \sim$ Poisson ($\lambda$)

$P(X > 200) = 0{,}01$

$P(X > 200) \approx P(N(\lambda, \lambda^{1/2}) > 200{,}5) = P(N(0,1) > (200{,}5-\lambda)/\lambda^{1/2}) = 0{,}01$

Con la tabla Normal(0,1),

$(200{,}5-\lambda)/\lambda^{1/2} \approx 2{,}33 \Rightarrow \lambda + 2{,}33\,\lambda^{1/2} - 200{,}5 = 0$

Resolviendo la ecuación de segundo grado,

$\lambda^{1/2} \approx 13 \Rightarrow \lambda \approx 169$

Por tanto, mensualmente tendrían que haber en promedio 169 reclamaciones y denuncias, para superar las 200 con probabilidad 1%.

## 4. Inferencia en poblaciones normales

**Problema 4.1.** El tiempo de transferencia (u.t.) de paquetes de un tamaño concreto, a través de la red, sigue una distribución N(m, $\sigma$=4).

a) Se extrae una muestra de tamaño 25. Calcula la probabilidad de que la diferencia en valor absoluto, entre la media poblacional y la media muestral sea mayor que 2 u.t..

b) Calcula el tamaño de muestra necesario, para que la probabilidad planteada en el apartado anterior, sea inferior al 0,1%.

**RESPUESTAS:**

a) Definición de la variable:

X= tiempo de transferencia

Datos:

$X \sim N(m, \sigma=4) \Rightarrow$

La media $\overline{X}$ de una muestra aleatoria simple de tamaño 25 es $N(m, \sigma=4/5) \Rightarrow$

$(\overline{X} - m) \sim N(0, \sigma=4/5)$

$P(|\overline{X} - m| > 2) = P(|N(0,4/5)| > 2) = P(N(0, 4/5) > 2) + P(N(0, 4/5) < -2) =$

$= 2P(N(0,4/5) > 2) = 2\,P(N(0,1) > 10/4) = 2\,P(N(0,1) > 2,5) =$ con la tabla de la N(0,1)

$= 2 \times 0,00621 = 0,01242 \Rightarrow$

1,242%

b) Datos:

$(\overline{X} - m) \sim N(0, \sigma=4/\sqrt{N})$ $\quad P(|\overline{X} - m| > 2) < 0,001$

$P(|\overline{X} - m| > 2) = P(|N(0,4/\sqrt{N})| > 2) = P(N(0, 4/\sqrt{N}) > 2) + P(N(0, 4\sqrt{N}/) < -2) =$

$= 2P(N(0, 4/\sqrt{N}) > 2) = 2\,P(N(0,1) > 2x\sqrt{N}/4) < 0,001$

A partir de la tabla de la N(0,1)

$2x\sqrt{N}/4 > 3,3 \Rightarrow N > (3,3 \times 2)^2 = 43,56$

Por tanto, el tamaño de muestra N tendrá que ser ≥ 44 unidades.

**Problema 4.2.** Calcular la probabilidad de que una $\chi^2$ con 20 grados de libertad sea mayor que 34,17.
a) A partir de la tabla de dicha distribución.
b) Aproximando con la distribución Normal.

**RESPUESTAS:**

a)
Datos: 20 grados de libertad

$P(\chi^2_{20} > 34,17) = 0,025$

b)

$\chi^2_{20}$ media = 20 desviación típica = $\sqrt{40}$

$P(\chi^2 > 34,17) \approx P(N(20, 6,32) > 34,17) = P(N(0, 1) > \frac{34,17-20}{6,32}) = P(N(0, 1) > 2,24) =$
= Con la tabla de la N(0,1) = 0,0125

**Problema 4.3.** La velocidad de ejecución de un procesador, sigue la distribución Normal. La varianza vale $\sigma^2 = 9$. Si se realizan 10 observaciones de esta variable, ¿cuánto vale la probabilidad de que la varianza de esa muestra sea mayor que 18?

**RESPUESTAS:**

Datos:

Velocidad ~ $N(m, \sigma^2 = 9)$

N= 10 observaciones

Se aplica el resultado $(N-1)\frac{s^2}{\sigma^2} \sim \chi^2$ con N-1 grados de libertad.

En este caso $9\frac{s^2}{9} \sim \chi^2$ con 9 grados de libertad

$P(s^2 > 18) = P(\chi^2_9 > 18)$

Buscando en la tabla de la distribución Chi-Cuadrado, en la fila de 9 grados de libertad, se tiene:

$P(\chi^2_9 > 19{,}023)=0{,}025$

$P(\chi^2_9 > 16{,}919)=0{,}05$

Por tanto, la probabilidad estará entre los dos valores:

$0{,}025<P(\chi^2_9 > 18)>0{,}05$

(NOTA: calculada con el programa Statgraphics da el valor $P(\chi^2_9 > 18)=0{,}035$).

**Problema 4.4.** Calcula un valor de la t de Student con 15 grados de libertad, que sea superado en valor absoluto con una probabilidad del 10%.

**RESPUESTAS:**

Datos: t de Student con 15 grados de libertad

$P(|t_{15}| > t) = 0{,}1$

$P(|t_{15}| > t)= P(t_{15} > t) + P(t_{15} < -t) = 2\,P(t_{15} > t) = 0{,}1$

$\Rightarrow P(t_{15} > t) = 0{,}05$

Con la tabla de la t de Student, en la fila de 15 grados de libertad, y la columna $\alpha/2=0{,}05$ $\Rightarrow t = 1{,}753$

**Problema 4.5.** Calcula dos valores de la t de Student con 12 grados de libertad, centrados en la media, entre los que se encuentre el 80% de la distribución.

**RESPUESTAS:**

Datos: t de Student con 12 grados de libertad

$P(-t < t_{12} < t)=0{,}8$

$\Rightarrow P(t_{12} > t) = 0{,}1$

Con la tabla de la distribución, y en la columna $\alpha/2=0{,}1 \Rightarrow t =1{,}356$.

Por tanto el intervalo del 80% de probabilidad es [-1,356, 1,356]

**Problema 4.6.** El tiempo de respuesta de un sistema informático, sigue la distribución Normal. Si se extraen dos muestras de $N_1=10$ y $N_2=18$ observaciones, calcula la probabilidad de que la varianza de la primera muestra resulte más de 2,5 el valor de la varianza de la segunda muestra.

**RESPUESTAS:**

Definición de la variable:
X = tiempo de respuesta

Datos:
X ~ Normal
Muestra 1 con 10 observaciones
Muestra 2 con 18 observaciones
$\frac{s_1^2}{s_2^2} \sim F_{9,17}$

$P(s_1^2 > 2{,}5\ s_2^2) = P(\frac{s_1^2}{s_2^2} > 2{,}5) = P(F_{9,17} > 2{,}5) \approx$ con la tabla de la distribución $F \approx 0{,}05$

**Problema 4.5.** La presentación oral en clase de estadística, de un alumno de ingeniería informática, fue evaluada por una muestra de compañeros, con las siguientes notas:

10 10 7 9 8 8 8 6,5 8,79 10 9 7 9,75 9 9,5 8 8,5 7,2 8,1 8 10

a) Calcula un intervalo de confianza para la nota media de evaluación m (nivel de confianza 95%). ¿Se puede admitir una nota media m=10 en la evaluación realizada por todos los compañeros de la asignatura? Justifica la respuesta.

b) Calcula un intervalo de confianza para la desviación típica σ de las notas (nivel de confianza 90%). ¿Es admisible una desviación típica σ=1?

c) Las notas de otro alumno de la asignatura, evaluadas por una segunda muestra de compañeros, en otra sesión de exposición oral de trabajos prácticos, fueron:

5 5 7 5 5 7 5 7 7 8 9 7 7 4 6,5 9 7 6,5 7 6,9 6,5 6

Estudia si se puede admitir que las dos varianzas poblacionales son iguales (riesgo de primera especie α=5%).

**RESPUESTAS:**

a)
Definición de la variable:
X = nota del alumno en la exposición oral, evaluada por sus compañeros

Datos:
N= 21
Media muestra $\overline{X} = 8,54$
Desviación típica muestra s= 1,0946

Nivel de confianza = 0,95 $\Rightarrow \alpha = 0,05$

Intervalo de confianza (al 95%) $[\overline{X} - t_{20}^{\alpha=0,05} \frac{s}{\sqrt{N}},\ \overline{X} + t_{20}^{\alpha=0,05} \frac{s}{\sqrt{N}}]$

$$[8,54 - 2,086\frac{1,0946}{\sqrt{21}},\ 8,54 + 2,086\frac{1,0946}{\sqrt{21}}]$$

El intervalo es: [8,04, 9,04]

Para estudiar si es admisible una nota media m = 10, en la evaluación realizada por todos los compañeros, se plantea el contraste de hipótesis:

$H_0 : m = 10$

El valor m=10 no está en el intervalo de confianza. Por tanto, con un nivel de confianza del 95%, no es admisible una media 10, sino inferior.

b) Datos: nivel de confianza para el intervalo 90%

$$\left[\sqrt{(N-1)\frac{s^2}{g_2}}, \sqrt{(N-1)\frac{s^2}{g_1}}\right] = \left[\sqrt{(21-1)\frac{1,0946^2}{31,41}}, \sqrt{(21-1)\frac{1,0946^2}{10,851}}\right] = [0,87,\ 1,49]$$

En la tabla $\chi^2_{20}$ con la columna $1 - \alpha/2 = 0,95$, $g_1 = 10,851$, y en la de $\alpha/2 = 0,05$, $g_2 = 31,41$

$H_0 : \sigma = 1 \in$ Intervalo de Confianza $\Rightarrow$ con un riesgo de primera especie del 10%, se puede aceptar para la dispersión $\sigma = 1$.

c)

Datos:

Segunda muestra, tamaño = 22

Varianza de la segunda muestra $s^2_2 = 1{,}65$

$$\left[\frac{s_1^2/s_2^2}{f_2}, \frac{s_1^2/s_2^2}{f_1}\right] = \left[\frac{1{,}198/1{,}65}{2{,}42}, \frac{1{,}198/1{,}65}{0{,}41}\right] = [\,0{,}299,\ 1{,}778\,]$$

Con la tabla F, y 20 y 21 grados de libertad, para $0{,}025 \Rightarrow f_2 = 2{,}42$, y $0{,}975 \Rightarrow f_1 = 0{,}41$.

$H_0$ : varianzas iguales $\Rightarrow\ \sigma_1^2 = \sigma_2^2 \Rightarrow \sigma_1^2/\sigma_2^2 = 1$

Como $\sigma_1^2/\sigma_2^2 = 1 \in$ Intervalo de Confianza, es admisible que las varianzas son iguales con un nivel de confianza del 95%.

**Problema 4.6.** En la gestión ambiental de un campus universitario, se registra la generación anual en Kg/persona, de residuos de aparatos eléctricos y electrónicos (RAEE). Se han observado los siguientes valores:

0,7 0,55 0,64 0,7 0,83 1,1 1,25 0,97 0,65

Calcula los intervalos de confianza para la media y la desviación típica de este indicador (nivel de confianza 99%)

**RESPUESTAS:**

Definición de la variable

X = residuos tipo RAEE generados anualmente (Kg/persona)

Datos:

N=9

Media muestral $\overline{X} = 0{,}82$

Desviación típica muestral s = 0,237

Varianza muestral $s^2 = 0{,}056$

Nivel de confianza 99%

Intervalo para la media m

$$[\overline{X} - t_8^{\alpha=0,01}\frac{s}{\sqrt{N}},\ \ \overline{X} + t_8^{\alpha=0,01}\frac{s}{\sqrt{N}}]$$

$$[0{,}82 - 3{,}355\frac{0{,}237}{\sqrt{9}},\ \ 0{,}82 + 3{,}355\frac{0{,}237}{\sqrt{9}}]$$

[ 0,56, 1,09]

Intervalo para la desviación típica σ

$$\left[\sqrt{(N-1)\frac{s^2}{g_2}}, \sqrt{(N-1)\frac{s^2}{g_1}}\right] = \left[\sqrt{(9-1)\frac{0{,}056}{21{,}995}}, \sqrt{(9-1)\frac{0{,}056}{1{,}344}}\right] = [0{,}14,\ 0{,}58]$$

En la tabla $\chi^2_8$ con la columna 1 - α/2 = 0,995, $g_1$ = 1,344, y en la de α/2 = 0,005, $g_2$=21,995.

## 5. Análisis de la Varianza (ANOVA)

**Problema 5.1.** Para comparar la velocidad de dos modelos de procesador, se han observado 9 valores con el A y otros 9 con el B. Los datos son:

| | |
|---|---|
| Procesador A | 51,6 52,1 51,5 53,1 51,5 50,6 52,5 52,2 53,1 |
| Procesador B | 60,0 58,1 58,2 60,6 59,3 58,3 58,1 58,5 58,6 |

a) ¿Difieren las velocidades medias de los modelos de procesador ($\alpha = 5\%$)?

NOTA: SCTotal=222,083, SCProcesador=210,125.

b) Calcula el intervalo de confianza para el cociente de varianzas, y estudia si hay diferencia significativa entre ellas ($\alpha = 1\%$).

**RESPUESTAS:**

a)
Datos:
SCTotal=222,083 SCProcesador=210,125

Riesgo de primera especie $\alpha = 5\%$

Grados de libertad totales 18-1=17

Grados de libertad del efecto del modelo de procesador 2-1=1

La SCResidual = SCTotal – SCProcesador = 222,083-210,125 = 11,958 con 17-1=16 grados de libertad.

La tabla resumen del ANOVA queda:

| O.Variabilidad | Sumas de Cuadrados | Grados de libertad | Cuadrados Medios | F-ratio |
|---|---|---|---|---|
| Procesador | 210,125 | 1 | 210,125 | 281,14 |
| Residual | 11,958 | 16 | 0,7474 | |
| Total | 222,083 | 17 | | |

La F de tabla con 1 y 16 grados de libertad para $\alpha$=5% resulta igual a 4,49.

Como F-Ratio=281,14 > F-Tabla $\Rightarrow$ el efecto de procesador sobre la velocidad media es significativo. Por tanto, difieren significativamente las dos velocidades medias.

b) Datos:

nivel de confianza = 99%

Varianza primera muestra $s^2_1 = 0,67$ grados de libertad=8

Varianza segunda muestra $s^2_2 = 0,83$ grados de libertad=8

Intervalo de confianza para el cociente de varianzas

$$\left[\frac{s_1^2/s_2^2}{f_2}, \frac{s_1^2/s_2^2}{f_1}\right] = \left[\frac{0,67/0,83}{7,5}, \frac{0,67/0,83}{0,13}\right] = [\,0,18,\ 3,57]$$

En la tabla de la distribución $F_{8,8}$, con los valores 1-α/2=0,995, y α/2=0,005, se obtienen $f_1 = 0,13$ y $f_2 = 7,5$.

$H_0$ : varianzas iguales $\Rightarrow$ $\sigma_1^2 = \sigma_2^2 \Rightarrow \sigma_1^2/\sigma_2^2 = 1$

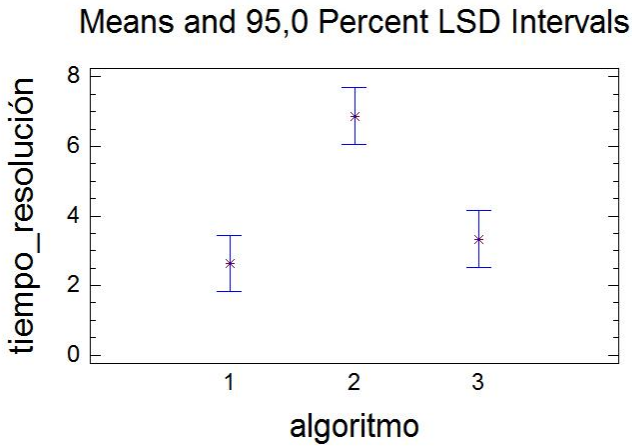
Como $\sigma_1^2/\sigma_2^2 = 1 \in$ Intervalo de Confianza, es admisible que las varianzas no difieren significativamente, con un nivel de confianza del 99%.

**Problema 5.2.** Se compara la rapidez de tres algoritmos de inversión de matrices. Los tiempos de resolución de este problema son:

| | |
|---|---|
| Algoritmo 1 | 3,6 3,9 1,9 3,9 1,7 0,8 |
| Algoritmo 2 | 6,3 6,5 8,7 7,2 4,6 7,9 |
| Algoritmo 3 | 3,5 1,8 2,3 5,3 3,6 3,5 |

a) Estudia con el ANOVA si el factor algoritmo influye significativamente sobre la media del tiempo de resolución.
Datos: SCTotal=88,2511, SCAlgoritmo=61,7911, riesgo de primera especie α=5%.

b) Interpreta el gráfico de intervalos LSD.




**RESPUESTAS:**

a)

Datos:
SCTotal = 88,2511
SCAlgoritmo = 61,7911

18 - 1 = 17 grados de libertad totales

3 – 1 = 2 grados de libertad del efecto de algoritmo

La SCResidual = SCTotal - SCAlgoritmo = 88,2511-61,7911=26,46 con 17-2=15 grados de libertad.

La Tabla resumen del ANOVA queda:

| O.Variabilidad | Sumas de Cuadrados | Grados de libertad | Cuadrados Medios | F-ratio |
|---|---|---|---|---|
| Algoritmo | 61,7911 | 2 | 30,8955 | 17,51 |
| Residual | 26,46 | 15 | 1,764 | |
| Total | 88,2511 | 17 | | |

La F de la tabla, con 2 y 15 grados de libertad para α=5% resulta igual a 3,68. Como F-Ratio=17,51 > F-Tabla, entonces el efecto de algoritmo sobre el tiempo medio de resolución es significativo.

b) En el gráfico de intervalos LSD se aprecia que con el algoritmo 2 el tiempo medio de resolución es significativamente mayor que con los algoritmos 1 y 3. Entre los algoritmos 1 y 3 no hay diferencias significativas (los intervalos se solapan).

**Problema 5.3.** Se han ensayado tres configuraciones de distribución de ficheros (A, B y C), combinadas con tres tamaños de espacio reservado (bajo, medio, alto). Hay dos repeticiones por condición operativa (la tabla adjunta da el promedio de los dos datos). Se ha medido el tiempo de acceso (u.t.):

| | **bajo** | **medio** | **alto** |
|---|---|---|---|
| **A** | 20 | 22 | 24,5 |
| **B** | 25 | 28 | 32 |
| **C** | 21 | 25 | 24 |

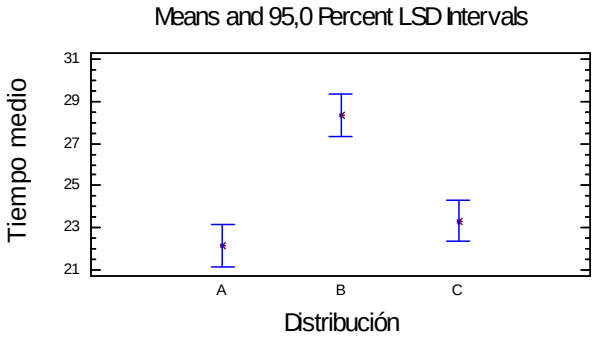
Algunos resultados del ANOVA son:

```
Analysis of Variance for Tiempo medio
-------------------------------------------------------------------------------
Source            Sum of Squares     Df    Mean Square    F-Ratio    F-tabla
-------------------------------------------------------------------------------
MAIN EFFECTS
 A:Espacio reservad   71,4444
 B:Distribución      128,78
INTERACTIONS
 AB                   15,55

RESIDUAL
-------------------------------------------------------------------------------
TOTAL                 236,278
-------------------------------------------------------------------------------
```

a) Completa la tabla ANOVA anterior, y estudia la significación de los efectos simples y de la interacción doble. Justifica los resultados con un riesgo de 1ª especie α=5%.

b) ¿Qué niveles del factor distribución difieren significativamente? Justifica la respuesta utilizando el gráfico LSD siguiente.




c) Representa el gráfico de medias según niveles del factor espacio reservado. ¿Cómo es el efecto de espacio reservado sobre el tiempo medio?

d) En un ANOVA, ¿qué miden las sumas de cuadrados?

**REPUESTAS:**

a)

3 - 1 = 2 grados de libertad del efecto del factor espacio reservado.

3 - 1 = 2 grados de libertad del efecto del factor distribución.

(3-1)x(3-1) = 4 grados de libertad de la interacción.

18-1 = 17 grados de libertad totales.

$SC_{residual}$ = SCtotal - $SC_{espacio\ reservado}$ - $SC_{distribución}$ - $SC_{espacio\ reservado\ x\ distribución}$ = 20,5

Con 17-2-2-4=9 grados de libertad

Analysis of Variance for Tiempo medio

| Source | Sum of Squares | Df | Mean Square | F-Ratio | P-Value |
|---|---|---|---|---|---|
| MAIN EFFECTS | | | | | |
| A:Espacio reservad | 71,4444 | 2 | 35,7222 | 15,68 | <0,05 |
| B:Distribución | 128,78 | 2 | 64,39 | 28,27 | <0,05 |
| INTERACTIONS | | | | | |
| AB | 15,55 | 4 | 3,8875 | 1,71 | >0,05 |
| RESIDUAL | 20,5 | 9 | 2,27778 | | |
| TOTAL | 236,278 | 17 | | | |

$F_{2,9}^{\alpha=0,05}$= 4, 26 $F_{4,9}^{\alpha=0,05}$= 3,63

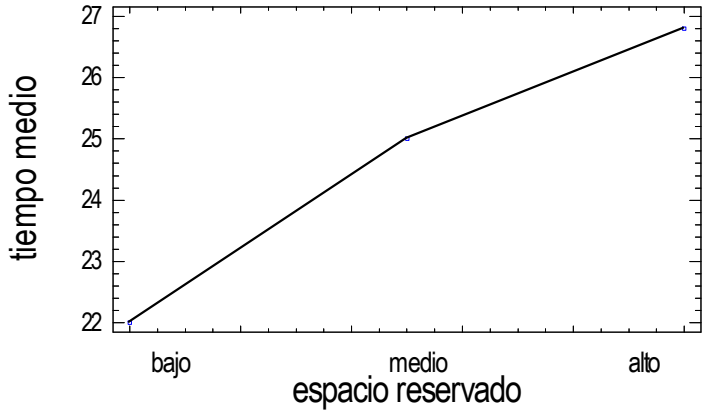
F-ratio de espacio reservado= 15,68 > 4,26 Efecto significativo

F-ratio de distribución = 28,27 > 4,26 Efecto significativo

F-ratio interacción = 1,71 < 3,63 Efecto no significativo

b) Se observa en el gráfico de intervalos LSD del factor distribución, que el tiempo medio de acceso difiere significativamente entre las distribuciones tipo A y B, siendo menor con tipo A. También difieren B y C, pero no A y C (en este tercer caso aparecen solapados los intervalos). La distribución con un tiempo de acceso medio significativamente superior es la B.

c) El gráfico con las medias observadas para cada nivel de espacio reservado es:




Se observa que cuando aumenta el espacio reservado, aumenta el tiempo medio de acceso de forma lineal.

d) La suma de cuadrados total está indicando la variabilidad de todos los datos del análisis. Se calcula con la suma de los cuadrados de las desviaciones de cada uno de los datos respecto de la media general. La suma de cuadrados de un efecto permite estimar la variabilidad que hay asociada a él. La suma de cuadrados residual se utiliza para estimar la variabilidad que hay en los datos con cada combinación de los factores. Se asume en el ANOVA, que esta variabilidad no depende de los niveles de los factores. Se calcula como la suma de los cuadrados de las desviaciones de cada dato respecto de la media de la condición operativa en que se ha obtenido.

**Problema 5.4.** En la gestión ambiental de una universidad, se registra la generación anual en Kg/persona, de residuos de aparatos eléctricos y electrónicos (RAEE). Se han observado los siguientes valores, en tres campus y cuatro periodos de tiempo:

| | Campus 1 | Campus 2 | Campus 3 |
|---|---|---|---|
| Periodo 1 | 0,18 | 0,77 | 0,70 |
| Periodo 2 | 0,68 | 0,47 | 0,64 |
| Periodo 3 | 0,27 | 0,19 | 0,55 |
| Periodo 4 | 0,31 | 0,57 | 0,69 |

Estudia con el ANOVA si hay diferencias significativas entre campus y/o periodos (α=5%). SCtotal = 0,496767; SCcampus = 0,162467; SCresidual = 0,217133.

**RESPUESTAS:**

Datos:

SCtotal = 0,496767 SCcampus = 0,162467 SCresidual = 0,217133

α=5%

SCperiodo= 0,496767 - 0,162467 - 0,217133 = 0,117167

Grados de libertad 4-1 = 3

Grados de libertad de campus = 3 - 1 = 2

Grados de libertad totales = 12 – 1 = 11

Grados de libertad residuales = 11- 3 – 2 = 6

La Tabla resumen del ANOVA queda:

| O.Variabilidad | Sumas de Cuadrados | Grados de libertad | Cuadrados Medios | F-ratio |
|---|---|---|---|---|
| Periodo | 0,117167 | 3 | 0,0812333 | 2,24 |
| Campus | 0,162467 | 2 | 0,0390556 | 1,08 |
| Residual | 0,217133 | 6 | 0,0361889 | |
| Total | 0,496767 | 11 | | |

F de tabla

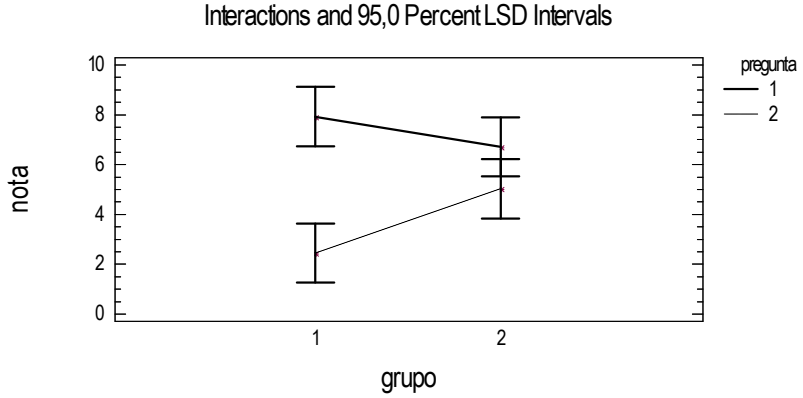
$F_{3,5}^{\alpha=0,05} = 4,76 > 2,24 \Rightarrow$ No hay diferencias significativas entre periodos

$F_{2,5}^{\alpha=0,05} = 5,14 > 1,08 \Rightarrow$ No hay diferencias significativas entre campus

**Problema 5.5.** Las notas de dos preguntas del primer parcial (descriptiva: 1, distribuciones discretas: 2), en dos grupos distintos, han dado como resultados: SCtotal = 489,588, SCgrupo = 5,18205, SCpregunta = 141,482, SCresidual = 303,404. Hay 44 datos en total.

a) Estudia si hay diferencias significativas en la nota media, según grupo y pregunta. Analiza la significación de la interacción entre estos dos factores. Riesgo de primera especie $\alpha = 5\%$.

b) ¿Qué indica el gráfico LSD de la interacción respecto a los dos efectos?




**RESPUESTAS:**

a) Datos:

SCtotal = 489,588, SCgrupo = 5,18205, SCpregunta = 141,482, SCresidual = 303,404

44 datos

Grados de libertad totales 44 - 1 = 43

Grados de libertad del efecto grupo 2-1 = 1

Grados de libertad del efecto pregunta 2-1 = 1

SCgrupo x pregunta = 489,588 - 5,18205 - 141,482 - 303,404 = 39,5202

Grados de libertad de la interacción (2-1) x (2-1) = 1

Analysis of Variance for nota

| Source | Sum of Squares | Df | Mean Square | F-Ratio | P-Value |
|---|---|---|---|---|---|
| MAIN EFFECTS | | | | | |
| A:grupo | 5,18205 | 1 | 5,18205 | 0,68 | >0,05 |
| B:pregunta | 141,482 | 1 | 141,482 | 18,65 | <0,05 |
| INTERACTIONS | | | | | |
| AB | 39,5202 | 1 | 39,5202 | 5,21 | <0,05 |
| RESIDUAL | 303,404 | 40 | 7,58509 | | |
| TOTAL | 489,588 | 43 | | | |

F de tablas,
$F_{1,\,30}^{\alpha=5\%} = 4{,}17$
$F_{1,\,50}^{\alpha=5\%} = 4{,}03$
Factor grupo: F-ratio = 0,68 < F-tabla $\Rightarrow$ No hay diferencia entre ambos grupos, en el promedio de nota.

Factor pregunta: F-ratio = 19,85 > F-tabla $\Rightarrow$ La nota media de las dos preguntas difiere significativamente.

Efecto interacción grupo x pregunta: F-ratio = 5,21 > F-tabla $\Rightarrow$ La interacción entre los dos factores es significativa $\Rightarrow$ la diferencia de las notas medias de las dos preguntas, cambia significativamente según el grupo.

b) En el grupo 1 es significativamente mayor la nota media en la pregunta de descriptiva. En el grupo 2 no hay diferencia significativa en la nota media de las dos preguntas, que es significativamente superior a la obtenida por el grupo 1 en descriptiva.

## 6. Regresión Lineal Simple

**Problema 6.1.** En una tienda de electrónica e informática, se ha registrado el precio (euros) y nº de unidades de distintos modelos de auriculares.

a) Define la población y la variable aleatoria en estudio.

b) La matriz de varianzas-covarianzas es:

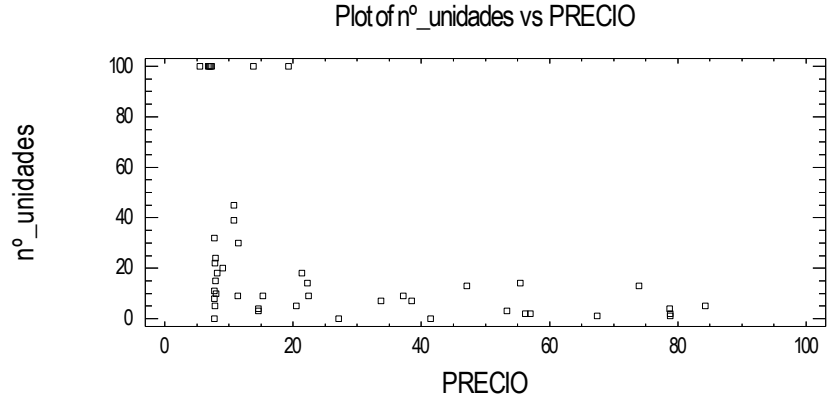
Covariances

| | n°_unidades | PRECIO |
|---|---|---|
| n°_unidades | 1155,74<br>( 45) | -363,201<br>( 45) |
| PRECIO | -363,201<br>( 45) | 616,769<br>( 45) |

Covariance
(Sample Size)

¿Cómo es la relación entre estas dos variables? Calcula el coeficiente de correlación.

c) El diagrama de dispersión de los datos es el siguiente. ¿Hay relación lineal entre precio y nº de unidades?



**RESPUESTAS:**

a) Población: modelos de auriculares

Variable aleatoria: bidimensional (precio, nº unidades disponibles en tienda)

b) La covarianza vale = -363,201, por lo que la relación es negativa.

El coeficiente de correlación lineal

$r = \text{covarianza}/(s_{\text{precio}}\ s_{\text{nºunidades}}) =$

$= -363{,}201/(1155{,}74 \times 616{,}769)^{1/2} = -0{,}4302$

c) En el gráfico de dispersión, se observa que no hay relación lineal.

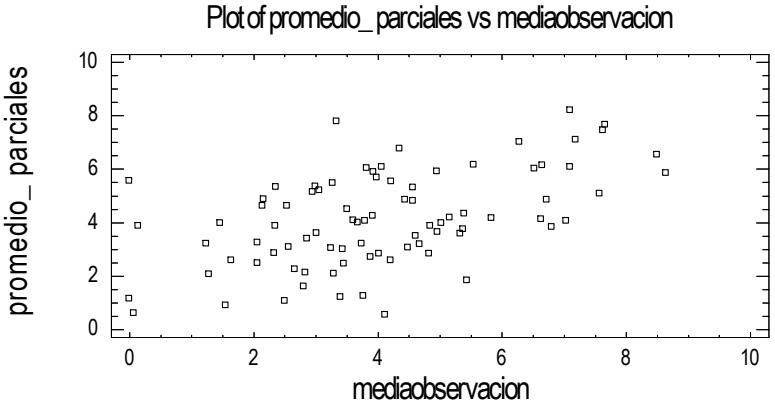
**Problema 6.2.** La matriz de varianzas-covarianzas de la variable bidimensional (nota de observación, nota promedio de parciales), de una muestra de alumnos de la asignatura Estadística de ingeniería informática.

```
Covariances
                    mediaobservacion    promedio_ parciales
-----------------------------------------------------------
mediaobservacion      3,84434               1,80919
                       (   84)              (   84)

promedio_ parciales   1,80919               3,12715
                       (   84)              (   84)
-----------------------------------------------------------
Covariance
(Sample Size)
```

a) ¿Cuánto vale el coeficiente de correlación lineal?

b) El diagrama de dispersión de la muestra es:




¿Cómo es la relación entre la nota media de observación y el promedio de los parciales?

¿Qué porcentaje de la variabilidad observada en el promedio de los parciales, está asociada a la nota de observación?

c) La estimación de la recta de regresión, para predecir el promedio de los parciales en función de la nota de observación, es:

```
Multiple Regression Analysis
------------------------------------------------------------
Dependent variable: promedio_ parciales
------------------------------------------------------------
                          Standard          T
Parameter    Estimate       Error       Statistic    P-Value
------------------------------------------------------------
CONSTANT         2,20454    0,383478
mediaobservacion 0,470612   0,0849653
------------------------------------------------------------
```

Escribe la recta estimada. Estudia la significación de la ordenada y la pendiente (α=5%).

Interpreta lo que representan en este caso.

**RESPUESTAS:**

a) El coeficiente de correlación lineal

r = covarianza/($s_{precio}$ $s_{nºunidades}$) = 1,80919/(3,84434x3,12715)$^{1/2}$ = 0,5218

b) Es una relación lineal positiva débil, tal y como se observa con el valor de r y con el gráfico.

%variabilidad = $r^2$ x 100 = $0,5218^2$ x 100 = 27,23%

c) La recta estimada es:

E(promedio_parciales/notaobservacion) = 2,20454 + 0,47612 notaobservacion

Nº datos = 84 (de la matriz de varianzas-covarianzas)

Grados de libertad residuales = 84 - 1 - 1 = 82

$t_{82}^{\alpha=0,05}$ entre 1,98 ($t_{120}^{\alpha=0,05}$) y 2 ($t_{60}^{\alpha=0,05}$)

Para la ordenada la |t-calculada| = |2,20454/0,383478| = 5,74882 > t-tabla

La ordenada es estadísticamente significativa. La estimación es 2,20, que representa en este problema, la nota promedio de los parciales de alumnos que tienen un cero en la nota de observación.

Para la pendiente la |t-calculada| = |0,470612/0,0849653| = 5,53887 > t-tabla

La pendiente es estadísticamente significativa. Su estimación es 0,47, y representa la diferencia en el promedio de los parciales entre alumnos con una unidad de diferencia en la nota de observación.

**Problema 6.3.** En el sistema de gestión ambiental de un campus universitario, se han registrado los kg/persona de residuos de papel y residuos de aparatos eléctricos y electrónicos (RAEE), generados anualmente.

El ajuste de regresión lineal simple, para predecir los RAEE en función de los residuos de papel generados es:

```
Regression Analysis - Linear model: Y = a + b*X
-------------------------------------------------------------
Dependent variable: RAEE
Independent variable: Papel
-------------------------------------------------------------
                      Standard        T
Parameter    Estimate    Error      Statistic     P-Value
-------------------------------------------------------------
Intercept    0,322623   0,20709      1,55789      0,1632
Slope        0,0854063  0,0338999    2,51937      0,0398
-------------------------------------------------------------
```

Analysis of Variance

| Source | Sum of Squares | Df | Mean Square | F-Ratio | P-Value |
|---|---|---|---|---|---|
| Model | 0,213467 | 1 | 0,213467 | 6,35 | 0,0398 |
| Residual | 0,235421 | 7 | 0,0336316 | | |
| Total (Corr.) | 0,448889 | 8 | | | |

Correlation Coefficient = 0,689599
R-squared = 47,5546 percent

a) ¿Cuánto valen las estimaciones de la ordenada y de la pendiente? ¿Son significativas? ($\alpha$ = 5%) ¿Qué representan?

b) ¿Cuánto vale el coeficiente de determinación? ¿Qué cuantifica?

c) Calcula el intervalo de probabilidad 90%, para los RAEE, generados por personas que generan anualmente 5 Kg de residuos de papel.

**RESPUESTAS:**

a) La estimación de la ordenada es 0,322623 kg/persona. No es significativa, ya que p-value = 0,1632 > $\alpha$ = 0,05. Esto indica que las personas que generan 0 kg de residuos de papel por año, generan en promedio 0 kg de RAEE por año.

La estimación de la pendiente es 0,0854063. Es significativa, ya que p-value = 0,0398 < $\alpha$ = 0,05. Se interpreta como el incremento en el promedio de RAEE generados, por unidad de incremento en los residuos de papel generados.

b) Coeficiente de determinación $R^2$= 47,55%.

El 47,55% de la variabilidad de los RAEE generados, está asociada al efecto lineal de los residuos de papel generados.

c)

(RAEE/ papel = 5) ~ N(0,32 + 0,085 x 5, $CMresidual^{1/2}$)

Intervalo de probabilidad 90%

[ (0,32 + 0,085 x 5) - Z x $CMresidual^{1/2}$, (0,32 + 0,085 x 5) + Z x $CMresidual^{1/2}$]

$[ (0,32 + 0,085 \times 5) - Z \times 0,034^{1/2},\ (0,32 + 0,085 \times 5) + Z \times 0,034^{1/2}]$

Z en la tabla de la N(0,1) para probabilidad $0,1/2=0,05 \Rightarrow Z \approx 1,64$

$[ (0,32 + 0,085 \times 5) - 1,64 \times 0,034^{1/2},\ (0,32 + 0,085 \times 5) + 1,64 \times 0,034^{1/2}] =$

[ 0,29, 1,21] Kg/persona

**Problema 6.4.** Con los datos de la encuesta, correspondientes a alumnos de Estadística de ingeniería informática, se ha realizado el siguiente ajuste:

```
Regression Analysis - Linear model: Y = a + b*X
----------------------------------------------------------
Dependent variable: PESO (kg)
Independent variable: ESTATURA-150 (cm)
Selection variable:
----------------------------------------------------------
                          Standard        T
Parameter    Estimate       Error      Statistic   P-Value
----------------------------------------------------------
Intercept    44,5067       3,49956      12,7178    0,0000
Slope         0,996        0,123023     8,09605    0,0000
----------------------------------------------------------
                  Analysis of Variance
----------------------------------------------------------
Source      Sum of Squares  Df  Mean Square F-Ratio P-Value
----------------------------------------------------------
Model          5791,32      1
Residual
----------------------------------------------------------
Total (Corr.)  13301,5     86

Correlation Coefficient = 0,65984
R-squared = 43,5388 percent
```

a) Completa la tabla del ANOVA del modelo, e indica las conclusiones de este análisis ($\alpha$= 5%).

b) Calcula la probabilidad de que el peso supere los 80 kg, en alumnos con estatura 170 cm.

**RESPUESTAS:**

a) La tabla completa es:

Analysis of Variance

| Source | Sum of Squares | Df | Mean Square | F-Ratio | P-Value |
|---|---|---|---|---|---|
| Model | 5791,32 | 1 | 5791,32 | 65,55 | 0,0000 |
| Residual | 7510,19 | 85 | 88,3551 | | |
| Total (Corr.) | 13301,5 | 86 | | | |

Como p-value < 0,05, la pendiente de la recta de regresión es significativa. La estatura tiene efecto lineal positivo sobre el peso medio, estimado = 0,996.

b) (peso/estatura=170) ~ N( 44,5 + 0,996 x (170-150), $88{,}35^{1/2}$)

(peso/estatura=170) ~ N(64,43, 9,39)

P((peso/estatura=170) > 80) = P(N(64,43, 9,39) > 80 ) = P(N(0,1) > (80-64,43)/9,39) =
= P(N(0,1) > 1,66) = 0,0485

Por tanto, la probabilidad vale 4,85%

# Bibliografía


Ajuntament de València. (2015). Anuari estadístic de la ciutat de València. València: CD-ROM, Ajuntament de València.

Capilla, C. (2016). Neural networks data mining in an air quality database. Proceedings de 8th International Congress on Environmental Modellling & Software, Toulouse, France, S. Savage, J.M. Sánchez-Pérez, Rizzoli, A. (Eds). http://www.iemss.org/society/index.php/iemss-2016-proceedings.

Consejo de Colegios de Ingeniería Informática. (2015). *Estudio nacional sobre la situación de los profesionales del sector de tecnologías de la información.* www.ccii.es.

Oficina de Estadística, Ajuntament de València. (2014). *Enquesta sobre equipament i ús de tecnologies d'informació i comunicació a les llars.* València: www.valencia.es/estadistica.

Peña, D. (1995). *Estadística, modelos y métodos. Vol 1, Fundamentos.* Alianza Universidad.

Romero Villafranca, Rafael; Zúnica Ramajo, Luisa. (2013). *Métodos estadísticos para ingenieros.* Valencia: Universidad Politécnica.

Statgraphics plus 5. (2000). *User Manual.* Maryland, USA: Manugistics, Inc.

Unitat de Medi Ambient, Universitat Politècnica de València (2016). Informe De revisión del sistema de gestión ambiental 2015. València: Universitat Politècnica de València.