Nombre y Apellidos:

CUESTIONARIO

1. Una muestra de datos cualitativos:
   a) Debe transformarse obligatoriamente en datos numéricos, asignando un valor a cada categoría.
   b) No tiene moda, pero sí el resto de parámetros o medidas de posición central.
   c) No tiene media, ni moda, ni varianza.
   d) No tiene media, ni varianza, ni rango.

2. ¿Qué medida estadística deberíamos utilizar para determinar si una persona de un colectivo sobre el que se ha observado la calificación obtenida en un examen está entre el 25 % de los mejores?
   a) El primer cuartil.
   b) El tercer cuartil.
   c) La moda.
   d) La mediana.

3. Las distribuciones ___________ corresponden al estudio, por separado, de cada una de las dos variables que componen una variable estadística bidimensional. (Señala la palabra que falta en esta frase):
   a) Absolutas.
   b) Condicionadas.
   c) Marginales.
   d) Relativas.

4. En una distribución bidimensional, ¿son compatibles los siguientes datos $S_{xy} = -4$, $\hat{y} = -3+2x$, $r = -0{,}9$?
   a) Sí ya que expresan una relación inversa.
   b) No ya que la covarianza tiene que estar entre -1 y 1.
   c) No ya que la pendiente de la recta de regresión debería ser también negativa.
   d) Necesitaría la otra recta de regresión para poder tomar una decisión.

5. Dados los sucesos A y B pertenecientes al mismo espacio de sucesos tales que $P(A) = 0{,}9$, $P(B) = 0{,}3$ y $P(A \cap B) = 0{,}27$. ¿Cuál de las siguientes opciones es correcta?
   a) La $P(A/B) = 0{,}8$
   b) Los sucesos A y B son incompatibles.
   c) Los sucesos A y B son complementarios.
   d) Los sucesos A y B son independientes.

6. En un grupo de 10 alumnos 7 han aprobado el test de estadística y 3 no. Elegimos al azar y sin reemplazamiento a 2 alumnos. ¿Cuál es la probabilidad de que los dos hayan aprobado el test?

   a) 0,49
   b) 0,466 . . .
   c) 0,42
   d) 0,54

7. Si Z representa una variable normal $N(0;1)$, ¿cuál de las siguientes afirmaciones es cierta?

   a) $P[Z \leq -0{,}67] = 1 - P[Z > -0{,}67]$
   b) $P[Z \leq -0{,}67] = 1 - P[Z > 0{,}67]$
   c) $P[Z \leq -0{,}67] = 1 - P[Z < 0{,}67]$
   d) $P[Z \leq -0{,}67] = 1 - P[Z < -0{,}67]$

8. En la inferencia Estadística, al conjunto de individuos o elementos en los que se desea estudiar alguna/s característica/s aleatoria/s, se denomina:

   a) Población
   b) Parámetro muestral
   c) Variable
   d) Muestra

9. En un intervalo de confianza, ¿cuál de las siguientes opciones es correcta?

   a) Es más pequeño cuanto menor es el nivel de confianza, para los mismos datos.
   b) Es más pequeño cuanto mayor es el nivel de confianza, para los mismos datos.
   c) Es más grande cuanto menor es el nivel de confianza, para los mismos datos.
   d) El tamaño no depende del nivel de confianza utilizado.

10. La probabilidad de error α de un contraste de hipótesis, es la probabilidad de:

   a) Aceptar H_0 cuando es verdadera H_1.
   b) Rechazar H_0 cuando es verdadera H_0.
   c) Equivocarse cuando se rechaza H_1.
   d) Equivocarse cuando se acepta H_0.

CUADRO DE RESPUESTAS

PREGUNTA	1	2	3	4	5	6	7	8	9	10
(a)										
(b)										
(c)										
(d)										

Nota: Cada respuesta correcta suma 0,20 y cada respuesta errónea resta 0,067.

Nombre y Apellidos:

EJERCICIOS

1. **La siguiente tabla recoge el tiempo de respuesta (en nanosegundos) de un circuito lógico en frío (X) y después de una hora de uso intensivo (Y), para un total de 8 máquinas:**

X	6	5	8	16	7	4	5	9
Y	4	8	11	9	11	6	9	6

   a) **(0.75 pts.) ¿Cuánto se estima que tardaría en responder un determinado circuito tras una hora de funcionamiento intensivo, si en frío tuvo un tiempo de respuesta de 10 nanosegundos?**

   b) **(0.5 pts) ¿Es fiable la estimación obtenida en el apartado anterior?**

SOLUCIÓN: Completamos la siguiente tabla de cálculos, donde *X=Tiempo en frio* e *Y=Tiempo tras una hora*. Para la realización de estos cálculos el modo estadístico de nuestra calculadora es una ayuda bastante buena.

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
6	4	36	16	24
5	8	25	64	40
8	11	64	121	88
16	9	256	81	144
7	11	49	121	77
4	6	16	36	24
5	9	25	81	45
9	6	81	36	54
$\sum(x_i) = 60$	$\sum(y_i) = 64$	$\sum(x_i^2) = 552$	$\sum(y_i^2) = 556$	$\sum(x_i \cdot y_i) = 496$

$$n = 8 \qquad S_x^2 = \frac{\sum(x_i^2)}{n} - \overline{x}^2 = \frac{552}{8} - 7{,}5^2 = 12{,}75$$

$$\overline{x} = \frac{\sum(x_i)}{n} = \frac{60}{8} = 7{,}5 \qquad S_y^2 = \frac{\sum(y_i^2)}{n} - \overline{y}^2 = \frac{556}{8} - 8^2 = 5{,}5$$

$$\overline{y} = \frac{\sum(y_i)}{n} = \frac{64}{8} = 8 \qquad S_{xy} = \frac{\sum(x_i \cdot y_i)}{n} - \overline{x} \cdot \overline{y} = \frac{496}{8} - 7{,}5 \cdot 8 = 2$$

a) Una vez hemos realizado los cálculos previos, vamos a construir el modelo de regresión lineal que permita estimar el valor Y cuando es conocido el valor de la variable X (en este caso X=10). En este caso vamos a generar la Recta de Regresión de Y sobre X:

$$\left.\begin{array}{l} b = \frac{S_{xy}}{S_x^2} = -\frac{2}{12{,}75} = 0{,}157 \\ \\ a = \overline{y} - b \cdot \overline{x} = 8 - 0{,}157 \cdot 7{,}5 = 6{,}8235 \end{array}\right\} \Rightarrow \quad \hat{y}(x) = a + b \cdot x = 6{,}8235 + 0{,}157 \cdot x$$

Sustituimos en la recta obtenida el valor x=10 y obtenemos el resultado que se pide:

$$\hat{y}(10) = 6{,}8235 + 0{,}157 \cdot 10 = \boxed{8{,}392}$$

b) La fiabilidad de un modelo de regresión se mide a través del Coeficiente de Determinación, que nos indica cómo de buenas son las predicciones que realiza el modelo. En primer lugar calculamos el coeficiente de correlación lineal:

$$r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{496}{\sqrt{12{,}75} \cdot \sqrt{5{,}5}} = 0{,}2388$$

El coeficiente de determinación sería $R^2 = r^2 = 0{,}057$ (5,7 %), valor muy bajo indicativo de un mal ajuste y de unas predicciones muy poco fiables. En este caso, la variable predictora o independiente (en este caso X) explica un 5,7 % de la variabilidad que tiene la variable respuesta o dependiente (en este caso Y).

2. El administrador de correo electrónico de una empresa quiere establecer un filtro para el correo tipo spam (mensajes no solicitados, no deseados o con remitente no conocido, habitualmente de tipo publicitario). Según sus datos el 40 % de los correos recibidos en la empresa son spam. Además, la palabra "gratis" aparece en el 85 % de los correos spam y en el 4 % de los que no son spam. Tomando como referencia estos datos se pregunta:

- **a) (0.5 pts.) Si un correo elegido al azar no incluye la palabra "gratis", ¿cuál es la probabilidad de que sea un correo spam?**
- **b) (0.5 pts.) Si un correo elegido al azar incluye la palabra "gratis", ¿cuál es la probabilidad de que no sea un correo spam?**
- **c) (0.5 pts.) Si elegimos aleatoriamente 15 correos, ¿cuál es la probabilidad de que 3 o 4 de ellos contengan la palabra "gratis"?**

SOLUCIÓN: Lo primero que hacemos es definir los sucesos que intervienen en este ejercicio:

S={Ser un correo tipo Spam}

G={Correo con la palabra Gratis}

Del enunciado pueden deducirse los siguientes datos:

$$P(S) = 0{,}4 \qquad P(G/S) = 0{,}85 \qquad P(G/\overline{S}) = 0{,}004$$

a) Planteamos para este apartado una probabilidad condicionada y aplicamos la fórmula de Bayes para su cálculo:

$$P\left(S/\overline{G}\right) = \frac{P\left(\overline{G}/S\right) \cdot P\left(S\right)}{P\left(\overline{G}\right)} = \frac{0{,}15 \cdot 0{,}4}{1 - 0{,}364} = 0.094$$

Previamente hemos calculado la $P(G)$ utilizando la fórmula de la probabilidad total:

$$P\left(G\right) = P\left(G/S\right) \cdot P\left(S\right) + P\left(G/\overline{S}\right) \cdot P\left(\overline{S}\right) = 0{,}85 \cdot 0{,}4 + 0{,}04 \cdot 0{,}6 = 0{,}364$$

b) Volvemos a aplicar Bayes:

$$P\left(\overline{S}/G\right) = 1 - P\left(S/G\right) = 1 - \frac{P\left(G/S\right) \cdot P\left(S\right)}{P\left(G\right)} = 1 - \frac{0{,}85 \cdot 0{,}4}{0{,}364} = 0.065$$

c) Definimos la variable aleatoria X="*Nº de correos con la palabra gratis en un grupo de 15 elegidos aleatoriamente*", que se ajusta a una distribución Binomial de parámetros $n = 15$ y $p = P(G) = 0{,}364$:

$$X \sim Bi(15; 0{,}364) \quad \Rightarrow \quad \begin{aligned} P[(X=3) \text{ o } (X=4)] &= p_3 + p_4 \\ &= \binom{15}{3}(0{,}364)^3(0{,}636)^{12} + \binom{15}{4}(0{,}364)^4(0{,}636)^{11} \\ &= 0{,}0961 + 0{,}1650 = \boxed{0{,}2611} \end{aligned}$$

3. Una empresa que presta servicios de internet asegura a todos sus usuarios, después de una profunda renovación y actualización de sus instalaciones, que han aumentado la velocidad de conexión de manera significativa. Seleccionados aleatoriamente una muestra de usuarios, la siguiente tabla recoge sus velocidades de conexión antes y después del cambio.

Vel. Antes	21	25	26	25	19	22	22	28	29	18	16	21	30	27
Vel. Después	26	22	30	32	24	16	18	30	29	21	20	26	33	32

Suponiendo normalidad en los datos:

a) **(0.5 pts.) ¿podemos confirmar lo que asegura esta empresa con un 1 % de significación?**

b) **(0.5 pts.) ¿qué tamaño muestral sería necesario, para poder construir un intervalo de confianza al 95 % del incremento medio de la velocidad de conexión, con un error de estimación inferior a 0,1 mbps.?**

SOLUCIÓN:

a) Las mediciones sobre la velocidad de conexión de los usuarios antes y después de hacer los cambios, corresponden a dos muestras dependientes o pareadas, por lo tanto trabajaremos con la muestra de las diferencias: $(x_2 - x_1)$:

Incremento Velocidad $(x_2 - x_1)$:	5	-3	4	7	5	-6	-4	2	0	3	4	5	3	5

Los datos muestrales sobre incremento de velocidad nos sirven para obtener (con la calculadora) los parámetros muestrales habituales: $\overline{d} = 2{,}143$ y $S_{c_d} = 3{,}92$. Con esta información podemos plantear el siguiente contraste de hipótesis:

$$\left.\begin{array}{l} H_0 \equiv \mu_d = 0 \\ H_1 \equiv \mu_d > 0 \end{array}\right\} \qquad t_{exp} = \frac{\overline{d} - 0}{S_{c_d}}\sqrt{n} = \frac{2{,}143}{3{,}92}\sqrt{14} = 2{,}05$$

La región crítica del contraste al 1 % de significación, teniendo en cuenta la distribución del estadístico t_{exp} y que el contraste planteado es unilateral derecho quedaría:

$$\text{R.C.} = \{t_{exp} > t_{1-\alpha}(n-1) = t_{0,99}(13) = 2{,}65\} \quad \Rightarrow \quad H_0$$

Por tanto no podemos rechazar H_0 para $\alpha = 0{,}01$, es decir, no podemos afirmar que se haya producido un incremento significativo de la velocidad de conexión para $\alpha = 0{,}01$.

b) Para determinar el tamaño muestral se aproxima la distribución t-student por la normal estándar y nos queda la siguiente expresión

$$n > \left[\frac{Z_{1-\alpha/2} \cdot S_{c_d}}{\epsilon}\right]^2 = \left[\frac{Z_{0,975} \cdot S_{c_d}}{0{,}1}\right]^2 = \left[\frac{1{,}96 \cdot 3{,}92}{0{,}1}\right]^2 = 5903{,}16 \qquad \Rightarrow \qquad \boxed{n = 5904}$$

4. **Una encuesta realizada por internet a 1500 usuarios reveló que 525 de ellos tenían preferencia por los ordenadores portátiles mientras que el resto mostraban sus preferencia por los ordenadores de sobremesa. A partir de estos datos y con la finalidad de estimar la proporción poblacional de internautas que preferían utilizar portátiles, se construyó el siguiente intervalo de confianza: (0,33 - 0,37).**

   a) **(0.75 pts.) ¿Cuál es el nivel de confianza del intervalo anterior?**

   b) **(0.5 pts.) A partir de estos datos, ¿es posible afirmar con un 1 % de significación que la proporción de internautas que prefieren los ordenadores portátiles es inferior al 40 %?**

SOLUCIÓN:

a) En este apartado conocemos un intervalo de confianza para la proporción poblacional de usuarios que prefieren ordenador portátil construido a partir de una muestra aleatoria de tamaño $n = 1500$. Nos piden el nivel de confianza $(1-\alpha)$ de dicho intervalo. En primer lugar escribimos la fórmula del intervalo de confianza para la proporción poblacional e igualamos con los datos de nuestro enunciado:

$$IdC_{1-\alpha}(\pi) = \left(p \pm z_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \right) = (0{,}35 \pm 0{,}02)$$

De esta igualdad se deduce en primer lugar que $p = 0{,}35$. Además si igualamos el segundo miembro de ambas expresiones tenemos que:

$$z_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} = z_{1-\alpha/2}\sqrt{\frac{0{,}35\cdot 0{,}65}{1500}} = 0{,}02 \qquad \Rightarrow \qquad z_{1-\alpha/2} = 1{,}62$$

Buscamos en la tabla de la distribución Normal estándar el valor $z = 1{,}62$ y comprobamos que corresponde con el percentil $1 - \frac{\alpha}{2} = 0{,}9774$. De modo que el nivel de confianza del intervalo debe ser $1 - \alpha = 0{,}8948$ o lo que igual el 89,48 %.

b) Para este apartado vamos a plantear un contraste para decidir si la proporción poblacional puede considerarse inferior al 40 %. Las hipótesis y el estadístico del contraste serán:

$$\left.\begin{array}{l} H_0 \equiv \pi = 0{,}4 \\ H_1 \equiv \pi < 0{,}4 \end{array}\right\} \qquad z_{exp} = \frac{p-\pi_0}{\sqrt{\frac{\pi_0\cdot(1-\pi_0)}{n}}} = \frac{0{,}35-0{,}4}{\sqrt{\frac{0{,}4\cdot 0{,}6}{1500}}} = -3{,}95$$

La región crítica del contraste al 1 % de significación, teniendo en cuenta la distribución del estadístico z_{exp} y que el contraste planteado es unilateral izquierdo quedaría:

$$\text{R.C.} = \{z_{exp} < Z_\alpha = Z_{0{,}01} = -2{,}326\} \quad \Rightarrow \quad H_1$$

Por tanto podemos rechazar H_0 para $\alpha = 0{,}01$, es decir, podemos afirmar con un margen de error del 1 %, que el porcentaje de internautas que prefieren el ordenador portátil es inferior al 40 %.