Grado en Ingeniería Informática

Estadística

EXAMEN FINAL

20 de junio de 2017

Apellidos y nombre:		
Grupo:	Firma:	
Marcar las casillas de los parciales presentados	P1 ☐	P2 ☐

Instrucciones

1. **Rellenar** la cabecera del examen: **nombre, grupo y firma**.
2. Responder a cada pregunta en la hoja correspondiente.
3. **Justificar todas las respuestas**.
4. No se permiten anotaciones personales en el formulario. Sobre la mesa sólo se permite el DNI, calculadora, útiles de escritura, las tablas y el formulario.
5. **No desgrapar** las hojas.
6. El examen consta de 6 preguntas, 3 correspondientes al primer parcial (50%) y 3 del segundo (50%). El profesor corregirá los parciales que el alumno haya señalado en la cabecera del examen. **En cada parcial, todas las preguntas puntúan lo mismo** (sobre 10).
7. Se debe **firmar** en las hojas que hay en la mesa del profesor **al entregar el examen**. Esta firma es el justificante de la entrega del mismo.
8. Tiempo disponible: **3 horas**

1. (1er Parcial) En el estudio sobre los servicios de telecomunicación de una provincia española se han estudiado 770 municipios. En cada municipio, se han evaluado, entre otras, las siguientes características en el año A:

- Nº de líneas ADSL en funcionamiento
- Edad media de los habitantes del municipio (años)
- Renta familiar media disponible por habitante (€), codificada como:
  - Hasta 7.200
  - Entre 7.200 y 8.300
  - Entre 8.300 y 9.300
  - Entre 9.300 y 10.200
  - Entre 10.200 y 11.300
  - Entre 11.300 y 12.100
  - Entre 12.100 y 12.700
- Población total del municipio

Tras realizar un estudio descriptivo se han obtenido los siguientes resultados:

Resumen Estadístico (Tabla 1)

	Edad media municipio	***Lineas ADSL municipio***	***Poblacion municipio***	**Tipo**
Recuento	770	770	770	
Promedio	41,4627	1075,84	10466,8	
Mediana		181,0	2761,0	
Desviación Típica	4,4481	4276,64	39302,0	
Coeficiente de Variación		397,517%	375,49%	
Mínimo	31,0		50,0	
Máximo		73274,0	699145,	
Rango		73274,0	699095,	
Cuartil Inferior	38,2	26,0	1012,0	
Cuartil Superior	44,4	736,0	7054,0	
Rango Intercuartílico	6,2		6042,0	
Asimetría estándar	1,9778		136,523	
Curtosis estándar	1,24895	953,325	1018,47	

Gráfico 1

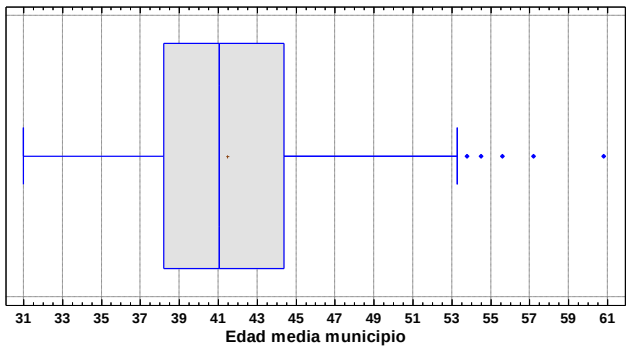


Gráfico 2

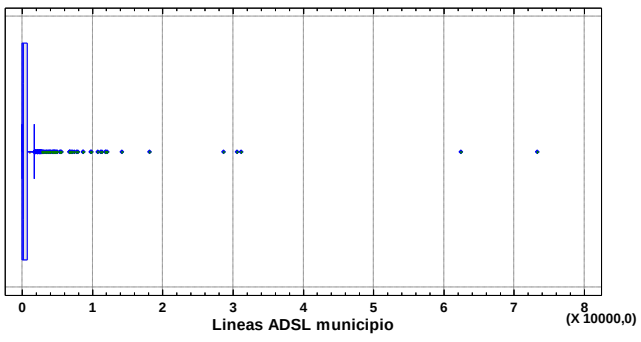


Tabla de Frecuencia para Renta familiar municipio (Tabla 3)

			Frecuencia	*Frecuencia*	*Frecuencia*
Clase	***Valor***	***Frecuencia***	***Relativa***	***Acumulada***	***Rel. acum.***
1	Hasta 7.200	106	0,1828	106	
2	Entre 7.200 y 8.300	232	0,4000	338	
3	Entre 8.300 y 9.300	152	0,2621	490	
4	Entre 9.300 y 10.200	70	0,1207	560	
5	Entre 10.200 y 11.300	17	0,0293	577	
6	Entre 11.300 y 12.100				
7	Entre 12.100 y 12.700	1	0,0017		

A la vista de los resultados anteriores (parámetros, tablas y gráficos), contesta razonadamente a las siguientes preguntas:

a) Rellena las celdas vacías (sombreadas) en la tabla "**Resumen Estadístico (Tabla 1)**" con los valores correspondientes e indica, en la misma tabla, de qué tipo es cada parámetro. ***(3 puntos)***

b) ¿De qué tipo son las variables estudiadas y qué dimensión tienen? ***(2 puntos)***

c) Determina qué parámetros de posición y dispersión serían los más adecuados para representar a las variables *Edad media municipio*, *Lineas ADSL municipio* y *Poblacion municipio* y por qué. ¿Cuál de ellas tiene una mayor dispersión? ***(2 puntos)***

d) Respecto a la renta familiar: ***(2 puntos)***

- Calcula el porcentaje de municipios de la provincia estudiada tienen una renta familiar inferior o igual a 12.100 €

- Calcula cuántos municipios de la provincia estudiada tienen una renta familiar Entre 11.300 y 12.100

e) ¿Cómo se denominan los gráficos 1 y 2? Explica qué ventajas o inconvenientes ofrecen respecto a los histogramas. ***(1 punto)***

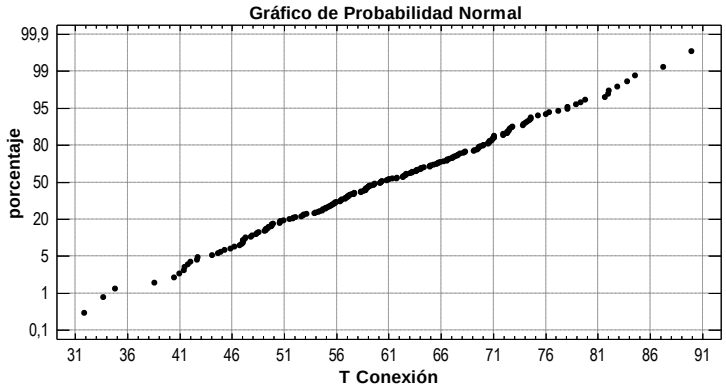
f) Indica qué representación gráfica hubiera sido la adecuada para describir la característica *Renta Familiar* de los municipios y dibújala de modo aproximado y señalando los elementos básicos de ésta. ***(1 punto)***

2. (1er Parcial) Un hotel en Bahía Tranquila tiene 120 habitaciones. En los meses de primavera, la ocupación hotelera diaria es aproximadamente del 75%. En caso de que el número de habitaciones reservadas con antelación supere el 85% del total de las habitaciones del hotel, la Dirección debe contratar personal extra para poder atender a los clientes.
Asumiendo que todos los días primaverales son igualmente "deseables" para hacer una reserva, se pide:

a) ¿Cuál es la probabilidad de que cierto día de primavera se ocupen todas las habitaciones? ***(2 puntos)***

b) ¿Cuál es la probabilidad, aproximadamente, de que cierto día de primavera se contrate personal extra? ***(4 puntos)***

c) Para mejorar el servicio de conexión WiFi, el hotel está planteándose adquirir un nuevo router. Para determinar las características del nuevo router se han tomado datos del tiempo de uso de la WIFI en cada una de las conexiones de los clientes. Los tiempos, representados sobre Papel Probabilístico Normal se muestran a continuación.



A partir del gráfico, se pide:

c.1) Calcula, aproximadamente, el tiempo medio de conexión. ***(1,5 puntos)***
c.2) ¿Qué porcentaje de las conexiones tienen un tiempo de uso entre 50 y 70 minutos? ***(2,5 puntos)***

3. (1er Parcial) La duración T (horas de funcionamiento hasta el fallo) de ciertos componentes electrónicos fluctúa aleatoriamente siguiendo una distribución exponencial. Se sabe que el 10% de las componentes duran menos de 1 año.

a) Calcula el tiempo medio de vida de los componentes mencionados. ***(3 puntos)***

b) Sabiendo que ya han transcurrido 3 años sin que se produzca ningún fallo, ¿cuál es la probabilidad de que en los siguientes 5 años un componente funcione correctamente? ***(3 puntos)***

c) Estos componentes se utilizan para el montaje de un dispositivo electrónico (D) que debe conectar dos bornes A y B. Se desea que el dispositivo tenga una fiabilidad de al menos 90% al año. Para garantizar esta fiabilidad se montan en paralelo N componentes del tipo estudiado. Sabiendo que las componentes presentan un funcionamiento independiente unas de otras, ¿cuánto debe valer como mínimo N? ***(4 puntos)***

4. (2º Parcial) Responde a cada una de las siguientes preguntas justificando convenientemente la respuesta.

a) Dada una población normal con desviación típica $\sigma=4$, extraemos una muestra de tamaño 20. ¿Cuál es la probabilidad de que la media muestral y la media poblacional difieran en menos de 1 unidad? ***(3 puntos)***

b) En cierto proceso se mide de forma rutinaria un parámetro de calidad. Se realizan ciertos cambios operativos en el proceso con el objetivo de mejorar la calidad. Para estudiar la efectividad de dichos cambios, se ha tomado una muestra aleatoria de 18 unidades, obteniéndose una media muestral de 21 y una varianza muestral de 2,3. Obtener un intervalo de confianza para la media poblacional, considerando $\alpha=0,01$. ¿Es razonable admitir que la media poblacional del parámetro es de 22 tras los cambios efectuados en el proceso? ***(3,5 puntos)***

c) ¿Es admisible asumir que la varianza poblacional del parámetro, después de los cambios operativos realizados, es de 8,0 unidades^2? ***(3,5 puntos)***

5. (2º Parcial) Se desea estudiar el efecto que la configuración (tres posibles: A, B y C) y el tamaño de memoria Caché (3 niveles: bajo, medio, alto) tienen sobre el rendimiento medio de un sistema informático. Cada tratamiento se ha ensayado dos veces.

a) Analiza qué efectos son estadísticamente significativos a partir del cuadro resumen del ANOVA (utiliza α=5%). Plantea todas las hipótesis que vayas a contrastar mediante el ANOVA en relación a los efectos estudiados.
Ten en cuenta que: SC_{total}=11039,2; SC_{config}=704,47; $SC_{caché}$=8390,4 ; $SC_{residual}$=690,005. ***(5 puntos)***

b) ¿Cuántas poblaciones se están comparando en este estudio? Asumiendo que se cumple la hipótesis de homocedasticidad, ¿cuánto vale la estimación de la varianza de cada una de las poblaciones? ***(2 puntos)***

c) En general, ¿qué información adicional a la proporcionada por la tabla resumen del ANOVA puede obtenerse a partir de la representación gráfica de los intervalos LSD? ***(3 puntos)***

6. (2º Parcial) Responde a los siguientes apartados A y B:

A).- Con el fin de estudiar el rendimiento de una base de datos se ha anotado a lo largo de 3 meses la carga media diaria del sistema, medida en nº de consultas por minuto (X) y el tiempo medio de respuesta en segundos (Y).

A partir de los datos reunidos durante el estudio se han calculado los siguientes parámetros: $\bar{X}= 3,5$ $S_X= 0,65$ $\bar{Y}=1,3$ $S_Y=0,6$ $r_{XY} = 0,9$

A partir de la información proporcionada, se pide:

a) Plantea la recta de regresión que permita obtener el tiempo medio de respuesta a partir del nº de consultas. ¿Qué significado tiene el valor de la pendiente de la recta?

(3 puntos)

b) Calcula el valor coeficiente de determinación. ¿Qué representa en la práctica este valor? ***(1 punto)***

c) Si cierto día la carga ha sido de 7 consultas por minuto, ¿cuál es la probabilidad de que el tiempo de respuesta sea mayor de 5 segundos?

(3 puntos)

B).- Durante un periodo de 5 años, se ha encontrado que el coeficiente de correlación entre el número de resfriados registrados semanalmente en una ciudad y la cantidad de cerveza vendida es -0,82.

B.1) Del estudio parece desprenderse que el consumo de cerveza ayuda a prevenir los resfriados ¿Qué comentarios te sugiere esta afirmación?

(1,5 puntos)

B.2) En el estudio se ha calculado también la correlación entre la cantidad de cerveza consumida durante la 1ª semana de agosto por 25 familias de esa ciudad y 25 de otra ciudad vecina. En este caso el coeficiente de correlación ha resultado positivo. ¿Qué comentarios te sugiere el cálculo de este coeficiente de correlación? ***(1,5 puntos)***

SOLUCIÓN

1a) Parámetros sobre Edad: Mediana = **41.1** (línea vertical dentro de la caja).
Coeficiente de variación = desv. típica / media = 4.448/41.463 = **0.107**
Máximo = **60.9** (aparece como un punto aislado en el gráfico).
Rango = máximo - mínimo = 60.9 - 31 = **29.9**

Parámetros sobre líneas ADSL: mínimo = máx - rango = 73,274 - 73,274 = **0**
Rango intercuartílico = $Q_3 - Q_1$ = 736 - 26 = **710**
Asimetría estándar: >> **2** (el valor exacto se desconoce, pero el gráfico 2 revela que la distribución es claramente asimétrica positiva).

Tipo de cada parámetro:

- Posición: media, mediana, mínimo, máximo, cuartil inferior y cuartil superior.
- Dispersión: desv. típica, coef. de variación, rango, rango intercuartílico.
- Forma: asimetría estándar, curtosis estándar.

El parámetro "recuento" no puede clasificarse en ninguna de estas categorías.

1b) Se conocen cuatro características de cada individuo de la población (municipio): nº de líneas ADSL, edad media de los habitantes, población total y renta familiar. Por tanto, se trata de una variable aleatoria **tetra**-dimensional:

- Líneas ADSL y población total: ambas son variables aleatorias cuantitativas que contienen valores discretos.
- Edad media: variable aleatoria cuantitativa que contiene valores en una escala continua, por ser la media de valores discretos.
- Renta familiar media por persona: es una variable continua, pero en este caso los valores están codificados en 7 categorías. Por tanto, puede considerarse como variable aleatoria cuantitativa codificada en un conjunto discreto de categorías.

1c) La distribución de "Edad media" es bastante simétrica (gráfico 1) y puede considerarse como una muestra aleatoria extraída de una población normal ya que el coeficiente de asimetría estándar está comprendido entre -2 y 2. En una distribución normal, el parámetros de posición más representativo es la media (que coincide con la mediana), y los parámetros de dispersión más representativos son la varianza y la desviación típica.

La distribución de "líneas ADSL" es claramente asimétrica a la vista del gráfico 2, al igual que la "población total" (ya que el coeficiente de asimetría estándar es 136, muy superior a 2). En estos casos, la mediana es más representativa que la media como parámetro de posición porque no está afectada por valores extremos. Por la misma razón, el rango intercuartílico es el parámetro de dispersión más representativo.

La variable "población" es la de mayor dispersión dado que su rango intercuartílico (6042) es muy superior que en el caso de ADSL (710) y de "edad" (6.2). Lo mismo sucede con la desviación típica.

1d1) Porcentaje de municipios con renta familiar ≤ 12100 € =
= 100·(770-1)/770 = 100·(1-0.0013) = **99.87%**

1d2) Nº de municipios con renta familiar entre 11300 y 12100 =
= 770 - (141+308+202+93+22+1) = 770 - 767 = **3**.

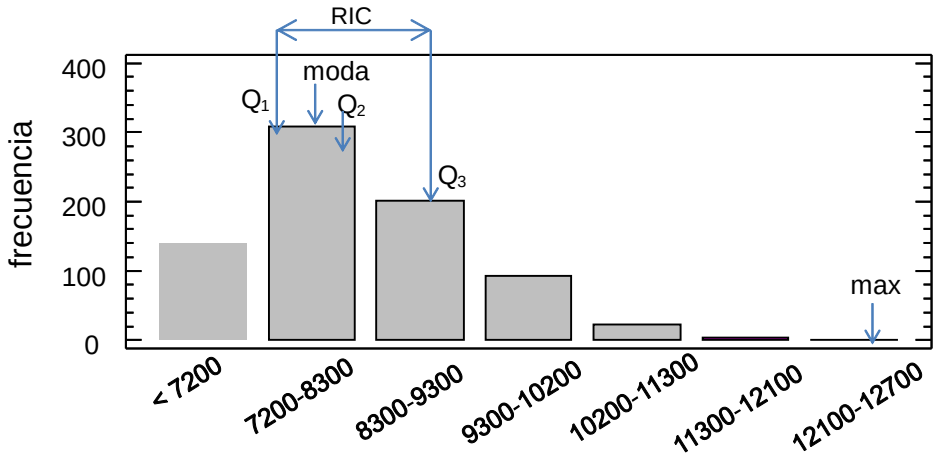
1e) Los gráficos 1 y 2 se denominan gráficos box-and-whisker o gráficos de caja y bigotes. Ventajas:

- Son útiles para visualizar valores extremos, lo cual es útil para diagnosticar datos anómalos.
- En caso de tener pocos valores, son más útiles que los histogramas a la hora de discutir la simetría o asimetría de la distribución.
- Visualizan parámetros importantes como la media, mediana y cuartiles, lo cual permite el cálculo directo del rango intercuartílico.
- Su forma no depende del número de intervalos que se elija, como sucede con los histogramas.
- La comparación de varios gráficos de caja-bigotes es sencilla (*multiple box-whisker plots*), lo cual no sucede con los histogramas.

Desventajas de los gráficos de caja-bigotes respecto a los histogramas:

- La escala vertical no aporta información. Por ello, no es posible saber el número total de valores, o el número de datos dentro de cada intervalo, de modo que no se pueden calcular frecuencias ni percentiles.
- Los histogramas permiten revelar ciertos patrones de variabilidad más claramente que los gráficos de caja-bigotes, por ejemplo distribuciones bimodales, o con datos truncados, o frecuencias anormales de valores concretos.

1f) Los datos de "renta familiar" están codificados en 7 categorías. Por ello, un **gráfico de barras** sería la representación más adecuada para describir la distribución de esta variable aleatoria, el cual se muestra a continuación. Se indica la posición aproximada de los parámetros básicos de distribución: primer cuartil, moda, mediana, tercer cuartil, rango intercuartílico (RIC) y máximo. En un histograma de frecuencias, todas las barras tienen la misma anchura, lo cual no sucede aquí. Por ello no es posible dibujar un histograma en este caso.



2a) La ocupación hotelera diaria del 75% no es la probabilidad de que el hotel esté totalmente ocupado un cierto día. Significa la probabilidad de que una habitación del hotel esté reservada: si se elige un día de primavera, 75% de las habitaciones estarán reservadas en promedio. Por tanto, la variable aleatoria X es define como "nº de habitaciones reservadas en un hotel con 120 habitaciones". Esta variable fluctúa entre 0 y 120 de modo que será una distribución Binomial: B (n=120, p = 0.75). La probabilidad de tener todas las habitaciones ocupadas se obtiene a partir de la función de probabilidad:

$$P(X=120)=\binom{120}{120}\cdot 0.75^{120}\cdot(1-0.75)^{0}=0.75^{120}= \mathbf{1.02\cdot 10^{-15}}$$

2b) X≈B (n=120; p=0.75); E(X) = n·p = 120·0.75 = 90

$$\sigma_x=\sqrt{n\cdot p\cdot(1-p)}=\sqrt{120\cdot 0.75\cdot 0.25}=\sqrt{22.5}=4.743$$

Dado que la varianza es superior a 9, la distribución puede aproximarse a una Normal: X ≈ N (m=90, σ=4.743). Si el número de habitaciones reservadas con antelación supera el 85% del total (es decir, 0.85·120 = 102), la Dirección debe contratar personal extra. Por tanto, la probabilidad de que se contrate personal extra equivale a calcular P(X>102):

$$P(X>102)\approx P\left[N(90;\ 4.74)>102.5\right]=P\left[N(0;\ 1)>\frac{102.5-90}{4.743}\right]=P\left[N(0;1)>2.635\right]=0.0042$$

2c.1) Dado que los puntos se ajustan bien a una línea recta, puede concluirse que los datos siguen una distribución normal. La escala vertical indica la probabilidad: P(X≤x). Según el gráfico, P(X≤60) = 50%, lo cual implica que 60 es la mediana. Dado que la distribución es normal, la media será cercana a **60**.

2c.2) Leyendo un 80% en la escala vertical, la horizontal cruza la línea de puntos aproximadamente en X = 70, lo cual implica que P(X≤70)=80%. De modo similar, resulta a partir del gráfico que P(X≤50) ≈ 15%. Por tanto:

$$P(X\in[50,\ 70])=P(X\le 70)-P(X\le 50)\approx 80\%-15\%\approx \mathbf{65\%}$$

3a) La variable aleatoria se define como T: tiempo (en horas) de funcionamiento hasta el fallo. Podemos definir otra variable X: tiempo (en años) de funcionamiento hasta el fallo.

X ≈exp (α); $P(X>1)=0.9=e^{-\alpha\cdot 1}$; ln 0.9 = -α ; α = 0.1054; E(X)=1/α = 9.49 años

E(T) = 9.49 años x 365 días/año x 24 horas/día = **83,143 horas**

3b) La distribución exponencial cumple la propiedad de falta de memoria:

P(X>8 / X>3) = P(X>5) = $e^{-0.1054\cdot 5}$ = **0.59**

Otra forma de resolver el problema es aplicando la probabilidad condicional:

$$P\left[(X>8)/(X>3)\right]=\frac{P\left[(X>8)\cap(X>3)\right]}{P(X>3)}=\frac{P(X>8)}{P(X>3)}=\frac{e^{-0.105\cdot 8}}{e^{-0.105\cdot 3}}=0.59$$

3c) Variable aleatoria Y: tiempo de funcionamiento del dispositivo hasta el fallo. Objetivo del problema: que la fiabilidad del dispositivo sea ≥ 0.9 al año. Según el concepto de fiabilidad, el requisito es: P(Y≥1) ≥ 0.9. Pero para cada componente individual, el 10% fallan antes de un año: P(X<1) = 0.1 → P(X≥1)= 0.9. Por tanto, solamente con un componente ya se cumple el requisito deseado, de modo que la solución es **N=1**.

4a) Tomando una nuestra aleatoria de tamaño n=20 de una población normal con σ=4, la distribución de la media muestral es:

$$\bar{x} \approx N\left(m; \sigma/\sqrt{n}\right) \approx N\left(m;\ 4/\sqrt{20}\right) \quad \rightarrow \quad \bar{x} - m \approx N\left(0;\ 4/\sqrt{20}\right)$$

¿Cuál es la probabilidad de obtener una media muestral que difiera menos de una unidad de la media poblacional? La diferencia es en valor absoluto:

$$P\left(\left|\bar{x} - m\right| < 1\right) = P\left[\left(\bar{x} - m\right) \in [-1;\ 1]\right] = 1 - 2 \cdot P\left[\left(\bar{x} - m\right) > 1\right] = 1 - 2 \cdot P\left[N\left(0;\ 4/\sqrt{20}\right) > 1\right] =$$

$$= 1 - 2 \cdot P\left[N\left(0;\ 1\right) > \sqrt{20}/4\right] = 1 - 2 \cdot P\left[N\left(0;\ 1\right) > 1.118\right] = 1 - 2 \cdot 0{,}131 = 0.736$$

4b) n=18, s^2=2.3, $\bar{x} = 21$; $t_{17}^{0.005} = 2.898$. Intervalo de confianza para m:

$$CI_m = \bar{x} \pm t_{n-1}^{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 21 \pm 2.898 \cdot \frac{\sqrt{2.3}}{\sqrt{18}} = 21 \pm 1.036 = \left[19.96;\ 22.04\right]$$

Dado que el valor 22 está dentro del intervalo obtenido, <u>es razonable</u> asumir un valor de 22 para la media poblacional (la hipótesis nula m=22 es aceptable).

4c) Con α=0.01, se obtiene de la tabla Gi cuadrado que: $P(\chi_{17}^2 > 35.718) = 0.005$. Además, $P(\chi_{17}^2 < 5.697) = 0.005$. Estos son los valores críticos empleados para obtener el intervalo de confianza (CI) de la varianza poblacional:

$$CI_{\sigma^2} = \left[\frac{(n-1)\cdot s^2}{g_2}; \frac{(n-1)\cdot s^2}{g_1}\right] = \left[\frac{17 \cdot 2.3}{35.718}; \frac{17 \cdot 2.3}{5.697}\right] = \left[1.095;\ 6.863\right]$$

Dado que el valor supuesto de H_0: σ^2 = 8 está fuera del intervalo, **no es aceptable** un valor de 8 para la varianza poblacional, considerando α=0.01.

5a) Hay 9 tratamientos (3 tipos de configuraciones combinadas con 3 tipos de memoria caché). Dado que cada tratamiento se ensayó dos veces, el nº total de valores (N) es 18. Grados de libertad (g.l.) totales = N-1 = 17; G.l. de cada factor = nº de variantes -1; G.l. interacción = 2·2 = 4. $SC_{interac}$ = SC_{total} - SC_{config} - SC_{mem} - SC_{resid}. Cuadrado medio: CM = SC / g.l. F-ratio = CM / CM_{resid}.

	SC	**g.l.**	**CM**	**F-ratio**
Configuración	**704.47**	2	352.235	$4.59 > (F_{2;9}^{0.05} = 4.26)$
Memoria	**8,390.40**	2	4,195.2	$54.72 > (F_{2;9}^{0.05} = 4.26)$
Interacción	1,254.325	4	313.58	$4.09 > (F_{4;9}^{0.05} = 3.63)$
Residual	**690.005**	9	46.667	
Total	**11,039.20**	17		

- Para configuración: H_0: $m_A = m_B = m_C$. Dado que el F-ratio es mayor que el valor crítico obtenido de la tabla F considerando α=0.05, se rechaza la hipótesis nula, lo cual implica que la media de al menos una de las configuraciones es distinta del resto a nivel poblacional. Por tanto, puede concluirse que el efecto simple de configuración resulta estadísticamente significativo.

- Para memoria caché: H_0: $m_{bajo} = m_{medio} = m_{alto}$. Dado que el F-ratio es mayor que el valor crítico de tablas, se rechaza H_0: el valor medio de al menos una de las memorias es distinta del resto a nivel poblacional. Por tanto, el efecto simple de memoria caché resulta estadísticamente significativo.

- Para la interacción doble: la hipótesis nula es que el efecto de la interacción doble es cero a nivel poblacional (es decir, el efecto sobre la variable respuesta de cambiar la configuración no depende de la memoria caché). Dado que el F-ratio es mayor que el valor crítico, se rechaza H_0: el efecto de la interacción doble resulta estadísticamente significativo.

5b) El número de poblaciones es igual al número de tratamientos = **9**, ya que el comportamiento medio de cada tratamiento puede caracterizarse con una media distinta. Dado que la varianza de todas las poblaciones se asume que es la misma (hipótesis de homocedasticidad), esta varianza puede estimarse como el cuadrado medio de los residuos obtenido en la tabla del ANOVA: CM_{resid} = **76.67**.

5c) Si la tabla del ANOVA revela que el efecto simple de un factor no es estadísticamente significativo, los intervalos LSD no aportan información adicional (el gráfico mostrará que todos los intervalos se solapan). En el caso de factores cualitativos con más de dos variantes, si el efecto de un factor resulta estadísticamente significativo según la tabla del ANOVA, puede concluirse que al menos una de las medias será distinta a nivel poblacional, pero es necesario visualizar los intervalos LSD para saber qué variantes son las que tienen una media distinta, que serán aquellas cuyos intervalos LSD no se solapen.

En caso de factores cuantitativos con más de dos niveles, el gráfico de intervalos LSD aporta pistas para interpretar la naturaleza (lineal o cuadrática) del efecto, si bien es necesario un análisis adicional por medio de regresión lineal para estudiar la significación estadística de la relación.

6Aa) Pendiente: $b = r \cdot s_y / s_x = 0.9 \cdot 0.6 / 0.65 = 0.8306$

Ordenada: $a = \bar{y} - b \cdot \bar{x} = 1.3 - 0.8306 \cdot 3.5 = -1.608$

Modelo de regresión: **Tiempo = -1.608 + 0.8306 · $N_{consultas}$**

Dado que el coeficiente de correlación es bastante alto, esta ecuación será útil para predecir el tiempo medio de respuesta en función del nº de consultas por minuto.

Interpretación de la pendiente: el valor 0.83 indica que la variable dependiente (tiempo de respuesta) aumentará 0.83 segundos en promedio por cada unidad de aumento del número de consultas.

6Ab) Coeficiente de determinación: $R^2 = 100 \cdot r^2 = 100 \cdot 0.9^2 =$ **81%**
Este coeficiente representa el porcentaje de variación de la variable Y explicado por la variabilidad de X. En este caso, el 81% de la variabilidad del tiempo de respuesta está explicado por el modelo, es decir, por la variable "nº de consultas".

6Ac) La distribución condicional del tiempo cuando x=7 es una distribución normal con los siguientes parámetros:

$$E(Y / X = 7) = -1.608 + 0.8306 \cdot 7 = 4.2062$$

$$\sigma(Y / X = 7) = \sigma_{resid} = \sqrt{s_Y^2 \cdot (1 - r^2)} = \sqrt{0.6^2 \cdot (1 - 0.9^2)} = \sqrt{0.0684} = 0.2615$$

$$P[(Y > 5)/(X = 7)] = P[N(4.206;\ 0.261) > 5] = P\left[N(0;1) > \frac{5 - 4.206}{0.2615}\right] = P[N(0;1) > 3.036] = 0.0012$$

6B.1) El coeficiente de correlación es bastante elevado, lo cual indica que un mayor consumo de cerveza vendida en la ciudad se corresponde con aquellas semanas en las que menos gente sufre resfriados. Esta correlación observada podría sugerir que un mayor consumo de cerveza previene los resfriados. Sin embargo, esta conclusión puede no ser cierta debido a que no es una relación de tipo causa - efecto. Se trata de un ejemplo de correlación espuria: *una relación matemática en la que dos o más variables no están relacionadas entre sí causalmente, aunque puede inferirse erróneamente que lo están, debido a la coincidencia o a la presencia de un cierto tercer factor desconocido, denominado "variable de respuesta común" o "factor de confusión"*. Más información en: https://en.wikipedia.org/wiki/Spurious_relationship

En este caso, el factor de confusión podría ser la distinta temperatura a lo largo del año: durante los meses más cálidos, los casos de gripe son lógicamente menores y, debido al calor, la gente acostumbra a beber más cerveza. Contrariamente, en los meses fríos, los casos de gripe son mucho mayores y la gente tiende a beber menos cerveza.

6B.2) Hay 25 familias de una ciudad y 25 familias distintas de otra ciudad. No hay relación entre estas familias, lo cual implica que los datos se estructuran como dos variables unidimensionales. En tal caso, no se puede realizar una regresión simple de las variables. La correlación positiva se obtiene por azar, y seguramente su valor no será demasiado alto.