Grado en Ingeniería Informática

Estadística

PRIMER PARCIAL

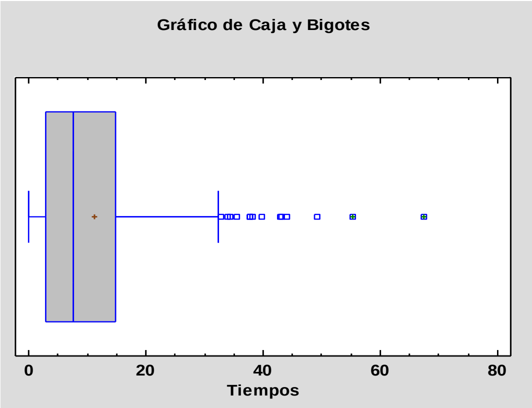
26 de marzo de 2019

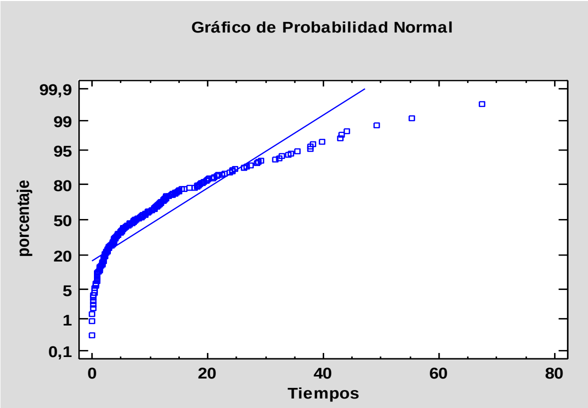
Apellidos, nombre:	
Grupo:	Firma:

Instrucciones

1. Rellenar la información de cabecera del examen.
2. Responder a cada pregunta en la hoja correspondiente.
3. Justificar todas las respuestas.
4. No se permiten anotaciones personales en el formulario.
5. No se permite tener teléfonos móviles encima de la mesa. Sobre la mesa sólo se permite el DNI, calculadora, útiles de escritura, las tablas y el formulario.
6. No desgrapar las hojas.
7. Todas las preguntas puntúan lo mismo (sobre 10).
8. 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.
9. **Tiempo disponible: <u>2 horas</u>**

1. Los tiempos hasta la rotura (en meses) de 200 componentes están representados en las dos gráficas siguientes:





Responde a las siguientes preguntas justificando convenientemente la respuesta.

a) Indica si la siguiente afirmación es verdadera o falsa (justifica la respuesta): "*La variable aleatoria objeto de estudio está definida en la población de los distintos meses en los cuales se mide el tiempo hasta la rotura*". *(2,5 puntos)*

b) Indica cuál es la utilidad, en general, de cada uno de los gráficos representados. *(3 puntos)*

c) Si los datos se expresan en otras unidades, ¿cambia el valor de los parámetros de posición y/o de dispersión? *(2 puntos)*

d) Comenta las principales conclusiones derivadas de los gráficos en relación a la pauta de variabilidad constatada en los datos. *(2,5 puntos)*

2. Se sabe que los alumnos matriculados en cierto grado de la UPV disponen de ordenadores portátiles de las marcas HPH, ADER y SANY (cada alumno tiene solamente un portátil). El 30% son HPH y el 15% son ADER. Tras realizar las averiguaciones oportunas, se ha obtenido que el 20% de los portátiles HPH, el 30% de los ADER y el 25% de los SANY fueron comprados en una tienda online.

a) Define los sucesos que es necesario considerar para resolver los cálculos de los siguientes apartados. ¿A qué sucesos corresponden cada uno de los porcentajes indicados en el enunciado? *(1 punto)*

b) Si se escoge un ordenador portátil al azar, ¿qué probabilidad hay de que haya sido comprado en la tienda online? *(2,5 puntos)*

c) Si se escoge un portátil al azar y resulta que fue comprado en la tienda online, ¿qué probabilidad hay de que sea HPH? *(3 puntos)*

d) Teniendo en cuenta la información anterior, si en una clase hubiesen 30 alumnos ¿qué probabilidad hay que 3 de ellos hayan comprado su ordenador en la tienda online? Define la variable utilizada, indicando cuál es su distribución y el valor de sus parámetros. *(3,5 puntos)*

3. El canal de la plataforma YouTube "Xes Music Official" recibe cada hora una media de 8 visitas. Si suponemos que las visitas recibidas en las distintas horas del día son independientes entre sí, responde a las siguientes preguntas:

a) Define la variable aleatoria objeto de estudio, indica su distribución y el valor de sus parámetros. ¿Sobre qué población está definida? *(2,5 puntos)*

b) Sabiendo que 8 es el valor medio de esta variable, ¿cuál es la probabilidad de que en una hora cualquiera elegida al azar se reciban exactamente 8 visitas? *(2,5 puntos)*

c) Teniendo en cuenta que el tiempo de cada visita sigue una distribución exponencial de mediana 3 minutos, se pide:

c.1) Calcula el porcentaje de visitas cuyo tiempo esté comprendido entre 3 y 5 minutos. *(3 puntos)*

c.2) Sabiendo que una visita elegida al azar lleva ya dos minutos, ¿qué probabilidad hay de que finalice antes de dos minutos más adicionales? *(2 puntos)*

4. Unos turistas que han venido varios días a las Fallas salen a pasear por la mañana, por la tarde y por la noche. Por la mañana caminan una media de 4,6 km con una desviación típica de 0,8 km; por la tarde caminan 7,8 km de media con una desviación típica de 0,6 km; y por la noche andan 4,4 km de media con una desviación típica de 1 km. Se asume un modelo normal en los tres casos y que las distancias recorridas son independientes entre sí.

a) Si se obtiene la variable aleatoria "*distancia total caminada en un día (mañana, tarde y noche) por un turista*", justifica la distribución de esta variable y obtén razonadamente el valor de sus parámetros. *(3 puntos)*

b) ¿Cuál es la probabilidad de que un turista elegido al azar haya caminado más de 17 km en un día? *(3 puntos)*

c) ¿A partir de qué longitud se puede considerar que un turista, a lo largo de un día, ha realizado un paseo anormalmente largo (que sólo se supera en el 3 por mil de los casos)? *(4 puntos)*

SOLUCIÓN

1a) Es falsa, ya que los individuos de la población son los componentes, y no los meses. Lo correcto sería decir: *"La variable aleatoria objeto de estudio está definida en la población de los distintos componentes en los cuales se mide el tiempo hasta la rotura"*.

1b) Utilidad de ambos gráficos: son herramientas gráficas de estadística descriptiva que permiten evaluar visualmente la pauta de variabilidad de los datos; en concreto, sobre su dispersión, parámetros de posición y sobre el grado de simetría de la distribución.

Utilidad del gráfico de caja y bigotes (box-whisker plot): además, al visualizar los valores extremos, permite discutir si alguno de ellos es un dato anómalo que requiere ser descartado. Es posiblemente la herramienta gráfica más útil para estudiar el grado de simetría de una distribución a partir de un conjunto reducido de datos. Por otra parte, facilita la identificación de los cuartiles. Otra de sus utilidades es que permite comparar dos conjuntos de datos mediante el gráfico box-whisker múltiple.

Utilidad del papel probabilístico normal: su principal utilidad es estudiar la normalidad de un conjunto de datos, es decir, si pueden considerarse como una muestra extraída aleatoriamente de una población con distribución normal. En ese caso, es posiblemente la herramienta gráfica más potente para estudiar la presencia de datos anómalos que no pertenezcan a la población y que posiblemente tengan que ser descartados.

1c) Si los datos se expresan en otras unidades (lo que implica ser multiplicados por una constante k), cambia proporcionalmente el valor de todos los parámetros de posición, que se multiplican por k. Los parámetros de dispersión también cambian excepto el coeficiente de variación: la desviación típica, el rango y el rango intercuartílico quedan multiplicados por k, y la varianza por k^2. Obviamente el parámetro no cambia si es cero.

1d) Las principales conclusiones derivadas de ambos gráficos son:
- Se observa una marcada asimetría positiva en la distribución de los datos. Por este motivo, tratándose de datos de tiempo con mínimo en cero, es posible que una distribución exponencial de mediana 7,5 modelice conveniente la pauta de variabilidad.
- En cuanto a los parámetros de posición: mediana = 7,5, media = 12, primer cuartil = 3, tercer cuartil = 15 (aproximadamente).
- En cuanto a la dispersión: los datos varían entre 0 y 67 (rango = 67) siendo el rango intercuartílico: 15 - 3 = 12 (aproximadamente).
- En el gráfico box-whisker se observan unos 12 datos extremos que quedan fuera del bigote derecho. No hay suficiente evidencia para afirmar que estos datos tengan que ser descartados por presentar algún tipo de anomalía que evidencie que no pertenecen a la misma población.

2a) Suceso A: el ordenador es de marca HPH; B: el ordenador es de marca ADER.
C: el ordenador es de marca SANY; T: el ordenador fue comprado en una tienda online.
Los porcentajes corresponden a las siguientes probabilidades:
0.3 = P(A); 0.15 = P(B); 0.2 = P(T/A); 0.3 = P(T/B); 0.25 = P(T/C).

2b) P(C)=1-0.3-0.15= 0.55; Aplicando el teorema de la probabilidad total se obtiene:
P(T) = P(A)·P(T/A) + P(B)·P(T/B) + P(C)·P(T/C) = 0.3·0.2 + 0.15·0.3 + 0.55·0.25 = **0.2425**

2c) Aplicando el teorema de Bayes se obtiene:

$$P(A/T) = \frac{P(A) \cdot P(T/A)}{P(T)} = \frac{0.3 \cdot 0.2}{0.2425} = 0.2474$$

2d) Variable aleatoria X: *nº de alumnos que han comprado su ordenador en la tienda online en una clase de 30 alumnos*. El valor máximo de esta variable discreta es 30, por lo cual sigue una distribución Binomial de parámetros n=30, p = 0.2425 (probabilidad obtenida del apartado 2b). Se pide calcular:

$P(X=3)=\binom{30}{3}\cdot 0.2425^3\cdot(1-0.2425)^{27}$ = 4060·0.0143·0.7575^{27} = **0.0321**

$$\binom{30}{3}=\frac{30!}{3!\cdot 27!}=\frac{30\cdot 29\cdot 28\cdot 27!}{6\cdot 27!}=5\cdot 29\cdot 28=4060$$

3a) Variable aleatoria X: *nº de visitas recibidas en el canal "Xes Music Official" en una hora*. En esta variable discreta, el valor mínimo es cero y el máximo no está acotado, por lo que puede modelizarse según una distribución de tipo Poisson de parámetro λ=8 (coincide con la media). Dado que se dispone de un dato cada hora, los individuos de la población son "horas": la población sería el conjunto de todas las horas de las cuales se mide el número de visitas recibidas en este canal de YouTube.

3b) Aplicando la función de probabilidad de la distribución de Poisson:

$P(X=8)=e^{-8}\cdot\frac{8^8}{8!}$ = **0.1396**

3c1) Siendo T la variable aleatoria, al ser la mediana de valor 3: P(T>3)=0.5=$e^{-\alpha\cdot 3}$; despejando de esta ecuación: α = - (ln 0.5)/3 = 0.231.

P(3<T<5) = P(T>3) - P(T>5) = 0.5 - $e^{-0.231\cdot 5}$ = 0.5 - 0.315 = 0.185 = **18.5%**

3c2) Si la visita lleva ya dos minutos, su valor finalmente será superior a 2. Nos piden calcular una probabilidad condicional, que puede resolverse aplicando la propiedad de falta de memoria *(p.f.m.)* de la distribución exponencial:

P(T<4 / T>2)= 1 - P(T>4 / T>2) = *(pfm)* = 1 - P(T>2) = 1 - $e^{-0.231\cdot 2}$ = 1- 0.63 = **0.37**

4a) Sean las variables X_1, X_2 y X_3 las distancias caminadas (km) por la mañana, tarde y noche, respectivamente, de modo que: $X_1 \approx N(4.6; 0.8)$; $X_2 \approx N(7.8; 0.6)$; $X_3 \approx N(4.4; 1)$.
La distancia total D caminada en un día será: $D = X_1+X_2+X_3$, seguirá una distribución normal por ser suma de variables normales independientes. Sus parámetros son:
Media: $m_D = m_{X1} + m_{X2} + m_{X3}$ = 4.6 + 7.8 + 4.4 = 16.8
Por ser independientes, $\text{varianza}_D = \text{var}_{X1} + \text{var}_{X2} + \text{var}_{X3} = 0.8^2 + 0.6^2 + 1^2 = 2$
Desviación típica = $\sqrt{2}$ = 1.414. Por tanto, **D ≈ N(16.8; 1.414)**

4b) P(D>17) = P[N(16.8; 1.414) >17] = P[N(0;1) > (17-16.8)/1.414] =
= P [N(0;1) > 0.14] = **0.444**

4c) La longitud *k* que nos piden deberá cumplir: P(D>*k*) = 0.003;
P[N(16.8; 1.414) >*k*] = 0.003; P[N(0;1) > (k-16.8)/1.414] = 0.003; a partir de la tabla de la normal tipificada se obtiene: (k-16.8)/1.414 = 2.75. Despejando, resulta: k = **20.69 km**