etsinf

Grado en Ingeniería Informática

Estadística

PRIMER PARCIAL

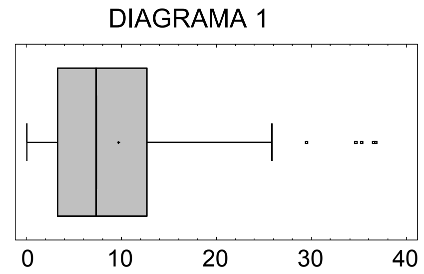
7 de abril de 2017

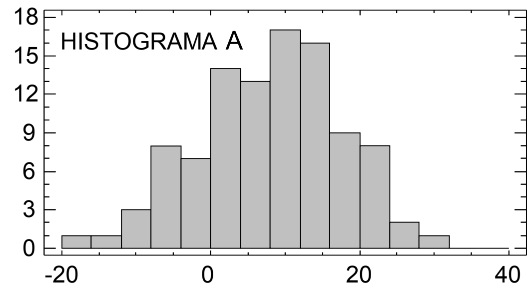
Apellidos, nombre:	
Grupo:	Firma:

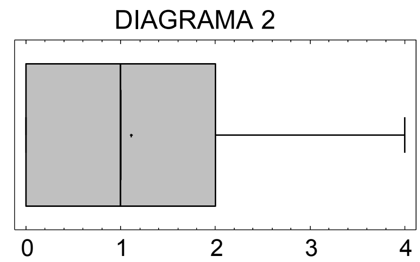
Instrucciones

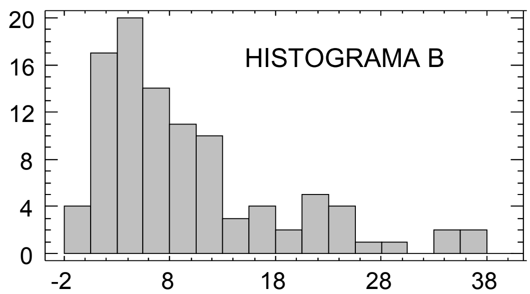
1. Rellenar la información de cabecera del examen.
2. Responder a cada pregunta en la hoja correspondiente.
3. Justificar todas las respuestas.
4. No se permiten anotaciones personales en el formulario.
5. No se permite tener teléfonos móviles encima de la mesa. Sobre la mesa sólo se permite el DNI, 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: **2 horas**

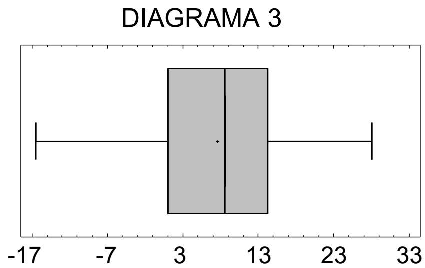
1. Se ha realizado un estudio descriptivo de tres muestras de **tamaño 100** correspondientes a distintos tipos de variables aleatorias. De cada grupo de datos se ha obtenido un histograma de frecuencias y un gráfico de caja y bigotes, los cuales se presentan a continuación de forma desordenada.

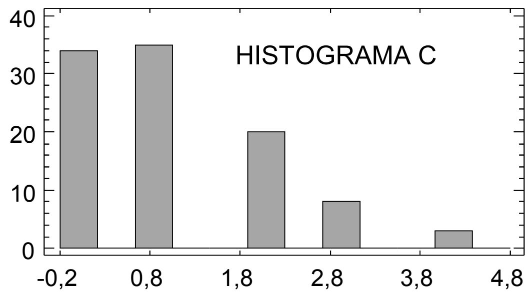












Además de los gráficos anteriores, también se dispone de los coeficientes de asimetría estandarizados (CAE) y de curtosis estandarizados (CCE) para cada uno de los tres conjuntos de valores, indicados en la tabla inferior.

a) Para cada fila de la tabla, selecciona una de las tres opciones de diagrama y de histograma. Indica también la correspondencia entre los gráficos anteriores mediante flechas. Justifica razonadamente la respuesta. *(6 puntos)*

Coef. de asimetría y curtosis est.	DIAGRAMA	HISTOGRAMA
CAE = 3,29556, CCE = 0,13803	1 - 2 - 3	A - B - C
CAE = -1,08673, CCE = -0,83587	1 - 2 - 3	A - B - C
CAE = 5,51562, CCE = 2,81815	1 - 2 - 3	A - B - C

b) Indica cuáles serían los parámetros más representativos para describir la posición y dispersión del conjunto de datos representado por el diagrama *Box & Whisker* 3. Calcula de forma aproximada, si es posible, su valor. Justifica todas tus respuestas. *(4 puntos)*

2. En cierta unidad hospitalaria se lleva a cabo un test clínico para detectar cierta patología congénita por medio de un chip de ADN cuyos datos se analizan con técnicas bioinformáticas. A partir de datos históricos, se sabe que esta patología la sufre el 4% de los enfermos ingresados en dicha unidad hospitalaria. Si un paciente padece realmente dicha patología, el test da positivo en el 95% de los casos. Pero si el paciente no la padece, el test da positivo (es decir, indica que existe la patología cuando realmente no es cierto) en el 2% de casos.

a) Define todos los sucesos involucrados en este problema. *(1 punto)*

b) Calcular el porcentaje de casos en los que el test ofrece un resultado positivo. *(3 puntos)*

c) Calcular el porcentaje de casos en los que el test ofrece un diagnóstico erróneo. *(3 puntos)*

d) En caso de que se realice el test con un paciente y éste resulte positivo, ¿qué probabilidad existe de que realmente padezca la patología? *(3 puntos)*

3. Cierta empresa de desarrollo de aplicaciones tiene constatado que sus programadores cometen en promedio 2 errores por página de código.

a) Calcular la probabilidad de que se cometan más de 6 errores en una página de código. *(4 puntos)*

b) En una aplicación con 50 páginas de código, ¿qué probabilidad hay de encontrar al menos una página con más de 6 de este tipo de errores? *(6 puntos)*

4. Una empresa fabrica equipos de resonancia magnética nuclear, de un determinado modelo, para hospitales. El precio de venta se negocia con cada cliente, resultando ser una variable aleatoria de media 500.000 €, con una desviación típica de 80.000 €. El coste de fabricación de cada uno de estos equipos consta de un término fijo de 200.000 €, y un coste variable cuya distribución es normal, de media 100.000 € y desviación típica 20.000 €. Se asume que las dos variables son independientes.

a) Indica cuál es la distribución de la variable aleatoria "coste total de fabricación", y calcula el valor de sus parámetros. *(3 puntos)*

b) Indica cuál es la distribución de la variable aleatoria "beneficio" (calculado como precio de venta menos coste total de fabricación), indicando el valor de sus parámetros. *(4 puntos)*

c) Calcula la probabilidad de que el beneficio obtenido por la venta de uno de estos equipos sea inferior a 230.000 €. *(3 puntos)*

5. El tiempo que cada usuario emplea en realizar una consulta en el servidor de una biblioteca fluctúa exponencialmente. Se sabe que el 25% de las consultas tardan menos de 5 minutos en finalizarse.

a) ¿Cuál es la probabilidad de que un usuario esté exactamente 5 minutos? *(1 punto)*

b) Calcula el tiempo medio que se tarda en atender una consulta. *(3 puntos)*

c) Cuando llega un usuario, encuentra que el servidor está ocupado por un usuario que lleva ya 7 minutos realizando una consulta. Calcula la probabilidad de que tenga que esperar más de cinco minutos adicionales. *(2 puntos)*

d) A lo largo de una mañana se realizan 20 consultas consecutivas. Calcula la probabilidad de que la duración total de ellas sea superior a 4 horas. *(4 puntos)*

,SOLUCIÓN

1a) La correspondencia entre coeficientes, diagrama e histograma se indica a continuación:

Coef. de asimetría y curtosis est.	DIAGRAMA	HISTOGRAMA
CAE = 3,29556, CCE = 0,13803	2	C
CAE = -1,08673, CCE = -0,83587	3	A
CAE = 5,51562, CCE = 2,81815	1	B

CAE = -1,09; CCE = -0,84: Al estar ambos comprendidos entre -2 y 2 pueden considerarse nulos a nivel poblacional, lo cual indica una muestra procedente de una distribución normal. Corresponde al **diagrama 3 e histograma A**, que son los más simétricos.

El **diagrama 1** se asocia con el **histograma B** por coincidir el rango de valores. Su coeficiente es CAE = 5,51 por presentar la mayor asimetría positiva de los tres, debido a que la media se acerca "bastante" al tercer cuartil, y aparecen puntos aislados (atípicos), lo cual justifica el mayor valor de CCE (2,81).

La muestra de valores discretos de 0 a 4 corresponde al **diagrama 2 e histograma C**, que también refleja asimetría positiva (CAE = 3,29 > 2) pero menos pronunciada que en el diagrama 1, pues la media es similar a la mediana. Este hecho junto a la ausencia de anomalías (datos extremos) proporciona un valor del CCE entre -2 y +2.

1b) Dado que el conjunto de datos representado por el diagrama Box & Whisker nº 3 puede considerarse como una muestra tomada de una población normal, los parámetros más representativos para describir la **posición** son la media (m = 7,5 que es el punto dibujado en el centro de la caja del diagrama 3), que coincide aproximadamente con la mediana (8,5: línea vertical dentro de la caja).

En una distribución normal, los parámetros de **dispersión** más representativos son la desviación típica (s) y la varianza (s^2):

a) Cálculo de la desviación típica: El intervalo comprendido entre $m \pm 2 \cdot s$ comprende el 95% de los datos. El histograma A se ha construido con 100 valores, de modo que se trata de encontrar el intervalo fuera del cual hay 5 valores (2 por debajo y 3 por encima, o al revés). A la vista de la figura, dicho intervalo está comprendido entre -12 y 24 y abarca 4 veces la desviación típica (2s por encima de m, y 2s por debajo). Por tanto: **s ≈** (24+12)/4 = **9**.

b) Cálculo de la varianza: $s^2 = 9^2 =$ **81** aproximadamente.

Nota: en una distribución normal, el intervalo intercuartílico es menos representativo que los dos anteriores. Su valor es: $Q_3 - Q_1 = 14{,}5 - 1 = 13{,}5$.

2a) Suceso A: el paciente ingresado en la unidad hospitalaria sufre la patología congénita. Suceso $\overline{A}$: el paciente no sufre la patología congénita.
Suceso B: el test ofrece un resultado positivo. Las probabilidades citadas en el enunciado son: $P(A) = 0{,}04$; $P(B/A) = 0{,}95$; $P(B/\overline{A}) = 0{,}02$.

2b) Aplicando el teorema de la probabilidad total se obtiene:
$P(B) = P(A)\cdot P(B/A) + P(\overline{A})\cdot P(B/\overline{A})$ = 0,04·0,95+0,96·0,02 = 0,0572 → **5,72%**

2c) El diagnóstico puede ser erróneo por dos motivos: el paciente no sufre la patología **y** el test sale positivo, <u>o bien</u> porque el paciente sufre la patología **y** el test no sale positivo. Escrito en forma de sucesos sería:

P (diag. Erróneo) = $P\left[(\overline{A}\cap B)\cup(A\cap\overline{B})\right] = P(\overline{A}\cap B) + P(A\cap\overline{B}) =$
$= P(\overline{A})P(B/\overline{A}) + P(A)\cdot P(\overline{B}/A) = 0{,}96\cdot 0{,}02 + 0{,}04\cdot 0{,}05 = 0{,}0212 \rightarrow$ **2,12%**

2d) Teorema de Bayes: $P(A/B) = \dfrac{P(A\cap B)}{P(B)} = \dfrac{P(A)\cdot P(B/A)}{P(B)} = \dfrac{0{,}04\cdot 0{,}95}{0{,}0572} = 0{,}664$

3a) Variable aleatoria X: nº de errores cometidos por página de código. Esta variable X sigue una distribución Poisson (λ) ya que el valor mínimo es 0 y el máximo no está acotado. En esta distribución, λ coincide con la media (λ=2).
$P(X > 6) = 1 - P(X \le 6) = 1 - 0{,}9955 =$ **0,0045**
El valor 0,9955 se obtiene a partir del ábaco de Poisson, levantando una vertical en λ=2, cortando la curva "6" y leyendo la probabilidad en el eje vertical.

3b) Variable aleatoria Y: nº de páginas que cumplen la condición "tener más de 6 errores" en una aplicación con 50 páginas de código. Esta variable sigue una distribución Binominal (n=50, p=0,0045) ya que el valor mínimo es 0 y el máximo es 50. La probabilidad del suceso "tener más de 6 errores" es la obtenida en el apartado anterior. Aplicando la función de probabilidad:
$P(Y \ge 1) = 1 - P(Y = 0) = 1 - \binom{50}{0}\cdot 0{,}0045^0 \cdot 0{,}9955^{50} = 1 - 0{,}9955^{50} =$ **0,203**

4a) CTF (coste total de fabricación) = CF (coste fijo) + CV (coste variable)
CF = 200.000; CV ≈ N (m=100.000; σ=20.000);
El coste fijo incrementa la media pero no afecta a la dispersión. Por tanto:
CTF ≈ N (m=300.000; σ=20.000) → El coste total de fabricación sigue una distribución normal de media 300.000 € y desviación típica 20.000 €.

4b) PV (precio de venta) ≈ N (m=500.000; σ=80.000).
No se especifica en el enunciado que su distribución sea normal, pero es lo más lógico ya que CV y CTF son variables normales.
Beneficio = PV - CTF.
Seguirá un modelo normal por ser la resta de variables normales. Sus parámetros, asumiendo independencia para el cálculo de la varianza, serán:
E(beneficio) = E(PV) - E(CTF) = 500.000 - 300.000 = 200.000
$\sigma^2(beneficio) = \sigma^2(PV - CTF) = \sigma^2(PV) + \sigma^2(CTF) = 80.000^2 + 20.000^2 = 6{,}8\cdot 10^9$
$\sigma(beneficio) = \sqrt{6{,}8\cdot 10^9} = 82.462$ → **Beneficio ≈ N (m=200.000, σ = 82.462).**

4c) $P[N(200.000;\ 82.462) < 230.000] = P\left[N(0;1) < \frac{230.000 - 200.000}{82.462}\right] =$

$= P[N(0;1) < 0{,}364] = 1 - 0{,}358 = \mathbf{0{,}642}$

5a) La probabilidad de que una variable continua tome exactamente un valor tiende a cero.

5b) Aplicando la función de distribución se obtiene:

$P(X < 5) = 0{,}25 = 1 - e^{-\alpha \cdot 5};\ e^{-\alpha \cdot 5} = 0{,}75;\ \alpha = -(\ln 0{,}75)/5 = 0{,}0575$

En una distribución exponencial, la media es la inversa del parámetro:

$m = 1/\alpha = 1/0{,}0575 = \mathbf{17{,}38\ min}$

5c) Nos piden una probabilidad condicional, que aplicando la propiedad de falta de memoria (pfm) de la distribución exponencial, y teniendo en cuenta que el valor 5 es el primer cuartil (dato del enunciado) queda del siguiente modo:

$P[(X > 12)/(X > 7)] \xrightarrow{p.f.m.} P(X > 5) = \mathbf{0{,}75}$

5d) En una distribución exponencial, la desviación típica coincide con la media. Aplicando el teorema central del límite, la suma de distribuciones exponenciales independientes (Y) tiende a una normal, cuya media será la suma de las medias y cuya varianza será la suma de las varianzas:

m=17,38; σ^2=17,38^2 = 302,1 ; $Y = X_1 + ... + X_{20}$

$E(Y) = E(X_1 + ... + X_{20}) = E(X_1) + ... + E(X_{20}) = 20 \cdot 17{,}38 = 347{,}6$

$\sigma^2(Y) = \sigma^2(X_1 + ... + X_{20}) = \sigma^2(X_1) + ... + \sigma^2(X_{20}) = 20 \cdot 17{,}38^2 = 6041{,}5$

$P(Y > 4 \cdot 60) = P\left[N\left(347{,}6; \sqrt{6041{,}5}\right) > 240\right] = P\left[N(0;1) > \frac{240 - 347{,}6}{\sqrt{6041{,}5}}\right] =$

$= P[N(0;1) > -1{,}384] = 1 - 0{,}0831 = \mathbf{0{,}917}$