Grado en Ingeniería Informática

Estadística

PRIMER PARCIAL

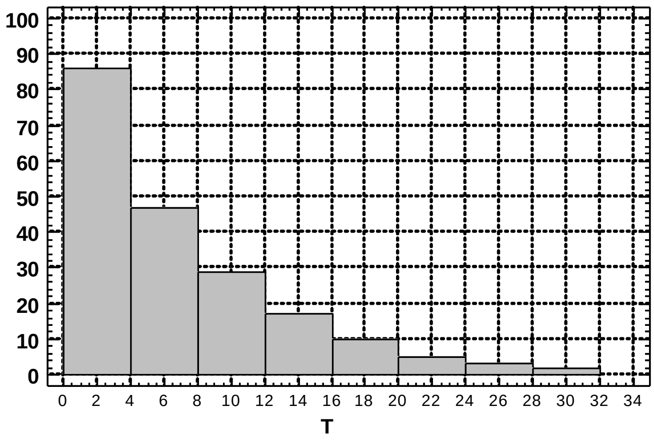
11 de abril de 2018

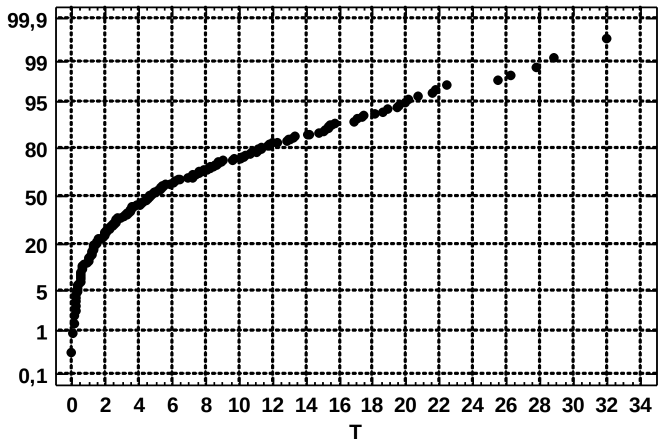
Apellidos, nombre:	
Grupo:	Firma:

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. El tiempo que tarda un algoritmo en procesar ciertas matrices de datos es una variable aleatoria. Se han obtenido 200 valores de tiempo (en milisegundos, ms) correspondientes a una muestra aleatoria de 200 matrices de tamaño muy distinto. Con estos datos se ha construido los siguientes gráficos:





Responde a las siguientes preguntas <u>justificando convenientemente</u> la respuesta.

a) El gráfico de la derecha es un papel probabilístico normal. ¿Cómo se llama el gráfico de la izquierda? Indica la utilidad de cada uno de ellos, y qué se representa en cada uno de los ejes. *(4 puntos)*

b) A la vista de las representaciones gráficas anteriores, indica a qué modelo teórico de distribución de probabilidad, de los que has estudiado, se asemeja el comportamiento del tiempo que tarda un algoritmo en procesar cada matriz.

(3 puntos)

c) El responsable de control de calidad considera que la media es el parámetro más representativo para describir la posición de estos datos, y que el recorrido intercuartílico es el más representativo para describir su dispersión. ¿Estás de acuerdo con esta afirmación? *(3 puntos)*

2. Una fábrica produce 10.000 envases diarios. La máquina A produce 3.000 de estos envases de los cuales el 1 % son defectuosos. El resto son producidos por la máquina B de los que se sabe que el 98 % son de buena calidad.

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

b) Determinar la probabilidad de que un envase elegido al azar sea defectuoso. *(4 puntos)*

c) Si se toma al azar un envase y resulta ser de buena calidad, ¿cuál es la probabilidad de haber sido producido por la máquina A? ¿Y por la máquina B? *(5 puntos)*

3. Unos cargadores USB universales se pueden empaquetar en cajas de 5 o de 10 unidades. Se sabe que la probabilidad de que cada cargador USB sea defectuoso es del 2%.

a) Si una empresa compra una caja de 5 cargadores USB universales y otra caja de 10 unidades, calcula la probabilidad de que en total haya menos de 2 cargadores defectuosos. *(4 puntos)*

b) Define la variable aleatoria implicada en el apartado anterior, y calcula su esperanza matemática. *(2 puntos)*

c) Esta misma empresa compra también cables para los cargadores, los cuales presentan un número medio de 0,5 defectos estéticos por metro. Si cada cargador necesita un cable de 2 metros, ¿qué probabilidad hay de que la empresa tome el cable de un cargador y encuentre como mínimo dos defectos estéticos? *(4 puntos)*

4. El tiempo de funcionamiento hasta el fallo de un determinado componente electrónico, medido en horas, sigue una distribución exponencial cuya mediana es de 69 horas.

a) Si a las 100 horas un componente todavía está en funcionamiento, ¿cuánto vale la probabilidad de que en total funcione más de 250 horas? *(3 puntos)*

b) Se tiene un dispositivo formado por 10 componentes de este tipo. El dispositivo está montado de manera que comienza a funcionar un componente hasta que se estropea y en ese preciso instante se pone a funcionar automáticamente el siguiente componente, y así sucesivamente hasta que se estropean los diez; en este momento el dispositivo deja de funcionar.

b.1) ¿Cómo se define en este caso la duración del dispositivo? Calcula la media y varianza de esta variable aleatoria. *(3 puntos)*

b.2) ¿Cuál es aproximadamente la probabilidad de que el dispositivo dure menos de 1.700 horas? *(4 puntos)*

5. La naviera GLOBAL LINE S.A. dispone de una flota de cargueros para transportar contenedores por diferentes rutas. La empresa está analizando el consumo de combustible según la ruta. Sabe que el consumo viene determinado por la velocidad y la longitud de la ruta mediante la siguiente expresión:

$$\text{Consumo}=0.2\cdot\text{Velocidad}+0.8\cdot\text{Longitud\_de\_ruta}$$

Además, la naviera ha estimado que la velocidad se comporta como una normal de media 100 unidades y desviación típica 10. La longitud de una ruta se modeliza como una normal de media 5.000 y desviación típica 200. Suponiendo que las variables velocidad y longitud de ruta son independientes, se pide:

a) ¿Qué probabilidad hay de que el consumo de una ruta escogida al azar sea superior a 4.340 unidades? *(7 puntos)*

b) Calcula el segundo cuartil de la variable aleatoria Consumo. *(3 puntos)*

SOLUCIÓN

1a) El gráfico de la izquierda se denomina histograma de frecuencias.
- En el histograma, en el eje de ordenadas (vertical) se representa en este caso la frecuencia absoluta, es decir, el número de valores comprendido en cada intervalo de valores de tiempo.
- En el papel probabilístico normal, en el eje de ordenadas se representa la frecuencia relativa acumulada (%): 100·P(X≤x), es decir, el porcentaje de valores menores o iguales a cada valor indicado en el eje horizontal.
- En ambos gráficos, en el eje de abscisas (horizontal) se representan los valores de la variable aleatoria T (tiempo requerido en procesar una matriz).
- Utilidad del histograma: permite visualizar gráficamente el patrón de variabilidad de los datos: nos hacemos una idea aproximada del valor mínimo, máximo, moda y mediana. Permite determinar si la distribución es simétrica o asimétrica, y detectar pautas atípicas en la distribución, por ejemplo: valores anómalos, distribuciones bimodales, datos truncados, etc.
- Utilidad del papel probabilístico normal: si los datos se ajustan aproximadamente a una recta puede concluirse que la muestra ha sido extraída de una población cuya distribución es normal. Es más potente que el histograma para este estudio, y también para detectar la presencia de datos anómalos.

1b) La forma del histograma indica que los valores más frecuentes son los cercanos a cero, y la frecuencia disminuye progresivamente. Esto se corresponde con la distribución exponencial. Por tanto, la variable T puede modelizarse según una distribución exponencial con mediana = 5 (percentil 50, obtenido a partir del papel probabilístico normal).

1c) No es correcto afirmar que la media sea el parámetro de posición más representativo, ya que está muy influenciada por los valores extremos, debido a que la distribución es muy asimétrica. Sería más correcto emplear la mediana, ya que no está influenciada por valores extremos o anómalos. Sin embargo, es correcto afirmar que el recorrido intercuartílico es el parámetro de dispersión más representativo, ya que indica el rango dentro del cual se encuentra el 50% de los valores, y no está afectado por los valores más elevados.

2a) Suceso A: el envase ha sido producido por la máquina A.
Suceso B (o bien $\overline{A}$): el envase ha sido producido por la máquina B.
Suceso D: el envase es defectuoso.

2b) P(A) = 3.000/10.000 = 0,3; P(B) = 1 - 0,3 = 0,7
P(D/A) = 0,01; P(D/B) = 1 - 0,98 = 0,02
Aplicando el teorema de la probabilidad total:
P(D) = P(A)·P(D/A) + P(B)·P(D/B) = 0,3·0,01 + 0,7 · 0,02 = **0,017**

2c) Aplicando el Teorema de Bayes se obtiene:

$$P\left(A/\overline{D}\right)=\frac{P(A)\cdot P\left(\overline{D}/A\right)}{P\left(\overline{D}\right)}=\frac{P(A)\cdot[1-P(D/A)]}{1-P(D)}=\frac{0.3\cdot(1-0.01)}{1-0.017}=0.3021$$

$$P\left(B/\overline{D}\right)=\frac{P(B)\cdot P\left(\overline{D}/B\right)}{P\left(\overline{D}\right)}=\frac{P(B)\cdot[1-P(D/B)]}{1-P(D)}=\frac{0.7\cdot(1-0.02)}{1-0.017}=0.698$$

Otra forma: $P\left(B/\overline{D}\right)=1-P\left(A/\overline{D}\right)=1-0.3021=0.698$

3a) Tomar primero una caja de 5 cargadores, luego otra de 10 y contar el número de defectuosos que hay en conjunto, es equivalente a tomar directamente una muestra de 15 cargadores. La variable aleatoria X (nº de cargadores defectuosos en una muestra de 15) será una distribución Binomial con n=15, p=0,02.

$$P(X<2)=P(X=0)+P(X=1)=\binom{15}{0}\cdot 0.02^0\cdot 0.98^{15}+\binom{15}{1}\cdot 0.02\cdot 0.98^{14}=$$

$$=0.98^{15}+15\cdot 0.02\cdot 0.98^{14}=0.7386+0.2261=0.965$$

Otra forma: puede aproximarse a una Poisson, pues se obtiene un valor similar:

$$X\approx Ps(\lambda=15\cdot 0.02=0.3)\ ;\quad P(X<2)=e^{-0.3}\cdot\frac{0.3^0}{0!}+e^{-0.3}\cdot\frac{0.3^1}{1!}=e^{-0.3}(1+0.3)=0.963$$

3b) Variable aleatoria X: número de cargadores USB universales defectuosos en una muestra total de 15 cargadores.
La esperanza matemática de una variable aleatoria es su valor medio, que en una distribución Binomial se calcula como: E(X) = m = n·p = 15·0,02 = **0,3**.

3c) Variable Y: nº de defectos estéticos en un metro de cable. El valor mínimo es cero y el máximo no está acotado, de modo que seguirá un modelo Poisson cuyo parámetro λ coincide con la media: Ps(λ=0,5).
Variable Z: nº de defectos estéticos en dos metros de cable: $Z=Y_1+Y_2$ (es decir, los defectos encontrados en el primer metro de cable más en el segundo metro), seguirá una distribución Ps ($\lambda_1+\lambda_2$), es decir, Ps (λ=1).

$$P(Z\geq 2)=1-P(Z=0)-P(Z=1)=1-e^{-1}\cdot\frac{1^0}{0!}-e^{-1}\cdot\frac{1^1}{1!}=1-e^{-1}(1+1)=0.264$$

Otra forma: P(Z≥2) = 1-P(Z≤1), lo cual se obtiene a partir del ábaco de Poisson, con λ=1, leyendo en la curva "1".

4a) P(T>69)=0,5; $e^{-\alpha\cdot 69}=0.5$; -69α = ln(0,5); α=-(ln0,5)/69 = 0,01005
Aplicando la propiedad de falta de memoria de la distribución exponencial:

$$P[(T>250)/(T>100)]=P(T>150)=e^{-0.01\cdot 150}=\mathbf{0{,}222}$$

4b1) Variable T_i: tiempo de funcionamiento (duración) del componente i.
La media será: E(T)=1/α = 1/0,01005 = 99,546. Coincide con la desv. típica.
La duración total del dispositivo, según se describe en el enunciado, se define como suma del tiempo de funcionamiento de los 10 componentes: T_{total} = $T_1+T_2+\ldots+T_{10}$. Estas variables son independientes entre sí.
Media = $E(T_1+\ldots+T_{10})$ = $E(T_1)+\ldots+E(T_{10})$ = $10\cdot E(T_i)$ = 10·99.55 = **995,5** horas

$$\sigma^2(T_{total})=\sigma^2(T_1+\ldots+T_{10})=\sigma^2(T_1)+\ldots+\sigma^2(T_{10})=10\cdot\sigma^2(T_i)=10\cdot 99.55^2=\mathbf{99094}\text{ horas}^2$$

La desviación típica es la raíz cuadrada: 314,8 horas.

4b2) Según el Teorema Central del Límite, el tiempo total será una distribución aproximadamente normal, por ser suma de variables aleatorias. Por tanto:

$$P(T_{total}<1700)=P[N(995.5;\ 314.8)<1700]=P[N(0;\ 1)<(1700-995.5)/314.8]=$$

$$=P[N(0;\ 1)<2.238]=1-0.0126=0.987$$

5a) Datos: Velocidad ~ Normal (m=100 σ=10); Longitud ruta ~ N (m=5000 σ=200)
C: consumo de combustible para una ruta cualquiera $\Rightarrow$ en el enunciado se indica que: C = 0.2 Velocidad + 0.8 Longitud_ruta;
C se comporta como una Normal al ser combinación de normales.
$E(C) = 0.2\ m_{veloc} + 0.8\ m_{long\_ruta} = 0.2 \cdot 100 + 0.8 \cdot 5000 = 4020$ unidades
Varianza de C $= 0.2^2 \cdot 10^2 + 0.8^2 \cdot 200^2 = 25604 \Rightarrow \sigma_{(Consumo)} = 160$
$P(C>4340) = P[N(4020; 160) > 4340] = P[N(0;1) > (4340-4020)/160] =$
$= P[N(0;1) > 2] =$ **0.0228**

5b) $P(C<Q_2)=0.50 \Rightarrow$ Como es una distribución normal, la mediana (segundo cuartil) coincide con la media. Por tanto, Q_2= **4020** unidades.