UOC - Primavera 2017 – Exámenes

# Examen 1 (10-06-17)

**Problema 1**. En un centro de documentación se quiere ver si las consultas de un determinado diario se hacen por Internet (desde mismo centro) o consultándolo en papel. Se toma nota del número de consultas durante 8 días y se obtienen los siguientes datos:

12, 6, 7, 3, 15, 10, 18, 5 (consultas en papel)

9, 3, 8, 8, 9, 8, 9, 18 (consultas electrónicas)

Se pide:

a) Encontrad el rango (o recorrido) en ambos casos y comentad su validez como medida de dispersión (en este caso).

b) Encontrad la media y los tres cuartiles de las dos muestras.

c) Haced un diagrama de caja para cada variable.

*Criterios de puntuación y valoración (sobre 10): a) 3 puntos; b) 4 puntos y c) 3 puntos.*

SOLUCIÓN:

a) En ambos casos, rango = número más grande - número más pequeño = 18 - 3 = 15

En este caso, si miramos las distribuciones de los datos de la distribución, tenemos:

Papel: 3, 5, 6, 7, 10, 12, 15, 18 y Electrónica: 3, 8, 8, 8, 9, 9, 9, 18

y vemos que la dispersión en el primer caso es más grande que en el segundo. Y como el rango no pone de manifiesto ninguna diferencia entre las dos distribuciones, tendríamos que decir que, en este caso, el rango no es una buena medida de dispersión.

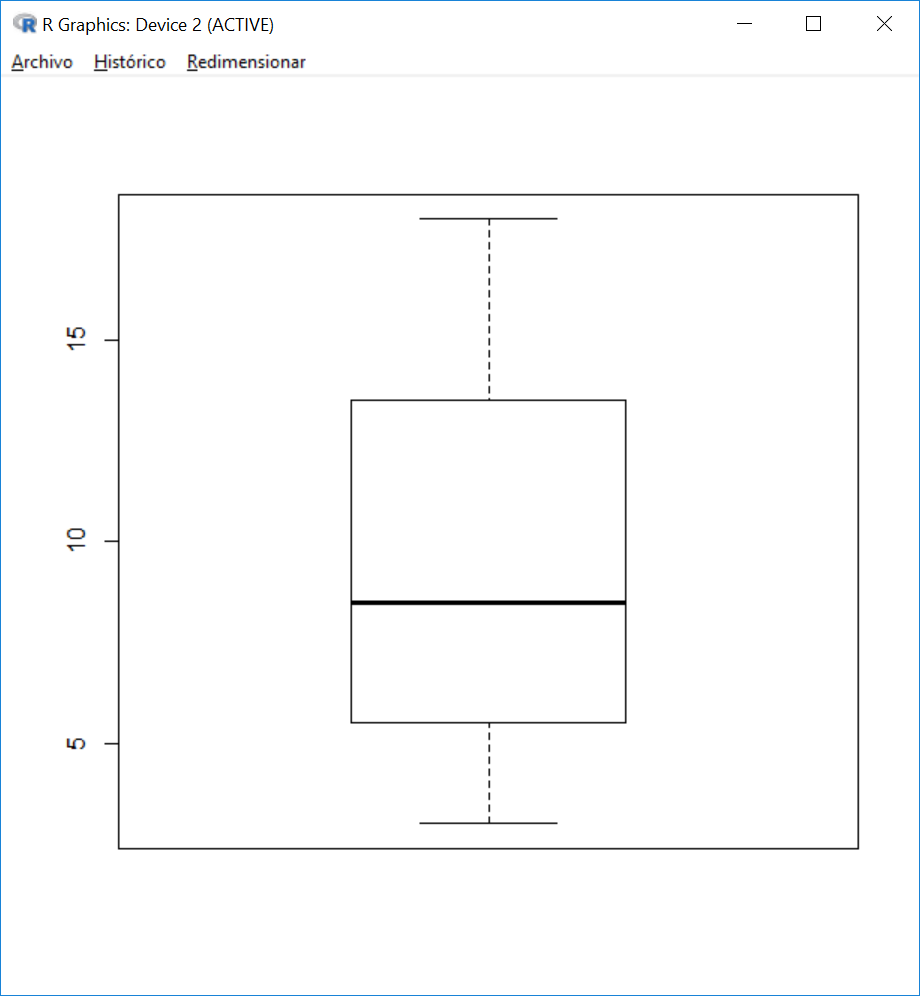
b) Media (papel)

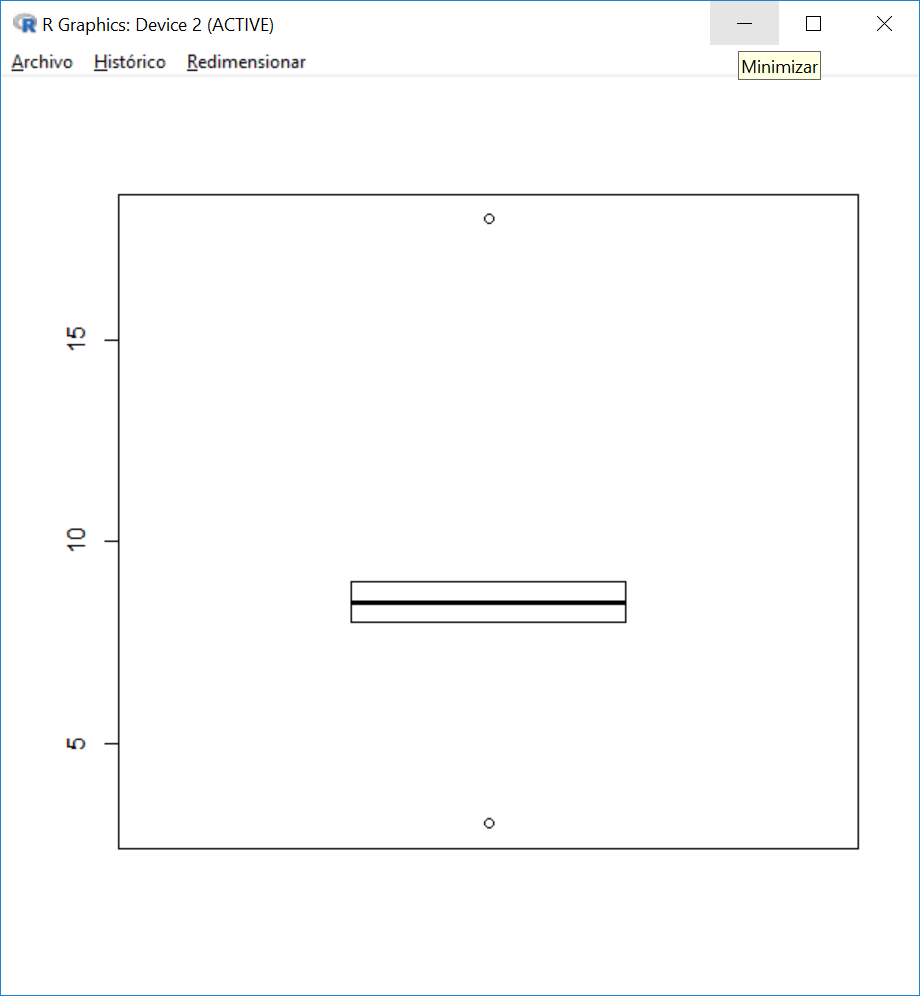


Mediana = (7+10)/2 = 8.5, Q1=5.5 Q3=13.5

Media (electrónico) = 9

Mediana = (8+9)/2 = 8.5 Q1=8 Q3=9

c) Boxplot de “Papel” Boxplot de “Electrónica”.



**Problema 2.**

La tabla siguiente presenta un resumen de las características solicitadas en 940 órdenes de compra de computadoras.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Memoria Adicional | |
|  |  | No | Sí |
| Procesamiento opcional de alta velocidad | No | 514 | 68 |
| Sí | 112 | 246 |

a) Completad la tabla de contingencia con los totales.

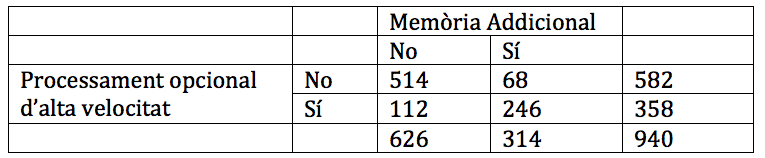
b) Calculad la probabilidad de que una orden de compra solicite un procesador de alta velocidad y no pida memoria adicional.

c) ¿Cuál es la probabilidad de que en una orden de compra se pida un procesador con alta velocidad sabiendo de que tiene memoria adicional?

d) ¿Los sucesos “pedir procesador de alta velocidad” y “pedir memoria adicional”, son independientes? ¿Por qué?

*Criterios de puntuación y valoración (sobre 10): a) 1 punto; b) 3 puntos), c) 3 puntos y d) 3 puntos.*

SOLUCIÓN:



b) P(SI Alta velocidad ∩ NO memoria adicional) =112/940= 0,1191=11,91%

c) P (SI Alta velocidad / memoria adicional)=246/314=0,7834= 78,34%

d) P(SI Alta velocidad)=358/940=0,38; P(SI memoria adicional)=314/940=0,33. Como P(SI Alta velocidad y SI memoria adicional)=246/940=0,26 es diferente de P(SI Alta velocidad)\* P(SI memoria adicional) decimos que los sucesos no son independientes.

**Problema 3**. Suponemos que el examen de una cierta asignatura lo han suspendido el 15% de los 200 estudiantes matriculados. Vamos escogiendo exámenes al azar y nos preguntamos:

a) ¿Cuál es la probabilidad de que el primer examen suspendido que encontremos sea el séptimo?

b) ¿Cuál es la probabilidad de que encontremos un examen **aprobado** estrictamente antes del tercer examen?

c) ¿Cuándo se espera que salga el primer examen suspendido?

En cada caso indicad cuál es la variable aleatoria, su tipo y los correspondientes parámetros.

*Criterios de puntuación y valoración (sobre 10): Se tienen que indicar las fórmulas y los cálculos realizados, así como los razonamientos. a) 4 puntos (1 punto distribución y parámetros, 3 puntos cálculos); b) 4 puntos (1 punto distribución y parámetros, 3 puntos cálculos); c) 2 puntos.*

SOLUCIÓN:

a) X=”Posición del primer examen suspendido”. Es geométrica de parámetro p=0.15. P(X=7)=0.856\*0.15=0.0566

b) Y=”Posición del primer examen aprobado”. Es geométrica de parámetro p=0.85. P(Y<=2)=1-0.152=0.9775

c) X=”Posición del primer examen suspendido”. Es geométrica de parámetro p=0.15. E(X)=1/0.15= 6.6667

**Problema 4**. Disponemos de una muestra de 300 exámenes de la UOC, con una nota media de 78 (sobre 100) y una desviación estándar de 20. Suponiendo que las notas siguen una distribución normal,

a) Calculad un intervalo de confianza para la media de las notas, con un nivel de confianza del 90%.

b) ¿Cuál tendría que ser el tamaño de la muestra para que la longitud del intervalo de confianza fuera inferior a 2?

Valores de probabilidades que os pueden ser útiles; si no encontráis exactamente el que necesitáis, utilizad el más cercano

|  |  |  |  |
| --- | --- | --- | --- |
| p(X>= x) | X~ N(0,1) | X~ t de Student con 299 grados de libertad | X~ t de Student con 300 grados de libertad |
| 0.01 | 2.326 | 2.339 | 2.339 |
| 0.025 | 1.96 | 1.967 | 1.968 |
| 0.05 | 1.645 | 1.65 | 1.65 |
| 0.005 | 2.575 | 2.592 | 2.592 |

*Criterios de puntuación y valoración (sobre 10): Se tienen que indicar las fórmulas y los cálculos realizados, así como los razonamientos. a) 6 puntos: planteamiento 2 puntos, cálculos 4 puntos; b) 4 puntos (planteamiento 2 puntos, cálculos 2 puntos).*

SOLUCIÓN

a) Es un intervalo de confianza de la media de una distribución normal con la desviación estándar desconocida=

b) Si la longitud es 2 entonces el margen de error es 1 y tenemos que =1089.12

**Problema 5**. Queremos contrastar si la media de mensajes recibidos por hora en un día en un servidor de correo es superior a 1500 mensajes. Escogemos una muestra aleatoria de 100 días, calculamos las medias del número de mensajes recibidos por hora obteniendo una media de 1529.971 mensajes con una desviación típica de 155.791.

Suponiendo que la variable que nos da la media del número de mensajes recibidos por hora es normal,

a) Estableced el contraste a realizar. Tenéis que indicar: si se trata de un contraste de una muestra o de dos muestras, de qué parámetro tenemos que hacer el contraste (media, proporción, etc.) y la hipótesis nula y la hipótesis alternativa.

b) ¿Cuál es el valor del estadístico de contraste?

c) A un nivel de significación de α=0.05, calculad el(los) valor(es) crítico(s).

d) ¿A qué conclusión llegaríais? No es suficiente afirmar si aceptamos o rechazamos la hipótesis nula, se tiene que razonar la conclusión en el contexto del problema.

*Criterios de puntuación y valoración (sobre 10): a) 3 puntos. Si indica si se trata de una muestra o dos, 1 punto. Si indica el parámetro, 1 punto. Si indica el contraste, 1 punto. b) 2 puntos. Indica la expresión del estadístico de contraste, 1 punto. Valor del mismo, 1 punto. c) 1 punto. d) 4 puntos: Indicar si se rechaza o se acepta, 1 punto. Razonamiento, 3 puntos.*

Valores de las tablas N(0,1) y la *t* de Student (utilizad los que consideréis necesarios, si necesitáis un valor que no es exacto, utilizad el valor más cercano)

|  |  |  |  |
| --- | --- | --- | --- |
| p(X >= x) | x (N(0,1)) | x t(99) | x t(100) |
| 0.01 | 2.326 | 2.365 | 2.364 |
| 0.025 | 1.96 | 1.984 | 1.984 |
| 0.05 | 1.645 | 1.66 | 1.66 |

SOLUCIÓN

a) Se trata de un contraste de una muestra y el parámetro es la media. El contraste a hacer es el siguiente:

donde representa la media de mensajes recibidos por hora en un día por el servidor.

b) El valor del estadístico de contraste es

c) El valor crítico será:

d) Como el valor del estadístico de contraste es mayor que el valor crítico, rechazaríamos la hipótesis nula y concluiríamos que efectivamente tenemos suficientes evidencias para afirmar que la media del número de mensajes por hora en un día es superior a 1500.

**Problema 6**

Queremos hacer un estudio para ver si el tiempo de ejecución de un determinado algoritmo de análisis de grafos depende o no linealmente del número de nodos del grafo. Se ha determinado el tiempo de ejecución en un determinado conjunto de grafos que tienen de 10 a 100 nodos. Hemos calculado la correspondiente recta de regresión y los resultados se muestran en la salida siguiente que nos ha dado R:

##   
## Call:  
## lm(formula = tiempo.ejecucion ~ nodos)  
##   
## Residuales:  
## Min 1Q Median 3Q Max   
## -49.462 -19.211 0.748 23.930 55.709   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.3274 6.9128 4.821 0.00000522 \*\*\*  
## nodos 5.2560 0.1099 47.825 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.32 donde 98 degrees of freedom  
## Multiple R-squared: 0.9589, Adjusted R-squared: 0.9585   
## F-statistic: 2287 donde 1 and 98 DF, p-value: < 0.00000000000000022

a) Encontrad la recta de regresión del tiempo de ejecución (en milisegundos) en función del número de nodos del grafo e interpretad los resultados.

b) Haced el contraste de hipótesis sobre la pendiente de la recta, con un nivel de significación de α=0.01 e interpretad los resultados. Tenéis que dar el contraste sobre la pendiente indicando la hipótesis nula y alternativa y el estadístico de contraste,

c) Calculad el coeficiente de determinación. Discutid si es un buen modelo.

*Criterios de puntuación y valoración (sobre 10): a) (3 puntos). Dar la recta de regresión, 1.5 puntos. Interpretación de los parámetros de la recta, 1.5 puntos. b) (5 puntos). Dar el contraste de hipótesis indicando la hipótesis nula y alternativa, 2 puntos. Dar el valor del estadístico, 3 puntos. c) (2 puntos). Dar el coeficiente de determinación, 1 punto. Interpretación del mismo, 1 punto.*

SOLUCIÓN

a) La recta de regresión es la siguiente:

El coeficiente es la pendiente de la recta de regresión y representa el aumento del tiempo de ejecución del algoritmo cuando el grafo aumenta en un nodo.

El coeficiente es la ordenada en su origen y no tendría interpretación real puesto que vendría a ser el tiempo de ejecución correspondiente a un grafo vacío.

b) El contraste sobre la pendiente de la recta es el siguiente:

donde es la pendiente de la recta de regresión. El estadístico de contraste vale: 4.82113 con un p-valor de 0.00001. Como el p-valor es más pequeño que el nivel de significación, podemos rechazar la hipótesis nula y afirmar que tenemos suficientes indicios para aceptar que la pendiente de la recta de regresión no es nula.

c) El valor del coeficiente de determinación vale 0.95891. Es un valor muy cercano a 1. Por lo tanto, podemos decir que la regresión es muy buena.

# Examen 2 (17-06-17)

**Problema 1**. Los siguientes datos corresponden a la cantidad de asignaturas superadas por 12 estudiantes de la UOC:

|  |  |
| --- | --- |
| *xi* | *ni* |
| 3 | 2 |
| 10 | 4 |
| 15 | 6 |

a) Calculad la media y la mediana.

b) Haced el diagrama de barras.

c) Calculad la desviación típica muestral.

*Criterios de puntuación y valoración (sobre 10): a) 4 puntos; b) 3 puntos; c) 3 puntos.*

SOLUCIÓN:

a)



Mediana = 12.5 (número par de datos, 12), ocupa el lugar 6 y 7 (10+15)/2 =12.5

b)

c) La desviación típica muestral valdrá:



**Problema 2**. Cuatro industrias (A, B, C, y D) producen microprocesadores para la fabricación de un determinado tipo de ordenador para coche. Producen respectivamente: A el 40%, B el 30%, C el 20% y D el 10% de la producción total. De los que produce A, el 5% son defectuosos, así como el 4% de los de B, el 2% de los de C y el 1% de los de D.

a) Representad el árbol de probabilidades.

b) Si seleccionamos al azar un vehículo dotado de ordenador a bordo, ¿cuál es la probabilidad de que su ordenador traiga el microprocesador defectuoso?

c) ¿Los sucesos “microprocesador fabricado por A” y “microprocesador defectuoso” son independientes? ¿Por qué?

*Criterios de puntuación y valoración (sobre 10): a) 4 puntos; b) 3 puntos y c) 3 puntos.*

SOLUCIÓN:

a) Consideraremos los siguientes sucesos y las probabilidades que nos proporciona el enunciado:

A = “Microprocesador fabricado por A”



B = “Microprocesador fabricado por B”



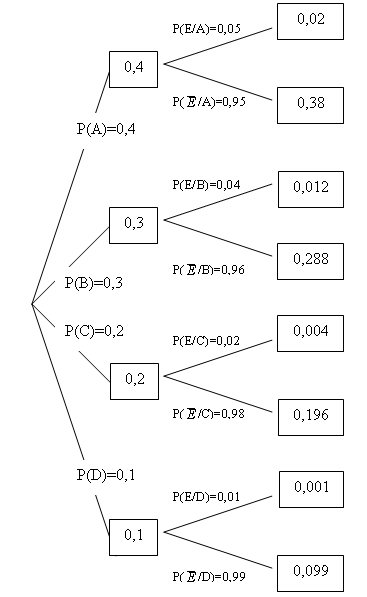
C = “Microprocesador fabricado por C”



D = “Microprocesador fabricado por D” 



E = “Microprocesador defectuoso”



b) La probabilidad que un ordenador traiga el microprocesador defectuoso es:



c) Los sucesos A y E no son independientes puesto que:



**Problema 3**. Suponemos que el examen de una cierta asignatura lo han suspendido el 15% de los 200 estudiantes matriculados. Escogemos una muestra aleatoria de 5 exámenes y nos preguntamos

a) ¿Cuál es la probabilidad de que entre ellos haya exactamente 4 aprobados? Indicad cuál es la variable aleatoria, su tipo y los correspondientes parámetros.

b) ¿Cuál es la probabilidad de que entre ellos haya más de un aprobado?

c) ¿Cuál es la varianza del número de exámenes aprobados cuando escogemos 5 exámenes al azar?

*Criterios de puntuación y valoración (sobre 10): Se tienen que indicar las fórmulas y los cálculos realizados, así como los razonamientos. a) 5 puntos (2 la distribución de la variable aleatoria y 3 la probabilidad); c) 3 puntos; d) 2 puntos.*

SOLUCIÓN

a) X=”número de exámenes aprobados si escogemos aleatoriamente 5 exámenes”. Binomial de parámetros n=5, p=0.85. P(X=4)=5\*0.854\*0.15=0.3915.

b) P(X>1)=1-P(X=0)-P(X=1)=1-0,155-5\*0.154\*0.85=0.9978

c) Var(X)=np(1-p)=0.6375

**Problema 4**. Un asistente de voz para móviles ha procesado 200 peticiones y ha procesado correctamente 180 de éstas.

a) Calculad un intervalo de confianza para la proporción de veces que el asistente procesa correctamente una petición, con un nivel de confianza del 99%.

b) ¿Cuál tendría que ser el tamaño de la muestra porque el margen de error fuera inferior a 0.01?

Valores de probabilidades que os pueden ser útiles; si no encontráis exactamente el que necesitáis, utilizad el más cercano

|  |  |  |  |
| --- | --- | --- | --- |
| p(X>= x) | X~ N(0,1) | X~ t de Student con 199 grados de libertad | X~ t de Student con 200 grados de libertad |
| 0.01 | 2.326 | 2.345 | 2.345 |
| 0.025 | 1.96 | 1.971 | 1.971 |
| 0.05 | 1.645 | 1.652 | 1.652 |
| 0.005 | 2.575 | 2.6 | 2.6 |

*Criterios de puntuación y valoración (sobre 10): Se tienen que indicar las fórmulas y los cálculos realizados, así como los razonamientos. a) 6 puntos: planteamiento 2 punto, cálculos 4 puntos; b) 4 puntos: planteamiento 2 puntos, cálculos 2 puntos.*

SOLUCIÓN

a) Es un intervalo de la proporción con por lo tanto tenemos.

b) Tenemos que , es decir, 5973

**Problema 5.** Queremos contrastar si el número de peticiones a un servidor es el mismo por la tarde y por la mañana en un día. Para hacer el estudio, escogimos aleatoriamente días e hicimos el contraste correspondiente con R. Los resultados fueron los siguientes:

##   
## Paired t-test  
##   
## data: servidor.mañana and servidor.tarde  
## t = -5.048, df = 89, p-value = 0.000002349  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -246.6335 -107.3136  
## sample estimates:  
## mean of the differences   
## -176.9736

a) Estableced el contraste a realizar. Tenéis que indicar si se trata de un contraste de una muestra o de dos muestras, de qué parámetro tenemos que hacer el contraste (media, proporción, etc.), y la hipótesis nula y la hipótesis alternativa.

b) ¿Cuál es el valor del estadístico de contraste?

c) ¿Cuál es el p-valor del contraste?

d) A un nivel de significación de α=0.05, ¿a qué conclusión llegaríais? No es suficiente afirmar si aceptamos o rechazamos la hipótesis nula, se tiene que razonar la conclusión en el contexto del problema.

*Criterios de puntuación y valoración (sobre 10): a) 4 puntos (una muestra o dos, 2 puntos; parámetro, 1 punto; contraste, 1 punto). b) 1 punto. c) 1 punto. d) 4 puntos (aceptar o no , 1 punto; conclusión en el contexto, 3 puntos).*

SOLUCIÓN

a) Se trata de un contraste de medias de dos muestras aparejadas. El contraste a realizar es el siguiente:

donde y representan las medias de las variables aleatorias que nos dan el número de peticiones al servidor por la mañana y por la tarde.

b) El valor del estadístico de contraste es -5.0479892.

c) El p-valor del contraste es 0.0000023.

d) Como el p-valor es menor que , rechazaríamos la hipótesis nula y concluiríamos que efectivamente tenemos suficientes evidencias para afirmar el número de peticiones por la mañana y por la tarde es diferente.

**Problema 6.** El número de procesadores y la memoria RAM de 5 ordenadores de un aula de informática es el siguiente:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RAM (Gb)** | 1 | 2 | 2 | 3 | 2 |
| **Número procesadores** | 2 | 4 | 4 | 8 | 3 |

a) Calculad la recta de regresión del número de procesadores en función de la RAM de cada ordenador.

b) Estimad el número de procesadores para un ordenador con Gb y medio de memoria RAM.

*Criterios de puntuación y valoración (sobre 10): a) (8 puntos). Cálculo de las medias, 2 puntos. Cálculo de las varianzas y covarianza, 2 puntos. Cálculo de la pendiente, 2 puntos. Cálculo de la ordenada en el origen, 2 puntos. b) 2 puntos.*

SOLUCIÓN

a) Primero calculamos las medias de las dos variables:

Las varianzas y la covarianza serán:

La recta de regresión será: , donde

La recta de regresión será, pues:

b) El número de procesadores estimado para un ordenador con Gb de RAM será:

# Examen 3 (21-06-17)

**Problema 1.** Los siguientes datos representan el tiempo en segundos que tardó un programa al ejecutarse en WINDOWS durante 12 veces:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2.5 | 7.1 | 5 | 8.5 | 7 | 8.1 | 2.3 | 7.1 | 4 | 8 | 6.6 | 5 |

1. Encontrad la media y los tres cuartiles.
2. Haced un diagrama de caja con estas observaciones.
3. Calculad la desviación típica poblacional.

*Criterios de puntuación y valoración (sobre 10): a) 4 puntos; b) 3 puntos y c) 3 puntos.*

SOLUCIÓN:

a) La media vale:



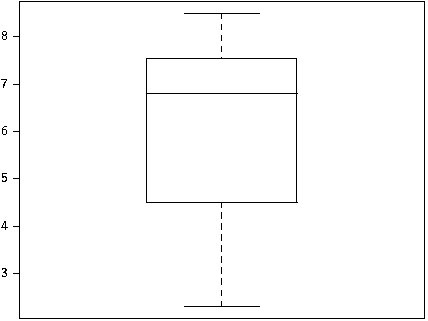
Para encontrar los cuartiles, primero ordenamos los datos: 2.3; 2.5; 4; 5; 5; 6.6; 7; 7.1; 7.1; 8; 8.1; 8.5.

El primero cuartil será la mediana de los 6 primeros datos: será la semisuma de los valores que ocupan los lugares 3 y 4: Q1=(4+5)/2=4.5.

La mediana será la semisuma de los valores que ocupan los lugares 6 y 7: Q2=(6.6+7)/2=6.8.

El tercer cuartil será la mediana de las 6 últimas datos: será la semisuma de los valores que ocupan los lugares 9 y 10: Q3=(7.1+8)/2=7.55.

b)



c) La desviación típica poblacional valdrá:



**Problema 2**. Hemos medido la velocidad de bajada (en KB/s) de 92 archivos alojados en un servidor en un determinado momento del día. Los resultados los hemos tabulado según la franja horaria (mañana, noche, tarde) y según la velocidad (alta, baja, media, muy alta):

V\_alta V\_baja V\_mediana V\_muy alta

Mañana 4 3 9 6

Noche 5 8 3 4

Tarda 13 13 11 13

a) Calculad la probabilidad de que la velocidad de bajada sea muy alta.

b) Calculad la probabilidad de que la bajada de un archivo sea con velocidad muy alta y se realice por la noche.

c) Si sabemos que el archivo se ha bajado por la tarde, calculad la probabilidad de que se haya bajado con velocidad alta.

d) ¿Los sucesos “franja horaria Noche” y “velocidad alta” son independientes? Razonad la respuesta.

*Criterios de puntuación y valoración (sobre 10): 2.5 puntos cada apartado.*

SOLUCIÓN:

a) Calculad la probabilidad que la velocidad de bajada sea muy alta.



b) Calculad la probabilidad que la bajada de un archivo sea muy alta y se realice por la noche.



c) Si sabemos que el archivo se ha bajado por la tarde, calculad la probabilidad que estos sean de velocidad alta.



d) ¿Los sucesos “franja horaria Noche” y “velocidad alta” son independientes? Razonar la respuesta.

Calculamos las probabilidades siguientes:



, ,



Como se verifica que

, no son independientes.



**Problema 3**. Una variable aleatoria tiene la siguiente función de masa de probabilidad:

|  |  |
| --- | --- |
| Valores possibles *X=x* | *P(X=x)* |
| 3 | 0.4 |
| 4 | A |
| 8 | B |

Sabiendo que su esperanza es 4,

a) Calculad A y B.

b) Calculad la función de distribución de X y la probabilidad de que X sea inferior o igual a 6.

c) Calculad la varianza de X.

*Criterios de puntuación y valoración (sobre 10): Se tienen que indicar las fórmulas y los cálculos realizados, así como los razonamientos. a) 4 puntos; b) 3 puntos (2 puntos la función de distribución, 1 punto la probabilidad); c) 3 puntos.*

SOLUCIÓN:

a) 0.4+A+B=1 y además 3\*0.4+4\*A+8\*B= 4, por lo tanto A=0.5, B=0.1

b) F(3)=0.4, F(4)=0.9, F(8)=1. P(X<=6)=0.4+A=0.4+0.5=0.9

c) Var= (3-4)2\*0.4+(4-4)2\*0.5+(8-4)2\*0.1=2

**Problema** 4. Los ingresos mensuales de una familia siguen una distribución normal de media 1000 euros y desviación típica 80 euros. Suponemos además que el ingreso de cada mes es independiente de los otros meses.

a) ¿Qué distribución siguen los ingresos anuales?

b) Calculad la probabilidad de que esta familia ingrese más de 12300 euros en un año

c) ¿Cuánto tendrían que gastar en un año para tener una probabilidad del 95% de no entrar en números rojos (es decir de no gastar más de lo que se ingresa)?

Valores de probabilidades que os pueden ser útiles; si no encontráis exactamente el que necesitáis, utilizad el más cercano.

|  |  |
| --- | --- |
| p(X>= x) | X~ N(0,1) |
| 0.0228 | 2 |
| 0.0359 | 1.8 |
| 0.05 | 1.645 |
| 0.0548 | 1.6 |
| 0.0876 | 1.4 |
| 0.1 | 1.2816 |
| 0.1151 | 1.2 |
| 0.1357 | 1.1 |
| 0.1587 | 1 |

*Criterios de puntuación y valoración (sobre 10): Se tienen que indicar las fórmulas y los cálculos realizados, así como los razonamientos. a) 4 puntos: planteamiento 2 puntos, cálculos 2 puntos; b) 3 puntos (planteamiento 1 punto, cálculos 2 puntos); c) 3 puntos (planteamiento 1 punto, cálculos 2 puntos).*

SOLUCIÓN:

a) Sigue una distribución normal de media 12\*1000=12000 y desviación típica.

b) Piden p(X>12300)= p(Z>(12300-12000)/277.128129211)= p(Z>1,082531755)=0.1395. Con las tablas del enunciado 0.1357.

c) Ahora piden p(X>g)=0.95, o el que es el mismo, p(Z>(g-12000)/ 277.128129211)=0.95 con lo que tenemos que (g-12000)/ 277.128129211=-1.645 y por lo tanto g=11544.1.

**Problema 5.** Queremos comprobar si el porcentaje de mensajes SPAM recibos por día en dos servidores de correo es el mismo o no. Escogemos una muestra aleatoria de 100 días, calculamos el porcentaje de mensajes SPAM recibos en los dos servidores. Nos dicen que al primer servidor han llegado mensajes SPAM y al segundo, mensajes SPAM.

a) Estableced el contraste a realizar. Tenéis que indicar si se trata de un contraste de una muestra o de dos muestras, de qué parámetro(s) tenemos que hacer el contraste (mediana(s), proporción(nes), etc.), y la hipótesis nula y la hipótesis alternativa.

b) ¿Cuál es el valor del estadístico de contraste?

c) A un nivel de significación del , ¿Cuál(es) es(son) lo(s) valor(es) crítico(s)?

d) A un nivel de significación de α=0.05, ¿a qué conclusión llegaríais? No es suficiente afirmar si aceptamos o rechazamos la hipótesis nula, se tiene que razonar la conclusión en el contexto del problema.

Valores de las tablas (utilizad los que consideráis necesarios, si necesitáis un valor que no es exacto, utilizad el valor más cercano)

|  |  |
| --- | --- |
| x | p(X >= x) |
| 2.06 | 0.0197 |
| 1.96 | 0.025 |
| 1.86 | 0.0314 |
| 1.76 | 0.0392 |
| 1.66 | 0.0485 |
| 1.56 | 0.0594 |
| 1.46 | 0.0721 |
| 1.36 | 0.0869 |
| 1.26 | 0.1038 |
| 1.16 | 0.123 |
| 1.06 | 0.1446 |
| 0.06 | 0.4761 |

*Criterios de puntuación y valoración (sobre 10): a) 3 puntos, si indica si se trata de una muestra o dos, 1 punto, si indica el parámetro, 1 punto, si indica el contraste, 1 punto, b) 2 puntos. Si indica la fórmula del estadístico, 1 punto. Valor del mismo, 1 punto. c) 1 punto d) 4 puntos (aceptar o no , 1 punto; conclusión en el contexto, 3 puntos).*

SOLUCIÓN:

a) Se trata de un contraste de proporciones de dos muestras independientes. El contraste a realizar es el siguiente:

donde i representan las proporciones de mensajes SPAM recibos en un día por el servidor 1 y por el servidor 2 respectivamente.

b) El valor del estadístico de contraste es

c) Los valores críticos del contraste son:

d) Como el valor del estadístico de contraste se encuentra entre los valores críticos, aceptaríamos la hipótesis nula y concluiríamos que NO tenemos suficientes evidencias para afirmar que la proporción de mensajes SPAM en los dos servidores NO sea el mismo.

**Problema 6**. Queremos estudiar si el porcentaje de mensajes SPAM recibos en un día en un determinado servidor de correo tiene relación con el día del mes en qué los mensajes son enviados. Por eso hemos estudiado durante dos meses el porcentaje de mensajes SPAM recibos durante los 60 días y los resultados han sido los siguientes:

##   
## Call:  
## lm(formula = porcentaje.spam ~ días)  
##   
## Residuales:  
## Min 1Q Median 3Q Max   
## -0.134565 -0.076360 -0.006697 0.068212 0.159768   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.084031 0.023317 3.604 0.000652 \*\*\*  
## días 0.020422 0.001313 15.548 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08806 donde 58 degrees of freedom  
## Multiple R-squared: 0.8065, Adjusted R-squared: 0.8032   
## F-statistic: 241.7 donde 1 and 58 DF, p-value: < 0.00000000000000022

a) Encontrad la recta de regresión del porcentaje de mensajes SPAM por día en función del día del mes e interpretad los resultados.

b) Haced el contraste de hipótesis sobre la pendiente de la recta, con un nivel de significación de α=0.01 e interpretad los resultados. Tenéis que dar el contraste sobre la pendiente indicando la hipótesis nula y alternativa, el estadístico de contraste, el p-valor y la conclusión a la que se llega.

c) Calculad el coeficiente de determinación. Discutid si es un buen modelo.

*Criterios de puntuación y valoración (sobre 10): a) (3 puntos). Dar la recta de regresión, 1.5 puntos. Interpretación de los parámetros de la recta, 1.5 puntos. b) (5 puntos). Dar el contraste de hipótesis indicando la hipótesis nula y alternativa, 1 punto. Dar el valor del estadístico, 1 punto. Dar el p-valor, 1 punto. Conclusión, 2 puntos. c) (2 puntos). Dar la expresión del coeficiente de determinación, 1 punto. Interpretación del mismo, 1 punto.*

SOLUCIÓN:

a) La recta de regresión es la siguiente:

El coeficiente es la pendiente de la recta de regresión y representa el aumento del porcentaje de mensajes SPAM en un día cuando el día del mes aumenta en un día.

El coeficiente es la ordenada en el origen y no tendría interpretación real puesto que vendría a ser el porcentaje de mensajes SPAM el día 0.

b) El contraste sobre la pendiente de la recta es el siguiente:

donde es la pendiente de la recta de regresión. El estadístico de contraste vale: 3.60378 con un p-valor de 0.00065. Como el p-valor es más pequeño que el nivel de significación, podemos rechazar la hipótesis nula y afirmar que tenemos suficientes indicios para aceptar que la pendiente de la recta de regresión no es nula.

c) El valor del coeficiente de determinación vale 0.8065. Es un valor cercano a 1. Por lo tanto, podemos decir que la regresión es moderadamente buena.