UOC
ESTUDIOS De INFORMÁTICA
Estadística Primavera 2014

*PEC – 2 GRADO*

# Fecha de propuesta: 23-3-2014. Fecha de entrega: 20-4-2014.

*Instrucciones*

- **El nombre del fichero que nos adjuntáis con la respuesta tiene que seguir la regla siguiente:** 3 primeras letras primer apellido + 3 primeras letras segundo apellido + "2". Por ejemplo, un estudiante que se diga Jorge Gratacos Pellicer tendría que enviar un fichero con el nombre: **GRAPEL2.doc (o la extensión que corresponda).**
- Sed breves. No es necesario escribir mucho.
- **No déis sólo el resultado de los ejercicios,** hace falta explicar el motivo de los razonamientos que uséis.
- Poned vuestro nombre completo dentro de la prueba.
- Tenéis que enviar la solución al **buzón de "Entrega de actividades".**
- Tiempo previsto: cinco horas para "realizar" la PEC y dos horas para pasarla al ordenador.
- Dad los resultados numéricos redondeando a dos decimales.

**Apellidos y Nombre.....**

**LA PEC CONSTA DE 4 CUESTIONES (A REALIZAR SIN R) Y DOS PROBLEMAS QUE REQUIEREN EL USO DE R..**

## CUESTIÓN 1 (12,5%)

El número de veces que un servidor se cuelga en un día sigue una variable aleatoria X con la siguiente función de probabilidad:

| Número de veces (ni) | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Probabilidad | 0,5 | 0,2 | 0,1 | 0,05 | *a* |

a) Decir si se trata de una distribución discreta o continua. Justificar la respuesta.
b) Hallar el valor de *a* para que la tabla anterior corresponda a una función de masa de probabilidad.
c) Hallar E(X) y Var(X).
d) Hallar la función de distribución de probabilidad, F(x).
e) ¿Cuál es la probabilidad de que en un día cualquiera el servidor se cuelgue 3 o más veces?

**Criterios de corrección.**

Todos los apartados valen lo mismo.

a) Si sólo decís el tipo de distribución sin justificarlo, 1.25% de la puntuación. Si lo justificáis, 2.5%.
b) Si imponéis bien la condición que tiene que verificar a pero os equivocáis en su cálculo, 1.25% de la puntuación. Si encontráis correctamente el valor de a, 2.5%.
c) Si calculáis correctamente E(X), 1.25% de la puntuación. Si os habéis equivocado en el cálculo de E(X) pero sois coherentes en el cálculo de Var(X), 1.25% de la puntuación. Si todo está bien, 2.5% de la puntuación.
d) Si os habéis equivocado en el cálculo de a de la apartado b) pero es coherente en el cálculo de la función de distribución, 1.25% de la puntuación. Restaremos 0.5% por cada valor incorrecto de los valores ni.

e) Si hay errores en los otros apartados (cálculo de a o de la función de distribución) pero sois coherentes en el cálculo de la probabilidad pedida, 2.5% de la puntuación total.

## CUESTIÓN 2 (12,5%)

La función de densidad de probabilidad del tiempo de vida (T, en horas) de un determinado dispositivo electrónico viene dada por:

$$f(t)=\begin{cases} a\left(t-\dfrac{1}{500}\cdot t^2\right), & 0\le t\le 500 \\ 0, & \text{altres valors de t} \end{cases}$$

a)Hallar el valor de *a*.
b)Calcular la probabilidad que el dispositivo dure más de 100 horas.
c)¿Cuál es la duración mediana del dispositivo (en horas)?

**Criterios de corrección.**

Los apartados a) y b) valen el 4.2% de la puntuación y el c), 4.1% de la puntuación.

## CUESTIÓN 3 (12,5%)

En un determinado servidor de correo, llegan de media 10 correos "SPAM" por hora.
a)Sea X la variable aleatoria "número de correos "SPAM" que llegan en 5 horas". ¿Qué tipo de distribución de probabilidad sigue la variable X? Dar su función de masa de probabilidad.
b)¿Cuál es la probabilidad que lleguen 8 o menos correos en una hora?
c)¿Cuál es la probabilidad que lleguen 250 o más correos "SPAM" en un día? Usar el Teorema Central del Límite y justificar su uso.

**Criterios de corrección.**

Los apartados a) y c) valen el 4.2% de la puntuación y el b), 4.1% de la puntuación.

## CUESTIÓN 4 (12,5%)

En cada uno de los apartados siguientes se tiene que indicar cuál es la variable aleatoria considerada, cuál es su distribución, así como todos los cálculos.

a)La probabilidad de que un ordenador que recibe un determinado mensaje "SPAM" quede infectado por un virus que trae el mensaje es 0.1. Suponemos que 10 ordenadores han recibido este mensaje, ¿cuál es la probabilidad de que más de la mitad queden infectados (la mitad incluida)?
b)La probabilidad de que un ordenador que recibe un determinado mensaje "SPAM" quede infectado por un virus que trae el mensaje es 0.1. ¿Cuál es la probabilidad que si escogemos los ordenadores un detrás el otro, los cuatro primeros no queden infectados?
c)Suponemos que el tiempo que pasa hasta que un ordenador queda infectado sigue una distribución exponencial de parámetro Hallar la probabilidad que pase

$$\lambda=\frac{1}{12}\ \text{minuts}^{-1}.$$

más de un cuarto de hora para que el ordenador quede infectado.

**Criterios de corrección.**

El apartado a) vale el 4.1% de la puntuación. Los apartados b) y c) valen el 4.2% de la puntuación.

a) Si decís cuál es el tipo de distribución de la variable, 0.5%. Si dáis los parámetros de la distribución de la variable, 1%. Si escribís la probabilidad pedida, 1%. Si calculáis la probabilidad pedida, 1.6%.
b) Si decís cuál es el tipo de distribución de la variable, 0.5%. Si dáis los parámetros de la distribución de la variable, 1%. Si escribís la probabilidad pedida, 1%. Si calculáis la probabilidad pedida, 1.7%.
c) Si escribís la probabilidad pedida, 1%. Si calculáis la probabilidad pedida, 3.2%.

## PROBLEMA 1 (25%) CON R

I. El 15% de los ordenadores llevan instalado el sistema operativo Linix. Consideramos una muestra de 100 ordenadores y sea X la variable aleatoria que nos dice el número de ordenadores con el sistema operativo Linix.

a) ¿Qué distribución sigue la variable X?
b) Calcular usando el R (R-Commander-> Distribuciones), la probabilidad que haya una cuarta parte o más de ordenadores con el sistema operativo Linix.
c) Calcular esta misma probabilidad de forma aproximada mediante la simulación. Generar una muestra de tamaño 500 de la variable X. Para hacerlo, ir a R-Commander -> Distribuciones -> Escogéis si discreta o continúa -> Escogéis la distribución -> Muestra de una distribución ...)
d) Usando la simulación anterior, calcular la probabilidad pedida en la apartado b). Para hacerlo, tenéis que usar la instrucción *table* de R. Para usar esta instrucción tenéis que escribir *table(condición)* y os dará los valores que cumplen la condición (*TRUE*) y los que no la cumplen (*FALSE*)
e) Calcular la media y la varianza de la muestra generada en el apartado c) y compararlas con los valores teóricos de la distribución X.

II. La vida media T de un ordenador con unas determinadas características sigue la distribución exponencial con una media de 5 años. Calcular usando R:

a) La probabilidad que un ordenador con estas características dure más de 3 años y medio.
b) Generar una muestra de 100 vidas medias de ordenadores con estas características.
c) Calcular la probabilidad pedida en el apartado a) usando esta muestra.
d) Calcular la media y la varianza de la muestra generada en la apartado b) y comparadlas con los valores teóricos de la distribución.

*Indicaciones: se tiene que entregar la salida del R con los razonamientos correspondientes.*

**Criterios de corrección.**

I. La parte I vale 12.5%. Los subapartados de la parte I valen 2.5% cada uno.
a) Si decís cuál es la distribución de X, 1%. Si dáis sus parámetros, 1.5%
b) Si escribís la probabilidad pedida, 1%. Si calculáis la probabilidad con R, 1.5%.
c) Si generáis la muestra con R, 2.5%.
d) Si calculáis aproximadamente la probabilidad pedida, 2%. Si decís si es o no una buena aproximación, 0.5%.
e) Si calculáis la media de la muestra, 0.5%. Si calculáis la varianza de la muestra, 0.5%. Si calculáis la media y la varianza de la variable X, 0.5% cada uno. Si decís si son o no buenas aproximaciones, 0.5%.

II. La parte II val 12.5%. Los subapardos de la parte II valen a) 3%, b) 3%, c) 3% y d) 3.5%

a) Si calculáis la probabilidad pedida con R, 3%.
b) Si generáis la muestra con R, 3%
c) Si calculáis aproximadamente la probabilidad pedida, 2.5%. Si decís si es o no una buena aproximación, 0.5%
d) Si calculáis la media de la muestra, 0.7%. Si calculáis la varianza, 0.7%. Si calculáis la media y la varianza de la variable T, 0.7% cada uno. Si decís si son o no buenas aproximaciones, 0.7%

## PROBLEMA 2 (25%) CON R

Vamos a comprobar el Teorema Central del Límite con la tabla de datos **iris** de R. Esta tabla de datos contiene las medidas de la longitud del pétalo y del sépalo de 150 flores de 3 especies: setosa, versicolor y virginica. Hacer ?iris para ver la descripción de la tabla de datos.

a) Generar 50 muestras con reposición de tamaño 100 de la variable "longitud del sépalo" (Sepal.Length) de la tabla de datos iris. Para hacerlo, tenéis que usar las instrucciones *sample* y **replicate** de R. Guardar los resultado en la variable **muestras**.
   Para usar replicate, tenéis que escribir **replicate(n, sample(...))** donde **n** representa las muestras que queréis hacer y **sample(...)** es para generar las muestras y se explica a continuación.
   Para usar **sample**, tenéis que escribir **sample(x,tamaño, replace=)** donde **x** es la variable de la que queréis extraer la muestra, **tamaño** es el tamaño de la muestra y **replace=TRUE** significa que la muestra es con reposición y **replace=FALSE** significa que la muestra es sin reposición.
   Una vez hayáis usado **replicate** y guardado el resultado en **muestras**, R os generará una matriz de 100 filas y 50 columnas (una por cada muestra). No hace falta que mostréis el resultado de **muestras**, basta dar la instrucción para generarlas.

b) Calcular las medias de las 50 muestras y guardáis los resultados en un vector llamado **medias**. Para hacerlo, tenéis que usar la instrucción de R **apply: apply(matriz,y,función),** donde matriz es la matriz de datos en la que se trabaja, **y=1** significa que aplicaremos la función por filas y **y=2** significa que la aplicaremos por columnas y **función** es la función que queremos aplicar a nuestra matriz de datos. La función de R que tenéis que usar es **mean**. Aquí sí que tenéis que mostrar, a parte de las instrucciones de R para generar las medias, el valor del vector **medias**.

c) Usando las instrucciones siguientes de R hacer el histograma de la curva normal:

```
h<-hist(medias,freq=F, col="red", xlab="Medias mostrales",main="Histograma con curva normal")
xfit<-seq(min(medias),max(medias),length=40)
yfit<-dnorm(xfit,mean=mean(medias),sd=sd(medias))
lines(xfit, yfit, col="blue", lwd=2)
```

d) Relacionar los cálculos anteriores con el Teorema Central del Límite. Concretamente, tenéis que comentar el gráfico que os sale en la apartado c), decir cuáles son la media y la desviación típica del vector **medias** y qué serien la media teórica y la desviación típica del vector **medias** que prevé el Teorema Central del Límite.

*Indicaciones: se tiene que entregar la salida del R con los razonamientos correspondientes.*

**Criterios de corrección.**

Todos los apartados valen lo mismo.

a) Si generáis la matriz **muestras**, 6.25%.
b) Si lo calculáis el vector **medias**, 6.25%.
c) Si hacéis el gráfico correctamente, 6.25%.
d) Si relacionáis el gráfico de la apartado c) con el Teorema Central del Límite, 1%. Si calculáis la media y la desviación típica del vector **medias**, 2%. Si decís a qué distribución se aproxima la longitud del sépalo según el Teorema Central del Límite, 1%. Si dáis los parámetros correctos de esta distribución, 2%. Si decís si las medias y las desviaciones típicas son parecidas o no, 0.25%.