Universidade Vigo
Departamento de Estatística e Investigación Operativa
E.Superior de Enxeñería Informática
Edificio Politécnico
Campus de Ourense
E-32004 Ourense
Tel. 986 387 000
http://esei.uvigo.es

Estatística Descritiva

Apelidos: **Nome:** **DNI:**

1. (*5 puntos*) Para el análisis de rendimiento de CPU's, se ejecutan en 3 ordenadores, equivalentes salvo la CPU, diferentes programas. La información contenida en el archivo *cpu.rendemento.txt*[1], se corresponde con el tiempo de ejecución (en segundos) de cada programa (variable *tempo*) y el correpondiente valor indicando a que CPU se corresponde (variable *cpu*). Para este conjunto de datos:

   Solución:
   ```
   datos <- read.table(file="rendemento.cpu.txt", header=TRUE, sep=";", dec=".")
   str(datos)
   head(datos)
   ```
   □

   a) Clasificar las variables estadísticas.

   Solución: $tempo: variable continua, valores positivos $cpu: variable cualitativa, valores 1,2,3 □

   b) Dar la distribución completa de frecuencias agrupando la distribución en 4 intervalos de longitud 1.5 segundos.

   Solución:
   ```
   tempo.agrup <- cut(datos$tempo, breaks=c(0,1.5,3,4.5,6))
   ni <- table(tempo.agrup)
   N <- sum(ni)
   fi <- ni/N
   Ni <- cumsum(ni)
   Fi <- Ni/N
   m.c <- 0.75 + 1.5*0:3
   amp <- 1.5
   cbind( m.c, ni, fi, Ni, Fi,hi=fi/amp)
   m.c  ni    fi   Ni    Fi        hi
   (0,1.5] 0.75 474 0.474  474 0.474 0.3160000
   (1.5,3] 2.25 424 0.424  898 0.898 0.2826667
   (3,4.5] 3.75  93 0.093  991 0.991 0.0620000
   (4.5,6] 5.25   9 0.009 1000 1.000 0.0060000
   ```
   □

   c) Calcular la media muestral y la mediana con la variable agrupada y sin agrupar.

   Solución:
   ```
   > mean(datos$tempo); median(datos$tempo)  # Media y mediana sin agrupar:
    [1] 1.688751
    [1] 1.594618
   > sum( ni*m.c)/N #Media agrupada:
    [1] 1.7055
    Mediana agrupada: rango mediano (1.5,3], valor elegido:
   1.5 + \frac{0.5-0.474}{0.424} * 1.5= 1.591981
   ```
   □

   d) Dar el histograma para la variable sin agrupar con: la densidad de frecuencia y los intervalos de longitud 1.5. A partir del gráfico anterior obtén la frecuencia relativa del intervalo [2,4.5].

   Solución:
   ```
   hist(datos$tempo, freq=FALSE, breaks=c(0,1.5,3,4.5,6))
   ```
   freq relativa de (2,3]: $fi[2] * 1/1.5$
   freq relativa de (3,4.5]: $fi[3]$
   freq relativa total: $fi[2] * 1/1.5 + fi[3] = 0.3756667$ □

[1] Descargar desde la url *https://dl.dropboxusercontent.com/u/29008031/cpu.rendemento.txt*

e) Dar un resumen numérico completo de los tiempos de computación para la CPU 2. Comenta los resultados obtenidos

Solución: Seleccionamos las filas y realizamos el resumen numérico:

```
> tt <- datos$tempo[ datos$cpu==2]
> summary(tt); sd(tt)
Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.001748 0.698100 1.337000 1.420000 2.045000 4.924000
[1] 0.8883224
```

...

f) Compara con un diagrama de cajas los tiempos en función de tipo de CPU. Extrae conclusiones

Solución: `boxplot( datos$tempo~ datos$cpu)`

...

g) Comparar el histograma de los tiempos de computación con la curva de referencia del coef de apuntamiento.
Para construir la curva de referencia ejecutar despues de construir el histograma: *curve(dnorm, from=-3, to=3,col="blue", lwd=2,lty=3,add=TRUE)*
Justificar en función de los gráficos el signo del coef de apuntamiento.

Solución:

```
library(fBasics)
dd <- scale(datos$tempo)
hist(dd, freq=FALSE)
curve( dnorm, from=-3, to=3,col="blue", lwd=2,lty=3,add=TRUE)
# El apuntamiento es menor que la curva normal por lo que el signo de la kurtosis es negativo
```

Universidade Vigo

Departamento de Estatística e Investigación Operativa

E.Superior de Enxeñería Informática
Edificio Politécnico
Campus de Ourense
E-32004 Ourense

Tel. 986 387 000
http://esei.uvigo.es

Estatística - Cálculo de Probabilidades

Apelidos: **Nome:** **DNI:**

1. (*1.5 puntos*) En un ordenador se ejecutan simultaneamente dos procesos. Sean A y B los conjuntos representando el primer/segundo proceso se termina antes de 60 segundos. Si $P(A \cup B) = 0.9$ y $P(A \cap B) = 0.5$, ¿cuál es la probabilidad de que exactamente un proceso termine antes de 60 segundos?

   Solución: La probabilidad que se pide es de la unión de los conjuntos $A \cap \bar{B}$ y $\bar{A} \cap B$, y

   $$P(A \cup B) = P(A \cap \bar{B}) + P(\bar{A} \cap B) + P(A \cap B)$$

   por lo tanto,

   $$P(A \cap \bar{B}) + P(\bar{A} \cap B) = 0.9 - 0.5 = 0.4$$

   □

2. (*1.5 puntos*) Describir el espacio probabilístico de la variable I representando el número de veces que se ejecuta **S** en el experimento correspondiente al loop genérico de C: **do {S} while (B);** .
   Nota: **B** es una variable lógica representando un experimento aleatorio genérico e independiente en cada realización.

   Solución: El espacio probabilístico es la terna $(\Omega, \sigma - \text{algebra}, P)$, donde $\Omega = \{1, 2, \ldots, i, \ldots\}$ donde el valor I representa el número de veces que se ejecutó **S**, por ejemplo si $I = 1$, quiere decir que el primer valor de **B** fué FALSE, si $I = 2$, quiere decir que la secuencia de valores de **B** fue *TRUE, FALSE,...*

   La $\sigma - \text{algebra} = P(\Omega)$, y la función de probabilidad es:

   a) $P(I = 1) = (1 - p)$,
   b) $P(I = 2) = p * (1 - p)$,
   c) $\ldots$,
   d) $P(I = i) = p^{i-1}(1 - p), \ldots$.

   con p la probabilidad de que $\mathbf{B} = TRUE$.

   □

3. (*2 puntos*) Un virus afecta a 1 de cada 500 ordenadores con un determinado antivirus. Se sabe que la probabilidad de que un usuario detecte un virus en su ordenador cuando está infectado es del 90 % y que el usuario cree que tiene un virus cuando no está infectado es del 0.01. Calcular la probabilidad de que un ordenador esté realmente infectado si el usuario cree que está infectado. ¿Por qué crees que la probabilidad resultante es tan baja?

   Solución: Sea *I=ordenador infectado* y *DT=virus detectado*, la probabilidad pedida es:

   $$P(I/DT) = \frac{P(DT/I)\,P(I)}{P(DT/I)\,P(I) + P(DT/\bar{I})\,P(\bar{I})} = \frac{0.9 * \frac{1}{500}}{0.9 * \frac{1}{500} + 0.01 * \frac{499}{500}} = \frac{0.9}{0.9 + 0.01 * 499} = 0.1528014$$

   La probabilidad es baja porque (aprox.) hay un elevado número de casos de falsas alarmas (de 4990 ordenadores se esperan 49), frente a alarmas positivas (de 5000 ordenadores casos se esperan infectar 10 y se esperan detectar 9).

   □