Apellidos: **Computadores** ☐

Nombre: **Software** ☐

Estadística. 2º examen parcial. 14-11-2013

Test (20 % de la nota del examen)

- Tiempo para esta parte del examen: 1 hora y 10 minutos.
- El test y la teoría se recogerán a los 45 minutos.
- En cada pregunta de test, una y sólo una de las respuestas (a), (b) y (c) es cierta. Poner la letra elegida o dejar en blanco.
- Calificación: acierto $= +1$, fallo $= -1/2$ y blanco $= 0$.

El número medio de veces que hay que lanzar un dado antes de obtener un resultado mayor que 4 es:

(a) 2*
(b) 5
(c) 7

☐

Solución
Se considera la variable aleatoria:

$$X: \text{“nº de lanzamientos de un dado antes de que el resultado sea mayor que 4”} \sim G\left(2/6\right).$$

Se trata de calcular la media de X, que vale:

$$\mu\left(X\right) = \frac{1}{p} - 1 = \frac{1}{1/3} - 1 = 2.$$

———— • ————

Sea $X \sim U\left(a, 2\right)$. Si $p\left(X < 1\right) = 1/3$, entonces a vale:

(a) 1/2*
(b) 1
(c) 3/2

☐

Solución
La probabilidad dada vale:

$$p\left(X < 1\right) = \frac{1-a}{2-a} = \frac{1}{3} \Rightarrow 3 - 3\,a = 2 - a \Rightarrow 1 = 2\,a \Rightarrow a = \frac{1}{2}.$$

———— • ————

La duración de un transistor, en años, es una variable aleatoria, X, con distribución exponencial. Si un transistor lleva 2 años funcionando, la probabilidad de que dure, como mucho, 4 años más vale:

(a) $p\left(X \leq 4 \mid X \geq 2\right)$
(b) $p\left(X \leq 6 \mid X \leq 2\right)$
(c) $p\left(X \leq 4\right)$*

☐

Solución
Se trata de calcular:

$$p\left(X \leq 2+4 \mid X \geq 2\right) = p\left(X \leq 4\right),$$

ya que la distribución exponencial carece de memoria.

Si $X \sim N\left(3,2\right)$ y $p\left(X > a\right) = 0.8997$, entonces el valor aproximado de a es:

(a) 0.44*
(b) 1.28
(c) 5.56

Solución
Se verifica:

$$p\left(X > a\right) = p\left(\frac{X-\mu}{\sigma} > \frac{a-3}{2}\right) = p\left(N\left(0,1\right) > \frac{a-3}{2}\right) = 0.8997 \Rightarrow$$
$$\Rightarrow p\left(N\left(0,1\right) \leq \frac{3-a}{2}\right) = 0.8997 \Rightarrow \frac{3-a}{2} \simeq 1.28 \Rightarrow a \simeq 3 - 2 \cdot 1.28 = 0.44.$$

Sean $X_1, \ldots, X_n \sim \text{Exp}\left(\beta\right)$ independientes. Para todo n, la variable aleatoria $X_1 + \cdots + X_n$ sigue una distribución:

(a) Exactamente $\text{Exp}\left(n\,\beta\right)$.
(b) Exactamente $\gamma\left(n,\beta\right)$*
(c) Aproximadamente $N\left(n/\beta, \sqrt{n}/\beta\right)$.

Solución
$X_i \sim \text{Exp}\left(\beta\right) = \gamma\left(1,\beta\right)$ son independientes y la distribución γ es reproductiva respecto de su primer parámetro:

$$X_1 + \cdots + X_n \sim \gamma\left(\overbrace{1+\cdots+1}^{n}, \beta\right) = \gamma\left(n,\beta\right).$$

El Teorema del Límite Central sólo es aplicable para valores grandes de n.

Teoría (10 % de la nota del examen) — Contestar al dorso

Sea X una variable aleatoria que sigue una distribución gamma de parámetros α y β. Demostrar que la media de X vale α/β.

Estadística. 2º examen parcial. 14-11-2013

Instrucciones:

- Tiempo para esta parte del examen: 1 hora y 10 minutos.
- Sólo se puede salir al servicio en casos excepcionales y previa autorización de un profesor.
- Las soluciones, las notas y la fecha de revisión del examen se publicarán en el espacio Moodle de la asignatura.

Problema 1 (35 % de la nota del examen)

Se considera un *canal binario simétrico*, por el que circulan bits de modo independiente.

(a) (2 puntos) En promedio, por el canal circulan 3 bits cada segundo. Calcular la probabilidad de que en 2 segundos circulen al menos 8 bits.

(b) (2 puntos) Cada bit llega a su destino correctamente, con probabilidad 0.8, o alterado por el ruido. Si se transmite un byte (8 bits), hallar la probabilidad de que menos de 3 bits lleguen alterados.

(c) (3 puntos) Se transmite un kilobit (2^{10} bits). Calcular la probabilidad de que entre 811 y 832 bits lleguen correctamente.

(d) (3 puntos) Por el canal circulan mensajes. El tamaño de un mensaje, en bits, es una variable aleatoria, T, cuya densidad aproximada es:

$$f(t) = \begin{cases} 0 & \text{si } t < 3 \\ 18/t^3 & \text{si } 3 \leq t \end{cases}$$

Hallar los tamaños mínimo y medio de un mensaje. ¿Cuanto vale la desviación típica del tamaño de un mensaje? Calcular la proporción de mensajes cuyo tamaño supera los 6 bits.

Solución

Apartado (a)

Se consideran las variables aleatorias:

$$X_i : \text{“número de bits que circulan en el segundo } i\text{”}, \quad i = 1, 2,$$
$$X : \text{“número de bits que circulan en 2 segundos”}.$$

Las variables X_i cuentan eventos por unidad de tiempo, con media constante 3, luego $X_i \sim P(3)$. La distribución de Poisson es reproductiva, luego $X = X_1 + X_2 \sim P(3+3) = P(6)$. Se trata de calcular:

$$p(X \geq 8) = \sum_{i=8}^{\infty} e^{-6} \frac{6^i}{i!} = 1 - \frac{2101}{7} e^{-6} \simeq 0.256$$
$$= 1 - p(X < 8) = 1 - p(X \leq 7) \simeq 1 - 0.744 = 0.256 \quad \text{(tablas)}.$$

Apartado (b)

Se considera la variable aleatoria:

$$Y : \text{“número de bits, en un byte, que llegan alterados”} \sim B(8, 0.2).$$

Esta variable cuenta éxitos, en pruebas independientes y con probabilidad de éxito constante. Se trata de calcular:

$$p(Y < 3) = p(Y \leq 2) = \sum_{i=0}^{2} \binom{8}{i} 0.2^i \, 0.8^{8-i} \simeq 0.7969 \quad (= \text{tablas}).$$

Apartado (c)

Se considera la variable aleatoria:

$$Z: \text{“número de bits, en un kilobit, que llegan correctamente”} \sim B\left(1024, 0.8\right).$$

Esta variable cuenta éxitos, en pruebas independientes y con probabilidad de éxito constante. Se trata de calcular:

$$p\left(811 \leq Z \leq 832\right) = \sum_{i=811}^{832} \binom{1024}{i} 0.8^i \, 0.2^{1024-i} \simeq 0.60376.$$

Para ello, se puede aproximar la distribución binomial por la normal, usando el Teorema del Límite Central:

$$Z \sim B\left(1024, 0.8\right) \approx N\left(1024 \cdot 0.8, \sqrt{1024 \cdot 0.8 \cdot 0.2}\right) = N\left(819.2, 12.8\right).$$

Tipificando, resulta:

$$\begin{aligned} p\left(811 \leq Z \leq 832\right) &= p\left(\frac{811 - 819.2}{12.8} \leq \frac{Z - 819.2}{12.8} \leq \frac{832 - 819.2}{12.8}\right) \simeq p\left(-0.64063 \leq N\left(0,1\right) \leq 1\right) = \\ &= p\left(N\left(0,1\right) \leq 1\right) - p\left(N\left(0,1\right) < -0.64063\right) = \\ &= p\left(N\left(0,1\right) \leq 1\right) - p\left(N\left(0,1\right) > 0.64063\right) = \\ &= p\left(N\left(0,1\right) \leq 1\right) - \left(1 - p\left(N\left(0,1\right) \leq 0.64063\right)\right) \simeq \\ &\simeq 0.8413 - 1 + 0.7389 = 0.5802 \quad \text{(tablas)}. \end{aligned}$$

Apartado (d)

Se considera la variable aleatoria:

$$T: \text{“tamaño de un mensaje (bit)”} \sim \text{Pareto}\left(2, 3\right).$$

Por su función de densidad, T sigue una distribución de Pareto de parámetros $\alpha = 2$ y $k = 3$. El tamaño mínimo de un mensaje es $k = 3$ bits. El tamaño medio de un mensaje es $\mu\left(T\right) = \dfrac{\alpha\, k}{\alpha - 1} = \dfrac{2 \cdot 3}{2 - 1} = 6$ bits. La desviación típica del tamaño de un mensaje no existe (o es infinita), porque la varianza de una distribución de Pareto sólo existe cuando $\alpha > 2$.

Se trata de calcular la probabilidad $p\left(T > 6\right)$. La función de distribución de T es:

$$F\left(t\right) := \begin{cases} 0 & \text{si } t < 3 \\ 1 - 9/t^2 & \text{si } 3 \leq t \end{cases}$$

Por tanto: $p\left(T > 6\right) = 1 - p\left(T \leq 6\right) = 1 - \left(1 - 9/6^2\right) = 1/4$.

•

Apellidos:		
Nombre:		Titulación:
Universidad Politécnica de Madrid ETSI SISTEMAS INFORMÁTICOS	**Estadística. Parte con ordenador.**	**14-11-2012**
	• **Tiempo: 40 minutos** • **Valor: 35%**	**MODELO A**

El anisaquis es un parásito que se cría en los peces y produce una enfermedad intestinal a los humanos que comen estos peces. Para estimar la proporción de peces infectados que llegan al mercado en España, se tomaron muestras del mismo número de peces de distintas variedades y tamaños en 130 lonjas distintas. Los datos obtenidos están recogidos en el fichero **noviembre2013.sf3.**

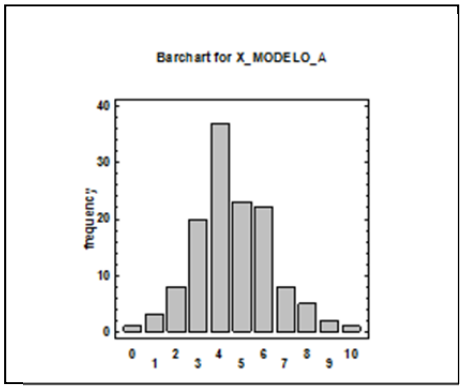
En primer lugar, se quiere estudiar la variable:

X_MODELO_A: Número de peces con anisaquis en cada muestra.

Se pide:

(a) (3.5 puntos) Ajustar un modelo de probabilidad a los datos correspondientes a la variable X, estimar sus parámetros a mano y contrastar la validez de la hipótesis efectuada. Indicar, justificando en cada caso la respuesta, el modelo elegido, los parámetros y el p-valor del contraste. Si no se resuelve el sistema para determinar los parámetros, hay que elegir razonadamente el p-valor.

A la vista de los datos de la muestra de la variable X_MODELO_A, se trata de una variable discreta. Si calculamos las principales medidas descriptivas y dibujamos el diagrama de barras podemos observar:



Summary Statistics for X_MODELO_A

Count	130
Average	4,6
Variance	3,07907
Coeff. of variation	38,1463%
Minimum	0
Maximum	10,0
Range	10,0
Skewness	0,333637

(1'5 puntos) Observamos que en la muestra aparecen once valores distintos 0,1,…,10. Además, el coeficiente de asimetría es positivo, lo que podría indicar que los datos proceden bien de una binomial con $p<0'5$, o bien de una Poisson con λ alto. Como la varianza es bastante menor que la media, nos decidimos por el modelo binomial.

(1 punto) Para estimar los parámetros, resolvemos el sistema:

$$\begin{cases} np = 4'6 \\ np(1-p) = 3'07907 \end{cases}$$

resultando $n = 13'91254035$ $p = 0'3306369565$. Para $n = 14$ resulta $p = 0'328571$

(1 punto) Realizamos el test correspondiente (Chi-square test) con Statgraphics, tomando $n = 14$.

Statgraphics ajusta a una binomial $B(14 \text{ (specified)}, 0'328571)$ con un $p\text{-}value = 0'56189$.

A la vista de los resultados admitimos que los datos proceden de una binomial con esos parámetros.

Para X_MODELO_B, resolviendo el sistema resulta: $n = 12'75817586$ $p = 0'2351433333$. Tomando $n = 13$ y ajustando a una $B(13\ (\text{specified}), 0'230769)$ resulta un $p\text{-}value =$ $0'636035$, luego admitimos que la variable sigue un modelo $B(13, 0'230769)$

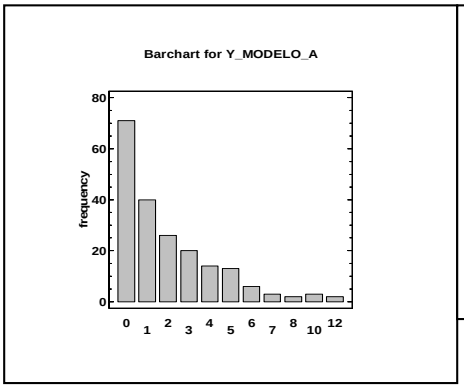
Hace unos años, se realizó una campaña para informar a los consumidores sobre la necesidad de congelar el pescado antes de consumirlo, sobre todo si se consume poco cocinado. Para medir el resultado de la campaña, entre las personas infectadas se han tomado datos de las siguientes variables:

Y_MODELO_A: Número de veces que se ha consumido pescado sin congelar antes de coger la infección.

Z_MODELO_A: Edad de las personas infectadas.

(b) (2 puntos) Ajustar un modelo de probabilidad a los datos correspondientes a la variable Y. Justificar la elección y estimar sus parámetros a mano. Indicar el modelo elegido, el valor del parámetro y el p-valor del contraste.

Y_MODELO_A, se trata nuevamente de una variable discreta. Si calculamos las principales medidas descriptivas y dibujamos el diagrama de barras podemos observar:



Summary Statistics for Y_MODELO_A

Count	200
Average	2,0
Variance	5,78894
Coeff. of variation	120,301%
Minimum	0
Maximum	12,0
Range	12,0
Skewness	1,6771

(1 punto) Observamos que en la muestra aparecen trece valores distintos 0,1,…,12. Además, el coeficiente de asimetría es positivo y el coeficiente de variación es muy alto, lo que junto con la gráfica sugiere un modelo geométrico. Para estimar a mano el valor de *p*, resolvemos la ecuación:

$$\frac{1-p}{p} = 2'0 \Rightarrow p = 0'3333$$

(1 punto) Statgraphics da directamente el valor del parámetro, luego la hipótesis es que Y es una geométrica con $p = 0'3333$. El p_valor que proporciona el Chi-square test es $p\text{-}value = 0{,}932731$, luego aceptamos la hipótesis.

Para Y_MODELO_B, la variable se ajusta a una distribución geométrica con $p = 0'27027$ y un $p\text{-}value = 0'805292$.

(c) (2 puntos) Se considera que una persona es temeraria si consume pescado sin congelar más de 10 veces y se considera que una persona es WERT (**W**alk-on-**E**dge **R**isk **T**aker) si consume pescado sin congelar más de 15 veces. Si se elige un individuo al azar entre los infectados y, utilizando los parámetros que proporciona Statgraphics:

Utilizando el parámetro estimado por Statgraphics, la variable sigue una distribución $G(0{,}3333)$.

a. (1 punto) Calcular la probabilidad de que sea temerario.

$P(G(0{,}3333) > 10) = 0{,}0115674$

b. (1 punto) Calcular la probabilidad de que sea temerario, pero sin llegar a ser WERT.

$$P(10 < G(0{,}3333) \le 15) = P(G(0{,}3333) \le 15) - P(G(0{,}3333) \le 10) =$$
$$= 0'997715 + 0'000761714 - (0'98265 + 0'00578282) =$$
$$= 0'01004389$$

O bien:

$$P(10 < G(0{,}3333) \le 15) = P(G(0{,}3333) = 11) + P(G(0{,}3333) = 12) + P(G(0{,}3333) = 13) +$$
$$+ P(G(0{,}3333) = 14) + P(G(0{,}3333) = 15) =$$
$$= 0'00385541 + 0'0025704 + 0'00171369 + 0'00114251 + 0'000761714 =$$
$$= 0'01004372$$

Una tercera posibilidad es:

$$P(10 < G(0{,}3333) \le 15) = F(16-) - F(11-) =$$
$$= 0'998476 - 0'988433 = 0'010043$$

Para el modelo B:

$$P(G(0{,}27027) > 10) = 0{,}0312453$$

$$P(10 < G(0'27027) \le 15) = P(G(0'27027) \le 15) - P(G(0'27027) \le 10) =$$
$$= 0'99114 + 0'00239459 - (0'957182 + 0'0115723) =$$
$$= 0'02478029$$

O bien:

$$P(10 < G(0'27027) \le 15) = P(G(0'27027) = 11) + P(G(0'27027) = 12) + P(G(0'27027) = 13) +$$
$$+ P(G(0'27027) = 14) + P(G(0'27027) = 15) =$$
$$= 0'00844466 + 0'00616232 + 0'00449683 + 0'00328147 + 0'00239459 =$$
$$= 0'02477987$$

Una tercera posibilidad es:

$$P(10 < G(0{,}3333) \le 15) = F(16-) - F(11-) =$$
$$= 0'993535 - 0'968755 = 0'024780$$

(d) (2.5 puntos) Estudiar si la variable que indica el número de veces que se ha comido pescado antes de infectarse depende linealmente de la edad de los consumidores. En la respuesta escribir la recta de regresión correspondiente y el coeficiente de correlación lineal. A la vista de los resultados:

a. (1'5 puntos) ¿Se puede admitir que el riesgo de infección por anisaquis aumenta con la edad? Razonar la respuesta.

Si calculamos la recta de regresión de Y sobre Z y el coeficiente de correlación lineal, obtenemos:

Y_MODELO_A = -3,8886 + 0,150488*Z_MODELO_A y Correlation Coefficient = 0,40667.

Para el modelo B:

Y_MODELO_B = 1,04429 + 0,0609613*Z_MODELO_B y Correlation Coefficient = 0,474688

Lo que indica que hay una relación lineal positiva, pero débil, entre las variables.

b. (1 punto) Estimar el número de veces que ha comido pescado sin congelar una persona infectada de 56 años.

Sustituyendo en la recta de regresión obtenemos:

$Y\_MODELO\_A = -3'8886 + 0,150488*56 = 4'5387$, es decir, se estima que habrá comido entre 4 y 5 veces pescado sin congelar antes de infectarse.

Para el modelo B:

$Y\_MODELO\_B = 1,04429 + 0,0609613*36 = 3'238958$