Estadística Computacional EST-46112-001

Examen Parcial (Examen para llevar)

Prof. Dr. León Berdichevsky Acosta

05 de octubre de 2017

Total: 100 puntos (40% de la Calificación Final).

Fecha de entrega: 10 de octubre antes de las 23:59 horas.

Instrucciones:

- El examen puede ser resuelto de manera individual o en parejas.
- El examen consta de 5 preguntas.
- Las respuestas a las preguntas deben ser claras y deben incluir procedimiento y código en R de manera ordenada y comentado.
- Las respuestas deben de enviarse en un solo archivo en formato *.html* o *.pdf*. El nombre del archivo debe ser "Examen_Parcial_*ClaveUnica*" con la clave única de un solo alumno.
- El código en R de todas las preguntas puede ser parte del archivo anterior o puede enviarse por separado en un archivo con el mismo nombre "Examen_Parcial_*ClaveUnica*.R" en formato de R Script *.R*.
- Los archivos deben incluir el (los) nombre(s) y clave(s) única(s) al inicio de los mismos.

1. Manipulación y Visualización de Datos en R

La base de datos `iris`, contenida en el paquete base de R, contiene las medidas en centímetros de algunos atributos de 50 flores de 3 especies del género *Iris*:

a. ¿Cumple la base de datos el principio de datos limpios? Justifique su respuesta.
b. En caso de que no cumpla el principio de datos limpios, limpie los datos. Imprima las primeras 6 líneas de los datos limpios (si ya estaban limpios entonces imprima las primeras 6 líneas de los datos originales).

Cada una de las siguientes preguntas requiere utilizar la base de datos con datos limpios del inciso b.:

c. ¿Cuántas observaciones y cuántas variables tiene la base de datos?
d. ¿Cuál es la clase atómica de cada una de las variables?
e. Filtre las flores de la especie (variable *Species*) *Setosa* e imprima las primeras 6 observaciones.

f. Ordene la base de datos de manera descendente con respecto a la variable *Petal.Length* e imprima las primeras 6 observaciones.
g. Cree una nueva variable en donde se muestre el atributo *Sepal.Length* en milímetros e imprima las primeras 6 observaciones.
h. Elimine las observaciones con valores faltantes en la variable *Sepal.Width* e indique el número de observaciones de la nueva base de datos.
i. ¿Cuál es la media de la variable *Petal.Width* para cada una de las especies (variable *Species*) de flores?
j. Realice una gráfica de dispersión de las variables x=*Sepal.Length* contra y=*Sepal.Width* en la que se distingan las diferentes especies (variable *Species*) por color o por forma de los puntos. La gráfica debe incluir título y nombres de los ejes horizontal y vertical.
k. Realice una gráfica de cajas de la variable *Petal.Length* en la que se distingan las diferentes especies (variable *Species*).

2. Espacio de Probabilidad y Variables Aleatorias

Considere un experimento que consiste en una carrera de caballos con tres caballos numerados del 1 al 3. Si no está permitido que dos o más caballos lleguen a la meta en la misma posición:

a. ¿Cuál es el espacio de resultados Ω del experimento?

Asumiendo que todos los elementos del espacio de resultados $\omega \in \Omega$ tienen la misma probabilidad $P(\omega)$ de ocurrir:

b. ¿Cuál es esta probabilidad $P(\omega)$?

Si A denota el evento en el que el caballo número 1 llega a la meta dentro de las primeras dos posiciones y B denota el evento en el que el caballo número 3 llega a la meta en la segunda posición…

c. ¿Cuáles son los elementos de los eventos A y B, respectivamente?
d. ¿Cuáles son los elementos del evento $A \cap B$?
e. ¿Cuáles son los elementos del evento $B \cup A$?
f. ¿Cuál es la probabilidad $P(B)$ de que ocurra el evento B?

Sea $X: \Omega \to \mathbb{R}$ una variable aleatoria que describe la posición en la que llegó a la meta el caballo número 2:

g. Liste los valores $X(\omega)$ que toma la variable X para cada uno de los elementos $\omega \in \Omega$.
h. ¿Cuál es la probabilidad $P(X = 1)$?

3. Probabilidad Condicional

Una inmobiliaria ha determinado que si X es el número de habitaciones de los departamentos que maneja y Y el número de lugares de estacionamiento, entonces la distribución conjunta de las variables X y Y se muestra en la siguiente tabla:

Número de Habitaciones X	Lugares de Estacionamiento Y					Total
	0	**1**	**2**	**3**	**4**	
1	0.06	0.12	0.02	0.00	0.00	0.20
2	0.03	0.18	0.21	0.00	0.00	0.42
3	0.01	0.09	0.11	0.07	0.00	0.28
4	0.00	0.02	0.05	0.02	0.01	0.10
Total	0.10	0.41	0.39	0.09	0.01	1.00

a. ¿Son X y Y variables independientes? Justifique su respuesta.
b. Calcule las probabilidades condicionales $P(Y|X = x)$ para $x = 1, 2, 3, 4$.
c. Verifique que $P(Y|X = x)$ satisface la segunda regla de probabilidad $\sum_{y=0}^{4} P(Y = y|X = x) = 1$ para $x = 1, 2, 3, 4$.
d. Calcule los valores esperados condicionales $E[Y|X = x]$ para $x = 1, 2, 3, 4$.
e. Grafique $g(x) = E[Y|X = x]$ para $x = 1, 2, 3, 4$.

4. Bootstrap

Se desea simular muestras de tamaño 20 de una distribución exponencial con tasa $\beta = 1$. El estadístico de interés $\hat{\theta}$ es el estimador de la media $\theta = \beta$. Siga el siguiente proceso:

i. Utilice la función `rexp()` (y la semilla 261285) para generar una muestra aleatoria de tamaño 20 de una distribución exponencial con tasa $\beta = 1$.
ii. Genere 2,000 muestras bootstrap y calcule intervalos de confianza con coeficiente de confianza de 95% para $\hat{\theta}$ usando 1) el método normal, 2) percentiles y 3) BC_a.
iii. Revise si el intervalo de confianza contiene el verdadero valor del parámetro θ; en caso de que no lo contenga registre si falló por la izquierda o falló por la derecha.

a. Repita el proceso descrito 500 veces y llena la siguiente tabla:

Método	% fallo izquierda	% fallo derecha	Cobertura (simulaciones)
Normal			
Percentiles			
BC_a			

La columna Cobertura es una estimación de la cobertura del intervalo basada en las simulaciones; para calcularla simplemente escriba el porcentaje de los intervalos que incluyeron el verdadero valor del parámetro. Recuerde usar la semilla.

b. Realice una gráfica de paneles: en cada panel mostrará los resultados de uno de los métodos (Normal, Percentil y BC_a), en el eje horizontal graficará el número de intervalos de confianza (1,2,…,500) y en el eje vertical graficará los límites de los intervalos, es decir, graficará 2 líneas (use la función `geom_line`): una corresponderá a los límites inferiores de los intervalos y la otra a los superiores.

5. Simulación de Variables Aleatorias

Una variable aleatoria X tiene una distribución binomial con parámetros n y p, esto es, $X \sim Binomial(n, p)$ si su función de masa de probabilidad es:

$$p_i \coloneqq P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \;\; i \in \{0, 1, \dots, n\}.$$

El objetivo es generar X a partir de una variable aleatoria U con distribución uniforme continua en el intervalo $(0,1)$, esto es, $U \sim Uniforme(0,1)$ utilizando el Método de Transformación Inversa Discreta. La clave para utilizar este método en el presente caso es seguir un procedimiento análogo al que se siguió en clase para la distribución Poisson:

a. Encuentre la relación de recurrencia entre p_{i+1} y p_i para $i \geq 0$.
b. Utilizando la relación de recurrencia del inciso a., escriba un algoritmo de 5 pasos que genere una variable aleatoria binomial con parámetros n y p mediante el Método de Transformación Inversa Discreta.
c. Escriba en R una función que implemente el algoritmo del inciso b. para $n = 10$ y $p = 0.3$.
d. Realice 10,000 simulaciones utilizando la semilla 221285 y reporte las primeras 5 simulaciones obtenidas.
e. Genere un histograma con las 10,000 simulaciones anteriores y compárelo con una distribución construida utilizando la función `dbinom` de R.