Inferencia Estadística

Tarea 3

09/09/2020

Escriba de manera concisa y clara sus resultados, justificando los pasos necesarios. Serán descontados puntos de los ejercicios mal escritos y que contenga ecuaciones sin una estructura gramatical adecuada. Las conclusiones deben escribirse en el contexto del problema. Todos los programas y simulaciones tienen que realizarse en `R`.

1. Resuelva lo siguiente:

   a) Let $X \sim Exponencial(\beta)$. Encuentre $P(|X - \mu_X| \geq k\sigma_X)$ para $k > 1$. Compare esta probabilidad con la que obtiene de la desigualdad de Chebyshev.

   b) Sean $X_1, \ldots, X_n \sim Bernoulli(p)$ y $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Usando las desigualdades de Chebyshev y Hoeffding, acote $P(|\bar{X}_n - p| > \epsilon)$. Demuestre que para n grande la cota de Hoeffding es más pequeña que la cota de Chebyshev. ¿En qué beneficia esto?

2. Sean $X_1, \ldots, X_n \sim Bernoulli(p)$.

   a) Sea $\alpha > 0$ fijo y defina

   $$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

   Sea $\hat{p_n} = n^{-1}\sum_{i=1}^n X_i$. Defina $C_n = (\hat{p_n} - \epsilon_n, \hat{p_n} + \epsilon_n)$. Use la desigualdad de Hoeffding para demostrar que

   $$P(C_n \text{ contiene a } p) \geq 1 - \alpha.$$

   Diremos que C_n es un $(1-\alpha)$-intervalo de confianza para p. En la práctica, se trunca el intervalo de tal manera de que no vaya debajo del 0 o arriba del 1.

   b) Sea $\alpha = 0.05$ y $p = 0.4$. Mediante simulaciones, realice un estudio para ver que tan a menudo el intervalo de confianza contiene a p (la cobertura). Haga esto para $n = 10, 50, 100, 250, 500, 1000, 2500, 5000, 10000$. Grafique la cobertura contra n.

   c) Grafique la longitud del intervalo contra n. Suponga que deseamos que la longitud del intervalo sea menor que 0.05. ¿Qué tan grande debe ser n?

3. Una partícula se encuentra inicialmente en el origen de la recta real y se mueve en saltos de una unidad. Para cada salto, la probabilidad de que la partícula salte una unidad a la izquierda es p y la probabilidad de que salte una unidad a la derecha es $1 - p$. Denotemos por X_n a la posición de la partícula después de n unidades. Encuentre $E(X_n)$ y $\text{Var}(X_n)$. Esto se conoce como una caminata aleatoria en una dimensión.

4. El siguiente conjuntos de datos contiene mediciones del diámetro de un agave, medido en decímetros, en distintas localizaciones no cercanas.

23.37	21.87	24.41	21.27	23.33	15.20	24.21	27.52	15.48	27.19
25.05	20.40	21.05	28.83	22.90	18.00	17.55	25.92	23.64	28.96
23.02	17.32	30.74	26.73	17.22	22.81	20.78	23.17	21.60	22.37

a) Escriba una función en `R` que calcule la función de distribución empírica para un conjunto de datos dado D. La función debe tomar como parámetros al valor x donde se evalúa y al conjunto de datos D. Utilizando esta función grafique la función de distribución empírica asociada al conjunto de datos de agave. Ponga atención a los puntos de discontinuidad. ¿Qué observa? **Nota**: Escriba la función mediante el algoritmo descrito en las notas de la clase; para este ejercicio no vale usar la funciones implementadas en `R` que hacen lo pedido.

b) Usando la desigualdad de Dvoretzky-Kiefer-Wolfowitz, escriba una función en `R` que calcule y grafique una región de confianza para la función de distribución empírica. La función debe tomar como parámetros al conjunto de datos que se usan para contruir la función de distribución empírica.

c) Escriba una función en `R` que determine la gráfica Q-Q normal de un conjunto de datos. La función debe tomar como parámetro al conjunto de datos y deberá graficar contra el percentil estandarizado de la normal. Para poder comparar el ajuste más claramente, la función además deberá ajustar en rojo a la recta $sx + \bar{x}$ (s =desviación estándar muestral y $\bar{x}$ =media muestral). Usando esta función, determine la gráfica Q-Q normal. ¿Qué observa? **Nota**: La misma del inciso a).

d) Escriba una función en `R` que determine el gráfico de probabilidad normal. La función debe tomar como parámetro al conjunto de datos. ¿Qué observa? **Nota**: La misma del inciso a).

e) ¿Los datos anteriores se distribuyen normalmente? Argumente.

5. En este ejercicio repasará la estimación de densidades.

   a) Escriba una función en `R` que estime una densidad por el método de kerneles. La función deberá recibir al punto x donde se evalúa al estimador, al parámetro de suavidad h, al kernel que se utilizará en la estimación y al conjunto de datos.

   b) Cargue en `R` al archivo "Tratamiento.csv", el cual contiene la duración de los períodos de tratamiento (en días) de los pacientes de control en un estudio de suicidio. Utilice la función del inciso anterior para estimar la densidad del conjunto de datos para $h = 20, 30, 60$. Grafique las densidades estimadas. ¿Cuál es el mejor valor para h? Argumente.

   c) En el contexto de la estimación de densidades, escriba una función en `R` que determine el ancho de banda que optimiza al ISE. Grafique la densidad con ancho de banda óptimo para el conjunto de datos de "Tratamiento.csv".

6. Cargue en `R` al conjunto de datos "Maíz.csv", el cual contiene el precio mensual de la tonelada de maíz y el precio de la tonelada de tortillas en USD. En este ejercicio tendrá que estimar los coeficientes de una regresión lineal simple.

   a) Calcule de forma explícita la estimación de los coeficientes via mínimos cuadrados y ajuste la regresión correspondiente. Concluya.

   b) Calcule de forma explícita la estimación de los coeficientes via regresión no-paramétrica tipo kernel (ver Nadaraya, E. A. (1964). "On Estimating Regression". Theory of Probability and its Applications. 9 (1): 141–2. doi:10.1137/1109020) y ajuste la regresión correspondiente. Concluya.

c) Compare ambos resultados. ¿Qué diferencias observa?

7. Demuestre que la fórmula de la densidad de la Beta integra 1.

8. En este ejercicio se comprobará que tan buena es la aproximación dada por las reglas empíricas para algunas de las distribuciones estudiadas en la clase. Considerese las distribuciones Unif($a = -3, b = 3$), Normal$(0, 1)$, Exponencial(2), Gamma($\alpha = 2, \beta = 1$), Gamma($\alpha = 3, \beta = 1$), Beta($\alpha = 2, \beta = 2$), Weibull($\alpha = 4, \beta = 1$) y Lognormal($\mu = 3, \sigma = 2$).

   a) Para cada una de las distribuciones anteriores, haga una tabla que muestre las probabilidades contenidas en los intervalos $(\mu - k\sigma, \mu + k\sigma)$, para $k = 1, 2, 3$. Utilice las fórmulas de las medias y varianzas contenidas en las notas para determinar μ y σ en cada caso. Puede usar `R` para determinar las probabilidades pedidas.

   b) En `R`, simule $n = 1000$ muestras de cada una de las distribuciones anteriores y calcule la media muestral $\bar{x}$ y la varianza muestral s^2 como se mencionó en la clase. En cada caso, calcule la proporción de observaciones que quedan en los intervalos $(\bar{x} - ks, \bar{x} + ks)$, para $k = 1, 2, 3$. Reporte sus hallazgos en una tabla como la del inciso anterior. ¿Qué tanto se parecen la tabla de este inciso y la del anterior?

Honors problems

1. *a*) Sea X una v.a. discreta con media finita y que toma valores en el conjunto 0, 1, 2, ... Demuestre que

$$E(X) = \sum_{k=1}^{\infty} P(X \geq k).$$

   b) Sea X una v.a. continua no-negativa con media finita, función de densidad f y función de distribución F. Demuestre que

$$E(X) = \int_0^{\infty} (1 - F(t))dt.$$

   c) ¿Cómo cambia la fórmula del caso anterior cuando el soporte de X es todo $\mathbb{R}$?

2. Sea X una v.a. continua con primer momento finito. Demuestre que la función $G(c) = E(|X - c|)$ $c \in \mathbb{R}$, se minimiza en $c = M(X)$ para $M(X)$ la mediana de X.

Entrega: 22/09/2020.