Tarea 7 - Análisis de datos
Giovanni Gamaliel López Padilla

Problema 2

Supongamos que $[1.15, 4.20]$ es un intervalo de 95 % de confianza para el promedio μ del número de televisiones por hogar en EE.UU. ¿Cómo interpretar eso? Clasifica cada uina de las siguientes frases como cierto o falso. Motiva tu respuesta.

a) **95 % de los hogares tienen entre 1.15 y 4.20 televisiones**

Esto es cierto, ya que podemos interpretar como una probabilidad el porcentaje de la confianza ya que esta es transformada a una distribución normal con μ y σ^2 de los datos dados. De tal manera que:

$$P(1.15 \leq X \leq 4.20) = \int_{1.15}^{4.20} \mathcal{N}(\mu, \sigma^2)dx = 95\,\%$$

b) **La probabilidad que μ esté entre 1.15 y 4.20 es de 95 %.**

Esto es falso, ya que una vez definido el intervalo, uno puede asegurar si el promedio se encuentra dentro o fuera de el.

c) **De 100 intervalos calculados de la misma manera, esperamos que 95 % contiene a μ.**

Es cierto si la muestra de los datos es diferente en cada calculo, ya que en este contexto, los límites del intervalo actuan como una variable aleatoria.

Problema 3

a) **Si $\hat{\theta}$ es un estimador insesgado para θ, entonces ¿$\hat{\theta^2}$ es un estimador insesgado para θ^2?**

Se tiene que para una muestra X basadas en una distribución con parámetro θ, y para el estimador $\hat{\theta}$, el sesgo esta definido como:

$$b(\theta) = E(\hat{\theta} - \theta)$$

Si $b(\theta) = 0$, entonces el estimador $\hat{\theta}$ es insesgado. Por lo tanto, calculando el sesgo para $\hat{\theta}^2$ y θ^2 es:

$$\begin{aligned} E(\hat{\theta}^2 - \theta^2) &= E\hat{\theta}^2 - E\theta^2 \\ &= E\hat{\theta}^2 - \theta^2 \end{aligned}$$

dado que $E\hat{\theta} = \theta$, entonces

$$\begin{aligned} E(\hat{\theta}^2 - \theta^2) &= E\hat{\theta}^2 - \theta^2 \\ &= E\hat{\theta}^2 - (E\theta)^2 \\ &= Var(\hat{\theta}) \end{aligned}$$

por lo que, el estimador $\hat{\theta}$ no necesariamente es insesgado.

b) Si $[\hat{\theta}_L, \hat{\theta}_R]$ es un intervalo de 95 % de confianza para θ, entonces ¿$[exp(\hat{\theta}_L), exp(\hat{\theta}_R)]$ es un intervalo de 95 % de confianza para $exp(\theta)$?

Esto es cierto, ya que el intervalo $[\hat{\theta}_L, \hat{\theta}_R]$ tiene la siguiente definición:

$$\frac{1}{n}\sum x_i - \frac{a}{\sqrt{n}} < \theta < \frac{1}{n}\sum x_i + \frac{b}{\sqrt{n}}$$

entonces, aplicando una función exponencial se tiene que:

$$exp\left(\frac{1}{n}\sum x_i - \frac{a}{\sqrt{n}}\right) < exp(\theta) < exp\left(\frac{1}{n}\sum x_i + \frac{b}{\sqrt{n}}\right)$$
$$exp(\hat{\theta}_L) < exp(\theta) < exp(\hat{\theta}_R)$$

Problema 4

El tiempo de ejecucción de un programa sigue una distribución normal. Para una muestra de tamaño 40 se obtiene que $\bar{x} = 32.2$s y $\sigma^2 = 3.1\ s^2$.

a) ¿Cuantas veces se debe ejecutar el programa para obtener un intervaloc de confianza de 95 % con un ancho menor que 2 s?

Se tiene que un intervalo de confianza es:

$$\left[\frac{1}{n}\sum x_i - \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}, \frac{1}{n}\sum x_i + \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}\right]$$

entonces, el ancho del intervalo es:

$$d = \frac{2Z_{\alpha/2}\sigma}{\sqrt{n}}$$

por ende:

$$n = \frac{4Z^2_{\alpha/2}\sigma^2}{d^2}$$

por lo tanto:

$$\begin{aligned} n &= \frac{4(1.959964)^2(3.1)^2}{2^2} \\ n &= 36.9164 \end{aligned}$$

b) En muchas situaciones, el interés no es tanto en el comportamiento promedio, si no en la variabilidad. Usando el hecho que para una muestra de una misma distribución normal con varianza σ^2

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

con $S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ y χ^2_{n-1} una distribución chicuadrada, deriva un intervalo de confianza de 95 % para la varianza.

Se tiene que

$$P(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}) = 1 - \alpha$$

entonces:

$$P\left(\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}}\right)$$

por lo tanto, el intervalo de confianza para la varianza es:

$$\left(\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}}\right)$$

Para un intervalo de confianza del 95 %, se tiene $\alpha = 0.05$. El intervalo de confianza para la varianza es:

$$\left(\frac{(n-1)S^2}{\chi_{n-1,0.025}}, \frac{(n-1)S^2}{\chi_{n-1,0.975}}\right)$$

Problema 5

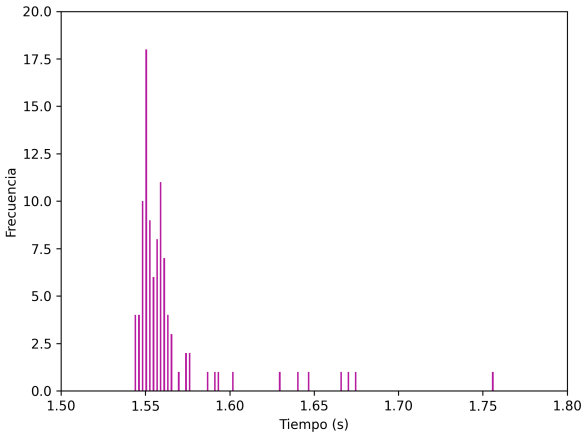
En este ejercicio usamos intervalos de confianza para entender mejor el desempeño de algoritmos. En general, el componente aleatorio puede entrar de dos maneras: en la dinámica del algoritmo o por los datos/parámetros de entrada. Calcula un intervalo de confianza de 95 % para el promedio del tiempoque los algoritmos quicksort y shellsort requieren para ordenar 10,000,000 números elegidos al azar de una distribución continua, basado en 100 corridas de cada algoritmo. Puedes usar la versión que está en R. Por ejemplo, para calcular el tiempode una corrida el código es:

```
system.time(x1 <- sort(x, method = "shell"), gcFirst = TRUE)[1]
system.time(x2 <- sort(x, method = "quick"), gcFirst = TRUE)[1]
```

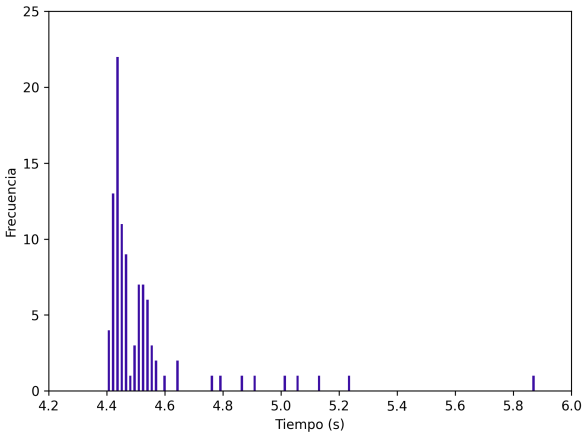
con GcFirst = FIRST se libera primero la memoria (en caso de que sea posible) Explica porque no importa de cual distribución se generan los números siempre y cuando que sean de una variable continua. Construye también un intervalo de confianza para la diferencia de sus tiempos de ejecución para un (mismo) conjunto.

La razón por la cual no importa la elección de la distribución es debido a que los algoritmo de ordenamiento realizaran una función semejante independientemente de la distribución de los

números generados. Otra razón que podemos dar de esta invarianza es el teorema del límite central, ya que toda distribución continua con varianza no nula y fínita se aproximará a una distribución normal. Para este ejercicio se generaron $10,000,000$ números aleatorios usando una distribución uniforme ($X \sim \mathcal{U}(0,1)$). Se crearon 100 sets de la cantidad de números aleatorios y se ordenaron usando los algoritmos quick sort y shell sort. En la figura 1 se representan las distribuciones del tiempo de ejecucción del algoritmo de quick sort (figura 1a) y shell sort (figura 1b).



(a) Distribución del tiempo de ejecucción del algoritmo quick sort.



(b) Distribución del tiempo de ejecucción del algoritmo shell sort.

Figura 1: Distribuciones del tiempo de ejecución para los diferentes algoritmos de ordenamiento.

Realizando la resta de los tiempos ejecución para un mismo set de datos genderdos se obtuvo la distribución mostrada en la figura 2.

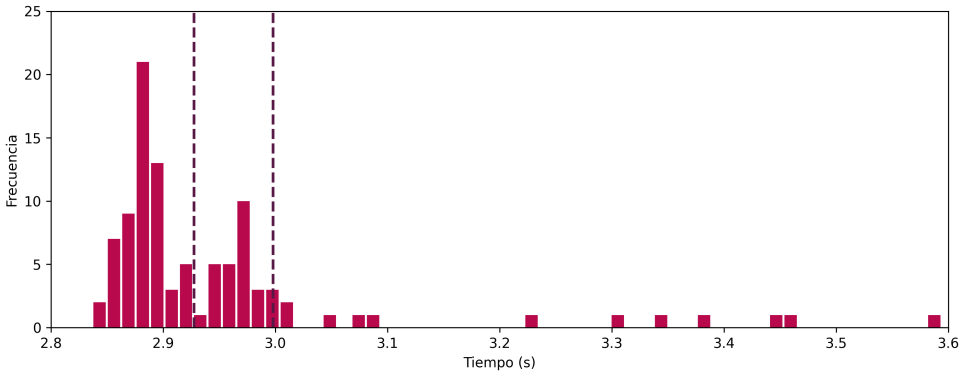


Figura 2: Distribución e intervalo de confianza de la diferencia en los tiempos de ejecución entre los algoritmos de ordenamiento.

El promedio, la varianza y el intervalo de confianza obtenidos de estos datos se encuentran en la tabla 1.

Promedio (μ)	**Varianza** (σ^2)	**IC**
2.96273	0.032194	[2.927563,2.997897]

Tabla 1: Promedio, varianza e intervalo de confianza obtenidos de la diferencia en los tiempos de ejecucción.

Problema 6

El pasado 27 de octubre la revista forbes publicó que según un sondeo entre 800 personas, 28.1 % de los mexicanos está muy de acuerdo con la reforma energética mientras 35.7 % está algo de acuerdo. Calcula un intervalo de 90 % de confianza para el porcentaje de la categoria muy de acuerdo.

Sea x_i una variable aleatoria tal que:

$$P(x_i) = \begin{cases} p & \text{si } x = 1 \\ 1-p & \text{si } x = 0 \end{cases}$$

es decir, $x \sim B(1,p)$. Entonces, si $X = x_1 + x_2 + x_3 + \cdots + x_n$, donde x_i es un individuo en la encuesta, se observa que X mide el número de individuos que estas muy de acuerdo con la reforma. Por lo tanto, $X \sim B(n,p)$. Definimos a $\hat{theta}$ como:

$$\hat{\theta} = \frac{X}{n}$$

el cual representa la proporcion de individuos que estan muy de acuerdo con la reforma. Obteniendo la media y varianza de $\hat{\theta}$, se obtiene que:

$$E(\hat{\theta}) = E\left(\frac{X}{n}\right) = \frac{1}{n}EX = p$$
$$Var(\hat{\theta}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{p(1-p)}{n}$$

Se tiene que

$$\frac{\hat{\theta} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$$

entonces:

$$P\left(-z_{\alpha/2}, \frac{\hat{\theta} - p}{\sqrt{\frac{p(1-p)}{n}}}, z_{\alpha/2}\right) = 1 - \alpha$$

por lo tanto, el intervalo de confianza es:

$$\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

para $\alpha = 0.1$ se tiene que $z_{\alpha/2} = 1.644854$. Entonces, el intervalo de confianza para el porcentaje de las personas que estan muy de acuerdo con la reforma es:

$$\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$\left(0.281 - (1.644854)\sqrt{\frac{0.281(1-0.281)}{800}}, 0.281 + (1.644854)\sqrt{\frac{0.281(1-0.281)}{800}}\right)$$

$$(0.2548603, 0.3071397)$$