Probabilidad y Estadística: un acercamiento vía problemas

(c) E. Cómer-Barragán, TecNM/ITT

Edición Experimental (**contenido parcial**). 23 Enero 2020. Se utilizó DrRacket y el lenguaje Scribble para conformar el presente documento. Para información sobre estas excelentes herramientas favor de visitar el sitio: racket-lang.org

Resumen: Se presenta una colección de problemas selectos (algunos algorítmicos, otros de carácter algebraico, geométrico, analítico o conceptual), los cuales resueltos de manera (preferentemente) secuencial, construyen y permiten obtener familiaridad con los temas principales de un Curso de Probabilidad y Estadística, congruente con el programa AEF-1052, en algunas de las carreras de ingeniería del Tecnológico Nacional de México, vigentes a la fecha de la presente edición.

1. ESTADÍSTICA DESCRIPTIVA

Problema 1.1

$\Pi_{1.1}$ Si X es una v.a. discreta, y se cuenta con datos no agrupados de una muestra de tamaño n, entonces el $\alpha \cdot 100\%$-cuantil (con $\alpha \in [0,1]$), se calcula como:

$$\tilde{x}_\alpha = \begin{cases} x_{(\lceil n\alpha \rceil)} & \text{si } n\alpha \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(n\alpha)} + x_{(n\alpha+1)}\right) & \text{si } n\alpha \in \mathbb{N} \end{cases}.$$

Recordemos que $Q1, Q2, Q3$ se calculan con $\alpha = 0.25, 0.50, 0.75$ respectivamente. Por otra parte la función techo $\lceil x \rceil$ corresponde al mínimo entero mayor o igual a x.

Para una v.a. discreta X, cuyo conjunto soporte S_X son los enteros entre 1 y 9 inclusive, se toma una muestra que resulta en los siguientes valores:

$$(7, 7, 1, 4, 5, 7, 6, 4, 6, 4).$$

Verificar que el promedio de dichos valores es 51/10, que su varianza es 329/90, y que su mediana es 11/2. Adicionalmente, verificar que el primer y tercer cuartil, son respectivamente: 4 y 7.

Observación: Los resultados dados actualmente por la función *median* y la función *quantile* de Racket, difieren un poco de los nuestros, para el caso en que el tamaño de la muestra es par. Esto es debido a que para no salir del conjunto soporte S_X, se selecciona sólo el elemento de la izquierda, y no el promedio de los dos elementos centrales, de los datos ordenados. Ambos resultados en cierto sentido son correctos, ya que no existe un concenso universal respecto al cálculo de cuantiles para el caso de variables discretas. Sin embargo para nuestros cálculos utilizaremos los promedios, conforme el recuadro $\Pi_{1.1}$.

Problema 1.2

Con los datos del Problema 1.1, favor de graficar el histograma correspondiente, y la FDAE (Función de Distribución Acumulada Empírica). Además con los resultados de dicho problema, graficar el diagrama de caja y alambre asociado a dichos datos.

$\Pi_{1.2}$ Para calcular la altura h_j de la clase C_j en el histograma, considerar $h_j = \frac{f_j}{d_j}$. Donde $f_j = \frac{n_j}{n}$. Además cada clase C_j se asocia con el intervalo $(e_{j-1}, e_j]$, y se cumple $d_j = e_j - e_{j-1}$.Este cálculo de h_j es importante cuando el ancho de las clases no es una constante.

Problema 1.3

Si una muestra de una v.a. discreta X con conjunto soporte $\{k \in \mathbb{N} | 1 \leq k \leq 30\}$ consiste de un valor α, y los valores:

$$(27, 4, 19, 29, 9).$$

Determine el valor de α para que la varianza sea igual a 1822/15.

$\Pi_{1.3}$ Si X es una v.a. continua, y se cuenta con datos agrupados de una muestra de tamaño n, con las clases $C_1, C_2, \cdots, C_k$, y sus correspondientes frecuencias absolutas son $n_1, n_2, \cdots, n_k$ (con $\Sigma_{j=1}^{k} n_j = n$), entonces el $\alpha \cdot 100\%$-**cuantil** (con $\alpha \in [0, 1]$), se calcula como: $\tilde{x}_\alpha = e_{q-1} + \frac{\alpha - F_{q-1}}{f_q} d_q$. Donde la clase C_q, cumple con: $F_{q-1} < \alpha$ y $F_q \geq \alpha$.

Problema 1.4

En base a la siguiente muestra de datos agrupados, asociados a una v.a. X continua, realizar el histograma y su FDAE correspondiente $F(x)$. Además, determinar el valor de $F(28)$.

j	C_j	n_j
1	[0,10]	5
2	(10,15]	12
3	(15,25]	21
4	(25,40]	11

Problema 1.5

Para los datos del Problema 1.4, calcular la media aritmética, mediana, moda, varianza y desviación estándar. *Sugerencia*. Ver las páginas 38 y 53 de (Heumann, et al. 2016).

Problema 1.6

Para los datos del Problema 1.4, calcular los cuartiles, y dibujar el diagrama de caja y alambre correspondiente.

2. FUNDAMENTOS DE LA TEORÍA DE PROBABILIDAD

Problema 2.1

¿Cuál es la probabilidad de que al acomodarse al azar para una fotografía normal (de frente, uno al lado del otro), siete niños de los cuales dos son gemelos y tres trillizos, los tres trillizos queden juntos? (nota: se asume que los demás niños son distinguibles entre sí).

$\Pi_{2.1}$ Se recomienda la lectura del documento **Conteo**, del Dr. Gary MacGillivray, así como el capítulo 5 de (Heumann, et al. 2016), este último accesible gracias a los servicios de CONRICyT.

Problema 2.2

¿Cuál es la probabilidad de que al formarse palabras al azar, con las letras de la palabra ABARBANEL, las dos Bs queden separadas?

Problema 2.3

¿Cuál es la probabilidad de que al seleccionar al azar un número entero de cuatro cifras entre 2222 y 7777 inclusive, el número seleccionado sea impar y no tenga cifras repetidas?

Problema 2.4

¿Cuál es la probabilidad de que al seleccionar al azar una solución entera positiva de la ecuación $x_1 + x_2 + 5x_3 = 2019$, se cumpla que $x_3 \geq 400$?

$\Pi_{2.2}$ Si $A_1, A_2, \cdots, A_m$ son conjuntos que forman una partición de Ω, entonces el **Teorema de Bayes**, nos dice que para un conjunto B contenido en Ω, se cumple: $P(A_i|B) = \dfrac{P(B|A_i)P(A_i)}{P(B)}$, donde $P(B) = \sum_{j=1}^{m} P(B|A_j)P(A_j)$.

Problema 2.5

¿Cuál es el valor de la probabilidad $P(\bar{A}|B)$, si se sabe que $P(A) = 0.15$, $P(\bar{B}|A) = 0.8$, y $P(B|\bar{A}) = 0.7$? (*Sug.* Construya un árbol binario, donde las ramas principales corresponden a probabilidades *a priori*, y las ramas secundarias corresponden a (las llamadas algunas veces) probabilidades *modelo*. La probabilidad buscada, se conoce como probabilidad *a posteriori*).Ver recuadro $\Pi_{2.2}$, y relacione su árbol binario (en este caso) con el Teorema de Bayes.

Problema 2.6

Para estimar la calidad de una prueba clínica diseñada para detectar un cierto tipo de virus, se aplica dicha prueba a 150 muestras de sangre, tomadas de pacientes cuyo estatus sobre dicha infección viral se conoce previamente. Si los resultados de la prueba son los indicados en la tabla anexa:

Resultado de prueba	Con virus	Sin virus
Positiva	70	10
Negativa	20	50

Determinar las siguientes probabilidades: (**a**) La prueba es positiva cuando el virus esta presente. (**b**) La prueba es positiva, cuando en realidad el virus esta ausente. (**c**) La prueba es negativa, cuando no hay virus.

Problema 2.7

Un sistema de comunicación de dos etapas (de estación transmisora a estación repetidora, y de ésta a estación receptora), se transmiten ceros y unos. Sin embargo debido a interferencias, la comunicación no es perfecta. En la primera etapa el 1% de los unos se convierten a ceros, y el 2% de los ceros se convierten a unos. En la segunda etapa, el 0.8% de los unos se convierten a ceros, y el 1.2% de los ceros se convierten a unos. Si se transmite un mensaje con 68% de unos, determine las siguientes probabilidades: (**a**) Se reciban unos; (**b**) se haya enviado uno, dado que se recibió uno; (**c**) se haya enviado cero, dado que se recibió cero; (**d**) se haya enviado cero, dado que se recibió uno.

Problema 2.8

¿Cuál es la probabilidad de falla para el sistema de componentes con redundancias, representado por la siguiente matriz de conectividad, sabiendo que la probabilidad de falla de cada componente es 0.002?

$$C = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

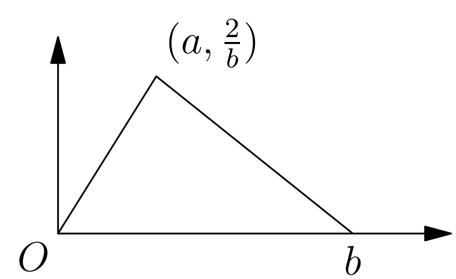
Observación: La matriz de conectividad utilizada es cuadrada, de dimensión $n+2$. Se asume que el primer elemento es un nodo de *entrada*, y el último un nodo de *salida*. Los n elementos intermedios representan los componentes $x_1, x_2, \cdots, x_n$ que conforman el sistema. Considerando índice 0, para el primer renglón (o columna), entonces si $c_{i,j} = 1$, hay una conexión dirigida del componente x_i al componente x_j, en otro caso, no hay conexión.

3. VARIABLES ALEATORIAS

Problema 3.1

Considere que X es una v.a. continua, con función de **densidad** de probabilidad $f(x)$ definida por la gráfica adjunta (con $f(x) = 0$ si $x \notin [0, b]$). Determinar una expresión para la varianza $Var[X]$, sabiendo que para su esperanza matemática se cumple: $E[X] = \dfrac{a+b}{3}$.

$\Pi_{3.1}$ Recordemos que: $Var[X] = E[X^2] - E^2[X]$ y que para una v.a. continua X, con fdp f(x) se cumple: $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$.



Observación: Denotaremos por $X \sim Tri0(a,b)$, para indicar que X es una v.a. contínua, con *fdp* como la indicada en este ejemplo. Nos reservaremos el nombre $Tri(a,b,c)$ para indicar una *fdp* triangular, con soporte $S_X = [a,b]$, y moda c.

Problema 3.2

Considere que $X \sim Tri0(2,9)$, determine: (**a**) $P(1 \leq X \leq 5)$; (**b**) El valor de X que es excedido por el 90% de los valores de X.

Problema 3.3

Considere que X es una v.a. discreta, con función de **masa** de probabilidad $f(x) = x\mathrm{e}^{\alpha x}$. Determine el valor de α, para que en efecto, $f(x)$ sea una *fmp*, con soporte $S_X = \mathbb{N}^{+}$.

Problema 3.4

Para la v.a. X del Problema 3.3, y con el valor de α encontrado: (**a**) Calcular $P(1 \leq X < 5)$; (**b**) ¿Qué valor de X se excede el 85% de las veces?

4. DISTRIBUCIONES DE PROBABILIDAD (SELECTAS)

Problema 4.1

Considere que X es una v.a. discreta, con la distribución indicada en cada inciso. Calcular en cada caso, la probabilidad indicada:

a). $P[X > 1]$ si $X \sim Binomial(n, 0.7)$ y $E[X] = \dfrac{35}{10}$.

b). $P[X \leq 9]$ si $X \sim Pascal(n, 0.8)$ y $Var[X] = \dfrac{35}{16}$.

c). $P[X > 2]$ si $X \sim Poisson(\lambda)$ y $E[X] = 0.4$

$\Pi_{4.1}$ Se recomienda utilizar la siguiente variante de la distribución de *Pascal* (conocida también como *Binomial negativa*). $X \sim Pascal(n,p)$ si su f.m.p. es: $f(x) = \binom{x-1}{n-1}(1-p)^{x-n}p^n$, con $S_X = \{n, n+1, \cdots\}$. Además $E[X] = \dfrac{n}{p}$, y $Var[X] = \dfrac{n(1-p)}{p^2}$.

Problema 4.2

Asuma que $X \sim \mathcal{N}(70, 25)$. Calcular:
a). $P(X > 80)$.
b). $P(65 < X < 75)$.

c). Determinar el valor x^* de X tal que el 25% de la "población" excede dicho valor.

5. ESTADÍSTICA APLICADA

Problema 5.1

Considere que los datos siguientes representan una muestra de los valores de una v.a. discreta X con distribución $Poisson(\lambda)$:

$$[3, 1, 2, 1, 1]\ .$$

Utilice el **método de máxima verosimilitud** (MMV), para determinar una estimación puntual $\hat{\lambda}$ del parámetro λ.

Problema 5.2

Si se asume que los datos del Problema 5.1, provienen de una distribución $Geom(p)$.
a). Determine (utilizando MMV) una estimación $\hat{p}$, del parámetro p.
b). ¿Cuál de las dos distribuciones $Poisson(\hat{\lambda})$, o $Geom(\hat{p})$, considera más apropiada para modelar los datos, y por qué?

$\Pi_{5.1}$ Recordamos que para el caso de un IC para la media de una población con varianza conocida σ, la longitud Δ de dicho IC es $2z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. De dicha relación se puede obtener el tamaño n de la muestra, dado Δ.Para el caso de varianza desconocida, se le invita a revisar (NIST/SEMATECH 2012) sección 7.2.2.2.

Problema 5.3

Utilizando un nivel de confianza del 95%, determinar un *intervalo de confianza*, para el promedio μ de la resistencia de un lote experimental de resistores, para los siguientes dos casos:
a). Si se conoce que la varianza de la población (debido al proceso de fabricación) es 2.56 Ω^2, y para una muestra de tamaño 20, el promedio de dicha muestra es 149.5 Ω.
b). Si no se conoce la varianza de la población, y para la muestra del inciso (a), el valor de la varianza de dicha muestra es 1.69 Ω^2.

Problema 5.4

Para el caso del Problema 5.3 inciso (a) previo,cuál debería haber sido el tamaño de la muestra, para que la longitud del *intervalo de confianza*, fuera igual a 2 Ω.(ver nota al margen $\Pi_{5.1}$)

Problema 5.5

Considere una prueba de hipótesis (PdH) donde $H_0: \ \mu = 220$ PENDIENTE

Problema 5.6

Considere una PdH para el promedio de una población de varianza desconocida. PENDIENTE

6. REGRESIÓN LINEAL Y CORRELACIÓN

Problema 6.1

Determinar los parámetros de un modelo lineal c.r.a dichos parámetros, que mejor se ajusten a los siguientes datos: PENDIENTE

Problema 6.2

Determinar los parámetros de un modelo inicialmente no lineal c.r.a dichos parámetros (pero linealizable), que mejor se ajusten a los siguientes datos: PENDIENTE

ANEXO A: ANTECEDENTES Y PRERREQUISITOS SELECTOS

Problema A.1

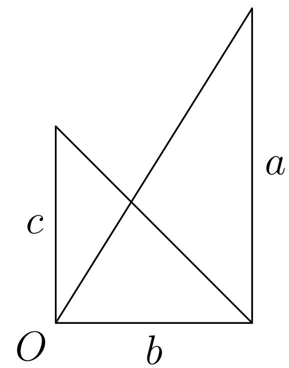
Calcular el valor de la integral

$$\int_{-\infty}^{\infty} \mathbf{e}^{-\frac{1}{2}\left(\frac{x-\alpha}{\beta}\right)^2} dx,$$

en términos de los parámetros $\alpha, \beta \in \mathbb{R}$.

Problema A.2

Determinar las coordenadas de la intersección de las diagonales, asumiendo que el punto O es el origen de coordenadas.



ANEXO B: RESPUESTAS [PARCIALES] A PROBLEMAS SELECTOS

Problema 1.3 ($\alpha = 30$).

BIBLIOGRAFÍA

Heumann, et al. Introduction to Statistics and Data Analysis. Springer, 2016. https://link.springer.com/book/10.1007/978-3-319-46162-5

NIST/SEMATECH. e-Handbook of Statistical Methods. NIST, 2012. https://www.itl.nist.gov/div898/handbook/index.htm

INDEX