Estadística e inferencia I

Primer Cuatrimestre — 2024

Práctica 0: Repaso [v0.1]

Probabilidad

1. Un estudio sobre la relación entre nivel de ingresos (A=alto, M=medio, B=bajo) y la preferencia por una de las tres grandes marcas de automóviles (Y,W,Z) da como resultado la siguiente tabla de probabilidades conjuntas.

	B	M	A	
Y	0.10	0.13	0.02	0.25
W	0.20	0.12	0.08	0.40
Z	0.10	0.15	0.10	0.35
	0.40	0.40	0.20	

Esta tabla muestra, por ejemplo, que $P(\text{ingreso bajo y preferencia } Y) = P(B \cap Y) = 0.10$, que $P(\text{ingreso bajo}) = P(B) = 0.40$ y que $P(\text{preferencia } Y) = P(Y) = 0.25$.

(*a*) Calcular las siguientes probabilidades condicionales:

$$P(W|A) \qquad P(M|Z) \qquad P(Y^c|M)$$
$$P(M|Y^c) \qquad P(M|W \cup Z) \qquad P(B \cup M|Z)$$

(*b*) ¿Cuál es la probabilidad de que una persona elegida al azar prefiera la marca Y o tenga un alto ingreso?

2. Supongamos que se descubre una nueva enfermedad, fatal pero muy rara: la tiene una de cada mil millones de personas. Para detectar esta enfermedad se desarrolló un test que tiene una tasa de falsos positivos de 1 en un millón. Es decir que si el resultado es positivo, la probabilidad de no tener esa enfermedad es de 1 en 1.000.000.

Nuestro héroe se hace el test y le da... ¡positivo! ¡un test tan preciso! A no desesperar:

(*a*) En el mundo hay aproximadamente 8 mil millones de personas: ¿cuántas personas aproximadamente sufren de esta enfermedad?

(*b*) ¿Y aproximadamentea cuántas personas en el mundo les da positivo?

(*c*) De los que salen positivos en el test, ¿cuántos tienen esta enfermedad?

(*d*) ¿Cuál es la probabilidad de que nuestro héroe tenga esa enfermedad rara?

(*e*) Formalizar las cuentas anteriores usando conceptos de probabilidad condicional y el Teorema de Bayes.

3. Calcular la esperanza $E(X)$ y la varianza $V(X)$ cuando

(*a*) X es una Bernoulli con parámetro $p \in (0, 1)$.

(*b*) X es una binomial con parámetros $p \in (0, 1)$ y $n \in \mathbb{N}$.

PDF y CDF

Sea X una variable aleatoria a valores reales. La *función de distribución acumulada*, o *CDF* por sus siglas en inglés, es $F(x) = P(X \leq x)$ para cada $x \in \mathbb{R}$. Si X es continua, la *función de densidad de probabildad*, o *PDF* por sus siglas en inglés, es la derivada de F.

4. Sea X una variable aleatoria con distribución $\mathcal{N}(5, 0.25)$. Calcular

(*a*) $P(4.75 \leq X \leq 5.50)$,

(*b*) $P(|X| > 5.25)$,

(*c*) el valor de c tal que $P(|X - 5| \leq c) = 0.90$,

(*d*) el 90-percentil de X.

5. La porción de memoria ocupada en un servidor de un sistema de terminales en red es una variable aleatoria continua X que toma valores entre 0 (sin carga) y 1 (carga completa). La PDF de X está dada por

$$f(x) = \begin{cases} 4x^3 & \text{si } 0 < x < 1; \\ 0 & \text{en caso contrario.} \end{cases}$$

(*a*) Hallar la mediana de la porción ocupada de memoria.

(*b*) Deducir la densidad de la variable que mide la porción de memoria que falta ocupar, es decir $Z = 1 - X$.

6. Una muestra de tabaco puede provenir de dos variedades distintas, I y II, con probabilidades 0.35 y 0.65 respectivamente. El contenido de nicotina es una variable aleatoria, cuya distribución es $\mathcal{N}(1.9, 0.16)$ en la variedad I y $\mathcal{N}(2.2, 0.09)$ en la variedad II.

(*a*) Hallar la probabilidad de que en una muestra elegida al azar el contenido de nicotina sea mayor o igual que 2.1.

(*b*) Hallar la probabilidad de que dado que el contenido de nicotina es mayor que 2.1, la muestra provenga de la variedad I.

7. Para cada una de las siguientes distribuciones graficar la PDF y la CDF. Calcular media, mediana, moda, error estándar y, si amerita, algunos percentiles.

(*a*) Normal estándar $\mathcal{N}(0, 1)$.

(*b*) Exponencial con parámetro $\lambda = 1/2$.

(*c*) Uniforme continua en el intervalo $(0, 1)$.

(*d*) Distribución t de Student con 10 grados de libertad, y con 100.

Teorema central del límite

8. Para cada n entre 1 y 3000, generar observaciones $x_1, \ldots, x_n$ de $X_1, \ldots, X_n$ vaiid con distribución exponencial con parámetro $\lambda = 1/2$, con λ a elección. Luego, obtener $\bar{x}_n$ como la media de estas observaciones, es decir,

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Graficar n vs x_n e interpretar.

9. Este ejercicio es una constatación experimental del Teorema Central del Límite.

(a) Considerar dos observaciones x_1 y x_2 de variables aleatorias X_1 y X_2 independientes con distribución exponencial con parámetro $\lambda = 1/2$ y guardar el promedio de ambas, es decir, $\bar{x}_2$. Repetir este proceso 1000 veces y a partir de los valores obtenidos, realizar un histograma.

(b) Aumentar a diez las variables promediadas, es decir, considerar ahora $n = 10$ observaciones de variables aleatorias independientes con la misma distribución del ítem anterior. Repetir este proceso 1000 veces y a partir de los valores obtenidos, realizar un histograma.

(c) Hacer histogramas para $n = 100$ y $n = 1000$. Encontrar parámetros μ y σ para que la CDF de una $\mathcal{N}(\mu, \sigma)$ se superponga lo mejor posible al histograma que corresponde a $n = 1000$.