Examen General de Estadística

Semestre 2022-1

Enero, 2021.

09:00-14:00 hrs

Instrucciones: Deberá responder todas las preguntas del examen justificando sus respuestas. Se requieren 4/6 preguntas para aprobar el examen. Tiempo máximo de examen: 5 horas.

1. Sea $X_1, X_2, \ldots, X_n$ una m.a. de $F_X(x)$ y sea $\widehat{F}_n(x)$ la función de distribución empírica, i.e.

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x).$$

Demuestra que

1.1. $\mathbb{E}(\widehat{F}_n(x)) = F_X(x)$.

1.2. $\mathrm{Var}(\widehat{F}_n(x)) = \dfrac{F_X(x)(1-F_X(x))}{n}$.

1.3. $\widehat{F}_n(x) \xrightarrow{P} F_X(x)$.

1.4. El estimador plug-in de un funcional estadístico lineal

$$T(F_X) = \int g(x)\mathrm{d}F_X(x),$$

esta dado por

$$T(\widehat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

[Valor: 4 puntos]

2. Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución $N(\mu, 1)$. Define $Y_i = \mathbb{1}(X_i > 0)$ y sea $\eta = P(Y_i = 1)$.

2.1. Encuentra el estimador máximo verosímil $\widehat{\eta}_n$ de η y demuestra que es un estimador consistente.

2.2. Obtén un intervalo asintótico del 95 % de confianza para η.

2.3. Si los datos no son normales, ¿el estimador $\widehat{\eta}_n$ es consistente? En caso de no serlo, ¿a qué converge en probabilidad?

[Valor: 4 puntos. 2.1 - 1 punto, 2.2. - 2 puntos, 2.3 - 1 punto]

3. Sea $X_1, X_2, \ldots, X_n$ una m.a. de una Bernoulli(p). Obtén un estimador insesgado de p^4 usando el Teorema de Rao-Blackwell.

[Valor: 3 puntos]

1. *Fundamentos*

(a) ¿Qué caracteriza el enfoque bayesiano de la inferencia estadística?

(b) Menciona tres diferencias entre los enfoques frecuentista y bayesiano de la estadística.

(c) Supongamos que interesa inferir la proporción θ de individuos en una población determinada que padecen de cierta enfermedad.

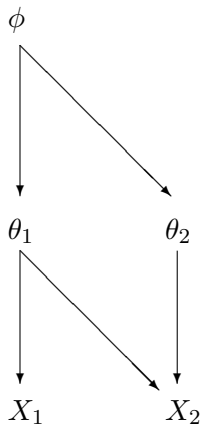
Para simplificar el problema, un grupo de epidemiólogas determinó inicialmente que esta prevalencia puede ser baja ($C_1 : \theta = 0{,}10$), media ($C_2 : \theta = 0{,}50$) o alta ($C_3 : \theta = 0{,}75$), y que $\Pr(C_1) = 0{,}20$, $\Pr(C_2) = 0{,}30$ y $\Pr(C_3) = 0{,}50$.

Posteriormente, las epidemiólogas tuvieron acceso a una muestra aleatoria, conformada por 100 individuos de esa población, en la cual observaron que 27 de ellos presentaban la enfermedad.

Tomando en cuenta esta información, ¿cuál es la probabilidad ahora de que la prevalencia sea baja? Calcula también la nueva probabilidad de que la prevalencia sea media. ¿Qué se puede decir sobre el valor de θ?

2. *Inferencia*

Considera el modelo jerárquico con la estructura descrita por la siguiente figura



y donde

$$\phi \sim N(m_0, 1/n_0), \quad (m_0 \in \mathbb{R}, n_0 > 0, \text{ conocidas});$$

condicional en ϕ, θ_1 y θ_2 son i.i.d. con

$$\theta_i \sim N(\phi, 1), \qquad i = 1, 2;$$

y, condicional en (θ_1, θ_2), X_1 y X_2 son independientes con

$$\begin{aligned} X_1 &\sim N\left(\frac{n_0 m_0 + \theta_1}{n_0 + 1}, \frac{1}{n_0 + 1}\right) \\ X_2 &\sim N\left(\frac{n_0 m_0 + \theta_1 + \theta_2}{n_0 + 2}, \frac{1}{n_0 + 2}\right). \end{aligned}$$

(a) Demuestra que la distribución marginal de cada una de las variables X_1 y X_2 (esto es, no condicional en θ_1, θ_2 o ϕ) es la misma que la de ϕ; es decir, $N(m_0, 1/n_0)$.

(b) Demuestra que la correspondiente correlación marginal entre X_1 y X_2 está dada por

$$\text{Corr}(X_1, X_2) = \frac{n_0 + 2}{(n_0 + 1)(n_0 + 2)}.$$

Hint para el inciso (a): Recuerda que la familia de distribuciones normales es conjugada para el modelo normal cuando la varianza de las observaciones es conocida, y toma en cuenta que $p(\mu) = \int p(\mu|y)p(y)\,dy$, con $p(y) = \int p(y|\tilde{\mu})p(\tilde{\mu})\,d\tilde{\mu}$.

3. *Teoría de decisiones*

El problema de "clasificación supervisada" es un problema de decisión estadístico. Supongamos que tenemos una muestra de n observaciones $(X_1, G_1), (X_2, G_2), \ldots, (X_n, G_n)$, donde cada $G_i \in \{1, 2, \ldots, K\}$ indica que la observación $X_i \in \mathbb{R}_+$ pertenece a uno de K grupos distintos. Supongamos ahora que

$$p(x|G = g) = \text{Gamma}(x|\alpha_g, \beta_g) \qquad (g = 1, 2, \ldots, K)$$

y que

$$\Pr(G = g) = 1/K,$$

donde $\alpha_g > 0$ y $\beta_g > 0$ son parámetros conocidos.

El problema de clasificación consiste en lo siguiente: dada una nueva observación $X_* = x_*$, determinar de cuál de los K grupos proviene. Desde el punto de vista estadístico, el problema es entonces predecir el valor de G_* asociado con X_*.

(a) Plantea y resuelve este problema como un problema de decisión, usando la función de pérdida

$$L(g_*, g) = 0 \qquad (\text{si } g_* = g)$$
$$L(g_*, g) = 1 \qquad (\text{si } g_* \neq g).$$

(b) ¿Cómo se resolvería este problema si los valores de los parámetros $\beta_1, \beta_2, \ldots, \beta_K$ fueran desconocidos?