Examen General de Estadística

Semestre 2019-1

Lunes 21 de enero de 2019

De 9:00 a 14:00 hrs

Instrucciones: Las tres primeras preguntas, correspondientes a Inferencia Estadística, son obligatorias. De las seis preguntas restantes, correspondientes a Inferencia Bayesiana y Modelos Lineales, responde solamente tres. Es decir, solamente debes resolver seis problemas: los tres primeros y otros tres a ser seleccionados de entre los últimos seis. Todos los problemas tienen el mismo valor.

Tiempo máximo de examen: 5 horas.

Inferencia Estadística

1. Sea T una variable aleatoria que toma valores en $\{0, 1, ..., 6\}$ y sean $g_1(t)$ y $g_2(t)$ dos densidades que corresponden a:

   - La primera, $g_1(t)$, es la probabilidad de que, de 5 monedas lanzadas al aire, el número de águilas observadas, T, sea igual a t.
   - La segunda, $g_2(t)$, es la probabilidad de que, al lanzar un dado numerado del 1 al 6, T, el número en la cara que quedó hacia arriba, sea igual a t.

   Como una convención denotaremos por $f(t;\theta_1)$ a $g_1(t)$ y, de manera similar, por $f(t;\theta_2)$ a $g_2(t)$. Así, sin ni siquiera especificar la naturaleza de θ_1 y de θ_2 (pueden ser números, letras o lo que quieras), el "espacio paramétrico" será $\Theta = \{\theta_1, \theta_2\}$, con sólo dos elementos.

   a) Supón que se observó $t = 1$. Exhibe la *estimación máximo-verosímil* de $\theta \in \Theta$.

   b) Repite el inciso anterior suponiendo ahora que se observó $t = 2$.

   c) Observa (esto es raro en planteamientos estadísticos) que hay dos posibles valores de t para los cuales, al hacer la estimación máximo-verosímil de θ, parece que podemos tener certeza de que la estimación se convierte en <u>identificación</u>. Di cuáles son esos dos valores y explica por qué hay identificación.

2. Para probar (contrastar) hipótesis en el caso paramétrico, un procedimiento muy utilizado (óptimo en ciertas circunstancias) es el del "cociente de verosimilitudes generalizadas" que consiste en obtener el cociente entre la máxima verosimilitud bajo H_0 (la hipótesis nula) y la máxima verosimilitud no restringida (o sea bajo $H_0 \cup H_1$), y se rechaza cuando el valor de ese cociente es pequeño.

   Considera el siguiente problema de "prueba de hipótesis":

   Para $x_1, x_2, \cdots, x_n$ una realización de variables independientes e idénticamente distribuidas según la función de distribución F, de la cual lo único que se sabe es que es discreta sobre el conjunto $\{0, 1, 2, \ldots\}$, se desea probar H_0: $F(\cdot) = F_0(\cdot; \theta)$. Aquí, la forma de F_0 es totalmente conocida (pero el parámetro θ es desconocido) y corresponde a una distribución que tiene probabilidades $p_{0,\theta}, p_{1,\theta}, \ldots$ en el soporte, con (obviamente) $p_{j,\theta} \geq 0$ para toda j y $\sum_{j \geq 0} p_{j,\theta} = 1$.

   Si Λ^* denota al cociente de verosimilitudes generalizadas, observa que el numerador es simplemente la verosimilitud generalizada bajo la hipótesis nula, que es $\prod_{j \in J_n} (p_{j,\hat{\theta}})^{N_j}$ donde: N_j es el número de observaciones (estrictamente positivas), J_n denota al conjunto de los índices observados (con N_j mayor que 0, o sea el número de x_i's iguales a j) y $\hat{\theta}$ es el estimador máximo-verosímil.

   a) Muestra que el denominador de Λ^*, resultante de maximizar sobre $H_0 \cup H_1$ es:
   $$\prod_{j \in J_n} (N_j/n)^{N_j},$$
   con lo cual Λ^* queda bien definido.

   Existe un concepto interesante en probabilidad y teoría de información que es la Divergencia Logarítmica de Kullback-Liebler, la cual mide "qué tan lejos" está la densidad discreta dada por unas $\{p_j\}$'s desde el punto de vista de la densidad dada por unas $\{q_j\}$'s (digamos). Está definida por
   $$D(p_j; q_j) = \sum_j q_j \log(q_j/p_j).$$

Se puede demostrar que es no-negativa y es cero si y sólo si las dos densidades son idénticas. Para estar bien definida, requiere que si $q_j>0$ entonces $p_j>0$ (o sea, el soporte de q está contenido en el soporte de p). Si éste no es el caso, la divergencia logarítmica no está definida.

Si para el problema de bondad de ajuste discreto se toman las q_j como las frecuencias empíricas $\frac{N_j}{n}$ y las p_j como las $p_{j,\hat{\theta}}$, sería absurdo tener una $\frac{N_j}{n}$ positiva con una $p_{j,\hat{\theta}}$ igual a cero (equivale a haber observado un j que bajo la H_0 tenía probabilidad estimada de cero; o lo que es lo mismo, eso llevaría automáticamente al rechazo de H_0).

b) Muestra que $-\frac{1}{n}\log(\Lambda^*)$ coincide con $D(p_{j,\hat{\theta}}; N_j/n)$ y que está bien definida (por la observación anterior). Nota que rechazar H_0 si Λ^* es pequeña equivale ahora a rechazar H_0 si la divergencia logarítmica mencionada es grande.

3. Considera dos muestras independientes (e independientes entre sí); la primera de tamaño n de una distribución normal $n(\mu_1, {\sigma_1}^2)$, y la segunda de tamaño m de una $n(\mu_2, {\sigma_2}^2)$. El problema clásico de comparación de medias es el de probar $H_0 : \mu_1 = \mu_2$ sin suponer que las varianzas son iguales; es el problema denominado "de Behrens-Fisher". Supón que la hipótesis nula es cierta; esto es, $\mu_1 = \mu_2 (= \mu)$ con μ desconocida.

   a) Describe la verosimilitud bajo H_0 e identifica una estadística suficiente. Demuestra que la estadística (de dimensión cuatro) formada por las dos medias muestrales y las dos varianzas muestrales, es suficiente minimal. Observa cómo contrasta esto con el hecho de que la dimensión del parámetro bajo H_0 es tres (una μ y dos ${\sigma^2}$'s).

   b) Muestra que la estadística suficiente minimal no es completa. Para ello exhibe una función de la estadística suficiente minimal que tenga esperanza cero siempre, pero que no sea la función nula.

Inferencia Bayesiana

1. En una auditoría, se están analizando los expedientes de n partidas presupuestarias distintas en busca de evidencia de fraude. Supongamos que el expediente que contiene la evidencia pertenece a la i-ésima partida con probabilidad π_i $(i = 1, \ldots, n)$. Supongamos también que, si el expediente pertenece a la partida i, entonces una extracción al azar entre todos los expedientes de esa partida descubrirá la evidencia con probabilidad ϕ_i.

   a) ¿Cuál es la probabilidad de **no** descubrir la evidencia en la i-ésima partida?

   b) Supongamos que una extracción al azar entre los expedientes de la i-ésima partida **no** descubrió la evidencia. Dado esto, ¿cuál es la probabilidad de que la evidencia esté en la i-ésima partida de todas formas?

   c) Demuestra que la probabilidad de que el expediente que contiene la evidencia pertenezca a la partida j (con $j \neq i$), dado que una extracción al azar entre los expedientes de la i-ésima partida **no** la descubrió, es
   $$\frac{\pi_j}{1 - \phi_i \pi_i} \qquad (j \neq i).$$

2. *Intercambiabilidad.*

   a) Sean $\{X_n : n \in \mathbb{N}\}$ variables aleatorias (infinitamente) intercambiables. Sea $m \in \mathbb{N}$ (con $m > 0$) y define $Y_n = X_{m+n}$ $\forall n \in \mathbb{N}$. Demuestra que, condicional en $X_1 = x_1, \ldots, X_m = x_m$, se tiene que $\{Y_n : n \in \mathbb{N}\}$ también son (infinitamente) intercambiables.

   b) Supón que $E(X_n^2) < \infty$ $\forall n \in \mathbb{N}$. Demuestra que
   $$\mathrm{Cov}(Y_i, Y_j \,|\, x_1, \ldots, x_m) \geq 0 \qquad (i, j \in \mathbb{N}).$$

3. Sea $X_1, X_2, \ldots, X_m$ una muestra de variables aleatorias i.i.d. de una distribución normal $N(x|\theta, 1)$ y supón que la distribución inicial de θ está dada por $p(\theta) = N(\theta|\mu, 1/\tau)$, donde τ denota al parámetro de precisión.

a) Encuentra el estimador bayesiano de θ utilizando la función de pérdida

$$L(a, \theta) = \left(a - \theta^2\right)^2.$$

b) Encuentra el estimador bayesiano de θ utilizando la función de pérdida

$$L(a, \theta) = \left(e^a - e^\theta\right)^2.$$

c) Habiendo resuelto los dos incisos anteriores, ¿cuál de las dos soluciones tiene más sentido para ti? Argumenta tu respuesta.

Modelos Lineales

1. Considera que las variables aleatorias Y, X_1 y X_2 tienen una distribución conjunta normal con medias μ_y, μ_1, μ_2 y una determinada matriz de varianzas-covarianzas (definida positiva) que puedes denotar como gustes.

   Con la información que ya tienes, describe lo mejor que te sea posible qué se quiere decir cuando se habla de los siguientes tres coeficientes de correlación:

   a) Coeficiente de correlación simple entre Y y (digamos) X_1; denótalo por $\rho(Y; X_1)$.

   b) Coeficiente de correlación parcial entre Y y X_1 habiendo condicionado en X_2, denotado usualmente por $\rho(Y; X_1|X_2)$.

   c) Coeficiente de correlación múltiple entre Y y el par (X_1, X_2), denotado por $\rho(Y; (X_1, X_2))$.

   Nota: Para b) necesitarás describir la distribución de (Y, X_1) dado que $X_2 = x_2$, la cual es una normal bivariada con un vector de medias que resulta depender linealmente de x_2 y con una cierta matriz de varianzas-covarianzas que estructuralmente depende de haber condicionado con X_2, pero que no depende del valor x_2.

2. En un modelo de regresión lineal normal

$$y_i = \underline{x}_i'\beta + e_i, \quad e_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n,$$

donde $\underline{x}_i^{'}$ es de la forma $(1, \underline{z}_i^{'})$ con $\underline{z}_i \in R^{p-1}$, se obtiene el ajuste correspondiente. Es decir, la matriz X es la matriz de $n \times p$ $(n > p)$ cuyas filas son las $\underline{x}_i^{'} = (1, \underline{z}_i^{'})$, y a partir de ella se obtienen $\hat{\beta}, \hat{\sigma}^2$ como cualquier solución a las ecuaciones normales, dando entonces un "modelo ajustado"

$$\hat{y} = \underline{x}'\hat{\beta},$$

en el que debemos tener cuidado de que la $\underline{x}$ en que se evalúa el modelo ajustado sea combinación lineal de las filas de X (pues, en caso contrario, no tenemos estimabilidad y mucho menos un estimador). Es obvio que podemos evaluarla, por ejemplo, en las $\underline{x}_i$'s y obtener las correspondientes $\hat{y}_i$'s. Define el residuo (pelón) como $\hat{e}_i = y_i - \hat{y}_i$ (podía haberse definido con los signos cambiados, pero eso es irrelevante).

Demuestra que en este modelo en particular (en el que la matriz X tiene como primera columna una columna de 1's) es cierto que

$$\sum_{i=1}^{n} \hat{e}_i = 0.$$

Sugerencia: Piensa en el vector residuo $\underline{\hat{e}}$ como la diferencia entre $\underline{y}$, el vector original de las observaciones, y el vector $\underline{\hat{y}}$, y pregúntate en qué subespacios están estos vectores.

Para ver que la suma de los residuos cero pudiera ocurrir con una matriz X que no contenga en su espacio columna a una columna de 1's, exhibe un ejemplo sencillo (mientras más sencillo mejor) en el que la suma de los residuos es cero pero X no contiene una columna de 1's en su espacio columna. Intenta un modelo lineal simple pero que pase por el origen (i.e., su α es cero).

3. Considera un modelo de regresión lineal normal

$$y_i = \underline{x}_i'\beta + e_i, \quad e_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n,$$

con $\underline{x}_i^{'} \in R^p$, conocidas. Para mayor facilidad, supón que X es de rango máximo.

Imagina que, en el futuro, se hará una observación Y_* (independiente de las n observaciones iniciales) dentro del mismo modelo, esto es:

$$Y_* = \underline{x}_*^{'}\beta + e_*,$$

y se desea obtener:

a) Un intervalo de confianza al 95 % para la media de la nueva observación Y_*. Exhibe la expresión para el intervalo.

b) Un intervalo de predicción para Y_* de contenido probabilístico 0.95. Exhibe la expresión.

c) Un intervalo que contenga por lo menos al $p\,\%$ de todas las observaciones que se hagan de la distribución $n(\underline{x}_*'\beta, \sigma^2)$, con una confianza del 95 %. Este tipo de intervalos se llaman 'intervalos de tolerancia'.