Examen General de Estadística
Semestre 2018-1
Jueves 11 de enero de 2018.
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 $X_1, \cdots, X_n$ una muestra aleatoria de una distribución uniforme en el intervalo $(0, \theta)$; esto es, con función de densidad

$$f(x;\theta) = 1/\theta, \quad 0 < x < \theta, \;\; \theta > 0.$$

   (a) Encuentra el estimador máximo-verosímil $\hat{\theta}$ e identifica su distribución; en particular exhibe su media y su varianza. Observa que lo "típico" en términos de resultados asintóticos para estimadores de máxima verosimilitud:

$$\sqrt{n}\{\hat{\theta} - \theta\} \rightarrow \psi$$

   cuando $n \to \infty$ (siendo ψ una variable aleatoria normal con varianza positiva) **no ocurre en este caso**. A este estimador $\hat{\theta}$ se le dice <u>super-eficiente</u>.

   (b) Corrige el estimador máximo-verosímil para hacerlo insesgado. Es decir, encuentra C para que $C\hat{\theta}$ sea insesgado.

   (c) Observa que $2\bar{X}$ es también un estimador insesgado para θ. Compara su varianza con la de $C\hat{\theta}$.

2. Considera una distribución exponencial negativa; esto es, con función de densidad

$$f(x;\theta) = \theta e^{-x\theta}, \quad x > 0, \quad \theta > 0.$$

Su función de distribución está dada por $F(x;\theta) = 1 - e^{-x\theta}$ para $x > 0$ y es cero si $x \leq 0$.

(a) Demuestra que si se quiere obtener la distribución exponencial truncada por arriba en 1 (también referida como condicionada, en este caso a $x \in (0,1)$), ésta tiene la función de distribución

$$F_c(x;\theta) = \frac{1 - e^{-x\theta}}{1 - e^{-\theta}},$$

obviamente para valores de x en (0,1), siendo cero si $x \leq 0$ y uno si $x \geq 1$. Esta distribución tiene una densidad (es continua) y además tiene la propiedad de que como función de distribución paramétrica está bien definida aún si $\theta < 0$.

(b) Demuestra que si $\theta \to 0$, la distribución converge a la $U(0,1)$. Por ello, como familia paramétrica para modelar distribuciones en el intervalo (0,1), resulta interesante pues las densidades son decrecientes si $\theta > 0$, crecientes si $\theta < 0$ y reconstruye la uniforme si $\theta = 0$.

(c) Considera $X_1, \cdots, X_n$ una muestra independiente de la distribución exponencial truncada y demuestra que $t(X) = \sum X_i$ es suficiente para θ; de hecho, demuestra que la distribución es miembro de la familia exponencial (con dimensión 1).

3. Considera el problema de comparación de medias de dos poblaciones normales con varianzas diferentes y desconocidas, donde la hipótesis nula a ser contrastada es que las medias son iguales (sin decir cuál es su valor común). Bajo esta hipótesis, se tienen entonces tres parámetros (la media común, y las dos varianzas). Sin embargo, la verosimilitud usando dos muestras de tamaños n y m arroja una estadística suficiente de dimensión cuatro, algo considerado anómalo en inferencia. Éste es el famoso problema de Behrens-Fisher.

Existe un planteamiento muy sencillo que lleva a otro problema anómalo. Considera una muestra de observaciones independientes e idénticamente distribuidas de una distribución $n(\theta, \theta^2)$, con θ un número real. Al describir la función de verosimilitud, aparece de manera inmediata que

$$T = (\sum X_i, \sum {X_i}^2)$$

es suficiente. El espacio paramétrico original es de dimensión 1 pero la dimensión de T es 2.

(a) Utiliza el resultado que dice que para dos muestras diferentes (en el soporte de la distribución), denotadas $\underline{x}$ y $\underline{y}$, si hacemos que las dos muestras sean equivalentes en verosimilitud (o sea, que den esencialmente la misma verosimilitud) y se sigue que $T(\underline{x}) = T(\underline{y})$, entonces T es suficiente minimal. Verifica que es el caso.

(b) Exhibe una función de T que tenga esperanza cero (para todo θ) pero que no sea cero, mostrando así que T no es completa. Este contexto es también uno patógico desde el punto de vista de la inferencia estadística.

Inferencia Bayesiana

1. Supongamos que tienes tu propia compañía de transporte, con un gran número de camiones, y que los camiones se descomponen de manera aleatoria a lo largo del tiempo. El número de descomposturas durante un intervalo de tiempo t sigue una distribución Poisson con media λt, donde el parámetro λ es la tasa diaria de descomposturas. Los valores posibles de λ son $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ y sus respectivas probabilidades iniciales son $\{0.10, 0.20, 0.30, 0.20, 0.15, 0.05\}$.

   (a) Encuentra la distribución final de λ si se descomponen 12 camiones en un periodo de seis días.

   (b) Calcula la probabilidad de que no se descomponga algún camión durante los siguientes seis días.

2. Supón que $\{X_n : n \in I\!N\}$ es una sucesión de variables aleatorias infinitamente intercambiables. Sea $m \in I\!N$ (con $m > 0$) y define $Y_n = X_{m+n}\ \forall\, n \in I\!N$. Demuestra que, condicional en $X_1 = x_1, \ldots, X_m = x_m$,

$\{Y_n : n \in I\!N\}$ también es una sucesión de variables aleatorias (infinitamente) intercambiables.

3. La paradoja de San Petersburgo resulta del siguiente juego: Una moneda honesta es lanzada tantas veces como sea necesario hasta que aparece un "águila". El jugador recibe entonces una recompensa de 2^n pesos, donde n es el número de lanzamientos que se requirieron para obtener la primera "águila". El valor esperado del juego es

$$\sum_{n=1}^{\infty} 2^n \Pr[n \text{ lanzamientos}] = \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} = \infty.$$

A pesar de esto, poca gente está dispuesta a pagar mucho dinero por jugar este juego, pues la probabilidad de ganar mucho es muy pequeña.

   (a) Sea c el costo (en pesos) por jugar el juego y sea $u(\cdot)$ una función que describe la utilidad del dinero. Encuentra la utilidad esperada de jugar el juego.

   (b) Sea

$$u(x) = \begin{cases} 100 & \text{si } x > 100, \\ x & \text{si } |x| \leq 100, \\ -100 & \text{si } x < -100. \end{cases}$$

   ¿Cuál es el valor máximo de c que estarías dispuesto a pagar por jugar este juego? Argumenta tu respuesta.

Modelos Lineales

1. Sea un modelo lineal simple

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \cdots, n,$$

donde los e_i son errores aleatorios no-correlacionados (de media cero y varianza constante σ^2) y α y β son parámetros desconcidos. Los números $x_1, \cdots, x_n$ son constantes conocidas.

El modelo anterior se "ajusta" obteniendo los estimadores de mínimos cuadrados para α y β, y en consecuencia para σ^2, como es bien conocido.

(a) Exhibe las expresiones para los estimadores $\hat{\alpha}$ y $\hat{\beta}$ y, suponiendo en adición normalidad, da la expresión para un intervalo de confianza para la combinación lineal $a\alpha + b\beta$ donde las constantes a y b son conocidas, de modo que la confianza sea del $100(1-\gamma)\,\%$. Para ello necesitas mencionar el estimador máximo-verosímil $\hat{\sigma}^2$ y referir a ciertos cuantiles (di cuáles y de qué distribución).

(b) Imagina una observación futura (denotada por Y^* e independiente de las previas) dentro del mismo modelo; esto es, con una distribución normal de media $\alpha + \beta x^*$ y varianza σ^2. Construye un intervalo (l, u), de probabilidad (ya no de confianza) tal que $P[Y^* \in (l, u)] = 1 - \gamma$. Indica con claridad qué cuantiles, y de qué distribución, son requeridos.

(c) Observa que a la media de la distribución de Y^* se le puede construir un intervalo de confianza (como en el inciso 1). Para la misma γ, este intervalo de confianza ¿contiene o es contenido por el intervalo de probabilidad obtenido en el inciso (b)?

2. Considera una clasificación cruzada de dos criterios sin interacción, o sea:

$$y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk},$$

con i =1,2, j =1,2,3 y $k = 1, \cdots, n_{ij}$, donde las n_{ij} son sencillitas, todas iguales a 1 excepto $n_{22} = 0$. En otras palabras, se tienen 5 observaciones nada más.

(a) De las tres cantidades paramétricas siguientes, di si son o no son estimables en el sentido de Gauss-Markov: $\tau_1 - \tau_2$, $\beta_1 - \beta_2$ y $\beta_2 - \beta_3$.

(b) Responde nuevamente el inciso anterior suponiendo que $n_{12} = 0$; es decir, ahora sólo se tienen 4 observaciones.

(c) Si en lugar de tener, como en (b), que $n_{12} = 0$, ahora suponemos que $n_{12} = 1$ y que $n_{13} = 0$, otra vez se tienen sólo 4 observaciones. ¿Qué respondes respecto a la estimabilidad de las tres cantidades?

En los tres casos argumenta tu respuesta.

3. Sea un modelo de regresión

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

donde las $\underline{x}_i^{'} \in R^p$ son conocidas.

Si la matriz X ($n \times p$, con $n > p$) cuyas filas son las $\underline{x}_i^{'}$ es de rango completo, entonces $\hat{\beta}$ es única y $\hat{\sigma^2}$ está bien definida también (son los estimadores máximo-verosímiles).

Si el rango de la matriz X no es máximo, entonces

(a) ¿Se pueden "estimar" todas y cada una de las p componentes del vector β?

(b) ¿Serán estimables (en el sentido de Gauss-Markov) las componentes del vector $X\beta$?

(c) La suma de cuadrados del error, definida como $||Y - X\hat{\beta}||^2$ (siendo $\hat{\beta}$ cualquier solución a las ecuaciones normales) ¿será única?

Argumenta tu respuesta en los tres incisos.