Estadística No Paramétrica

Clase 13: Jackknife

Joaquin Cavieres G.

Ingeniería en Estadística

Facultad de Ciencias, Universidad de Valparaíso

Introducción

Introducción

Literatura

- Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Plans, SIAM.
- Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. Monographs on statistics and applied probability, 57, 1-436.
- Manly, B. F. (2006). Randomization, bootstrap and Monte Carlo methods in biology (Vol. 70). CRC press.

Adicionales

- Shao, J. & Tu, T. (1995) The Jackknife and Bootstrap, Springer-Verlag.
- Efron, B. & Tibshirani, R.J. (1993) An Introduction to the Bootstrap, Chapman & Hall.

Resumen de conceptos vistos

Un parámetro de interés, llamemosle θ, es una función de una función de probabilidad F (p.d.f) tal que:

$$\theta = s(F)$$

con media

$$\theta = \mathbb{E}_F(x) = \int xF(x)dx = \mu_F$$

y varianza

$$\theta = \mathbb{E}_F[(x-\mu_F)^2] = \int (x-\mu_F)^2 F(dx)dx = \sigma_F^2$$

Estadísticos: Media

Un estadístico $\hat{\theta}$ es una función de la muestra o de la función de distribución $\hat{F}$:

$$\hat{\theta} = s(\hat{F})$$

(o en la forma notacional como los vimos en la clase anterior (Bootstrap)):

$$\hat{\theta} = f(E_n),$$

por lo tanto si consideramos a $\hat{\theta}$ como la media, entonces:

$$\begin{aligned}\hat{\theta} = s(\hat{F}) &= \int xs\hat{F}(x)dx \\ &= \int x1/n\sum_{i=1}^{n}\eta(x-x_i)dx = 1/n\sum_{i=1}^{n}x_i \\ &= f(E_n) = \bar{x}\end{aligned}$$

Estadístico: Varianza

Si $\hat{\theta}$ es la varianza, entonces:

$$\hat{\theta} = \int (x - \bar{)}^2 \hat{F}(x) dx$$

$$1/n \sum_{i=1}^{n} (x_i - \bar{x}))^2$$

$$= \sigma^2$$

Plug-in estimador

El plug-in estimador de un parámetro, con $\theta = s(F)$, es:

$$\hat{\theta} = s(\hat{F}),$$

en donde $\theta = s(F)$ es el estimador de la función de densidad de probabilidad de F a través de la funcion $s(\cdot)$ de la distribución empírica $\hat{F}$. Así:

- $\bar{x}$ es el plug-in estimador de μ_F
- $\hat{\sigma}^2$ es el plug-in estimador de σ_F^2

Precisión en las estimaciones

¿Que tan confiables son las estimaciones sobre $\hat{\theta}$?

Precisión en las estimaciones

¿Que tan confiables son las estimaciones sobre $\hat{\theta}$?

La idea principal es centrarse en la distribución de probabilidad de $\hat{\theta}$

Precisión en las estimaciones

¿Que tan confiables son las estimaciones sobre $\hat{\theta}$?

La idea principal es centrarse en la distribución de probabilidad de $\hat{\theta}$

Cantidades de interés adicionales

- Error estándar
- Intervalos de confianza
- Sesgo (*bias*)

Error estándar de $\hat{\theta}$

Mide la precisión de un estimador en la distribución (función) de una población.

$$se(\hat{\theta}) = \sqrt{\mathbb{V}_F(\hat{\theta})}$$

Ejemplo: error estándar para $\bar{x}$:

$$se_F(\bar{x}) = \mathbb{V}_F(\bar{x})^{1/2} = \sigma_F/\sqrt{n}$$

Ejemplo: Si tenemos una muestra aleatoria $X_1, ..., X_n$, usando el plug-in estimador podemos determinar el error estándar como:

$$\hat{se}(\hat{\theta}) = \hat{se}_{\hat{F}}(\hat{\theta}) = \mathbb{V}_{\hat{F}}(\hat{\theta})^{1/2}$$

así, el error estándar de $\bar{x}$ es:

$$\hat{se}(\bar{x}) = \hat{\sigma}/\sqrt{n}$$

Ver ejemplo en R

Método de Jackknife

Método de Jackknife

Uno de los problemas que pueden generars en terminos teóricos con el Bootstrap es que las muestras son generadas desde $\hat{F}$ y no de F, pero es posible muestrear o remuestrar exactamente desde F?

Mirar diferentes subconjuntos de nuestra muestra original equivale a muestrear sin reemplazo las observaciones $X_1, ..., X_n$ para obtener remuestras (ahora llamadas submuestras) de tamaño m. Esto nos lleva al concepto de submuestreo y al método de Jackknife.

Método de Jackknife

Definición

El método de Jackknife cálcula muestras dejando fuera una observación X_i desde $X_1, ..., X_n$

$$\boldsymbol{X}_i = (X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n),$$

donde:

- La dinensión de la muestra Jackknife $(\boldsymbol{X}_i)$ es $m = n - 1$.
- Existen n diferentes muestras Jackknife: $\{\boldsymbol{X}_i\}_{i=1...n}$
- No es necesario un método para calcular las n muestras Jackknife.

Método de Jackknife

Para la replicación i-ésima $\hat{\theta}_i$ del estadístico $\hat{\theta} = s(\boldsymbol{X})$:

$$\hat{\theta}_i = s(\boldsymbol{X}_i), \;\; \forall \;\; i = 1, ..., n$$

Mediante Jackknife de la media sería:

$$\begin{aligned} s(\boldsymbol{X}_i) &= \frac{1}{n-1} \sum_{j \neq i} x_j \\ &= \frac{(n\bar{x} - x_i)}{n-1} \\ &= \bar{x}_i \end{aligned}$$

Jackknife para el error estándar

- Calcular n submuestras mediante Jackknife $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ desde $\boldsymbol{X}$
- Evaluar n replicaciones de Jackknife mediante $\hat{\theta}_i = s(\boldsymbol{X}_i)$
- Calcular el error estándar:

$$\hat{se}_{\text{jack}} = \left(\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{.} - \hat{\theta}_i)^2\right)^{1/2}$$

donde $\hat{\theta}_{.} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$

Jackknife para el error estándar

Observaciones

- El termino $(n-1)/n$ es mucho más grande que el $1/B-1$ del Bootstrap.
- Este factor de necesario ya que la desviación en Jackknife $(\hat{\theta}_i - \hat{\theta}_{\cdot})^2$ tiende a ser más pequeña que el del Bootstrap $(\hat{\theta}^*_b - \hat{\theta}^*_{\cdot})^2$

Jackknife para el sesgo (*Bias*)

- Calcular n submuestras mediante Jackknife $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ desde $\boldsymbol{X}$
- Evaluar n replicaciones de Jackknife mediante $\hat{\theta}_i = s(\boldsymbol{X}_i)$
- Calcular el sesgo:

$$\hat{\text{Bias}}_{\text{jack}} = (n-1)(\hat{\theta}_{\cdot} - \hat{\theta})$$

donde $\hat{\theta}_{\cdot} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i$

Jackknife para el sesgo (*Bias*)

Observaciones

- El termino $(n-1)$ se agrega a la ecuació si comparamos con el Bootstrap estimación del sesgo.
- $\hat{\theta} = \bar{x}$ es insesgado, por tanto la varianza es $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$

Jackknife para el sesgo (*Bias*) corregido

Pseudovalores

Si queremos estimar un parámetro θ con el estimador $\hat{\theta}$, entonces:

$$PV(\boldsymbol{X}_i) = n\hat{\theta} - (n-1)\hat{\theta}_i,$$

donde $PV(\boldsymbol{X}_i)$ es llamado el i-ésimo pseudovalor

Se espera entonces que $PV(\boldsymbol{X}_i) \approx n\theta - (n-1)\theta = \theta$, así cada pseudovalor puede ser visto como un estimador de θ.

Media de los pseudovalores

$$\hat{\theta}_{\text{jack}} = \hat{\theta} - \widehat{\text{Bias}}_{\text{jack}} = n\hat{\theta} - (n-1)\hat{\theta}.$$

donde $\hat{\theta}_{\text{jack}} = \overline{PV}$, que a su vez es el bias corregido del estimador Jackknife

Bootstrap vs Jackknife

Relación

- Cuando el n es pequeño entonces es recomendable utilizar Jackknife
- El Jackknife es una aproximación del Boostrap

Limitaciones del método Jackknife

- Jackknife puede fallar si $\hat{\theta}$ no es una función suave (*smooth*). Por ejemplo, un pequeño cambio en los datos puede producir cambios en el estadístico.
- Un ejemplo de un estadístico que no es smooth es la mediana.
- Jackknife no es un buen método de estimación para estimar percentiles.
- Lo anterior no ocurre con el método de Bootstrap.