El Bootstrap y sus aplicaciones

Edgar Acuna

Departamento de Ciencias Matematicas

UPR-Mayaguez

Motivacion

Talvez la parte más importante de la Estadística es usar los datos de la muestra que se ha recolectado para sacar conclusiones acerca de la población de donde procede la muestra. Este proceso es llamado Inferencia Estadística. Por ejemplo, supongamos que a 8 personas que acaba de tener un ataque cardiaco se le tomado una muestra de su colesterol y se obtuvo los siguientes resultados

233 259 215 322 289 220 276 299

El promedio o media de la muestra resulta ser 264.125. Pero uno está interesado realmente en caracterizar el colesterol del total de personas que tiene ataques cardiacos. Las cantidades que se usan para caracterizar una población son llamados parámetros y se representan por θ. Las cantidades que se calculan usando la muestra tomada y que se espera reflejen el comportamiento del parámetro se llaman estimados o estadísticos y se representan por $\hat{\theta}$. Formalmente, $\hat{\theta}$ =T(X1,……Xn) donde (X1,…..,Xn) representa la muestra aleatoria y T es una función optima que se usa para estimar θ y es llamado el estimador.

Notar que el valor del estimado varía con la muestra tomada. Métodos para hallar la función T son tratados en cursos de Estadística Matemática. Uno de los parámetro que mas interesa es la media poblacional entonces el estimador óptimo a usar es la media muestral $\bar{X}$.

Por ejemplo, $T(\boldsymbol{x}) = \bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} x_i}{n}$,$T(x) = s^2 = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{n-1}$

El proceso clásico de estimación requiere asumir una forma distribucional, generalmente Normal o Gaussiana para la población. Sin embargo, hoy en día la orientación es usar la data también para estimar esta forma distribucional en lugar de asumirla. Las técnicas que hacen esto caen en el área que es llamada Estadística Noparamétrica.

Muestras Aleatorias

Consideremos una población de N unidades $U_1,U_2,\ldots,U_N$, cada una de las cuales es igualmente probable de ser seleccionada en cualquier una extracción al azar que se haga.
Una muestra aleatoria de tamaño n es una colección de n unidades $u_1,u_2,\ldots u_n$ seleccionadas al azar de la población. Básicamente lo que se selecciona son n enteros $j_1,j_2,\ldots j_n$ entre 1 al N con igual probabilidad 1/N, estos enteros definen los índices de las unidades seleccionadas en la muestra. En principio, los enteros $j_1,j_2,\ldots j_n$ se pueden repetir y se dice que la muestra es con reemplazo. Si se desea que sean todos distintos entonces la muestra se dice que es sin reemplazo. Sin embargo, cuando el tamaño de muestra es bien pequeño comparado con la población hay una probabilidad muy baja de que haya elementos repetidos en la muestra. En Bootstrapping las muestras que se usan son con reemplazo.

Sea x_i las mediciones de interés para la unidad u_i en la muestra. Sea $x=(x_1,x_2,\ldots,x_n)$ las mediciones observadas en toda la muestra. Por otro lado, sea $X=(X_1,X_2,\ldots X_N)$ el conjunto de mediciones de todas las unidades de la población o simplemente la población. Entonces, x será llamada una muestra aleatoria de X.

Función de distribución Empírica

La función de distribución (acumulativa) Empírica para una muestra aleatoria $X_1,\ldots X_n$ con función de distribución F_X está dada por

$$\hat{F}_n(t) = \frac{\#\{i,\ \ 1 \le i \le n : X_i \le t\}}{n} = \frac{\#\{X_i \le t\}}{n}$$

Para cada t, $\hat{F}_n(t)$ es un estadístico que da la frecuencia relativa de los valores en la muestra que son menores o iguales que t.

La función de distribución empírica sirve como un estimador no-paramétrico de la función de distribución F_X de una variable aleatoria. La definición puede extenderse al caso multivariado.

El error estandar de la media muestral

Asumamos que la variable aleatoria X tiene una distribución F con valor esperado $\mu_F = E_F(X)$ y con varianza $\sigma_F^2 = VAR_F(X) = E_F[(x-\mu_F)^2]$

Si se toman la muestra aleatoria $X_1, \ldots . X_n$, entonces la media de la muestra

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ tiene media $E(\bar{X}) = \mu_F$ y varianza $VAR(\bar{X}) = \frac{\sigma_F^2}{n}$.

La prueba se basa en la linealidad del valor esperado y en la independencia de las variables aleatorias X_i's. La forma de la distribución de la media muestral es desconocida.

El error estándar de la media muestral representado por $se_F(\bar{X})$, o simplemente $se(\bar{X})$, es la raíz cuadrada de la varianza de $\bar{X}$. Esto es,

$$se_F(\bar{X}) = \sigma_F / \sqrt{n}$$

Usualmente, la varianza poblacional σ^2 no es conocida, y será estimada por la varianza muestral s^2. Así, $\frac{s}{\sqrt{n}}$ es llamado el error estándar estimado, que en muchos lados es simplemente llamado error estándar ya que su valor puede ser siempre cuantificado de la muestra. El error estándar de la media y de un estimado en general da una buena idea de su precisión.

Existen muchos estimadores en estadística, para los cuales es dificil calcular su varianza.
Por ejemplo, la mediana es un parámetro cuya varianza no es fácil de calcular. Asimísmo, existen parámetros para los cuales es difícil calcular intervalos de confianza y hacer pruebas de hipótesis ya que no existe una expresión explicita para la distribución de los estimadores y/o de la prueba estadística.

La idea de Bootstrapping, introducido por Efron en 1979, es tomar muchas muestras con reemplazamiento y del mismo tamaño de la muestra original. De esta manera se genera variabilidad del estimador y una estimación de la distribucion empirica que a su vez es un estimado de la verdadera función de distribución.
Esto permitiría hacer inferencia con estimadores dificiles o sin tener que hacer suposiciones acerca de la población. El Bootstrapping es una de las técnicas de estadística noparamétrica.

Intervalos de Confianza usando Boostrapping-Metodo de los percentiles

La idea aquí es estimar la función de distribución F del estimador $\hat{\theta}$ usando Bootstrapping. Luego, el intervalo de confianza del $100(1-\alpha)$ para θ será simplemente

$$[\hat{F}^{-1}(\alpha/2),\ \hat{F}^{-1}(1-\alpha/2)]$$

donde $\hat{F}^{-1}(\alpha/2)$ representa el percentil del $\alpha/2$ de la distribución del estimador, es decir un valor tal que la probabilidad acumulada hasta dicho valor sea $\alpha/2$, y $\hat{F}^{-1}(1-\alpha/2)$ representa el percentil del $1-\alpha/2$.

El Algoritmo Bootstrap para estimar errores estándar

1 Seleccionar B muestras bootstrap independientes, $\mathbf{x}^{*1},\mathbf{x}^{*2},\ldots,\mathbf{x}^{*B}$ cada una consistente de n valores extraídos con reemplazo de la muestra original $\mathbf{x}=(x_1,\ldots.x_n)$.

2 Evaluar el estadístico en cada muestra bootstrap. Esto es

$\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$ para b=1,2,…..B.

3. Estimar el error estándar $se_F(\hat{\theta})$ por la desviación estándar de las repeticiones del estadístico en las B muestras bootstrap. Es decir, por

$$\hat{se}_B(\hat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}^*(b) - \bar{\hat{\theta}}^*)^2}{B-1}}$$

con $\bar{\hat{\theta}}^* = \dfrac{\sum_{b=1}^{B}\hat{\theta}^*(b)}{B}$

Intervalo de Confianza usando Bootstrap estudentizado

La idea aquí es sustituir los percentiles de la distribución t por los percentiles de la distribución de los valores bootstrapeados

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{se(\hat{\theta}^*)}$$

donde $\hat{\theta}^*$ y $se(\hat{\theta}^*)$ son el valor del estimador en cada muestra bootstrapeada y valor estimado del error del estimador en dicha muestra bootstrapeada respectivamente; $\hat{\theta}$ es el estimado en la muestra original. Si se está tratando de estimar la media el error estándar estimado sería $s(\hat{\theta}^*)/n^{1/2}$, pero en el caso de otros estimadores tales como la mediana habría que aplicar otro boostrapping para estimar este error estándar.

Luego, el Intervalo de confianza del $100(1-\alpha)\%$ por bootstrapping estudentizado estará dado por

Intervalo de Confianza usando Bootstrap estudentizado-I

$$(\hat{\theta}+\hat{F}_t^{-1}(\alpha/2)*\ \tilde{se},\ \hat{\theta}+\hat{F}_t^{-1}(1\text{-}\alpha/2)*\tilde{se})$$

donde $\hat{F}_t^{-1}$ $(\alpha/2)$ representa el percentil del $\alpha/2$ de la distribución de t*, es decir un valor tal que la probabilidad acumulada hasta dicho valor sea $\alpha/2$, y $\hat{F}_t^{-1}$ $(1\text{-}\alpha/2)$ representa el percentil del $1\text{-}\alpha/2$. Aqui, $\tilde{se}$ representa el error estándar del estimador de la muestra tomada. Nuevamente, si es la media no es problema, pero si es otro estimador habría que aplicar bootstrapping o Jackknife (a ser visto más adelante) para estimarlo.