Datos Categóricos: Clase 5

Juan Carlos Correa

21 de marzo de 2022

Bootstrap

- Esta metodología fue propuesta por Efron en los años 70.
- Esta técnica hace parte de las técnicas que hacen uso intenso del computador.
- Los métodos bootstrap pueden ser aplicados tanto en casos donde hay modelos probabilísticos bien definidos como donde no los hay.
- La idea es remuestrear los datos originales, bien sea directamente o vía algún modelo ajustado, para crear datos replicados, a partir de los cuales la variavilidad de las cantidades de interés puedan ser determinadas sin tener que recurrir a extensos desarrollos analíticos (que pueden ser muy tediosos o largos o a veces imposibles de hacer)

- El boostrap se puede utilizar para construir intervalos de confianza de muchos parámetros de interés: medias, proporciones, perecentiles centrales tales como la mediana, varianzas, cocientes de parámetros, por ejemplo coeficiente de variación, coeficientes de correlación, etc. El parámetro de interés por θ (puede ser un vector).

- Se pueden construir intervalos bootstrap

  - No Paramétricos: En estos no se asume un modelo probabilístico que genera los datos.

  - Paramétrico: En este caso se asume un modelo probabilístico particular en el cual se desconocen parámetros y debemos estimarlos a partir de la muestra. Es posible construir un intervalo bootstrap para una distribución completemente especificada y este intervalo se usa en pruebas de hipótesis como una región de aceptación.

Intervalos Bootstrap No Paramétrico

- Sea $X_1, X_2, \cdots, X_n$ una m.a. de la población F.
- Idea: Se asume que la distribución empírica F_n representa muy bien la verdadera distribcuión desconocida F.
- Saque muchas, digamos M, muestras de tamaño n de la distribución empírica, esto es equivalente a sacar muestras de tamaño n con reemplazo de la muestra.
  - $X_1^{(1)}, X_2^{(1)}, \cdots, X_n^{(1)}$ y calcule $\hat{\theta}^{(1)}$.
  - $X_1^{(2)}, X_2^{(2)}, \cdots, X_n^{(2)}$ y calcule $\hat{\theta}^{(2)}$.
  - $\cdots$
  -
  - $X_1^{(M)}, X_2^{(M)}, \cdots, X_n^{(M)}$ y calcule $\hat{\theta}^{(M)}$.
- Haga el histograma de $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \cdots, \hat{\theta}^{(M)}$. Esta es la distribución muestral bootstrap del estimador.
- Calcule los percentiles $\hat{\theta}_{(\alpha/2)}$ y $\hat{\theta}_{(1-\alpha/2)}$. Estos corresponden a los límites inferior y superior del intervalo de confianza para θ.

Pruebas de hipótesis bootstrap

Las pruebas bootstrap son una alternativa para las pruebas asintóticas que son utilizadas con frecuencia en el análisis de modelos para tablas de datos.
Davison y MacKinnon (2007) señalan que para un tamaño muestral fijo las pruebas bootstrap son mejores que las asintóticas en el sentido de cometer errores de orden más pequeños.
Si los resultados asintóticos (valores p) son similares a los resultados bootstrap, entonces podemos tener cierta garantía de los resultados asintóticos, pero esto no es cierto si los resultados son muy diferentes.

Descripción de una prueba bootstrap

Sea T un estadístico de prueba (por ejemplo el G^2 que se usa en tablas de contingencia) y sea $\hat{T}$ el valor calculado del estadístico a partir de la muestra de tamaño n.
Asumimos que el estadístico T es asintóticamente pivotal.
Para una prueba que rechaza H_0 cuando $\hat{T}$ está en la cola superior (estilo pruebas χ^2), el verdadero valor-p de $\hat{T}$ es $1 - F(\hat{T})$, donde F es la función de distribución acumulada de T bajo H_0.
Si no conocemos la verdadera F, podemos a menudo estimarla usando el bootstrap. El procedimiento es así:

- Generamos B muestras bootstrap.
- A cada muestra le calculamos un estadístico bootstrap T_j^* para $j = 1, \cdots, B$.
- Calculamos la distribución empírica de estos valores T_j^*, llamada la distribución bootstrap, denotada por $\hat{F}_B^*(T)$, entonces $\hat{F}_B^*(\hat{T})$ y será un estimador de $F(\hat{T})$.

Entonces el valor-p bootstrap es

$$\hat{p}^*\left(\hat{T}\right) = 1 - \hat{F}_B^*(\hat{T}) = \frac{1}{B}\sum_{j=1}^{B} I\left(T_j^* > \hat{T}\right)$$

que es el porcentaje de muestras bootstrap para las cuales T_j^* es mayor que $\hat{T}$. Para una prueba de nivel α, rechazamos H_0 cuando $\hat{p}^*\left(\hat{T}\right) < \alpha$.

Como un ejemplo consideremos la siguiente tabla que presenta el mes de nacimiento de los estudiantes de pregrado de la Universidad Nacional-Sede Medellín, también aparece la probabilidad estimada de nacer en cada mes y la probabilidad teórica asumiendo uniformidad, o sea asumiendo que una persona extraída al azar tiene igual probabilidad de haber nacido en cualquier día del año. Observe que los meses tienen diferentes probabilidades teóricas debido a que los números de días por mes no es constante.

Mes	Nacimientos	Frecuencia Relativa	Probabilidad Teórica
Enero	856	0.09069	0.08493
Febrero	716	0.07586	0.07671
Marzo	740	0.07840	0.08493
Abril	721	0.07639	0.08219
Mayo	803	0.08507	0.08493
Junio	751	0.07956	0.08219
Julio	790	0.08370	0.08493
Agosto	830	0.08793	0.08493
Septiembre	830	0.08793	0.08219
Octubre	801	0.08486	0.08493
Noviembre	789	0.08359	0.08219
Diciembre	812	0.08603	0.08493

Las pruebas tradicionales basadas en el estadístico χ^2 de Pearson y la prueba LRT producen los siguientes resultados:

```
> chisq.test(table(mes),p=esp)

        Chi-squared test for given probabilities

data:  table(mes)
X-squared = 18.4984, df = 11, p-value = 0.07071

> chisq.test(table(mes),p=esp,simulate.p.value=T)

Chi-squared test for given probabilities with simulated p-
value (based
on 2000 replicates)

data:  table(mes)
X-squared = 18.4984, df = NA, p-value = 0.06847

prueba.multinomial(table(mes),esp)
$G2
[1] 18.55022

$valor.p
[1] 0.06966068
```

Página www
Página de Abertura
Contenido
Página 11 de 100
Regresar
Full Screen
Cerrar
Abandonar

```
chi.cal<-function(x,p.esp){
  nXp.esp<-sum(x)*p.esp
  chi.c<-sum((x-nXp.esp)^2/nXp.esp)
  chi.c
}

 Nboot<-1000
 mes<-c(856,716,740,721,803,751,790,830,830,801,789,812)
 dist.obs<-mes/sum(mes)
 esp<-c(31,28,31,30,31,30,31,31,30,31,30,31)/365
 muestra.b<-rmultinom(Nboot,length(mes),dist.obs)
 res<-apply(muestra.b,2,chi.cal,dist.obs)

 plot(density(res))
 points(xx<-seq(from=0.01,to=40,length=100),
        dchisq(xx,11),type='l',col='red')

 quantile(res,probs=c(0.05,1:9/10,0.95))
        5%       10%       20%       30%       40%       50%       6
 5.498160  5.950277  7.687522  8.181504  9.398423
10.000479 11.145016 12.401174
      80%       90%       95%
14.038471 16.381451 18.847421
>

 qchisq(0.95,11)
[1] 19.67514

 chi.cal(mes,esp)
[1] 18.49845
```

Si consideramos la prueba LRT, podemos calcular la distribución simulada así:

```
> prueba.multinomial2<-function(observado,prob.teoricas){
+  if(length(observado)!=length(prob.teoricas))stop('Longitudes dife
+  observado<-ifelse(observado==0,0.5,observado)
+  G2<--2*sum(observado*log(prob.teoricas/(observado/sum(observado))
+  G2
+
+ }
>
> res<-apply(muestra.b,2,prueba.multinomial2,esp)
> quantile(res,probs=0.95)
     95%
19.79915
> qchisq(0.95,11)
[1] 19.67514
>
> G2.obs<-prueba.multinomial2(table(mes),esp)
> mean(ifelse(G2.obs-res>0,0,1))
[1] 0.077
```

Página www
Página de Abertura
Contenido
Página 13 de 100
Regresar
Full Screen
Cerrar
Abandonar

la siguiente prueba considera el estadístico de prueba definido como $T = \max |\hat{p}_i - p_{i,H_0}|$. La distribución asintótica no es fácil de calcular y por lo tanto la hallamos vía simulación.

```
# Prueba basada en la máxima distancia entre lo observado y espera

otra.prueba<-function(x,p.esp){
 res<-max(abs(x-p.esp*sum(x)))
 res
}

res<-apply(muestra.b,2,otra.prueba,esp)
quantile(res,probs=0.95)
     95%
76.19178

max.obs<-otra.prueba(table(mes),esp)
mean(ifelse(max.obs-res>0,0,1))
[1] 0.232
```

Medidas de Asociación en Tablas 2×2

Medidas basadas en el coeficiente de correlación

Asumamos que la primera columna (fila) toma el valor cero y la segunda columna (fila) toma el valor de uno.
Utilicemos la siguiente notación:
$\mu_f = \pi_{2+}$
$\mu_c = \pi_{+2}$
$\sigma_f^2 = \mu_f(1-\mu_f) = \pi_{1+}\pi_{2+}$
$\sigma_c^2 = \mu_c(1-\mu_c) = \pi_{+1}\pi_{+2} cov_{fc} = \pi_{22} - \mu_f\mu_c = \pi_{22} - \pi_{2+}\pi_{+2}$
Entonces el coeficiente de correlación es

$$\rho = \frac{\pi_{22} - \pi_{2+}\pi_{+2}}{\sqrt{\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2}}}$$

- El ρ es invariante ante cambios de filas y columnas
- Cambia solo de signo si intercambiamos solo las filas o columnas
- $\rho = 0$ is las variables son independientes
- Si $\pi_{12} = \pi_{21} = 0$, entonces $\rho = 1$
- Si $\pi_{11} = \pi_{22} = 0$, entonces $\rho = -1$

```
rho.c<-function(x){
 x<-matrix(x,ncol=2,byrow=2)
 p22<-x[2,2]/(N<-sum(x))
 p1m<-(x[1,1]+x[1,2])/N
 p2m<-1-p1m
 pm1<-(x[1,1]+x[2,1])/N
 pm2<-1-pm1
 r<-(p22-p2m*pm2)/sqrt(p1m*p2m*pm1*pm2)
 return(r)
 }

niños<-c(79,202,57,138)

rho.c(niños)
[1] -0.01215828

muestra.b<-rmultinom(2000,sum(niños),niños/length(niños))
res<-apply(muestra.b,2,rho.c)
quantile(res,probs=c(0.025,0.975))
       2.5%        97.5%
-0.10155227  0.07971674

plot(density(res))
```

Medidas basadas en la χ^2 de Pearson

- El estadístico chi-cuadrado no es una buena medida del grado de asociación entre dos variables.
- Pero el amplio uso de este estadístico ha propiciado la creación de medidas de asociación basadas en él.
- Cada una de estas medidas intenta minimizar la influencia del tamaño muestral y de la del número de celdas de la tabla. Además se pretende establecer límites, usualmente entre cero y uno, a estas medidas para darle comparabilidad a diversas tablas.

Aunque pueden estas medidas ser difíciles de interpretar y carecer de interpretación probabilística y por lo tanto no se recomiendan (Upton).
Para una tabla 2×2 es fácil verificar que la chi-cuadrada de Pearson es

$$\chi^2 = \frac{N(ad - bc)^2}{k_1 k_2 n_1 n_2}$$

El coeficiente ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Para aquellas tablas en las cuales una dimensión sea mayor que 2, puede no estar entre 0 y 1 ya que el valor de la chi-cuadrado puede ser mayor que el tamaño muestral.

El Coeficiente de Contingencia

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Esta medida fue sugerida por Pearson. Está confinada al rango 0 y 1, pero puede no alcanzar el límite superior del intervalo. Por ejemplo, para tablas 4×4, el máximo valor de es 0.87.

V de Cramér

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

donde k es el mínimo entre el número de filas y el de columnas de la tabla. El estadístico V de Cramér puede alcanzar el máximo 1 para cualquier tabla. Si una de las dimensiones de la tabla es 2, entonces V y ϕ son idénticas.

El Coeficiente de Tschuprov

$$T = \sqrt{\frac{\chi^2}{N\sqrt{(I-1)(J-1)}}}$$

El Estadístico G^2

El estadístico G^2 está basado en la razón de verosimilitud, y es tal vez la medida de ajuste que más sirve en el análisis de datos categóricos, dadas sus propiedades.

$$G^2 = 2\sum_i \sum_j n_{ij}\left[\log(n_{ij}) - \log(e_{ij})\right]$$

Bajo el supuesto de independencia tenemos en una tabla bidimensional 2×2 y bajo el esquema de muestreo multinomial $\pi_{ij} = \pi_{i+} \times \pi_{+j}$

$$P\left(N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}\right) =$$

$$\frac{n_{++}!}{n_{11}!n_{12}!n_{21}!n_{22}!}\pi_{11}^{n_{11}}\pi_{12}^{n_{12}}\pi_{21}^{n_{21}}\pi_{22}^{n_{22}}$$

El estadístico de la razón de verosimilitud es $LR = L(\hat{\omega})/L(\hat{\Omega})$, que en nuestro caso y sabiendo que el estimador de π_{ij} es $\hat{\pi}_{ij} = n_{ij}/n_{++}$ en el caso general y bajo el modelo de independencia es $\hat{\pi}_{ij} = n_{i+}/n_{++} \times n_{+j}/n_{++}$.

Recordemos que $-2\log(LR)$ se distribuye asintóticamente con grados de libertad datos por $dim(\Omega) - dim(\omega)$. Por lo tanto

$$LR = \frac{\frac{n_{++}!}{n_{11}!n_{12}!n_{21}!n_{22}!}\left(\frac{n_{1+}}{n_{++}}\frac{n_{+1}}{n_{++}}\right)^{n_{11}}\left(\frac{n_{1+}}{n_{++}}\frac{n_{+2}}{n_{++}}\right)^{n_{12}}\left(\frac{n_{2+}}{n_{++}}\frac{n_{+1}}{n_{++}}\right)^{n_{21}}\left(\frac{n_{2+}}{n_{++}}\frac{n_{+2}}{n_{++}}\right)^{n_{22}}}{\frac{n_{++}!}{n_{11}!n_{12}!n_{21}!n_{22}!}\left(\frac{n_{11}}{n_{++}}\right)^{n_{11}}\left(\frac{n_{12}}{n_{++}}\right)^{n_{12}}\left(\frac{n_{21}}{n_{++}}\right)^{n_{21}}\left(\frac{n_{22}}{n_{++}}\right)^{n_{22}}}$$

$$LR = \frac{(e_{11})^{n_{11}}(e_{12})^{n_{12}}(e_{21})^{n_{21}}(e_{22})^{n_{22}}}{(n_{11})^{n_{11}}(n_{12})^{n_{12}}(n_{21})^{n_{21}}(n_{22})^{n_{22}}}$$

donde $e_{ij}\frac{n_{i+}n_{+j}}{n_{++}}$ es el valor esperado de la celda $i-j$. Tomado logaritmo, tomando el signo negativo y multiplicando por dos tenemos

$$G^2 = -2\log(LR) = \sum_i \sum_j n_{ij}\log\left(\frac{e_{ij}}{n_{ij}}\right)$$

o también se puede expresar como

$$G^2 = 2\log(LR) = \sum_i \sum_j n_{ij}\log\left(\frac{n_{ij}}{e_{ij}}\right)$$

El Q de Yule

El Q de Yule es una medida de asocición que ha resistido el paso del tiempo. Se define como

$$Q = \frac{ab - cd}{ab + cd}$$

Si n_{++} es razonablemente grande, la distribución de Q es normal, con varianza

$$\frac{1}{4}(1 - Q^2)^2(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d})$$

El rango de Q es $(-1, 1)$, con los puntos extremos corespondiendo a asociación completa (positiva o negativa) y con 0 como no asociación.

A continuación presentamos una función en R que permite calcular estas medidas de asociación para una tabla 2×2 y la aplicamos al ejemplo del primer capítulo sobre destreza manual y sexo.

```
medidas.de.asociación.2x2<-function(a,b,c,d){

k1<-a+b
k2<-c+d
n1<-a+c
n2<-b+d
N<-n1+n2

chi<-N*(a*d-b*c)^2/(k1*k2*n1*n2)
phi<-sqrt(chi/N)
C<-sqrt(chi/(chi+N))
V<-phi
T<-phi
Q<-(a*b-c*d)/(a*b+c*d)

list(chi2=chi,phi=phi,C=C,V=V,T=T,Q=Q)

}
```

```
> medidas.de.asociación.2x2(79,202,57,138)
$chi2
[1] 0.07036408

$phi
[1] 0.01215828

$C
[1] 0.01215738

$V
[1] 0.01215828

$T
[1] 0.01215828

$Q
[1] 0.3396575

>
```

Prueba de Simetría de McNemar

La prueba de simetría Chi-cuadrado de McNemar para tablas de contingencia cuadradas. Es apropiada en experimentos con muestras pareadas. Aquí se consideran respuestas de N sujetos en la muestra "antes" y "después" de algún evento, por ejemplo la aplicación de un tratamiento.

La prueba chi-cuadrada de Pearson es fácil de mostrar está dada por

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Si pensamos en el problema de las parejas de casados en Medellín, tenemos la tabla siguiente

	No se casaría	Sí se casaría
No se casaría	13	12
Sí se casaría	25	97

```
> library(ctest)
> mcnemar.test(matrix(c(13,12,25,97),ncol=2,byrow=T))

        McNemar's Chi-squared test with continuity correction

data:  matrix(c(13, 12, 25, 97), ncol = 2, byrow = T)
McNemar's chi-squared = 3.8919, df = 1, p-value = 0.04852


> mcnemar.test(matrix(c(13,12,25,97),ncol=2,byrow=T),correct=F)

        McNemar's Chi-squared test

data:  matrix(c(13, 12, 25, 97), ncol = 2, byrow = T)
McNemar's chi-squared = 4.5676, df = 1, p-value = 0.03258
```

La prueba nos indica que no hay simetría en la tabla, esto es, la insatisfacción de uno de los cónyugues no es la misma si se trata de mujeres o de hombres.