INFERENCIA NO PARAMÉTRICA

Estadística II*

January 26, 2020

Abstract

Se estudiará una introducción a las técnicas de inferencia no paramétrica. La Sección 1 es una introducción. En la Sección 2 se cubre Bondad de Ajuste, referente a la unidad XX y en la Sección 3, Independencia de Atributos, unidad XX del programa de la materia. En ambas secciones se cubrirán ejemplos en el software estadístico R. La Sección 4 es un anexo.

*Curso a cargo de la profesora Silvia Vietri en FCE-UBA. Se agradece la colaboración de Fiona Franco Churruarín en la elaboración de esta nota de clase.

1 *Introducción*

Hasta ahora, se han estudiado métodos de inferencia paramétrica. A grandes rasgos, el procedimiento consistía en hacer pruebas acerca de los *parámetros* de una distribución poblacional –que se suponía conocida– a partir de información provista por una muestra aleatoria. Por ejemplo, se intentaba estimar la media o la varianza poblacional, una proporción poblacional, o el cociente de varianzas. En este contexto, era deseable que las técnicas inferenciales fueran robustas para que vulneraciones de los supuestos de la distribución no alterasen significativamente los resultados de las estimaciones o inferencias. Es decir, en el paradigma de la inferencia paramétrica es deseable que los resultados que se obtengan no varíen mucho cuando la distribución subyacente a los datos no sea la que se está suponiendo.

Por el hecho de partir de una especificación errónea del modelo estadístico que siguen los datos, se corre el riesgo de tomar una decisión equivocada a partir de una inferencia errónea. Ahora bien ¿qué significa esto? Un ejemplo concreto –correspondiente a Bondad de Ajuste– se da en los mercados financieros. Es muy común utilizar la distribución normal para modelar los retornos de un activo financiero (una acción, un bono, el tipo de cambio entre distintas monedas, entre otros). Es conocido que la distribución normal está centrada en su media y que entre tres desvíos estándar a derecha y a izquierda de su media acumula 99.73% de probabilidad. Sin embargo, es un hecho estilizado dentro de la literatura de finanzas muchos activos financieros registran retornos de más de 3 desvíos estándar más frecuentemente que lo que una distribución normal implica. Así, de estar utilizando un modelo de retornos financieros que se base en la distribución normal podría estar subestimándose la probabilidad de ocurrencia de eventos 'extremos' en los retornos financieros. Para poner un ejemplo numérico: una bono puede tener un rendimiento esperado de 5% del capital invertido al plazo de un año, con un desvío estándar del 1%, de modo tal que con probabilidad del 99% los retornos están entre 2% y 8%. Sin embargo, se han registrado varios eventos en los que se esperaban retornos como éstos y que, ex post, se registre un retorno sobre el capital invertido negativo.[1]

Para intentar evitar situaciones como ésta, se puede recurrir a la utilización de contrastes no paramétricos. Éstos no hacen ningún supuesto acerca de la distribución de probabilidad de la población (llamada *distribución subyacente*) de la que surge la muestra aleatoria.

[1]Este ejemplo puede servir para ejemplificar la inadecuación de la distribución normal para modelar retornos de activos financieros, aunque el lector debería considerar que esta cuestión representa una rama de la literatura de macro-finanzas muy amplia. Algunos de los tópicos (no exhaustivos de esta literatura) que pueden ser de utilidad para el lector interesado en estos temas son los siguientes: en lo que respecta a las fallas que la distribución normal, una de las referencias más conocidas es Nassim Taleb (2007); sobre predicción de retornos de activos, véase la Hipótesis de Mercados Eficientes; para los vínculos entre retornos de activos y macroeconomía véase la literatura de *consumption-based asset pricing.*

Las técnicas no paramétricas pueden categorizarse –no excluyentemente– en las siguientes categorías:

- **Técnicas no paramétricas en sentido estricto:** no aparece ningún tipo de hipótesis acerca de un determinado parámetro θ de una población, sino que el proceso se basa en un estadístico sin referencia a ningún parámetro poblacional
- **Métodos de distribución libre:** el estadístico utilizado presenta una distribución de probabilidad que no depende de la distribución de probabilidad de la población de la que se ha extraído la muestra que suministra información al estadístico.

Las técnicas no paramétricas también difieren de los métodos estudiados anteriormente en que no suelen utilizar directamente la información muestral de los valores de la variable que es objeto de estudio, sino que se utiliza la frecuencia con la que aparecen dichos valores en la muestra, o su posición, orden o rango en una muestra ordenada. Para ejemplificar esto, recuérdese que en el cálculo de los estadísticos t se utilizan directamente los valores de las observaciones, y estos estadísticos también dependían de parámetros poblacionales (o de las estimaciones de los mismos a partir de los datos muestrales).[2]

A fines prácticos, se estará utilizando el procedimiento desarrollado en la unidad de Test de Hipótesis, aunque los estadísticos de contraste a utilizar no utilizarán estimadores de parámetros poblcionales.

Otra diferencia entre ambas técnicas –y para subrayar el hecho de que una no es superior a la otra, sino que la utilización de una u otra técnica depende de las características de los datos que se tienen y de la pregunta de investigación– es que los contrastes no paramétricos son algo menos eficientes que los contrastes paramétricos cuando la población tiene una distribución normal, en el sentido de que poseen menor potencia[3] a un mismo nivel de significación α, aunque son más eficientes cuando la distribución de la población no es normal.

En esta nota de clase sólo se explicarán contrastes basados en una sola muestra, aunque los hay de más de una muestra. Dentro de los tests de una muestra, sólo se estudiarán los contrastes de bondad de ajuste (Sección XX) y de independencia de atributos (Sección XX). Sin embargo, la inferencia no paramétrica no se agota en sólo estas pruebas, sino que también existen otros procedimientos — sólo

[2]Piense en el caso en el que sea desea testear si una hipótesis de una media poblacional es igual a cero ¿qué información muestral utiliza para calcularlo?

[3]Recuérdese que esto es la probabilidad de no cometer error de tipo II, o de tomar una decisión correcta, rechazar la hipótesis nula cuando ésta es falsa. $1 - \beta = P[\text{Rechazar}H_0/H_0\text{Falsa}]$

a fines ennumerativos, algunos de estos son: pruebas de aleatoreidad, contrastes de localización y contrastes de asociación.

El objetivo puntual de esta nota de clase será explicar los contrastes utilizados en (i) bondad de ajuste y (ii) independencia de atributos.

2 Bondad de Ajuste

2.1 Notación y algunas definiciones

- h: número de clases de observaciones, indizados por $i = 1, \ldots, h$.
- k: número de parámetros a estimar en un test
- Frecuencia absoluta: número de veces que aparece un determinado valor de una variable x_0 en la muestra, se denota por n_i, y debe cumplirse que

$$\sum_{i=1}^{h} n_i = n$$

  siendo n el número total de observaciones o tamaño de la muestra.
- Frecuencia absoluta acumulada: Número de observaciones iguales o inferiores a un cierto valor de la variable considerada, se representa por N_i, siendo

$$\begin{aligned} N_1 &= n_1 \\ N_2 &= n_1 + n_2 \\ N_3 &= n_1 + n_2 + n_3 \\ &\ldots \\ N_h &= n_1 + n_2 + \ldots + n_h \end{aligned}$$

- $f_0(x)$: Función de densidad teórica
- $f(x)$: Función de densidad empírica o de la muestra

3 Bondad de Ajuste

Cuando se habla de *bondad de ajuste*, se hace referencia a contrastes estadísticos que permiten probar qué tan bien un modelo estadístico se ajusta a los datos muestrales. El objetivo de estos contrastes es el de verificar si una muestra que contiene un conjunto de observaciones se puede suponer provienente de una población con cierta distribución de probabilidad que tiene una función de densidad teórica $f_0(x)$. Se llama $f(x)$ función de densidad de la muestra o función de densidad empírica.

La hipótesis nula H_0 de estos contrastes es que el ajuste de la información de la muestra o función de densidad empírica $f(x)$ a un modelo estadístico o función de densidad $f_0(x)$ es *bueno*, entendiéndose como que la muestra no provee evidencia para rechazar esta hipótesis nula. La hipótesis alternativa H_1 es el opuesto de este enunciado, justamente, que el ajuste de los datos muestrales al modelo teórico $f_0(x)$ *no es bueno*, es decir, que éstos proveen información para decir que los datos *no* se ajustan a esa función de densidad planteada.

$$\text{Test general}: \begin{cases} H_0 : f(x) = f_0(x) \\ H_1 : f(x) \neq f_0(x) \end{cases}$$

3.1 Test Ji-Cuadrado

La hipótesis nula de este contraste es la recién explicitada. Puede aplicarse tanto a distribuciones poblaciones discretas como continuas, y con observaciones ordinales o nominales.

- Si la variable aleatoria a ajustar es **discreta**, los datos se agrupan en una serie simple.

  Ejemplo 3.1. Considérese el caso de un dado de seis caras. Cada vez que se lanza el dado, hay seis posibles resultados. Suponga que se repite el experimento 304 veces. Los datos se ordenarían de la siguiente manera.

Cara del dado (x_i)	Frecuencia (n_i)
⚀	53
⚁	51
⚂	48
⚃	43
⚄	57
⚅	52
Total (n)	304

- Si la *v.a.* a ajustar es **continua**, los datos se agrupan en una serie de intervalos.

El estadístico de prubeba a utilizar será el *estadístico Ji-Cuadrado de Pearson*, que sigue una distribución Ji-Cuadrado con $h - k - 1$ grados de libertad.[4]

$$\sum_{i=1}^{h} \frac{(n_i - n_i^e)^2}{n_i^e} \sim \chi^2_{h-k-1} \tag{1}$$

¿Cómo puede este estadístico de contraste proveer información para probar qué tanto se ajustan los datos a una distribución? Nótese que, en escencia, habiendo clasificado los datos muestrales en un número de categorías, se están compararado las frecuencias observadas con las que se esperaría que ocurrieran si los datos surgieran de una población que sigue la función de densidad teórica.

Una forma matemática de introducir cierta medida de "acumulación" es la suma de términos cuadráticos, que es siempre mayor o igual a cero. De esta forma, estamos tomando la diferencia entre la frecuencia observada n_i y la esperada n_i^e. La frecuencia esperada se calcula como $n_i^e = np_i$, siendo $p_i = \widehat{p(x_i)}$ la probabilidad de obtener la observación x_i bajo el modelo de probabilidad teórico $f_0(x)$.

Esta diferencia puede ser positiva o negativa, luego, tomando el cuadrado de esa diferencia, se obtiene una cantidad iguall a positiva o cero, de modo tal que no se cancelen diferencias positivas con diferencias negativas. El estadístico sería cero sólo en el caso en que todas las frecuencias observadas sean iguales a las esperadas.

Luego, el cuadrado de la diferencia se divide por la frecuencia esperada para obtener una medida de si las diferencias son grandes o pequeñas respecto de lo que debieramos observar.

Una cuestión no menor es que para poder aplicar el test debe cumplirse la siguiente condición[5]:

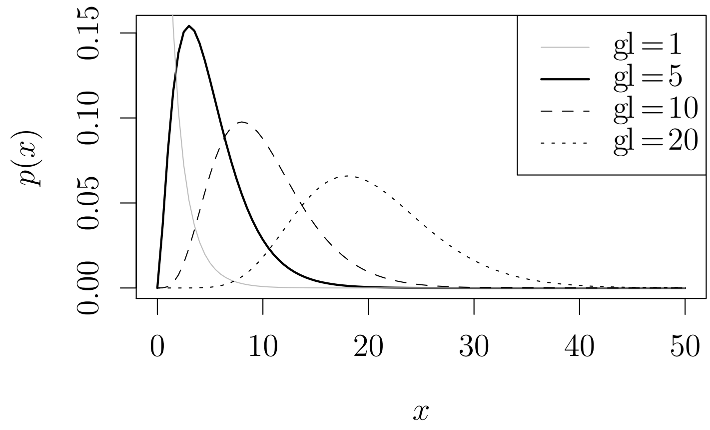
$$n_i^e \geq 5 \quad \text{ó} \quad 5h \leq n$$

De no cumplirse, deberán reagruparse las clases de modo tal que todas las frecuencias esperadas de cada clase sean mayores a 5. Esto generará, a su vez, una reducción de los grados de libertad del estadístico. Es también aconsejable, a fines de obtener una mejor aproximación, que todas las clases se construyan de forma que sus probabilidades sean aproximadamente iguales.

Recuérdese la forma de la distribución Ji-Cuadradado. Esta posee un sesgo hacia la derecha, más notorio a menos grados de libertad.

[4]Demostración en el Anexo.

[5]Esto se debe a que la distribución del estadístico Ji-Cuadrado es asintótica y no exacta.

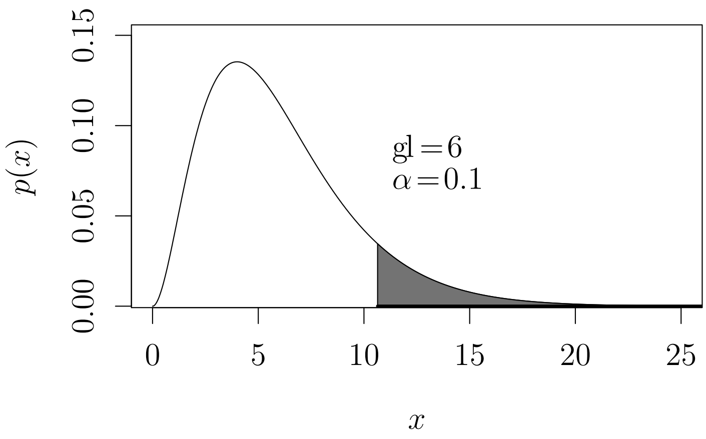


La pregunta relevante ahora es ¿de qué modo provee este estadístico evidencia para rechazar o no la hipótesis nula H_0? Obsérvese que el numerador del estadístico de prueba es una medida de la discrepancia entre lo observado y lo esperado, elevada al cuadrado para evitar que diferencias positivas y negativas se compensen entre sí, y penalizando más a discrepancias más grandes.[6] Entonces, a mayores diferencias, mayores valores del estadístico, lo que implica que se estará en un punto "más a la derecha" del dominio de la distribución Ji-Cuadrado. ¿Qué implica esto para la decisión estadística?

Si se decide utilizando un nivel de significación $\alpha = P[\mathrm{R}H_0/H_0\mathrm{V}]$ dado, típicamente $\alpha = 0.10$, entonces, *ceteris paribus* los grados de libertad, habrá más chances de que se esté dentro de la región crítica, es decir, de que haya evidencia estadística de que las distribución empírica $f(x)$ es significativamente diferente de la distribución teórica $f_0(x)$, permitiendo rechazar la hipótesis nula H_0.

Por otro lado, recuérdese que el *p-valor* es el mínimo valor de α para el cual se puede rechazar H_0 para la muestra elegida. La regla de decisión en este caso era

[6] La diferencia al cuadrado es, ciertamente, más manejable que el valor absoluto de la diferencia



Ejemplo 3.2. Sea X una variable aleatoria, y k el número de parámetros a estimar. Se quiere probar si X sigue una distribución normal, cualesquiera sean sus parámetros. Se plantea el siguiente test.

$$H_0 : X \sim \mathcal{N}(\mu,\, \sigma)$$
$$H_1 : X \nsim \mathcal{N}(\mu,\, \sigma)$$
$$k = 2$$

Los parámetros del test pueden o no estar prefijados: se puede testear si la muestra tiene una distribución normal, de Poisson, u otra, o si sigue una distribución con parámetros específicos. Esto afecta al estadístico a través de k. Cuando se estiman parámetros, se reduce el número de grados de libertad, generando una región crítica mayor a un mismo nivel de significación.

En este caso, se está testeando si la muestra que surge de X sigue una distribución normal, cualesquiera sean sus parámetros. No se podrá rechazar la hipótesis nula si la distribución es $X \sim \mathcal{N}(0, 1)$, $X \sim \mathcal{N}(0, 2)$, o $X \sim \mathcal{N}(20, 4)$, aunque si se podrá en casos como $X \sim Bi(m, p)$ o $X \sim P(\lambda)$.

En este caso, se está testeando si la muestra que surge de X sigue una distribución normal con media μ_0, y cualquier desvío estándar σ. Es decir, tanto para $X \sim \mathcal{N}(0, 1)$ como para $X \sim \mathcal{N}(0, 1)$ no se podrá rechazar H_0 pero si para una distribución $X \sim \mathcal{N}(20, 4)$. Un tercer ejemplo es testear con todos los parámetros que caracterizan la distribución, como puede ser $X \sim \mathcal{N}(0, 1)$.

Ejemplo 3.3.

$$\begin{gathered} H_0 : X \sim \mathcal{N}(\mu_0, \sigma) \\ H_1 : X \nsim \mathcal{N}(\mu_0, \sigma) \\ k = 1 \end{gathered}$$

Ejemplo 3.4.

$$\begin{gathered} H_0 : X \sim \mathcal{N}(\mu_0, \sigma_0) \\ H_1 : X \nsim \mathcal{N}(\mu_0, \sigma_0) \\ k = 2 \end{gathered}$$

4 Independencia de Atributos

4.1 Test Ji-Cuadrado para independencia de atributos

Falta armar esto

4.2 Test de Kolmogorov-Smirnov

Otro test

$$\varepsilon = \Big| F_i - F_0 \Big| \tag{2}$$

Ejemplo 4.1. Reinhart & Rogoff en su libro de 2009 *This Time is Different* utilizan contrastes de Kolmogorov-Smirnov para ...

5 Independencia de atributos

6 Anexo

6.1 Deducción y demostración de la distribución del estadístico Ji-Cuadrado de Pearson

6.2 Código de los gráficos en R

Distribución Ji-Cuadrado

```
x <- seq(1:50)
```

```r
curve( dchisq(x, df=5), col='black', lwd = 2,
       from = 0, to = 50,
       ylab = 'p(x)',
       main = 'Distribucion Ji-Cuadrado')
curve( dchisq(x, df=1), col='gray', add=TRUE)
curve( dchisq(x, df=10), col='black', add=TRUE, lty = 2)
curve( dchisq(x, df=20), col='black', add=TRUE, lty = 3 )

legend('topright',
      expression(gl==1, gl==5, gl==10, gl==20),
      lwd = c(1, 2, 1, 1),
      lty = c(1,1, 2, 3),
      col = c('grey','black', 'black', 'black'))
```

Región crítica del test Ji-Cuadrado

```r
#Graficando la region critica
#Seleccion de grados de libertad
df <- 6
#Seleccion de error tipo I
alpha <- 0.10
#Defino valor del eje de las abcisas (sera un valor del
   estadistico) a partir del cual rechazo H0
rc <- qchisq(1 - alpha, df)
#Lo que haremos para graficar la region critica sera
   construir un poligono con la forma de la distribucion
#Defino el eje x (tecnicamente deberia seguir hasta infinito,
    pero a fines del grafico se corta antes)
x <- seq(0,30,0.01)
#Defino la region critica en el eje
z <- seq(rc,30,0.01)
#Defino el "lado" superior del poligono, que sera una curva
   con la forma de la distribucion en el segmento que quiero
p <- dchisq(z, df)
#redefino los vectores de lados para la funcion poligono
z <- c(z,30,rc)
p <- c(p,min(p),min(p))
#Grafico distribucion Ji-Cuadrado
plot(x,dchisq(x, df),type="l",
     ylab="$p(x)$",xlab="$x$",
     ylim=c(0.005,0.15), xlim = c(0,25))
#ylim solo esta para que quede el eje x bien cruzado con el
   cero
#Agrego al grafico el poligono
polygon(z,p,col="gray45")
#Agrego una linea que define la region critica
segments(x0=rc, y0=0, x1=30, y1=0, lwd=3)
#Agrego leyenda
```

```
legend('center',
       expression(gl==6, '$\\alpha$'==0.10),
```

Cara	\$n_i#	$p(x_i)$	n_i^e	$n_i - n_i^e$	$(n_i - n_i^e)^2$	$(n_i - n_i^e)^2/n_i^e$
As	70	0.167	100.0	-30.00	900.00	9.00
K	115	0.167	100.0	15.00	225.00	2.25
Q	122	0.167	100.0	22.00	484.00	4.84
J	98	0.167	100.0	-2.00	4.00	0.04
Rojo	85	0.167	100.0	-15.00	225.00	2.25
Negro	110	0.167	100.0	10.00	100.00	1.00
h=6	n=600	**1**				$\varepsilon_M = 19.380$