Ejercicios

Estadística No Paramétrica

Joaquin Cavieres G.

1. Introducción

Generación de números aleatorios

Las siguientes funciones nos permiten generar números aleatorios en R:

```
# x = es un vector del cual se quieren muestrear
# size = el número de muestras
# replace = Debería el muestreo hacerse con remplazo?
# FALSE = sólo una vez, TRUE = multiples veces
sample(x = 1:10, size  = 5) # sin reemplazo
```

```
## [1] 9 5 2 6 4
```

```
sample(x = 1:5, size = 10, replace = TRUE) # con reemplazo
```

```
##  [1] 3 5 4 5 2 2 4 3 2 2
```

```
sample(x = c("C", "S"), # Valores de la moneda
       size = 5,        # 5 lanzamientos
       replace = TRUE)  # Muestrando con reemplazo
```

```
## [1] "C" "S" "C" "C" "S"
```

Generación desde una distribución Gaussiana

```
# 5 muestras desde una distribución Gaussiana con media 0 y desvest 1
rnorm(n = 5, mean = 0, sd = 1)
```

```
## [1]  2.6447745  0.9664382  0.5476309  0.2799596 -1.2656079
```

```
# 1000 muestras desde una distribución Gaussiana con media 5 y desvest 0.5
rnorm(n = 10, mean = 5, sd = 0.5)
```

```
##  [1] 5.442822 5.696979 5.470590 5.604883 5.508456 4.984813 4.671682 4.293488
##  [9] 3.860028 5.028089
```

Generación desde una distribución Uniforme

```
# 5 muestras desde una distribución Uniforme con mínimo 0 y máximo 1
runif(n = 5, min = 0, max = 1)
```

```
## [1] 0.1286182 0.5268423 0.4382104 0.2212438 0.8703212
```

Nota: Si queremos dejar un ejercicio o simulación reproducible en `R` entonces podemos utilizar la función `set.seed()` para fijar una "semilla". Así por ejemplo:

```
set.seed(100)
rnorm(n = 5, mean = 0, sd = 1)
```

```
## [1] -0.50219235  0.13153117 -0.07891709  0.88678481  0.11697127
```

```
rnorm(n = 5, mean = 0, sd = 1)
```

```
## [1]  0.3186301 -0.5817907  0.7145327 -0.8252594 -0.3598621
```

Función de distribución empírica (`v.a` continua)

```
n = 1000
y = rnorm(n, mean=0, sd=1)
mean(y)
```

```
## [1] 0.02323841
```

```
var(y)
```

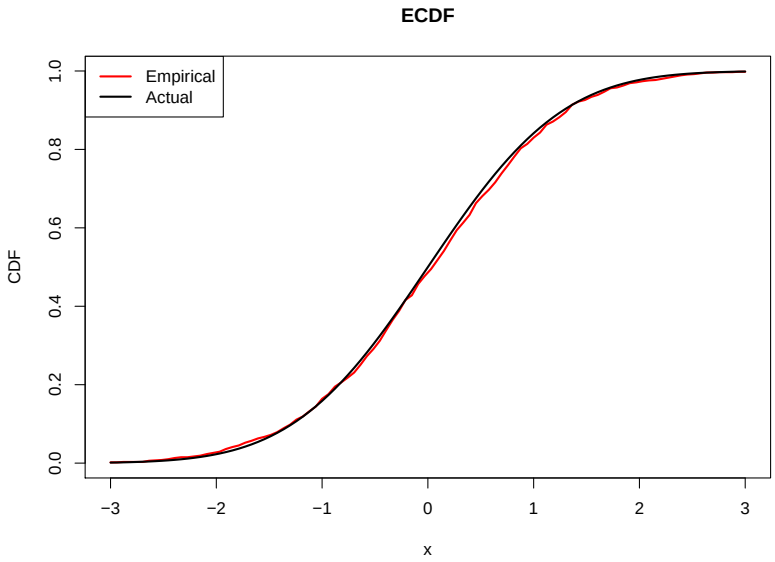
```
## [1] 1.066474
```

```
x = seq(-3,3,length=100)
ecdf.fun = ecdf(y)     # Crea la ECDF
class(ecdf.fun)        # Función
```

```
## [1] "ecdf"     "stepfun"  "function"
```

graficámos la función:

```
plot(x, ecdf.fun(x), lwd=2, col="red", type="l", ylab="CDF", main="ECDF")
lines(x, pnorm(x), lwd=2)
legend("topleft", legend=c("Empirical", "Actual"), lwd=2, col=c("red","black"))
```

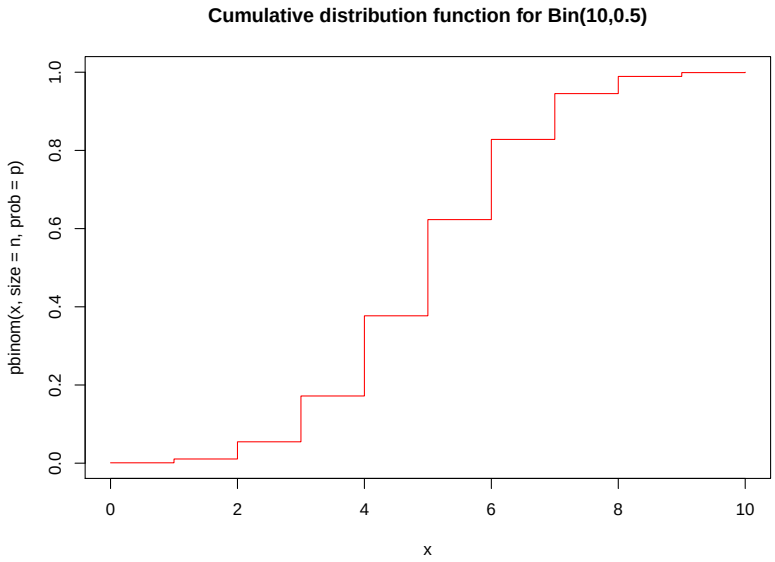


Función de distribución empírica (v.a discreta)

```
# Discreta
n = 10
p = 0.5
dbinom(1, size = n, prob = p)
```

```
## [1] 0.009765625
```

```
x <- 0:n
plot(x, pbinom(x, size = n, prob = p), type="s",
     main = "Cumulative distribution function for Bin(10,0.5)", col="red")
```



Test de Hipótesis

Una hipótesis es una declaración sobre algún fenómeno que aún no ha sido probado. La prueba de hipótesis es una versión más formal de esta declaración pero utilizando pruebas estadísticas. Hay otras formas de probar hipótesis (si cree que mañana va a llover entonces puede esperar al día siguiente para comprobarlo), pero nos centraremos solo en los métodos que nos brindan las estadísticas.

La prueba de hipótesis nos sirve de ayuda para analizar si existen relaciones entre diferentes fenómenos o variables. ¿Existe una relación entre la alimentación de un niño y su peso ? ¿Existe una relación entre el alcoholismo y la hipertensión? ¿Existe una relación entre que obtenga buenas notas y que los días estén nublados? Todas esas son relaciones que podemos probar mediante la prueba de hipótesis.

La prueba de hipótesis se parece mucho al trabajo de detective en cierto modo (o al menos a la forma en que se supone que se gestionan los justificantes penales). ¿Cuál es la presunción con la que comenzamos en el sistema legal? Todo el mundo se presume inocente hasta que se pruebe más allá de toda duda razonable su culpabilidad. En el contexto de la estadística, llamaríamos hipótesis nula a la presunción de inocencia. La hipótesis nula establece cuál es nuestro estado inicial de conocimiento, el cual es que no hay relación entre dos cosas, fenómenos o variables. Por tanto, hasta que sepamos que una persona no es inocente, es inocente. Hasta que sepamos que hay una relación, no hay relación. Generalmente se escribe como $\boldsymbol{H}_0$, $\boldsymbol{H}$ para reconocer la "hipótesis" y 0 como el punto partida.

$$\boldsymbol{H}_0 \quad : \quad \text{El acusado es inocente}$$

Si nuestras pruebas no refutan la hipótesis nula entonces se mantendrá (no se rechazará). Debemos proporcionar pruebas para refutarlo, por tanto, es tarea de los fiscales o investigadores probar la hipótesis alternativa (s) que se ha(n) propuesto. Podemos tener múltiples hipótesis alternativas las cuales podemos escribirlas como $\boldsymbol{H}_1$, $\boldsymbol{H}_2$, etc.

Según Fisher (uno de los estadísticos más importantes de la historia) "**una hipótesis nula nunca se prueba ni se establece, pero posiblemente si se refuta en el curso de la experimentación**".

Entonces, en nuestro caso, no importa si el abogado defensor prueba que el defendido es inocente, puede ayudar, pero eso no es lo importante, lo que importa es si el fiscal prueba la culpabilidad del acusado. A menos que demostremos que nuestra hipótesis alternativa ($\boldsymbol{H}_1$ es correcta, no podemos rechazar la hipótesis nula ($\boldsymbol{H}_0$). Entonces, estamos intentando refutar la hipótesis nula para confirmar la alternativa que hemos propuesto. Si no lo hacemos, entonces hemos fallado en rechazar $\boldsymbol{H}_0$, no la hemos probado, hemos fallamos en rechazarla.

Ejemplo 1

En 2004 unos investigadores quisieron probar el impacto de que los comerciales de televisión podrían tener sobre los votantes jóvenes a acudir a emitir sus votos. Para probar el impacto de los comerciales de televisión eligieron 43 mercados de televisión (similares a las ciudades pero un poco más grandes) que verían los comerciales varias veces al día y seleccionaron otros mercados de televisión similares que no verían el comercial. De esa manera pudieron observar si ver el comercial tuvo algún impacto en la cantidad de jóvenes de 18 y 19 años que realmente votaron en las elecciones presidenciales de 2004.

¿Cuál sería la hipótesis nula?

$\boldsymbol{H}_0$: Los comerciales de televisión no aumentaron las tasas de votación de los jóvenes de 18 y 19 años

¿Como podriamos refutar la hipótesis nula?

$\boldsymbol{H}_1$: Los comerciales de televisión aumentaron las tasas de votación de los jóvenes de 18 y 19 años

Analicemos los datos "`RockTheVote`" disponibles en la librería `pcl` en `R`.

```
library(pscl)
library(pander)
data("RockTheVote")

pander(summary(RockTheVote))
```

Cuadro 1: Table continues below

strata	treated	r	n
Min. : 1.00	Min. :0.0000	Min. : 21.0	Min. : 30.0
1st Qu.:10.00	1st Qu.:0.0000	1st Qu.: 83.0	1st Qu.:159.0
Median :20.00	Median :0.0000	Median :109.0	Median :226.0
Mean :20.02	Mean :0.4941	Mean :151.1	Mean :280.8
3rd Qu.:30.00	3rd Qu.:1.0000	3rd Qu.:194.0	3rd Qu.:370.0
Max. :40.00	Max. :1.0000	Max. :718.0	Max. :990.0

p	treatedIndex
Min. :0.2570	Min. : 1.00
1st Qu.:0.4752	1st Qu.:10.00
Median :0.5324	Median :21.00
Mean :0.5304	Mean :20.87
3rd Qu.:0.5946	3rd Qu.:31.00
Max. :0.7804	Max. :42.00

`treated` es una variable numérica dicotómica, es decir, 1 si el mercado de televisión vio los comerciales y 0 si no vio los comerciales. La media aquí indica que el 49,41 % de los mercados de televisión fueron tratados y el resto no fue tratado. En un experimento, los investigadores crearon un grupo de tratamiento (aquellos que vieron los comerciales) y un grupo de control (quienes no vieron los comerciales) con el fin de probar la diferencia.

`r` es el número de jóvenes de 18 y 19 años que votaron en las elecciones de 2004. El mercado de televisión promedio tenía 151 votantes jóvenes registrados que votaron efectivamente en las elecciones.

`n` es el número de votantes registrados entre las edades de 18 y 19 en cada mercado de televisión.

`p` es el porcentaje de votantes registrados entre las edades de 18 y 19 que votaron en la elección, lo que significa que podría calcularse dividiendo `r` por `n`.

Las variables `Strata` y `treatedIndex` no son importantes en este ejemplo. Se eligieron los diferentes mercados de televisión porque eran similares, por lo que hay un mercado que vio el comercio y otro mercado similar que no lo hizo. Entonces, para reafirmar nuestra hipótesis, tenemos la intención de probar si estar en un mercado de televisión que vio comerciales alentando a los jóvenes ir a votar (tratados) afectó las tasas de votación (`p`). La hipótesis nula que intentamos rechazar es que no existe relación entre los tratados y `p`.

Entonces, ¿qué debemos hacer para probar la hipótesis de que estos comerciales de televisión aumentaron las tasas de votación?

Nosotros no sabemos cuál es la población de votantes de 18 y 19 años, pero tenemos un grupo de control, que asumimos que representa a todos los jóvenes de 18 y 19 años. Suponemos que el grupo tratado es una muestra aleatoria de la población de 18 y 19 años por lo que deberían tener exactamente las mismas tasas de votación que todos los demás jovenes entre 18 y 19 años. Sin embargo, vieron los comerciales, así que si hay una diferencia entre los dos grupos podemos atribuirla a los comerciales. Por lo tanto, podemos probar si la tasa media de votación entre los mercados de televisión que fueron tratados con los comerciales difiere significativamente de los que no vieron los comerciales.

Comencemos entonces calculando la tasa de votación media para los dos grupos, los mercados de televisión tratados (vieron comerciales) y el grupo de control (no vieron comerciales). Podemos hacerlo usando el comando `subset()` para dividir `RockTheVote` en dos conjuntos de datos, en función de si el mercado de la televisión estaba en el grupo tratado o no.

```
treatment <- subset(RockTheVote, treated==1)
control <- subset(RockTheVote, treated==0)
mean(treatment$p)
```

```
## [1] 0.5451407
```

```
mean(control$p)
```

```
## [1] 0.5160555
```

La tasa de voto promedio entre los jóvenes de 18 y 19 años para los mercados de televisión que vieron los comerciales es .545 o 54.5 %, y el promedio para los mercados de televisión que no fueron tratados es .516 o 51.6 %. Interesante, la media difiere entre las dos muestras, sin embargo, deberíamos

esperar alguna variación entre las medias ya que estamos tomando diferentes muestras. La pregunta entonces es si la media del grupo de tratamiento difiere significativamente de la media del grupo de control.

Significancia estadística

La significancia estadística es importante. Gran parte de las ciencias sociales se basa en la significancia estadística. Como hemos comentado, las medias de las muestras diferirán un poco de la media de la población y esas medias diferirán en cierto número de desviaciones estándar. Esperamos que la mayoría de los datos caigan dentro de dos veces la desviación estándar por encima (positivamente) o por debajo de la media (negativamente), y que muy pocos caerán más allá de estas magnitudes.

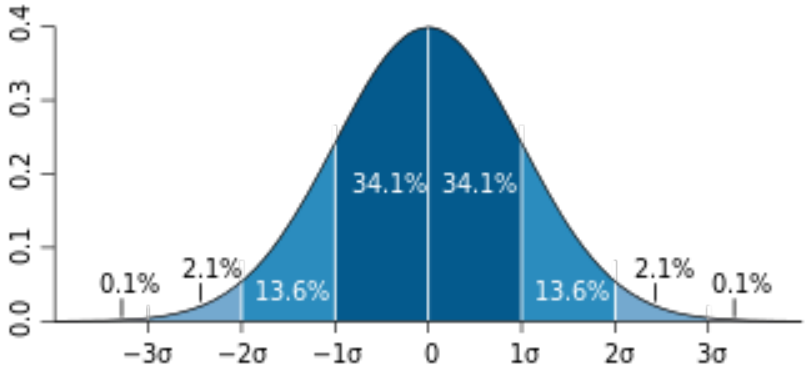


Figura 1: Sacado de Wikipedia

El porcentaje de los datos cae dentro de 1 desviación estándar por encima y por debajo de la media (derecha e izquierda respectivamente). Eso es en ambos lados, por lo que un total del 68.2 por ciento de los datos cae entre una desviación estándar por debajo de la media y una desviación estándar por encima de la media. El 13,6 por ciento de los datos se encuentra entre 1 y 2 desviaciones estándar. En total, esperamos que el 95,4 por ciento de los datos estén dentro de dos desviaciones estándar, ya sea por encima o por debajo de la media

Eso significa, para decirlo de otra manera, que la probabilidad de que la media de una muestra tomada de una población esté dentro de 2 desviaciones estándar es .954, y la probabilidad de que caiga más de la media es solo .046. Eso es bastante improbable. Entonces, si la media del grupo de tratamiento cae más de 2 veces la desviación estándar con respecto de la media del grupo de control, significa que es una muestra extraña o que no es de la misma población.

T-Test

La forma más sencilla de evaluar esto es con lo que se conoce como una prueba t: Esta prueba analiza rápidamente las medias de dos grupos y determina cuántas eveces la desviación estándar están separadas de ellas. Se puede usar una prueba t para probar si una muestra proviene de una determinada población o si dos muestras difieren significativamente. La mayoría de las veces se utiliza para probar si dos muestras son diferentes, generalmente con el objetivo de comprender si alguna política, intervención o rasgo hace que dos muestras sean diferentes y se espera atribuir esa diferencia a lo que estamos probando.

Interpretar la prueba t correctamente es muy importante pero implementarla es muy sencillo mediante el comando `t.test()` en `R`. Entre paréntesis ingresamos los dos conjunto de datos y la variable de interés. Aquí nuestros dos conjuntos de datos se denominan tratamiento y control y la variable de interés es `p`.

```
library(pander)
pander(t.test(treatment$p, control$p))
```

Cuadro 3: Welch Two Sample t-test: `treatment$p` and `control$p`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
1.354	83	0.1794	two.sided	0.5451	0.5161

Veamos ahora "a mano" lo que `R` ha hecho.

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

Figura 2: Sacado de Wikipedia

`x1` y `x2` son las medias de los dos grupos que estamos comparando. En este caso, llamaremos con "1" el grupo de tratamiento y "2" el grupo de control.

```
x1 <- mean(treatment$p)
x2 <- mean(control$p)
x1-x2
```

```
## [1] 0.02908512
```

`s1` y `s2` son las desviaciones estándar para el grupo de tratamiento y control.

```
s1 <- sd(treatment$p)
s2 <- sd(control$p)
```

y `n1` y `n2` son los números de observaciones de la muestra para ambos grupos.

```
n1 <- nrow(treatment$p)
n2 <- nrow(control$p)
```

finalmente hacemos el cálculo:

```r
(mean(treatment$p)-mean(control$p))/
  sqrt(((sd(treatment$p)*sd(treatment$p))
        /nrow(treatment)) + ((sd(control$p)*sd(control$p))/nrow(control)))
```

```
## [1] 1.354094
```