INFERENCIA ESTADÍSTICA

2022-12-13

Alumnos

- Iker Gonzalez
- Miguel Albertí

Ejercicio 1

Enumere cuales son les elementos clave de una prueba de hipótesis y su significado.

Una prueba de hipótesis tiene los siguientes elementos:

- **Hipótesis nula (H0):** Esta hipótesis comúnmente se utiliza para representar el estado actual de las cosas.
- **Hipótesis alternativa (H1):** En contraposición a la H0 nos encontramos H1, esta es la hipótesis en la cual tenemos interés y la cual queremos ver si podemos verificar.
- **Test estadístico:** El test estadístico es una variable aleatoria porque depende de los datos muestrales, hay que saber la distribución de probabilidad o de referencia. Este test es el que nos posibilita la comparación entre H0 y H1
- **Reference distribution:** Es la distribución de H0 en el caso de que sea cierta.
- **p-value**:Nos indica si el test statistic es un valor normal para H0 o no. P-value elevado hace que aceptemos H0 y, por el contrario, un p.value pequeño quiere decir que el test statistic es un valor extremo para esta distribución de referencia, por tanto, H1 debe ser la correcta. Se puede asegurar que como más pequeño sea el p-value más podremos aceptar la hipótesis alternativa.

Ejercicio 2

Un elemento clave en toda prueba de hipótesis es la distribución de referencia del estadístico de la prueba. Diga de que maneras podemos conocer (o aproximar) la distribución de referencia para la prueba de hipótesis de querer ver si una media de una muestra es igual o no a un cierto valor nominal. Especifique como obtendría o cual sería esta distribución de referencia.

A través de un historico de datos, o en caso de no tener acceso a dicho historico, see puede aproximar utilizando una muestra aleatoria calculando su respectiva media y desviación estandar que se utilizarán como parametros de referencia poblacional y con ellos se obtendrá una distribucion de referencia aproximada para realizar la comparativa.

Ejercicio 3

Un amigo me ofrece un piso por valor de 8MPts. Dejando de lado todos los otros factores intervinientes en la decisión de compra de un piso y solo teniendo en cuenta su precio, puedo considerar que se trata de una buena ocasión para comprar?. Que suposiciones necesito hacer para resolver el problema. (Para resolver esta pregunta utilice los datos utilizados en la clase, disponibles en el fichero: preu_3hab.r).

Se debe asumir que el precio es la única referencia a tomar en cuenta en la decisión sin importar las características de los pisos (tamaño, habitaciones, ubicación, superficie, etc).

Los datos históricos de precios son representativos del mercado y son asimétricos siguiendo una distribución normal.

```
bcn_flat_prices <- c(7.80, 12.60, 15.96, 12.75, 13.50, 8.25, 8.25, 15.05, 13.83, 20.61,
                     16.46, 18.00,  15.40, 16.25, 13.39, 25.99, 16.64, 11.03, 16.28, 23.40,
                     14.81, 17.60, 20.21, 18.04, 18.70, 11.10, 15.60, 8.83, 15.05, 16.15,
                     15.00, 11.10, 21.08, 28.66, 21.25, 20.47, 25.53, 10.89, 15.01, 11.78,
                     14.82,  12.17, 14.56, 16.96, 15.50, 14.43, 14.43, 12.71, 31.66, 15.75,
                     15.75, 11.69, 14.52, 17.35, 22.57, 20.00, 13.80, 13.68, 12.61, 19.00,
                     24.61, 16.80, 16.72, 23.95, 16.11, 19.41, 13.99, 16.48, 13.20,  13.47,
                     13.63, 14.76, 16.93, 31.31, 12.81, 21.81, 20.82, 35.77, 15.54, 12.62,
                     13.91, 21.18, 13.72, 12.00, 19.89, 16.46, 32.70, 22.73, 15.51, 16.26,
                     28.70, 18.90, 25.75, 16.89, 13.99, 13.99, 28.22,  20.79, 16.81, 20.25,
                     22.31, 24.03, 15.65, 27.28, 12.60, 17.55, 25.60, 29.44, 27.44, 24.00,
                     30.10, 10.14, 11.31, 11.20, 15.73, 16.90, 25.40, 18.56, 17.55, 21.75,
                     21.68, 24.60, 46.32, 42.12, 21.70, 21.42, 22.31, 31.35, 10.00, 10.92,
                     15.72, 17.10, 15.39, 14.79, 14.11, 14.36, 14.53, 15.11, 18.00, 12.08,
                     12.53, 17.32, 23.96, 22.60, 6.21, 14.54, 16.59, 16.59, 18.26, 19.42,
                     10.66, 16.01, 16.14, 16.00, 10.15, 16.46, 16.20, 14.28, 12.10, 8.40,
                     11.10, 7.52, 13.44, 12.48, 14.72, 8.58, 14.87, 14.35, 13.12, 15.40,
                     15.00, 15.91, 15.20, 14.51, 22.36, 29.29, 22.70, 10.06, 13.89, 14.24,
                     17.64, 14.62, 16.23, 14.60, 15.99, 13.65, 15.69, 22.44, 7.14, 6.76, 11.02,
                     10.65, 12.00, 12.96, 10.71, 11.38, 12.63, 11.11, 11.50, 15.61, 14.80, 15.50,
                     14.51, 14.95, 13.57, 14.80, 11.84, 13.87, 14.43, 16.83, 14.43, 21.12, 22.62,
                     12.75, 13.92, 11.05, 10.56, 13.32, 22.75, 12.68, 17.75)

price_friend_recomendation = 8
number_of_flats = length(bcn_flat_prices)
df_bcn_flat_prices = data.frame(bcn_flat_prices)
number_of_cheaper_flats = sum(df_bcn_flat_prices <= price_friend_recomendation)
p_value = number_of_cheaper_flats / number_of_flats
p_value
```

```
## [1] 0.02262443
```

Tenemos una probabilidad del 2% para encontrar un piso más barato que el que nos ofrece nuestro amigo, deberíamos adquirir el piso ya que será difícil encontrar pisos más baratos.

Ejercicio 4

Sabemos que la media de los precios de los pisos (de 3 habitaciones) en l'Eixample es de (μ= 16.81) y su desviación tipo es de (=) 5.91. Suponiendo que la muestra obtenida en el ejercicio

1 (7.80, 12.60, 15.96, 13.50, 8.25, 31.29, 16.46) es aleatoria. ¿Podemos asegurar de que se trata de una muestra de pisos de l'Eixample?

Vamos a establecer unas hipotesis para poder descubirir si la muestra recogida se trata de pisos de la Eixample.

- H0: muEixample = muMuestra
- H1: muEixample != muMuestra

```
mu_eixample = 16.81
random_sample = c(7.80, 12.60, 15.96, 13.50, 8.25, 31.29, 16.46)

mean_random_sample = mean(random_sample)
mean_random_sample
```

```
## [1] 15.12286
```

Podemos ver que la media poblacional 16.81 y la muestral=15.12 no son muy parecidas por tanto habrá que seguir investigando para saber si se trata de una muestra de pisos de la Eixample.

```
number_random_sample = length(random_sample)
s_random_sample = sd(random_sample)

t <- sqrt(number_random_sample)*(mu_eixample - mean_random_sample)/s_random_sample
pt(t, df=number_random_sample-1)
```

```
## [1] 0.7039106
```

Al obtener un valor de 0,7 podemos aceptar H0 y rechazar H1, es decir, la muestra si que pertenece a la Eixample.

Ejercicio 5

El siguiente año, los precios de una muestra aleatoria de pisos de 3 hab. en l'Eixample han sido 13.57 14.80 22.36 29.29 22.70. Puedo afirmar de que no ha habido cambio de precio entre los dos años?

Para ver si han cambiado de precio vamos a establecer dos hipótesis:

- H0: x1 - x2 = 0
- H1: x1 - x2 > 0

```
random_sample_year_1 = c(7.80, 12.60, 15.96, 13.50, 8.25, 31.29, 16.46)
mean_sample_year_1 = mean(random_sample_year_1)
n1 = length(random_sample_year_1)

random_sample_year_2 = c(13.57, 14.80, 22.36, 29.29, 22.70)
mean_sample_year_2 = mean(random_sample_year_2)
n2 = length(random_sample_year_2)
```

```r
# t_test
s_pool = (n1-1)*var(random_sample_year_1) + (n2-1)*var(random_sample_year_2) / (n1+n2-2)
s_pool = sqrt(s_pool)
t <- (mean_sample_year_2 - mean_sample_year_1)/(s_pool*sqrt((1/n1)+(1/n2)))
pt(t, df=(n1+n2-2), lower.tail=F)
```

```
## [1] 0.324696
```

Aunque tengamos la desviación tipo, la muestra que tenemos es inferior a 30, por tanto, hemos decidido utilizar la t-student.

El resultado obtenido nos indica que aceptamos H0 y descartamos H1.

Ejercicio 6

Calcular el p_valor en para la misma prueba del problema anterior, usando el método de permutaciones.

```r
precio_pisos= c(random_sample_year_1, random_sample_year_2)
dif_per <- NULL
for (i in 1:1000) {rnd <- sample(1:12,5);
                   dif_per[i] = mean(precio_pisos[rnd])-mean(precio_pisos[-rnd])}

sum(dif_per>=1.30)/length(dif_per)
```

```
## [1] 0.399
```

Usando el método de las permutaciones aceptamos H0, es decir, no ha habido un cambio de precio.

Ejercicio 7

Sabemos que la probabilidad de compra de un producto en el canal internet es de 0.02. En un mes se han conectado 2300 visitantes, de los cuales 94 han comprado nuestro producto, puedo pensar que ha habido un incremento en la probabilidad de compra por internet?

- H0: p_true - p_hat = 0
- H1: p_true - p_hat > 0

```r
p_true = 0.02
n = 2300
x = 94
p_hat = x / n

p_hat
```

```
## [1] 0.04086957
```

```
pbinom(x, n, p_true, lower.tail=F)
```

```
## [1] 1.041736e-10
```

```
pnorm(p_hat,mean=p_true,sd=sqrt(p_true*(1-p_true)/n),lower.tail=F)
```

```
## [1] 4.368528e-13
```

El P-value obtenido de 4.368528e-13 es muy bajo, por lo cual podemos rechazar H0, indicando que efectivamente ha habido un incremento en la probabilidad de compra por internet.

Ejercicio 8

Por otro lado, en el mismo mes de la pregunta anterior se ha lanzado una campaña de márqueting directo con un target preseleccionado de 1000 clientes potenciales, obteniendo una respuesta positiva, esto es la compra del producto, en 56 casos. ¿Podemos afirmar que la tasa de respuesta obtenida en el target preseleccionado es mejor que la obtenida por internet?.

En primer lugar estableceremos una hipótesis:

- H0: p_directo > p_online
- H1: p_directo < p_online

```
n1 = 2300
p1_hat = 94 / n1
n2 = 1000
p2_hat = 56 / n2

z = (p1_hat - p2_hat)/sqrt((p1_hat*(1-p1_hat)/n1) + (p2_hat*(1-p2_hat)/n2))
pv <-pnorm(z,lower.tail=F)
list(zval=z, pval=pv)
```

```
## $zval
## [1] -1.809633
##
## $pval
## [1] 0.9648237
```

Observamos que el p_value es bastante elevado y, por tanto, podemos afirmar H1, es decir, que el grupo preseleccionado es mejor que él online. El valor Z obtenido de -1.809633 indica que la probabilidad de compra online está por debajo de la probabilidad de compra del target preseleccionado.

Ejercicio 9

Un día me encuentro con un amigo al que hace tiempo que no veía, va acompañado por un hijo varón. Me dice pero que tiene dos hijos. Cuál es la probabilidad de que su otro hijo sea también varón.

Suponiendo que la probabilidad de tener una niña o un niño son iguales, podemos decir que la probabilidad que tenían para que el primer hijo fuese varón era de 1/2 o 50%. Para saber qué probabilidad hay para que

su segundo hijo también sea varón volvemos a tener un 50% de probabilidades porque en el espacio muestral solo se pueden dar dos casos:

niño y niña

niño y niño

Ya que, sabemos que el primero es un niño y el orden en este caso sí que importa.

Ejercicio 10

Hace mucho tiempo, cuanto la televisión era en blanco y negro, empezó en TVE un programa concurso de gran éxito, se llamaba "Un, dos, tres, responda otra vez". Su primer presentador fue el gran Kiko Ledgard. Una situación típica en dicho programa era cuando al concursante se le ofrecían tres puertas, detrás de una sola de las cuales había el premio. El concursante escogía una de las puertas, y entonces Kiko Ledgard abría una de las dos puertas no escogidas en donde NO había el premio y preguntaba al concursante si quería cambiar de opción (problema de Monty Hall en honor de su creador). ¿Cuál es la mejor opción para el concursante, mantenerse en su primera opción o cambiar de puerta?.

En un primer momento, el concursante se encuentra ante tres puertas de las cuales no tiene ninguna información, teniendo en ese momento una probabilidad de 1/3 de acertar la puerta que tiene el premio. Una vez el presentador abre una de las puertas, la puerta abierta pasa a tener una probabilidad de 0 y para la puerta que no ha seleccionado el concursante pasa a tener una probabilidad de 2/3, por tanto, la respuesta es si, si debería cambiar de puerta.