ANÁLISIS DE UNA MUESTRA DE LA BASE DE DATOS ESTUDIANTIL DE LA UNIVERSIDAD NACIONAL DE COLOMBIA

Autor: JUAN PABLO CHAMORRO BOLAÑOS
CC: 1004580770

La base de datos con la que le corresponde trabajar, se obtiene como una muestra aleatoria de una gran base de datos. La base original corresponde a la información de 465 estudiantes de la Universidad Nacional Sede Medellín. Dicha base contiene las variables: **GENERO** (HOMBRE o MUJER), **ESTATURA** (en cm. del estudiante), **MASA** (en kgr), **ESTRATO** y **FUMA** (SI o NO).

- **¿Es la estatura promedio inferior a 170 cm?**

  Para determinar si la estatura promedio de los estudiantes es inferior a 170 centímetros, se debe realizar una prueba de hipótesis de cola izquierda para la media. Se define la media poblacional como μ, la media muestral como $\overline{x}$, la desviación estándar muestral como S y al tamaño de la muestra como n que es igual a 80. Se plantea el siguiente juego de hipótesis:

$$\begin{cases} H_0 : \mu & = 170 \\ H_1 : \mu & < 170 \end{cases}$$

  Dado que se cuenta con un tamaño de muestra superior a 30, se usa el teorema del límite central, teniendo así por estadístico de prueba

$$Z_c = \sqrt{n}\frac{\overline{x} - 170}{S} \sim n(0,1); \text{ Bajo } H_0 \text{ cierta.}$$

  Teniendo en cuenta que $\overline{x} = 168.45,\ S = 8.880216$ se obtuvo que $Z_c = -1.561181$ y por tanto un valor p obtenido como $P(Z < Z_c)$ igual a 0.05924, por lo que no se rechaza la hipótesis nula puesto que el valor p no es inferior a el nivel de significancia usual del 5%, aunque se recomienda obtener más datos para realizar esta prueba ya que dar una conclusión con un valor p tan cercano al nivel de significancia es delicado.

  Por tanto, no se tiene evidencia muestral suficiente para concluir el hecho de que la estatura media de los estudiantes es inferior a 170 centímetros.

```r
n <- dim(datos)[1]

#pregunta 1
estaturaprom <- mean(datos$ESTATURA) #estatura promedio muestral
estaturadesvi <- sd(datos$ESTATURA) #desviacion estandar muestral de la estatura
muo1 <- 170 #valor de mu bajo H0
Zc1 <- sqrt(n)*(estaturaprom-muo1)/estaturadesvi #valor del estadístico de prueba
pvalue1 <- pnorm(Zc1) #valor p de la prueba
```

- **¿La masa de los hombres es mayor que la de las mujeres?**

  Para probar que la masa de los hombres es mayor que la de las mujeres se debe proceder con una prueba de hipótesis para la diferencia de medias de cola derecha, definiendo al tamaño de muestra de los hombres, al tamaño de muestra de las mujeres, a la masa promedio de los hombres, a la masa promedio de las mujeres, a la masa promedio muestral de los hombres, a la masa promedio muestral de las mujeres, a la varianza muestral de la masa de los hombres y a la varianza muestral de la masa de las mujeres como $n_H,\ n_M,\ \mu_H,\ \mu_M,\ \overline{x}_H,\ \overline{x}_M,\ S_H^2,\ S_M^2$ respectivamente. Los valores de los estadísticos muestrales (medias y varianzas) para hombres y mujeres fueron de 43, 64.35581 y 37, 61.89459 respectivamente.

  Con toda esta información se plantea el siguiente juego de hipótesis.

$$\begin{cases} H_0 : \mu_H - \mu_M = 0 \\ H_1 : \mu_H - \mu_M > 0 \end{cases}$$

  Teniendo en cuenta que ambos tamaños de muestra son superiores a 30, se acude al teorema del límite central, lo cual lleva al siguiente estadístico de prueba.

$$Z_c = \frac{\overline{x}_H - \overline{x}_M}{\sqrt{\frac{S_H^2}{n_H} + \frac{S_H^2}{n_H}}} \sim n(0,1);\ \text{Bajo } H_0 \text{ cierta.}$$

  Cuyo valor fue 1.08123, llevando a un valor p (calculado como $P(Z > Z_c)$) de

```r
#pregunta2
hombres <- datos[datos$GENERO == "HOMBRE",] #filtrando por hombres
mujeres <- datos[datos$GENERO == "MUJER", ] #filtrando por mujeres
n1 <- dim(hombres)[1]
n2 <- dim(mujeres)[1]
meanhombres <- mean(hombres$MASA) #masa media muestral para los hombres
varhombres <- var(hombres$MASA) #varianza muestral masa de los hombres
meanmujeres <- mean(mujeres$MASA) #masa media muestral para las mujeres
varmujeres <- var(mujeres$MASA) #varianza muestral masa de las mujeres
Zc2 <- (meanhombres - meanmujeres)/sqrt(varhombres/n1 + varmujeres/n2) #valor del estadístico de prueba
pvalue2 <- pnorm(Zc2, lower.tail = F) #valor p de la prueba
```

0.1397974, el cual es mayor que el nivel de significancia escogido del 5%. Por ende, no se rechaza la hipótesis nula, lo que indica que no existe evidencia muestral suficiente para concluir que la masa promedio de los hombres es mayor a la masa promedio de las mujeres.

- **¿La proporción de fumadores es inferior a 10%?**

  Para dar respuesta a la pregunta de si la proporción de fumadores es inferior al 10% se debe realizar una prueba de hipótesis de cola izquierda para la proporción. Definiendo la proporción y la proporción muestral como p, $\hat{p}$ respectivamente. El valor para la proporción muestral luego de ser calculada es de 0.05.

  Se realiza el planteamiento de la respectiva prueba de hipótesis.

$$\begin{cases} H_0 : P = 0.1 \\ H_1 : P < 0.1 \end{cases}$$

  Con el tamaño de muestra igual a 80 se usa el estadístico de prueba
  $Z_c = \sqrt{n}\dfrac{P - 0.1}{0.1 \times 0.9} \sim n(0,1);\ \text{Bajo } H_0 \text{ cierta.}$

  Dicho estadístico de prueba tiene un valor igual a -1.490712, lo cual conlleva a un valor p de 0.06801856, lo que lleva a no rechazar la hipótesis nula y concluir que no hay evidencia muestral suficiente para afirmar el hecho de que el porcentaje de personas que fuman es inferior al 10%.

```
#pregunta3
fuma <- mean(datos$FUMA == "SI") #proporcion de fumadores
po <- 0.1 #valor de po
Zc3 <- sqrt(n)*(fuma - po)/sqrt(po*(1-po)) #valor del estadistico de prueba
pvalue3 <- pnorm(Zc3) #valor p de la prueba
```

- **¿La proporción entre el estrato y la cantidad de estudiantes es similar a la siguiente tabla?**

Estrato	1	2	3	4	5	6
Proporción	0.07	0.15	0.38	0.255	0.08	0.065

se consideran los siguientes elementos: Proporción de estudiantes pertenecientes al i-ésimo estrato $p_i,\ 1 \leq i \leq 5$.

Frecuencia observada de estudiantes pertenecientes al i-ésimo estrato $N_i\ 1 \leq i \leq 5$.

Dicho esto, se quiere probar si la distribución de los estudiantes es estadísticamente igual a la de la tabla, por lo que se considera el siguiente juego de hipótesis.

$$\begin{cases} H_0 : F(x) & = F_0(x) \\ H_1 : F(x) & \neq F_0(x) \end{cases}$$

Donde $F_0(x)$ es la distribución de los estratos dada en el enunciado de la pregunta.
Se tienen las siguientes frecuencias observadas.

Estrato	1	2	3	4	5	6
Frecuencia	11	31	27	7	2	2

Luego, se calcula el estadístico de prueba; el cual es

$$\chi_0^2 = \sum_{i=1}^{6} \frac{(N_i - np_i)^2}{np_i} \sim \chi_5^2;\ \text{Bajo } H_0 \text{ cierta}$$

Realizados los cómputos, el estadístico de prueba obtuvo un valor de 49.46693, lo cual conlleva a un valor p igual a 1.781339e-09, lo cual desemboca en el rechazo de la hipótesis nula, es decir hay evidencia muestral suficiente para descartar que la distribución de los estratos es diferente a la asumida como cierta bajo H_0.

```
#pregunta4
estratosbase <-table(datos$ESTRATO) #frecuencias observadas
estratosdoc <- c(0.07, 0.15, 0.38, 0.255, 0.08, 0.065) #proporciones bajo H0
chio2 <- sum((estratosbase - n*estratosdoc)^2/(n*estratosdoc)) #valor del estadistico de prueba
gl <- length(estratosdoc) - 1 #grados de libertad
pvalue4 <- pchisq(chio2, gl, lower.tail = F)  #valor p de la prueba
```

*Nota: como siempre se tuvo tamaños de muestra mayores o iguales a 30, se acudió al teorema del límite central por simplicidad.