Pruebas de Hipótesis

Marzo 13, 2024

Prof. Sergio Béjar

Departamento de Estudios Políticos, CIDE

Objetivos

1. Introducción al paquete `t.test`
2. Introducir a los estudiantes a las pruebas de hipótesis con más de una muestra

Función `qt()` para calcular el valor de *t*

`qt()` calcula el valor de *t* para un determinado intervalo de confianza y grados de libertad de la siguiente forma:

- *t = qt(p = confianza del intervalo + (1-confianza intervalo)/2, df= , lower.tail = TRUE)*

```
qt(p = 0.95 + 0.05/2, df = 15, lower.tail = TRUE)
```

```
## [1] 2.13145
```

Función `t.test()`

Para ver la aplicación vamos a usar la base de datos llamada `sleep` (del paquete `datasets`)
`sleep` es una data frame que contiene 20 observaciones sobre 2 variables:

- **extra:** incremento numérico en horas de sueño.
- **group:** grupo del paciente.

Función `t.test()`

Preparamos los datos:

```
## llamamos la base de datos
sleep <- sleep

## calculamos la media de "extra"
mean(sleep$extra)
```

```
## [1] 1.54
```

Función `t.test()`

Antes de usar la función `t.test` calculemos paso a paso el valor de t en R.

```
mediaExtra <- mean(sleep$extra)
sdExtra <- sd(sleep$extra) # Desviación estándar de extra
seExtra <- sdExtra/(sqrt(20)) # Error estándar de extra
```

```
valor_t <- (mediaExtra - 3)/seExtra    ## fórmula para calcular t
valor_t
```

```
## [1] -3.235668
```

Función `t.test()`

Ahora calculamos el p-value :

```
2 * (1 - pt(abs(valor_t), df = 20 - 1))
```

```
## [1] 0.004351604
```

El valor calculado es menor a p = .05. Por lo tanto aceptamos la hipothesis alternativa H_a.

Función `t.test()`

`t.test` es una función que calcula una *prueba-t* en vectores de datos que pueden ser de una o dos muestras. Todo el procedimiento que hicimos en las 2 diapositivas anteriores lo podemos reducir a una línea de código.

Función `t.test()`

```r
t.test(sleep$extra, mu = 3)
```

```
##
##  One Sample t-test
##
## data:  sleep$extra
## t = -3.2357, df = 19, p-value = 0.004352
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  0.5955845 2.4844155
## sample estimates:
## mean of x
##      1.54
```

Función `z.test()`

Para calcular el valor de z podemos utilizar la función `z.test` del paquete `BSDA`

```
library(BSDA) ## carga el paquete
data <- c(26, 25, 10, 34, 30, 23, 28, 29, 25, 27)  ## creo un vector de
z_test <- z.test(data, mu = 24,sigma.x=10) ## función z.test
```

Función `z.test()`

```r
print(z_test) ## imprimo resultados
```

```
##
##  One-sample z-Test
##
## data:  data
## z = 0.53759, p-value = 0.5909
## alternative hypothesis: true mean is not equal to 24
## 95 percent confidence interval:
##  19.50205 31.89795
## sample estimates:
## mean of x
##      25.7
```

Pruebas de Hipótesis para 2 Muestras Independientes

Hasta el momento todas las pruebas de hipótesis que hemos realizado han sido para una sola muestra (i.e. estamos comparando μ con $\bar{x}$).
En ocasiones, sin embargo, nos puede interesar investigar si las medias de dos muestras **independientes** son significativamente diferentes la una de la otra.
¿Qué podemos hacer al respecto?

- Si las observaciones en las muestras independientes se distribuyen normalmente, podemos calcular una *prueba-t* de *independencia*

Pruebas de Hipótesis para 2 Muestras Independientes

Si asumimos que la medida poblacional x esta normalmente distribuida con media μ_x y varianza σ_x^2, que la media poblacional de y se distrubuye normalmente con media μ_y y varianza σ_y^2, la fórmula para calcular el valor t es:

$$t = \frac{\bar{x} = \bar{y}}{e.s.(\bar{x} - \bar{y})} \tag{1}$$

Pruebas de Hipótesis para 2 Muestras Independientes

donde,

$$e.s.(\bar{x} - \bar{y}) = \frac{\sigma_x}{\sqrt{n_x}} + \frac{\sigma_y}{\sqrt{n_y}} \tag{2}$$

Pruebas de Hipótesis para 2 Muestras Independientes

Ejemplo: La siguiente tabla muestra estadísticas descriptivas de duración de gobiernos por tipo de gobierno (i.e. moyoritario o minoritario).

Tipo de Gobierno	# Observaciones	Media de Duración	D. E.
Mayoritario	124	930.5	466.1
Minoritario	53	674.4	421.4
Combinados	177	853.8	467.1

Pregunta: ¿Hay diferencia significativa en la duración de gobiernos mayoritarios y minoritarios?

Pruebas de Hipótesis para 2 Muestras Independientes

Información de la tabla anterior

```
may_bar <- 930.5 ## media gob. mayoritario
min_bar <- 674.4 ## media gob. minoritario
ds_may <- 466.1 ## desviacion estándar moyorit.
ds_min <- 421.4
n_may <- 124
n_min <- 53
```

Pruebas de Hipótesis para 2 Muestras Independientes

Calculamos el error estándar de (X - Y):

```
es_may_min <- sqrt((ds_may^2/n_may) + (ds_min^2/n_min)) ## ver fórmula
es_may_min
```

```
## [1] 71.43205
```

Ahora calculamos el valor de *t*:

```
t_calc <- (may_bar - min_bar)/es_may_min ## ver fórmula 1
t_calc
```

```
## [1] 3.585226
```

Para calcular grados de libertad usamos las observaciones de la muestra más pequena (g.l. en este caso = 52)

Pruebas de Hipótesis para 2 Muestras Independientes

El valor de t para una $p = .05$ es **1.676**

```
qt(p = 0.95, df = 52, lower.tail = TRUE)
```

```
## [1] 1.674689
```

```
2*pt(-abs(t_calc), df = 52) # p dos colas
```

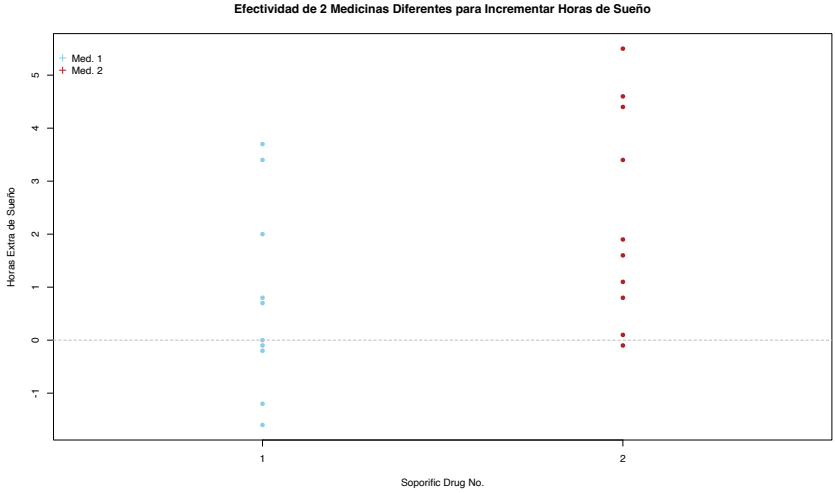
```
## [1] 0.0007429797
```

Regresamos al Estudio del Sueño

Vamos a asumir que el tiempo extra de sueño de cada individuo cuando usa la medicina se distribuye normalmente. Entonces, lo que nos interesa saber es si la diferencia en las horas promedio de sueño para cada medicina, esto es μ_1 y μ_2, es diferente de cero. En otras palabras, ¿Es una medicina mejor que la otra para incrementar el tiempo promedio de sueño?
Formalmente:

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_a : \mu_1 - \mu_2 \neq 0$

Regresamos al Estudio del Sueño



Regresamos al Estudio del Sueño

```
pander(t.test(extra ~ group, data = sleep, mu = 0,
              alternative = "two.sided",
              conf.level = 0.95),
              caption="Prueba t para muestras independientes",
              split.table=Inf)
```

Table 2: Prueba t para muestras independientes

Test statistic	df	P value	Alternative hypothesis	mean in group 1	mean in group 2
-1.861	17.78	0.07939	two.sided	0.75	2.33

Regresamos al Estudio del Sueño

¿Cuál es nuestra conclusión si p estimado es $< .05$?

Pruebas de Hipótesis para 2 Muestras NO Independientes

Ahora asumamos que las muestras No son independientes. Comparamos datos para la misma población en diferentes puntos en el tiempo. Por ejemplo, tus calificaciones en los examenes 1 y 2 del semestre.

Ejemplo: El Profe quiere saber si hay diferencias en el aprovechamiento de los estudiantes entre el exámen 1 y el exámen 2.

Pruebas de Hipótesis para Muestras NO Independientes

```
estudiante <- c("es1", "es2", "es3", "es4", "es5", "es6", "es7", "es8",
examen1 <- c(99, 98, 67, 68, 70, 71, 72, 88, 75, 83)
examen2 <- c(94, 93, 62, 63, 65, 66, 67, 83, 70, 76)
examen <- data.frame(estudiante, examen1, examen2)
examen
```

```
##    estudiante examen1 examen2
## 1         es1      99      94
## 2         es2      98      93
## 3         es3      67      62
## 4         es4      68      63
## 5         es5      70      65
## 6         es6      71      66
## 7         es7      72      67
## 8         es8      88      83
## 9         es9      75      70
```

Función `var.test()`

Primero checamos si la varianza entre los dos grupos es diferente o no.

```
pander(var.test(examen$examen1, examen$examen2),
       caption = "Prueba F de Varianza")
```

Table 3: Prueba F de Varianza (continued below)

Test statistic	num df	denom df	P value	Alternative hypothesis
1.009	9	9	0.9895	two.sided

ratio of variances
1.009

Función `var.test()`

El valor de $p > \alpha = .05$. Por lo tanto ACEPTAMOS H_0. Es decir, la diferencia entre las varianzas de 1 y 2 es 0 con el 95% de confiabilidad.

Función `t.test()` Muestras NO Independientes

Las hipótesis en este caso son:

- H_0 : No hay diferencia en las calificaciones de los estudiantes entre los exámenes 1 y 2.
- H_a : Hay diferencia en las calificaciones de los estudiantes entre los exámenes 1 y 2.

Función `t.test()` Muestras NO Independientes

```r
t.test(examen$examen1, examen$examen2,
               paired = TRUE,
               var.equal = TRUE)
```

```
##
##  Paired t-test
##
## data:  examen$examen1 and examen$examen2
## t = 26, df = 9, p-value = 8.884e-10
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  4.747569 5.652431
## sample estimates:
## mean difference
##             5.2
```

Función `t.test()` Muestras NO Independientes

La conclusión es que hay diferencias significativas en la calificaciones de los estudiantes en los exámenes 1 y 2.
¿Por qué?

Conclusión

- Estamos listos para el primer exámen del semestre.

Table of Contents