Lectura 5

Inferencia con medias muestrales

Prueba z

Como ya adelantamos, la prueba Z es adecuada para inferir acerca de las medias con una o dos muestras, aunque aquí solo veremos el primer caso. Para poder usarla, debemos **verificar el cumplimiento** de algunas condiciones, muchas de las cuales están asociadas al modelo normal que conocimos en el capítulo anterior:

- La muestra debe tener al menos 30 observaciones. Si la muestra tiene menos de 30 observaciones, se debe conocer la varianza de la población.
- Las observaciones deben ser independientes, es decir que la elección de una observación para la muestra no influye en la selección de las otras.
- La población de donde se obtuvo la muestra sigue aproximadamente una distribución normal.

Esta prueba resulta adecuada si queremos **asegurar o descartar** que la media de la población tiene un cierto **valor hipotético**

Ejemplo de tabla

Obs.	Utilidad [M$]	Obs.	Utilidad [M$]	Obs.	Utilidad [M$]	Obs.	Utilidad [M$]
1	19,33	6	22,22	11	22,55	16	29,68
2	29,37	7	31,26	12	20,69	17	29,27
3	29,14	8	26,92	13	24,68	18	26,72
4	32,10	9	31,40	14	28,74	19	27,08
5	25,04	10	17,66	15	26,85	20	20,62

Script 5.1: prueba Z para una muestra.

```r
library(TeachingDemos)
library(ggpubr)

# Ingresar los datos.
muestra <- c(19.33, 29.37, 29.14, 32.10, 25.04, 22.22, 31.26, 26.92,
             31.40, 17.66, 22.55, 20.69, 24.68, 28.74, 26.85, 29.68,
             29.27, 26.72, 27.08, 20.62)

# Establecer los datos conocidos.
desv_est <- 2.32
n <- length(muestra)
valor_nulo <- 20

# Crear gráfico Q-Q para verificar la distribución de la muestra,
datos <- data.frame(muestra)

g <- ggqqplot(datos, x = "muestra", color = "SteelBlue")
print(g)

# Verificar distribución muestral usando la prueba de normalidad
# de Shapiro-Wilk.
normalidad <- shapiro.test(muestra)
print(normalidad)

# Fijar un nivel de significación.
alfa <- 0.01

# Calcular la media de la muestra.


cat("\tPrueba Z para una muestra\n\n")
media <- mean(muestra)
cat("Media =", media, "M$\n")

# Calcular el estadístico de prueba.
Z <- (media - valor_nulo) / desv_est
cat("Z =", Z, "\n")

# Calcular el valor p.
p <- 2 * pnorm(Z, lower.tail = FALSE)
cat("p =", p, "\n")

# Hacer la prueba Z con R.
prueba <- z.test(media, mu = valor_nulo, alternative = "two.sided",
                 stdev = desv_est, conf.level = 1-alfa)

print(prueba)
```

Prueba t

Aunque la prueba t no opera bajo el control de supuesto de normalidad, aún así requiere verificar algunas condiciones para poder usarla:

1) La observaciones son independientes entre sí
2) Las observaciones provienen de una distribución cercana a la normal

->No limitan el tamaño de la muestra para que sea mayor a 30, e sdecir permite su uso para muestras pequeñas

Ejemplo de tablas

Obs.	t [ms]	Obs.	t [ms]	Obs.	t [ms]
1	411,5538	6	388,6731	11	418,1169
2	393,2753	7	430,0382	12	408,4110
3	445,8905	8	469,4734	13	463,3733
4	411,4022	9	409,5844	14	407,0908
5	498,8969	10	442,0800	15	516,5222

Script 5.2: prueba t para una muestra.

```r
library(ggpubr)

# Cargar los datos.
tiempo <- c(411.5538, 393.2753, 445.8905, 411.4022, 498.8969,
            388.6731, 430.0382, 469.4734, 409.5844, 442.0800,
            418.1169, 408.4110, 463.3733,407.0908, 516.5222)

# Establecer los datos conocidos.
n <- length(tiempo)
grados_libertad <- n - 1
valor_nulo <- 500


# Verificar si la distribución se acerca a la normal.
g <- ggqqplot(data = data.frame(tiempo),
              x = "tiempo",
              color = "steelblue",
              xlab = "Teórico",
              ylab = "Muestra",
              title = "Gráfico Q-Q muestra v/s distr. normal")

print(g)

# Fijar un nivel de significación.
alfa <- 0.025

# Calcular el estadístico de prueba.
cat("\tPrueba t para una muestra\n\n")
media <- mean(tiempo)
cat("Media =", media, "M$\n")
desv_est <- sd(tiempo)
error <- desv_est / sqrt(n)
t <- (media - valor_nulo) / error
cat("t =", t, "\n")
```

```r
# Calcular el valor p.
p <- pt(t, df = grados_libertad, lower.tail = TRUE)
cat("p =", p, "\n")

# Construir el intervalo de confianza.
t_critico <- qt(alfa, df = grados_libertad, lower.tail = FALSE)
superior <- media + t_critico * error
cat("Intervalo de confianza = (-Inf, ", superior, "]\n", sep = "")

# Aplicar la prueba t de Student con la funcuón de R.
prueba <- t.test(tiempo,
                 alternative = "less",
                 mu = valor_nulo,
                 conf.level = 1 - alfa)

print(prueba)
```

Prueba t para dos muestras pareadas

Ejemplo de tablas

instancia	t_A [ms]	t_B [ms]	dif [ms]
1	436,5736	408,5142	28,0594
2	470,7937	450,1075	20,6862
3	445,8354	490,2311	-44,3957
4	470,9810	513,6910	-42,7100
5	485,9394	467,6467	18,2927
6	464,6145	484,1897	-19,5752
7	466,2139	465,9334	0,2805
8	468,9065	502,6670	-33,7605
9	473,8778	444,9693	28,9085
10	413,0639	456,3341	-43,2702
11	496,8705	501,1443	-4,2738
12	450,6578	471,7833	-21,1255
13	502,9759	441,1206	61,8553
14	465,6358	544,1575	-78,5217
15	437,6397	447,8844	-10,2447
16	458,8806	432,4108	26,4698
17	503,1435	477,1712	25,9723
18	430,0524	482,4828	-52,4304

instancia	t_A [ms]	t_B [ms]	dif [ms]
19	438,5959	458,2536	-19,6577
20	439,7409	474,9863	-35,2454
21	464,5916	496,0153	-31,4237
22	467,9926	485,8112	-17,8186
23	415,3252	457,4253	-42,1001
24	495,4094	483,3700	12,0394
25	493,7082	510,7131	-17,0049
26	433,1082	467,5739	-34,4657
27	445,7433	482,5621	-36,8188
28	515,2049	453,5986	61,6063
29	441,9420	385,9391	56,0029
30	472,1396	548,7884	-76,6488
31	451,2234	467,2533	-16,0299
32	476,5149	494,7049	-18,1900
33	440,7918	451,9716	-11,1798
34	460,1070	522,3699	-62,2629
35	450,1008	444,1270	5,9738

```r
# Cargar los datos.
instancia <- seq(1, 35, 1)

t_A <- c(436.5736, 470.7937, 445.8354, 470.9810, 485.9394,
         464.6145, 466.2139, 468.9065, 473.8778, 413.0639,
         496.8705, 450.6578, 502.9759, 465.6358, 437.6397,
         458.8806, 503.1435, 430.0524, 438.5959, 439.7409,
         464.5916, 467.9926, 415.3252, 495.4094, 493.7082,
         433.1082, 445.7433, 515.2049, 441.9420, 472.1396,
         451.2234, 476.5149, 440.7918, 460.1070, 450.1008)

t_B <- c(408.5142, 450.1075, 490.2311, 513.6910, 467.6467,
         484.1897, 465.9334, 502.6670, 444.9693, 456.3341,
         501.1443, 471.7833, 441.1206, 544.1575, 447.8844,
         432.4108, 477.1712, 482.4828, 458.2536, 474.9863,
         496.0153, 485.8112, 457.4253, 483.3700, 510.7131,
         467.5739, 482.5621, 453.5986, 385.9391, 548.7884,
         467.2533, 494.7049, 451.9716, 522.3699, 444.1270)

diferencia <- t_A - t_B

# Verificar si la distribución se acerca a la normal.
normalidad <- shapiro.test(diferencia)
print(normalidad)

# Fijar un nivel de significación.
alfa <- 0.05

# Aplicar la prueba t de Student a la diferencia de medias.
prueba_1 <- t.test(diferencia,
                   alternative = "two.sided",
                   mu = valor_nulo,
                   conf.level = 1 - alfa)

print(prueba_1)

# Otra alternativa puede ser aplicar la prueba t de Student
# para dos muestras pareadas.
prueba_2 <- t.test(x = t_A,
                   y = t_B,
```

```
                    paired = TRUE,
                    alternative = "two.sided",
                    mu = valor_nulo,
                    conf.level = 1 - alfa)

print(prueba_2)
```

Prueba t para muestras independientes

En este caso, la prueba t se usa para comparar las medias de dos poblaciones en que las observaciones con que se cuenta no tienen relación con ninguna de las otras observaciones, ni influyen en su selección, ni en la misma ni en la otra muestra. En este caso la inferencia se hace sobre la diferencia de las medias: $\mu_1 - \mu_2 = d_0$, donde d_0 es un valor hipotético fijo para la diferencia. Usualmente se usa $d_0 = 0$, en cuyo caso las muestras podrían provenir de dos poblaciones distintas con igual media, o desde la misma población. Para ello, la prueba usa como estimador puntual la diferencia de las medias muestrales $(\overline{x}_1 - \overline{x}_2)$. Así, el estadístico T en este caso toma la forma de la ecuación 5.3.

$$T = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{SE_{(\overline{x}_1 - \overline{x}_2)}} \tag{5.3}$$

Al usar la distribución t Student para la diferencia de medias, se deben cumplir los siguientes requisitos:

1) Cada una de las muestras cumple con las condiciones para usar la distribución t
2) Las muestras son independientes entre sí

Ejemplo de tablas

Anticuerpos [mg/ml]	
Vacuna A	Vacuna B
6,04	5,32
19,84	3,31
8,62	5,68
13,02	5,73
12,20	4,86
14,78	5,68
4,53	2,93
26,67	5,48
3,14	6,10
19,14	2,56
10,86	7,52
13,13	7,41
6,34	4,02
11,16	
7,62	

Tabla 5.4: Concentración de anticuerpos de los pacientes vacunados.

Script 5.4: prueba t para dos muestras independientes.

```r
library(ggpubr)

# Cargar los datos.
vacuna_A <- c(6.04, 19.84, 8.62, 13.02, 12.20, 14.78, 4.53, 26.67,
              3.14, 19.14, 10.86, 13.13, 6.34, 11.16, 7.62)

vacuna_B <- c(5.32, 3.31, 5.68, 5.73, 4.86, 5.68, 2.93, 5.48, 6.10,
              2.56, 7.52, 7.41, 4.02)

# Verificar si las muestras se distribuyen de manera cercana
# a la normal.
normalidad_A <- shapiro.test(vacuna_A)
print(normalidad_A)
normalidad_B <- shapiro.test(vacuna_B)
print(normalidad_B)

# Fijar un nivel de significación.
alfa <- 0.01

# Aplicar la prueba t para dos muestras independientes.
prueba <- t.test(x = vacuna_A,
                 y = vacuna_B,
                 paired = FALSE,
                 alternative = "greater",
                 mu = 0,
                 conf.level = 1 - alfa)

print(prueba)

# Calcular la diferencia entre las medias.
media_A <- mean(vacuna_A)
media_B <- mean(vacuna_B)
diferencia <- media_A - media_B
cat("Diferencia de las medias =", diferencia, "[mg/ml]\n")
```