# Actividad8_pruebas_hipotesis

Rodolfo Jesús Cruz Rebollar

2024-08-25

## Problema 1. Enlatados

Los pesos de 21 latas de duraznos empacados elegidas al azar fueron:

Peso de las latas: 11, 11.6, 11.6, 11.7, 10.9, 11.6, 12, 11.2, 11.5, 12, 12, 11.4, 11.2, 10.8, 10.5, 11.8, 12.2, 10.9, 11.8, 11.4, 12.1

Por estudios anteriores se saber que población del peso de las latas se distribuye normalmente.

Si a los dueños no les conviene que el peso sea menor, pero tampoco mayor a 11.7, prueba la afirmación de que el verdadero peso de las latas es de 11.7 con un nivel de confianza de 0.98 haciendo uso de los datos obtenidos en la muestra.

### Paso 1. Hipótesis

$$H_0: \mu = 11.7$$

$$H_1: \mu \neq 11.7$$

¿Cómo se distribuye $\bar{X}$?

- X se distribuye como una normal.
- n < 30
- No conocemos sigma.

Entonces, la distribución muestral es una t de Student.

### Paso 2. Regla de decisión

- Nivel de confianza es de 0.98 ($1 - \alpha = 0.98$)
- Nivel de significancia es de 0.02 ($\alpha = 0.02$)

Se necesita encontrar a cuántas desviaciones estándar está lejos el valor frontera.

```
# Cantidad de datos de la muestra

n = 21

# Nivel de significación de la prueba de hipótesis
```

```r
alfa = 0.02

# Calcular valor frontera del estadístico t para definir la regla de
decisión

t_f = qt(alfa / 2, n - 1)

# el valor frontera está en valor absoluto porque representa la distancia
a la que el
# estadístico t tiene que estar de la media especificada en la hipótesis
nula, ya sea
# por encima o pot debajo de la misma, para rechazar la hipótesis nula h0

cat("t frontera = ", abs(t_f))

## t frontera =  2.527977
```

*Regla de decisión*

Rechazo $H_0$ si:

- $|t_e| > 2.53$
- valor p < 0.02

## Paso 3. Análisis del resultado

- $t_e$: Número de desviaciones estándar al que $\bar{X}$ se encuentra lejos de $\mu = 11.7$
- Valor p: Probabilidad de obtener lo que se obtuvo en la muestra o un valor más extremo

*Estadístico de prueba*

```r
# Datos del peso de 21 latas

X = c(11, 11.6, 11.6, 11.7, 10.9, 11.6, 12, 11.2, 11.5, 12, 12, 11.4,
11.2, 10.8, 10.5, 11.8, 12.2, 10.9, 11.8, 11.4, 12.1)

# Media de la variable aleatoria X (peso promedio de las latas)

xb = mean(X)

# Desviación estándar del peso de las latas

s = sd(X)

# Media referente al verdadero peso promedio de las latas

miu = 11.7
```

```r
# Valor estimado del estadístico de prueba t

te = (xb - miu) / (s / sqrt(n))

# Mostrar el valor estimado para el estadístico de prueba t

cat("te = ", te)
```

```
## te =  -2.068884
```

```r
# Calcular el valor p de la prueba en base a la estimación del
estadístico de prueba t

# Con 20 grados de libertad (21 - 1)

valorp = 2 * pt(te, n - 1)

# Mostrar valor p de la prueba

cat("valor p = ", valorp)
```

```
## valor p =  0.0517299
```

### Paso 4. Conclusión

Comparar: Regla de decisión vs Análisis del resultado

Entonces:

- $|t_e| = 2.07 < 2.53$ -> No RH0
- valor p = 0.05 > 0.02 -> No RH0

En el contexto del problema, las latas de durazno tienen el peso requerido, ya que en ambos casos de la regla de decisión, no se rechaza $H_0$, lo cual implica que el peso de las latas es efectivamente de 11.7 ($\mu = 11.7$).

*Más fácil para análisis del resultado*

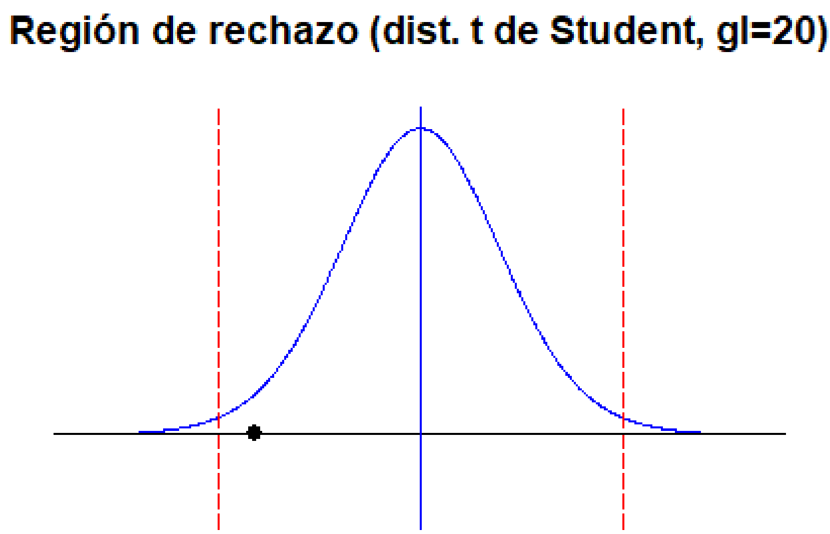
```r
# Alternativa t.test para realizar prueba de hipótesis con t de student

# mu=11.7 -> hipótesis nula
# "two.sided" -> prueba de 2 colas
# conf.level -> nivel de confianza (1 - significancia)

t.test(X, mu = 11.7, alternative = "two.sided", conf.level = 0.98)
```

```
##
##  One Sample t-test
##
```

```
## data:  X
## t = -2.0689, df = 20, p-value = 0.05173
## alternative hypothesis: true mean is not equal to 11.7
## 98 percent confidence interval:
##  11.22388 11.74755
## sample estimates:
## mean of x
##  11.48571
```

### Gráfico de la regla de decisión y el estadístico de prueba

```
# Se grafica una secuencia de números para x que abarque 4 desviaciones
estándar alrededor de la media (se ejemplifica con la t de student) con
su respectivo valor de y:

sigma = sqrt((n - 1) / (n - 3))

x = seq(-4*sigma,4*sigma,0.01)

y = dt(x,n-1)

plot(x, y, type="l", col="blue",
     xlab="", ylab="", ylim=c(-0.1,0.4), frame.plot=FALSE,
     xaxt="n",yaxt="n",main="Región de rechazo (dist. t de Student,
gl=20)")

# Para indicar la zona de rechazo (se ejemplifica con dos colas) y la
media (para la t de Student la miu = 0):

abline(v=t_f,col="red",lty=5)

abline(v=-1*t_f,col="red",lty=5)

abline(h = 0)

abline(v = 0, col = "blue", pch = 19)

points(miu,0,col="blue",pch=19)

# Para dibujar el estadístico de prueba, insertar un punto o una recta
punteada de diferente color a la zona de rechazo, para insertar un punto:

points(te, 0, pch=19, cex=1.1)
```




## Problema 2: La decisión de Fowle Marketing Research, Inc.

Fowle Marketing Research, Inc., basa los cargos a un cliente bajo el supuesto de que las encuestas telefónicas (para recopilación de datos) pueden completarse en un tiempo medio de 15 minutos o menos. Si el tiempo es mayor a 15 minutos entonces se cobra una tarifa adicional. Compañías que contratan estos servicios piensan que el tiempo promedio es mayor a lo que especifica Fowle Marketing Research Inc. así que realizan su propio estudio en una muestra aleatoria de llamadas telefónicas y encuentran los siguientes datos:

Tiempo: 17, 11, 12, 23, 20, 23, 15, 16, 23, 22, 18, 23, 25, 14, 12, 12, 20, 18, 12, 19, 11, 11, 20, 21, 11, 18, 14, 13, 13, 19, 16, 10, 22, 18, 23

Por experiencias anteriores, se sabe que $\sigma = 4$ minutos. Usando un nivel de significación de 0.07, ¿está justificada la tarifa adicional?

### Paso 1. Hipótesis

$H_0: \mu = 15$ minutos

$H_1: \mu > 15$ minutos

- Se conoce el valor de la desviación estándar $\sigma$.
- El tamaño de la muestra n es mayor a 30 elementos (35).

- X: tiempo en minutos que las encuestas telefónicas tardan en completarse.
- Nivel de significancia: $\alpha = 0.07$

Dado que en este problema se conoce el valor de la desviación estándar $\sigma$ y además el tamaño de la muestra n es de 35 datos, lo cual es mayor a 30 datos, es necesario utilizar una distribución normal para resolver este problema, misma que tendrá como parámetros: $\mu = 15$ y $\sigma = 4$, motivo por el cual, el estadístico estimado $\bar{X}$ se distribuye de la siguiente manera para este problema en particular:

$$\bar{X} \sim N\left(\mu = 15, \sigma = \frac{4}{\sqrt{(35)}}\right)$$

## Paso 2. Regla de decisión

Para definir la regla de decisión para este problema, es necesario considerar que el estadístico de prueba a emplear para la prueba de hipótesis es $Z_0$, mismo que se calcula por medio de la siguiente expresión:

$$Z_0 = \frac{\bar{x_0} - \mu}{\sigma_{\bar{x}}}$$

```
# Guardar datos de tiempo en una lista

tiempo = c(17, 11, 12, 23, 20, 23, 15, 16, 23, 22, 18, 23, 25, 14, 12,
12, 20, 18, 12, 19, 11, 11, 20, 21, 11, 18, 14, 13, 13, 19, 16, 10, 22,
18, 23)

# Definir media miu (en base a H0)

miu = 15

# Definir desviación estándar sigma_xbarra (del estadístico estimado)

sigma_barra = 4 / sqrt(length(tiempo))

# Definir valor de x0 barra para calcular Z0

x0_barra = qnorm(0.07, miu, sigma_barra)

cat("x0 barra = ", x0_barra, "\n")

## x0 barra =  14.00218

# Calcular el valor del estadístico de prueba Z0

# Tomar valor absoluto de Z0 ya que es una distancia y las distanias no
pueden
# ser negativas
```

```
Z0 = abs((x0_barra - miu) / sigma_barra)

cat("Estadístico Z_0 = ", Z0)

## Estadístico Z_0 =  1.475791
```

En base a los dos resultados previos, la regla de decisión para este problema es la siguiente:

**Regla de decisión**

Se rechaza $H_0$ si:

- $|Z| > 1.475791$
- valor p < 0.07 (nivel de significancia de la prueba)

## Paso 3. Análsis del resultado

$Z$: medición de la lejanía a la que la media muestral $x^*$ está lejos de la media $\mu$ bajo el supuesto de que $H_0$ es verdadera.

Valor p: probabilidad de que se obtenga ya sea el valor del estadístico $x^*$ o algún otro valor extremo, al suponer que $H_0$ es verdadera.

```
# Calcular la media muestral para calcular a su vez el estadístico de
# prueba de la muestra

time_mean = mean(tiempo)

# Usar la media de los datos de tiempo para calcular el valor del
estadístico de prueba
# de la muestra (Z) y compararlo con el valor frontera del estadístico
(Z0)

est_z = (time_mean - miu) / sigma_barra

cat("Z = ", est_z, "\n")

## Z =  2.95804

# Calcular el valor de probabilidad correspondiente al hecho de que la
variable
# aleatoria X (tiempo que tardan las encuestas telefónicas en
completarse) tome
# un valor por encima de la media muestral time_mean

cat("P(X > xbarra*) = ", pnorm(time_mean, miu, sigma_barra, lower.tail =
FALSE))
```

```
## P(X > xbarra*) =  0.00154801
```

En este caso, dado que el valor p es igual al complemento de la probabilidad de que el tiempo que tardan las encuestas telefónicas en completarse sea inferior al tiempo promedio en el que dichas encuestas se completan, por lo cual, el valor p de la prueba en pocas palabras es igual a la probabilidad de que el tiempo que demora completar las encuestas telefónicas sea superior a 15 minutos, por lo que considerando lo anterior, el valor p es igual a 0.00154801, mientras que por otro lado, la estimación del estadístico de prueba para la muestra analizada arroja un valor de 2.95804.

## Paso 4. Conclusión

A manera de conclusión, retomando la regla de decisión planteada anteriormente, se tiene que:

Se rechaza $H_0$ en caso de que:

- $|Z| > 1.475791$ -> 2.95804 > 1.475791 -> Rechazar $H_0$
- valor p < 0.07 -> 0.00154801 < 0.07 -> Rechazar $H_0$

Motivo por el cual, tomando en cuenta el resultado de la estimación del estadístico de prueba para la muestra (Z), además del valor p obtenido en la prueba de hipótesis, es posible observar que en cuanto al estadístico Z para la muestra, éste mismo tiene un valor de 2.95804 lo cual es superior a la cota máxima permitida que es de 1.475791, motivo por el cual, bajo éste criterio no se rechaza $H_0$. Además, en cuanto al valor p de la prueba de hipótesis, éste fue de 0.00154801, lo cual es inferior al nivel de significancia de 0.07, por lo tanto, bajo éste otro criterio también se rechaza $H_0$, por lo cual, dado que mediante ambos criterios la hipótesis nula $H_0$ se rechaza, entonces se puede concluir que en base a la regla de decisión completa (incluyendo ambos criterios antes descritos), $H_0$ se rechaza, lo cual implica que dentro del contexto del problema a resolver, el tiempo promedio que los usuarios tardan en completar las encuestas telefónicas es mayor a 15 minutos ($\mu > 15$), por lo cual, se concluye que bajo un nivel de significación de 0.07, la tarifa adicional sí está justificada.

**Gráfico de la regla de decisión y el estadístico de prueba**

```r
# Actualizar el valor de n con el total de datos del problema 2

n = length(tiempo)

# Vector de números desde -4sigma hasta 4 sigma con incremento de 0.01

x = seq(-4 * sigma_barra, 4 * sigma_barra, 0.01)

# Evaluar valores de x en la función de densidad de la distribución
normal

# Estandarizar el valor del parámetro estimado xbarra restandole la media
```

```r
miu y
# dividiendo el resultado entre la desviación estándar de

y = dnorm(x, sigma_barra)

# Graficar valores de x y su correspondiente valor de densidad,
involucrando una
# distribución normal con media 15 y desviación estándar de 4/sqrt(35)

plot(x, y, type="l", col="blue", xlab="", ylab="",
     ylim = c(0, 0.4), xlim = c(-3, 3), frame.plot=FALSE, xaxt = "n",
yaxt="n",
     main="Región de rechazo (dist normal N(0, 4/sqrt(35)))")

# Para indicar la zona de rechazo (se ejemplifica con dos colas) y la
media para la
# normal es de 15

abline(v = Z0 + sigma_barra, col = "red", lty = 5)

abline(v = -Z0 + sigma_barra, col = "red", lty = 5) # Test de dos colas

abline(v = 0 + sigma_barra, col = "blue", pch = 19)

abline(h = 0)

# Para dibujar el estadístico de prueba, insertar un punto o una recta
punteada de diferente color a la zona de rechazo, para insertar un punto:

points(est_z, 0, pch = 19, cex = 1.1)
```

## Región de rechazo (dist normal N(0, 4/sqrt(35)))