Prueba de Hipótesis

5/3/2021

Prueba de Hipótesis

El método científico es un proceso con el cual se investiga de forma sistemática las observaciones, se resuelven problemas y se prueban hipótesis. Como parte del método científico la propuesta de una hipótesis y luego su comprobación, son temas bien definidos, y a pesar de la incertidumbre asociada al problema es posible cuantificar el error de la conclusión planteada por la hipótesis.

Los pasos del método científico son: Plantear un problema a resolver, Colectar una serie de observaciones, formular una o más hipótesis, probar dichas hipótesis y declarar las conclusiones. Aqui exploraremos herramientas estadísticas que nos puede ayudar el paso de prueba de hipótesis. Una hipótesis se puede definir de la siguiente manera: Una explicación tentativa que cuenta con un conjunto de hechos que pueden ser probados con una investigación posterior.

A su vez los pasos para poder efectuar una prueba de hipótesis son los siguientes:

1) Establecer las hipótesis en base a lo que se pretende verificar. Fijar H_0 y H_1
2) Definir la prueba adecuada. Buscar el estadístico del test que bajo la hipótesis nula tenga una distribución conocida
3) Definir el nivel de riesgo o precisión que lleva a determinar la región crítica
4) Seleccionar una muestra de tamaño n, para la cual el estadístico de contraste tome un valor numérico (valor experimental del estadístico de contraste
5) Adoptar la decisión sobre el rechazo o no de H_0

Como en el práctico de estimación las veremos como realizar contraste de hipótesis a traves de pruebas para los siguientes estadísticos:

- media μ,
- proporción p,
- varianza σ^2,
- diferencia de medias $\mu_1 - \mu_2$ para muestras independientes y dependientes (o pareadas),
- diferencia de proporciones $p_1 - p_2$, y
- cociente de varianzas σ_1^2/σ_2^2.

Prueba de hipótesis para μ con muestras grandes

Ejemplo autos

Se afirma que los autos particulares recorren en promedio más de 20000 kilómetros por año pero hay quienes dicen que este promedio es en realidad menor. Para probar esta afirmación se pide a una muestra de 100 propietarios de autos seleccionada de manera aleatoria que lleven un registro de los kilómetros que recorren en un año.

De la muestra se obtienen los siguientes datos:

Media = 19500 DE = 3900

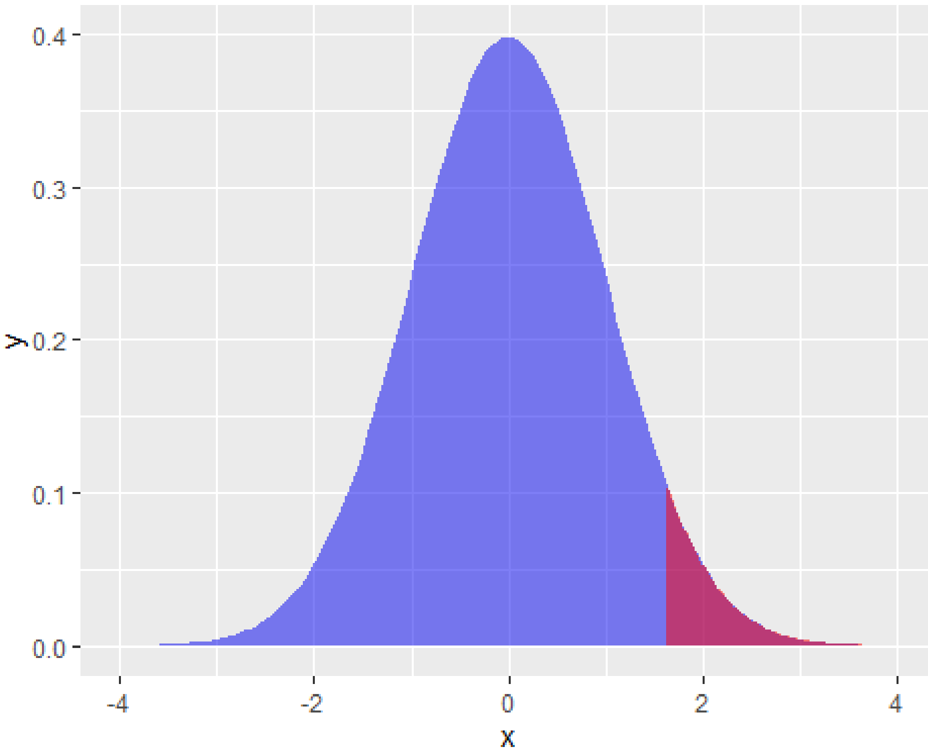
En este problema un planteo podría ser:

$$H_0: \mu \geq 20000 \quad km$$

$$H_1: \mu < 20000 \quad km$$

Vamos a hacer paso a paso los cálculos para obtener los resultados deseados, a continuación las instrucciones para calcular el estadístico y su valor-p.

```
library(tidyverse)

xbarra <- 19500  # Datos del problema
desvio <- 3900   # Datos del problema
n <- 100         # Datos del problema
mu <- 20000      # Media de referencia
est <- (xbarra - mu) / (desvio / sqrt(n))
est  # Para obtener el valor del estadístico

## [1] -1.282051

xvals <- seq(-4,4,0.1)# 120:200

plotdata <- data.frame(x = xvals, y = dnorm(xvals, 0, 1))

ggplot(data = data.frame(x = seq(-4,4,0.1)), aes(x))+

  stat_function(fun = dnorm, args =list(mean = 0, sd = 1),
                geom = "area",fill="blue2",alpha=0.5)+

  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), xlim = c(1.64,
4),
                geom ="area", fill = "red",alpha=0.5)
```




```
pnorm(est)  # Para obtener el valor-P

## [1] 0.09991233
```

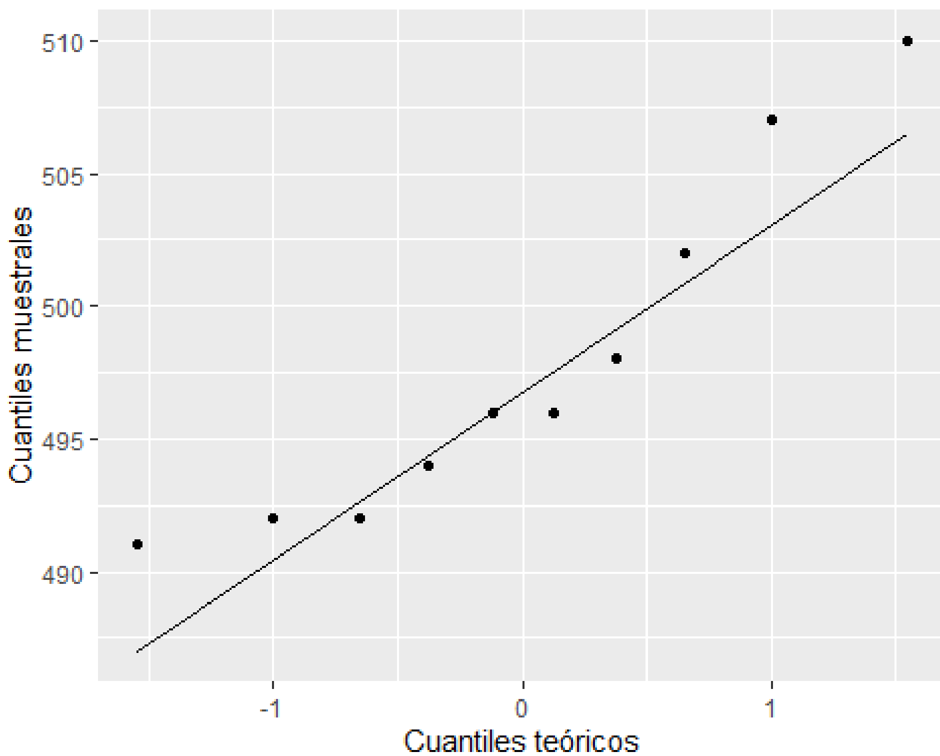
Como el valor-P es mayor que el nivel de significancia 5%, no hay evidencias suficientes para pensar que es menor el recorrido anual promedio de los autos.

Prueba de hipótesis para μ

Ejemplo Cafe

Para verificar si el proceso de llenado de bolsas de café con 500 gramos está operando correctamente se toman aleatoriamente muestras de tamaño diez cada cuatro horas. Una muestra de bolsas está compuesta por las siguientes observaciones: 502, 501, 497, 491, 496, 501, 502, 500, 489, 490.

```
contenido <- as.data.frame(c(510, 492, 494, 498, 492,
               496, 502, 491, 507, 496))
colnames(contenido) <- c("gramos")

ggplot(contenido, aes(sample=gramos)) +
  geom_qq()+
  geom_qq_line()+
  xlab('Cuantiles teóricos')+
  ylab('Cuantiles muestrales')
```




```
mean(contenido$gramos)

## [1] 497.8

sd(contenido$gramos)

## [1] 6.545567
```

¿Está el proceso llenando bolsas conforme lo dice la envoltura? Use un nivel de significancia del 5%.

El planteo de hipótesis se puede resumir así:

$$H_0: \mu = 500 \quad gr$$

$$H_1: \mu \neq 500 \quad gr$$

La prueba de hipótesis se puede realizar usando la función `t.test` por medio del siguiente código.

```
t.test(contenido, alternative='two.sided',
       conf.level=0.95, mu=500)

##
##  One Sample t-test
##
## data:  contenido
## t = -1.0629, df = 9, p-value = 0.3155
```

```
## alternative hypothesis: true mean is not equal to 500
## 95 percent confidence interval:
##  493.1176 502.4824
## sample estimates:
## mean of x
##     497.8
```

Como el valor-P es 30% y mayor que el nivel de significancia 5%, no se rechaza la hipótesis nula, es decir, las evidencias no son suficientes para afirmar que el proceso de llenando no está cumpliendo con lo impreso en la envoltura.

Prueba de hipótesis para la proporción p

Ejemplo quitamanchas

Existen varias pruebas para estudiar la propoción p de una distribución binomial, a continuación el listado de las más comunes.

1. Prueba de Wald,
2. Prueba X^2 de Pearson,
3. Prueba binomial exacta.

Prueba de Wald

Esta prueba se recomienda usar cuando se tiene un tamaño de muestra n suficientemente grande para poder usar la distribución normal para aproximar la distribución binomial.

En esta prueba el estadístico está dado por

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

donde $\hat{p}$ es la proporción muestral calculada como el cociente entre el número de éxitos x observados en los n ensayos y p_0 es el valor de referencia de las hipótesis. El estadístico z tiene distribución $N(0,1)$ cuando $n \to \infty$.

Un fabricante de un quitamanchas afirma que su producto quita 90% de todas las manchas. Para poner a prueba esta afirmación se toman 200 camisetas manchadas de las cuales a solo 174 les desapareció la mancha. Pruebe la afirmación del fabricante a un nivel $\alpha = 0.05$.

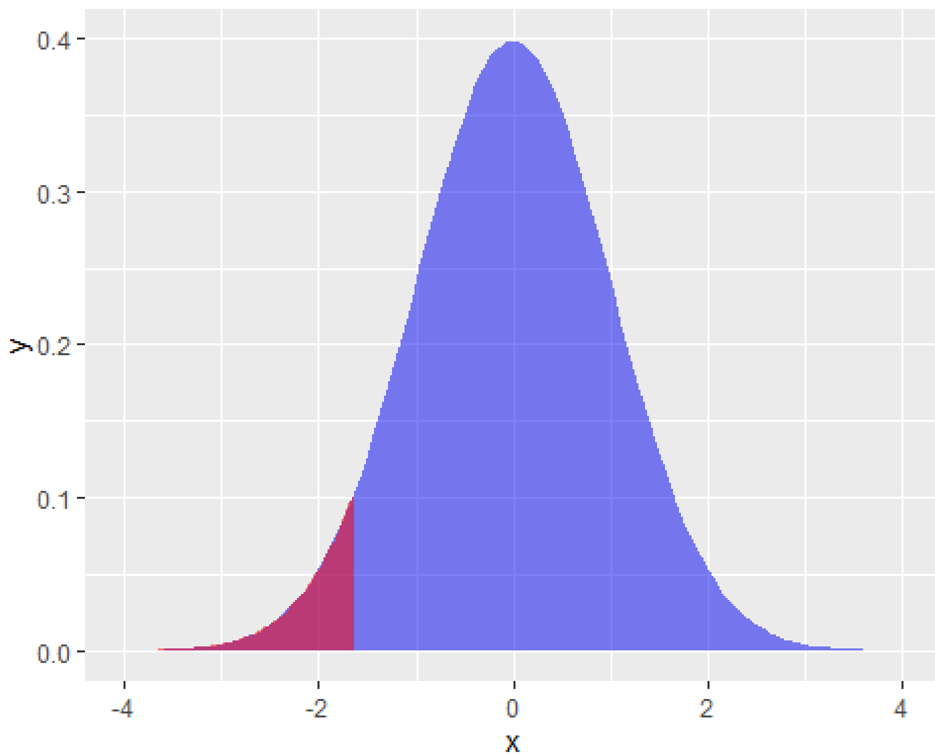
En este problema interesa probar lo siguiente:

$$H_0: p = 0.90$$

$$H_1: p < 0.90$$

Del anterior conjunto de hipótesis se observa que el valor de referencia de la prueba es $p_0 = 0.90$. De la información inicial se tiene que de las $n = 200$ pruebas se observó que en $x = 174$ la mancha desapareció, con esta información se puede calcular el estadístico z así:

```
z <- (174/200 - 0.90) / sqrt(0.90 * (1 - 0.90) / 200)
z  # Para obtener el valor del estadístico

## [1] -1.414214

ggplot(data = data.frame(x = seq(-4,4,0.1)), aes(x))+

  stat_function(fun = dnorm, args =list(mean = 0, sd = 1),
                geom = "area",fill="blue2",alpha=0.5)+

  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), xlim = c(-1.64, -
4),
                geom ="area", fill = "red",alpha=0.5)
```



Para obtener el valor-P de la prueba debemos tener en cuenta el sentido en la hipótesis alternativa $H_1: p < 0.90$, por esa razón el valor-P será $P(Z < z)$ y para obtenerlo usamos el siguiente código

```
pnorm(q=z, lower.tail=TRUE)  # Para obtener el valor-P

## [1] 0.0786496
```

Prueba X^2 de Pearson

Para realizar la prueba X^2 de Pearson se usa la función `prop.test`.

```
ggplot(data = data.frame(x = seq(0,300)), aes(x))+
```

```
  stat_function(fun = dchisq, args =list(df=199),
                geom = "area",fill="blue2",alpha=0.5)
```

Los argumentos a definir dentro de `prop.test` para hacer la prueba son:

- `x`: número de éxitos en la muestra.
- `n`: número de observaciones en la muestra.
- `alternative`: tipo de hipótesis alterna. Los valores disponibles son `"two.sided"` cuando la alterna es $\neq$, `"less"` para el caso $<$ y `"greater"` para $>$.
- `p`: valor de referencia de la prueba.
- `correct`: valor lógico para indicar si se usa la corrección de Yates.
- `conf.level`: nivel de confianza para reportar el intervalo de confianza asociado (opcional).

Volvemos al ejemplo del quitamanchas

$$H_0: p = 0.90$$

$$H_1: p < 0.90$$

La forma de usar la función `prop.test` para realizar la prueba se muestra a continuación.

```
prop.test(x=174, n=200, p=0.9, alternative='less',
          conf.level=0.95, correct=FALSE)

## 
##  1-sample proportions test without continuity correction
## 
## data:  174 out of 200, null probability 0.9
## X-squared = 2, df = 1, p-value = 0.07865
## alternative hypothesis: true p is less than 0.9
## 95 percent confidence interval:
##  0.0000000 0.9042273
## sample estimates:
##    p 
## 0.87
```

Como el valor-P (con valor de 0.07865 pero repotado en la salida como 0.08) es mayor que α no se rechaza la hipótesis nula y se concluye que no hay evidencias suficientes para rechazar la hipótesis nula.

Prueba binomial exacta

Para realizar la prueba binomial exacta se usa la función `binom.test` que tiene la siguiente estructura.

```
binom.test(x, n, p = 0.5,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95)
```

Los argumentos a definir dentro de `binom.test` para hacer la prueba son:

- `x`: número de éxitos en la muestra.
- `n`: número de observaciones en la muestra.
- `alternative`: tipo de hipótesis alterna. Los valores disponibles son `"two.sided"` cuando la alterna es $\neq$, `"less"` para el caso $<$ y `"greater"` para $>$.
- `p`: valor de referencia de la prueba.
- `conf.level`: nivel de confianza para reportar el intervalo de confianza asociado (opcional).

Otro ejemplo

En una polleria asegura que 90% de sus órdenes se entregan en menos de 10 minutos. En una muestra de 20 órdenes, 17 se entregaron dentro de ese lapso. ¿Puede concluirse en el nivel de significancia 0.05, que menos de 90% de las órdenes se entregan en menos de 10 minutos?

En este problema interesa probar lo siguiente:

$$H_0: p = 0.90$$

$$H_1: p < 0.90$$

La forma de usar la función `binom.test` para realizar la prueba se muestra a continuación.

```
binom.test(x=17, n=20, p=0.9, alternative="less")

##
##  Exact binomial test
##
## data:  17 and 20
## number of successes = 17, number of trials = 20, p-value = 0.3231
## alternative hypothesis: true probability of success is less than 0.9
## 95 percent confidence interval:
##  0.0000000 0.9578306
## sample estimates:
## probability of success
##                   0.85
```

Como el valor-P (reportado como 0.3 pero con valor de 0.3231) es mayor que α no se rechaza la hipótesis nula y se concluye que no hay evidencias suficientes para rechazar la hipótesis nula.

Prueba de hipótesis para el cociente de varianzas σ_1^2/σ_2^2

Para realizar este tipo de prueba se puede usar la función `var.test`.

Ejemplo tratamiento porotos

Se realiza un estudio para comparar dos tratamientos que se aplicarán a porotos crudos con el objetivo de reducir el tiempo de cocción. El tratamiento T1 es a base de bicarbonato de sodio, el T2 es a base de cloruro de sodio o sal común. La variable respuesta es el tiempo de cocción en minutos. Los datos se muestran abajo. ¿Son las varianzas de los tiempos iguales o diferentes? Usar $\alpha = 0.05$.

T1: 76, 85, 74, 78, 82, 75, 82.

T2: 57, 67, 55, 64, 61, 63, 63.

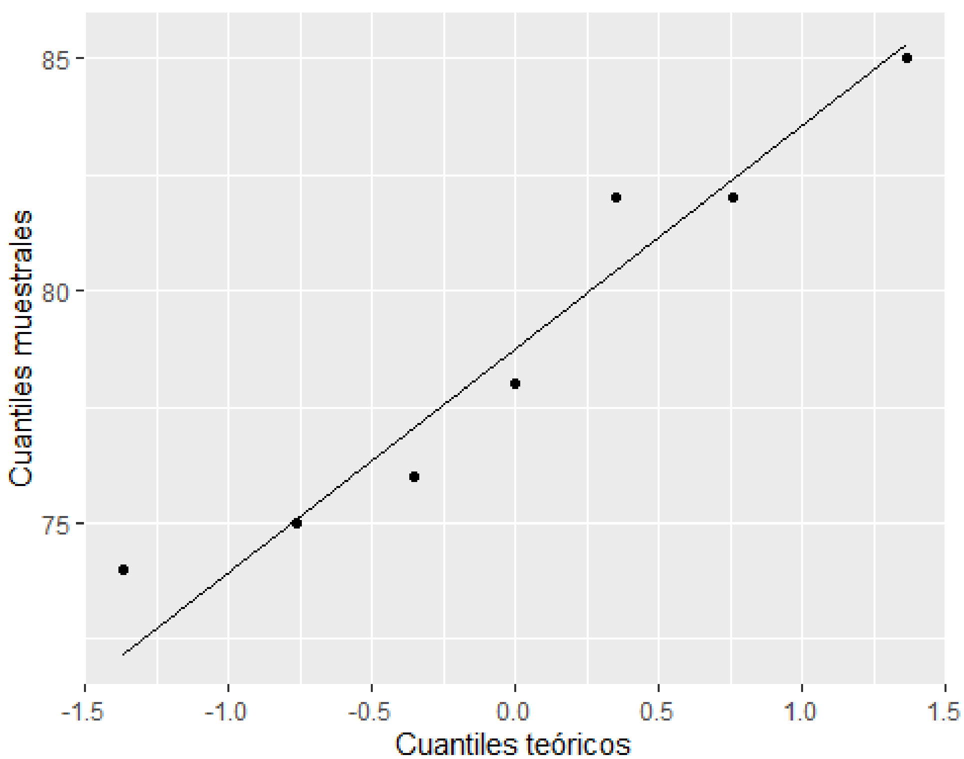
En este problema interesa probar si las varianzas poblacionales son iguales o no, por esta razón el cociente de $\sigma_{T1}^2/\sigma_{T2}^2$ se iguala al valor de 1 que será el valor de referencia de la prueba.

$$H_0: \sigma_{T1}^2/\sigma_{T2}^2 = 1$$

$$H_1: \sigma_{T1}^2/\sigma_{T2}^2 \neq 1$$

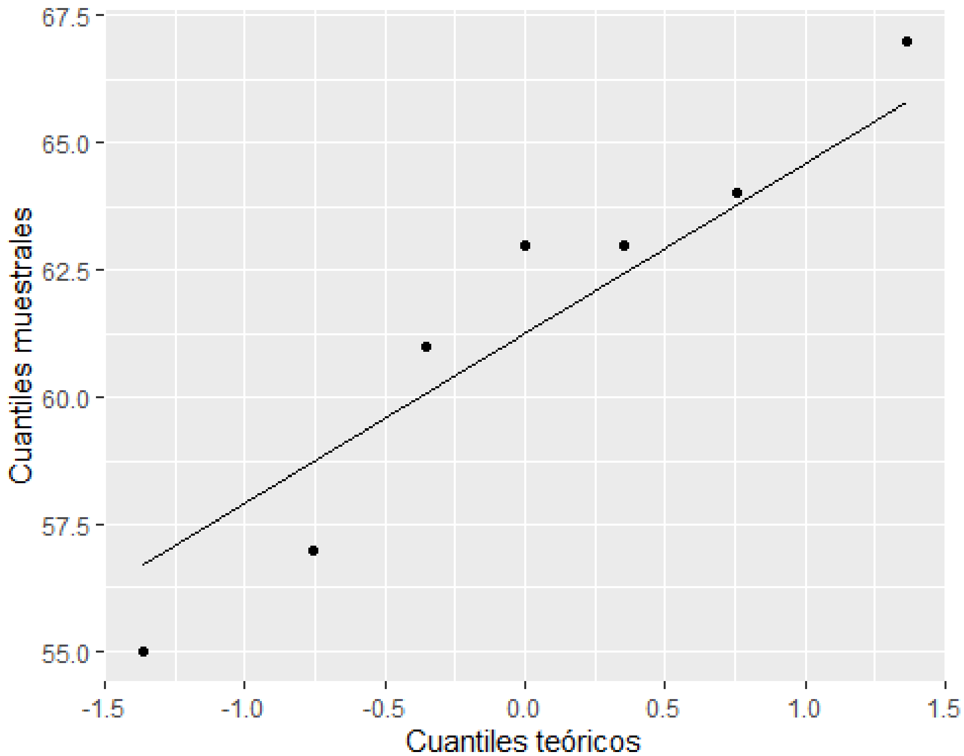
Para ingresar los datos se hace lo siguiente:

```
porotos <- as.data.frame(cbind("T1"=c(76, 85, 74,78, 82, 75, 82), "T2"=c(57,
67, 55, 64, 61, 63, 63)))

ggplot(porotos,aes(sample=T1)) +
  geom_qq()+
  geom_qq_line()+
  xlab('Cuantiles teóricos')+
  ylab('Cuantiles muestrales')
```




```
ggplot(porotos,aes(sample=T2)) +
  geom_qq()+
```

```
  geom_qq_line()+
  xlab('Cuantiles teóricos')+
  ylab('Cuantiles muestrales')
```



Una alternativa para probar normalidad: Tambien podemos plantear una prueba de hipótesis la prueba de normalidad Kolmogorov-Smirnov

```
require(nortest)
lillie.test(porotos$T1)$p.value

## [1] 0.520505

lillie.test(porotos$T2)$p.value

## [1] 0.3952748
```

Del QQplot mostrado en la Figura y las pruebas de normalidad se observa que se puede asumir que las poblaciones son normales.

La función `var.test` se puede usar para probar H_0, a continuación el código para realizar la prueba.

```
var.test(porotos$T1, porotos$T2, null.value=1, alternative="two.sided",
         conf.level=0.95)

##
##  F test to compare two variances
##
```

```
## data:  porotos$T1 and porotos$T2
## F = 1.011, num df = 6, denom df = 6, p-value = 0.9897
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1737219 5.8838861
## sample estimates:
## ratio of variances
##           1.011019
```

Como el valor-P es 0.9897 (reportado como 1 en la salida anterior), muy superior al nivel α de significancia 5%, se puede concluir que las varianzas son similares.

Ejemplo As

El arsénico en agua potable es un posible riesgo para la salud. Un artículo reciente reportó concentraciones de arsénico en agua potable en partes por billón (ppb) para diez comunidades urbanas y diez comunidades rurales. Los datos son los siguientes:

Urbana: 3, 7, 25, 10, 15, 6, 12, 25, 15, 7

Rural: 48, 44, 40, 38, 33, 21, 20, 12, 1, 18

¿Son las varianzas de las concentraciones iguales o diferentes? Usar $\alpha = 0.05$.

En este problema interesa probar:

$$H_0: \sigma^2_{Urb}/\sigma^2_{Rur} = 1$$

$$H_1: \sigma^2_{Urb}/\sigma^2_{Rur} \neq 1$$

Para ingresar los datos se hace lo siguiente:

```
As <- as.data.frame(cbind("urb"= c(3, 7, 25, 10, 15, 6, 12, 25, 15, 7) ,
                          "rur"= c(48, 44, 40, 38, 33, 21, 20, 12, 1, 18)))
```

Primero se debe explorar si las muestras provienen de una población normal

```
lillie.test(As$urb)$p.value

## [1] 0.5522105

lillie.test(As$rur)$p.value

## [1] 0.6249628
```

La función `var.test` se puede usar para probar H_0, a continuación el código para realizar la prueba.

```
var.test(As$urb, As$rur, null.value=1, alternative="two.sided",
         conf.level=0.95)

##
##  F test to compare two variances
```

```
##
## data:  As$urb and As$rur
## F = 0.24735, num df = 9, denom df = 9, p-value = 0.04936
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.06143758 0.99581888
## sample estimates:
## ratio of variances
##          0.2473473
```

Como el valor-P es 0.0493604 (reportado como 0.05 en la salida anterior) y es menor que el nivel de significancia $\alpha = 0.05$, se puede concluir que las varianzas no son iguales.

Prueba de hipótesis para la diferencia de medias $\mu_1 - \mu_2$ con varianzas iguales

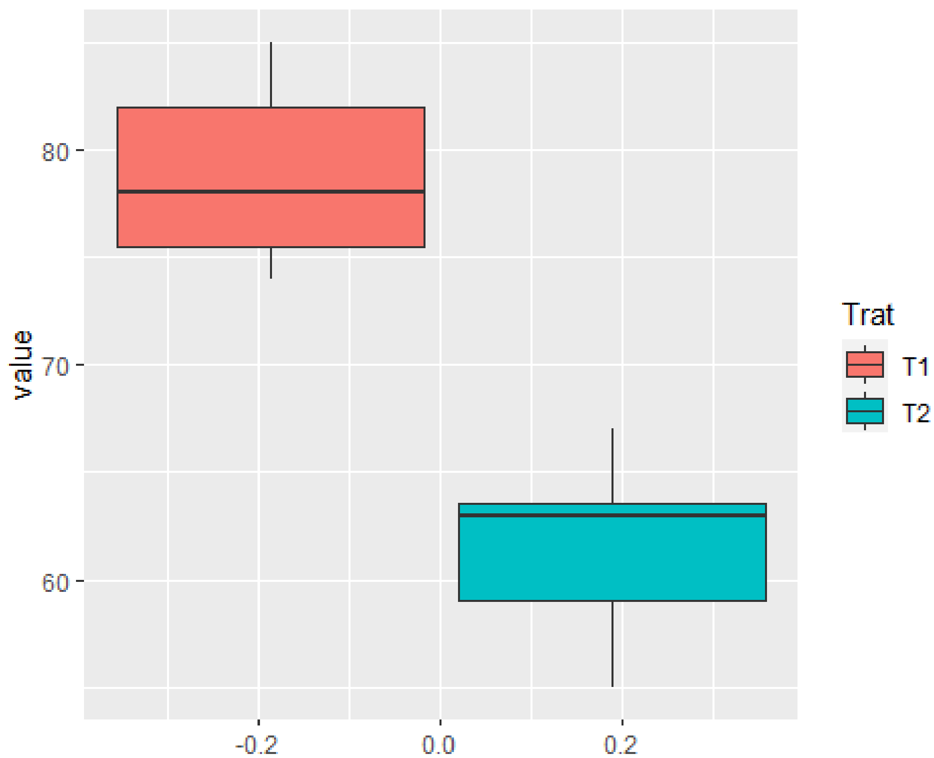
Para realizar este tipo de prueba se puede usar la función `t.test` que tiene la siguiente estructura.

Ejemplo porotos

Retomando el ejemplo de los fríjoles, ¿existen diferencias entre los tiempos de cocción de los fríjoles con T1 y T2? Usar un nivel de significancia del 5%.

Primero se construirá un boxplot comparativo para los tiempos de cocción diferenciando por el tratamiento que recibieron.

```
porotos %>% pivot_longer(cols=c("T1","T2"), names_to = "Trat") %>%
  ggplot(aes(y=value, fill=Trat)) +
  geom_boxplot()
```



Se muestra el boxplot, de esta figura se observa que las cajas de los boxplot no se solapan, esto es un indicio de que las medias poblacionales, μ_1 y μ_2, son diferentes, se observa también que el boxplot para el tratamiento T1 está por encima del T2.

En este problema interesa estudiar el siguiente conjunto de hipótesis.

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

$$H_1: \mu_1 - \mu_2 \neq 0$$

El código para realizar la prueba es el siguiente:

```
t.test(x=porotos$T1, y=porotos$T2, alternative="two.sided", mu=0,
       paired=FALSE, var.equal=TRUE, conf.level=0.97)

##
##  Two Sample t-test
##
## data:  porotos$T1 and porotos$T2
## t = 7.8209, df = 12, p-value = 4.737e-06
## alternative hypothesis: true difference in means is not equal to 0
## 97 percent confidence interval:
##  11.94503 22.91212
## sample estimates:
## mean of x mean of y
##  78.85714  61.42857
```

De la prueba se obtiene un valor-P muy pequeño, por lo tanto, podemos concluir que si hay diferencias significativas entre los tiempos promedios de cocción con T1 y T2, resultado que ya se sospechaba al observar el boxplot

¿Cómo cocinamos los porotos?

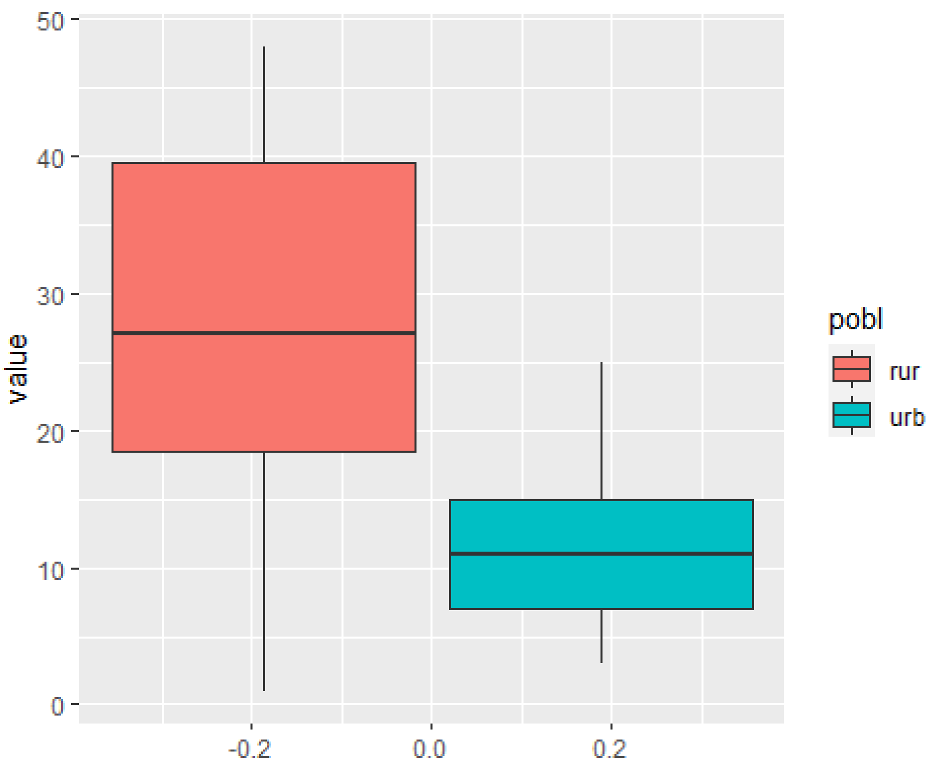
Prueba de hipótesis para la diferencia de medias $\mu_1 - \mu_2$ con varianzas diferentes

Ejemplo As

Retomando el ejemplo de la concentración de arsénico en el agua, ¿existen diferencias entre las concentraciones de arsénico de la zona urbana y rural? Usar un nivel de significancia del 5%.

Primero hacemos un boxplot comparativo para las concentraciones de arsénico diferenciando por la zona donde se tomaron las muestras.

```
As %>% pivot_longer(cols=c("urb","rur"), names_to = "pobl") %>%
  ggplot(aes(y=value, fill=pobl)) +
  geom_boxplot()
```



n este problema interesa estudiar el siguiente conjunto de hipótesis.

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

$$H_1: \mu_1 - \mu_2 \neq 0$$

El código para realizar la prueba es el siguiente:

```r
t.test(x=As$urb, y=As$rur, alternative="two.sided", mu=0,
       paired=FALSE, var.equal=FALSE, conf.level=0.95)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  As$urb and As$rur
## t = -2.7669, df = 13.196, p-value = 0.01583
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -26.694067  -3.305933
## sample estimates:
## mean of x mean of y 
##      12.5      27.5
```

De la prueba se obtiene un valor-P pequeño, por lo tanto, podemos concluir que si hay diferencias significativas entre las concentraciones de arsénico del agua entre las dos zona. La zona que presenta mayor concentración media de arsénico en el agua es la rural.