10.Pruebas Comunes

Paula Vargas Pellicer

08/03/2022

Dos muestras

Comparar dos varianzas (Prueba F Fisher)

Para encontrar si dos varianzas son distintas hay que dividir la varianza mayor entre la varianza menor. Obviamente, si las varianzas son iguales, la tasa será 1. Para que sea significativamente diferente, la tasa deberá ser significativamente más grande que 1 (porque la varianza más grande va arriba, en el numerador). ¿Cómo distinguimos un valor significativo de la tasa de varianza de uno no significativo? La respuesta, es buscar el valor crítico de la razón de varianza.

```
#Usamos "ToothGrowth" como base de datos
Mis_datos<- ToothGrowth
```

Queremos probar la equidad de varianzas entre dos grupos (OJ y VC) de la columna "supp" La prueba F es muy sensible a la suposición de normalidad de datos, por lo que hay que comprobar que sí lo sean usando la prueba de Shapito-Wilk y una gráfica de Q-Q plot

```
library(ggpubr)
shapiro.test(Mis_datos$len)
ggqqplot(Mis_datos$len)
```

Ahora sí:

```
res.Ftest<- var.test(len~supp, data=Mis_datos)
res.Ftest
```

Comparar dos medias, prueba Student-t

Muestras independientes

Supongamos que hemos medido el peso de 100 individuos: 50 mujeres (grupo A) y 50 hombres (grupo B). Queremos saber si el peso medio de las mujeres (mA) es significativamente diferente de la de los hombres (mB).

Creamos los datos

```
# Datos en dos vectores
peso_mujer <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5)
peso_hombre <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
# Crear base de datos
```

```
mi_data <- data.frame(
                grupo = rep(c("Mujer", "Hombre"), each = 9),
                peso = c(peso_mujer,  peso_hombre)
                )
```

Visualizamos

```
library(dplyr)
group_by(mi_data, grupo) %>%
  summarise(
    count = n(),
    mean = mean(peso, na.rm = TRUE),
    sd = sd(peso, na.rm = TRUE)
  )
library("ggpubr")
ggboxplot(mi_data, x = "grupo", y = "peso",
          color = "grupo", palette = c("#00AFBB", "#E7B800"),
         ylab = "Peso", xlab = "Grupos")
```

Comprobar que los datos están distribuidos normalmente y que las varianzas son iguales

```
# Shapiro-Wilk
shapiro.test(mi_data$peso)

res.ftest <- var.test(peso ~ grupo, data = mi_data)
res.ftest
```

Ahora sí, prueba *t* de muestras independientes

```
res <- t.test(peso ~ grupo, data = mi_data, var.equal = TRUE)
res
```

Si quisieras probar si el peso promedio de los hombres es menor que el peso promedio de las mujeres, entonces:

```
t.test(peso ~ grupo, data = mi_data, paired = TRUE,
        alternative = "less")
```

Muestras dependientes

```
#Formato ancho
library(tidyr)
data("mice2", package = "datarium")
head(mice2, 3)
#tansforma a formato largo
mice2.long <- mice2 %>%
  gather(key = "group", value = "weight", before, after)
head(mice2.long, 3)
```

Podemos sacar algunas estadísticas básicas

```
mice2.long %>%
  group_by(group) %>%
  get_summary_stats(weight, type = "mean_sd")
```

Y ahora sacar la prueba de *t*

```
res <- t.test(weight ~ group, data = mice2.long, paired = TRUE)
res
```

Y visualizar los datos

```
bxp <- ggpaired(mice2.long, x = "group", y = "weight",
         order = c("before", "after"),
         ylab = "Weight", xlab = "Groups")
```

Comparar dos medianas

Muestras emparejadas, Prueba Wilcoxon

Supongamos que aplicamos una prueba de matemáticas en una clase de 12 estudiantes al comienzo de un semestre y que aplicamos una prueba similar al final del semestre exactamente a los mismos estudiantes. Tenemos los siguientes datos:

```
dat2 <- data.frame(
  Beginning = c(16, 5, 15, 2, 14, 15, 4, 7, 15, 6, 7, 14),
  End = c(19, 18, 9, 17, 8, 7, 16, 19, 20, 9, 11, 18)
)

dat2
dat2 <- data.frame(
  Time = c(rep("Before", 12), rep("After", 12)),
  Grade = c(dat2$Beginning, dat2$End)
)
dat2
```

Visualizamos los datos

```
# Reordenamos el tiempo
dat2$Time <- factor(dat2$Time,
  levels = c("Before", "After")
)

ggplot(dat2) +
  aes(x = Time, y = Grade) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```

Las calificaciones antes y después son iguales o distintas

```
test <- wilcox.test(dat2$Grade ~ dat2$Time,
  paired = TRUE
)

test
```

Muestras independientes, prueba Mann-Withney-Wilcoxon

Para la prueba de Wilcoxon con muestras independientes, supongamos que queremos probar si las calificaciones en el examen de estadística difieren entre estudiantes mujeres y hombres Hemos recopilado calificaciones de 24 estudiantes (12 niñas y 12 niños):

```
dat <- data.frame(
  Sex = as.factor(c(rep("Girl", 12), rep("Boy", 12))),
  Grade = c(
    19, 18, 9, 17, 8, 7, 16, 19, 20, 9, 11, 18,
    16, 5, 15, 2, 14, 15, 4, 7, 15, 6, 7, 14
  )
)

dat
#Visualizar
ggplot(dat) +
  aes(x = Sex, y = Grade) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```

¿Cómo puedes ver si hay normalidad de los datos?

Dado que no la hay, podemos continuar con esta prueba no paramétrica

```
test <- wilcox.test(dat$Grade ~ dat$Sex,
  alternative = "less"
)

test
```

¿Por qué ponemos "less"? la respuesta la encuentras basándote en las graficas de cajas

Prueba Binomial de proporciones

Compara una proporción de muestra con una proporción hipotética. La prueba tiene las siguientes hipótesis nula y alternativa:

H0: $\pi = p$ (la proporción poblacional π es igual a algún valor p)

HA: $\pi \neq p$ (la proporción poblacional π no es igual a algún valor p)

La prueba también se puede realizar con una alternativa de una cola de que la verdadera proporción de la población es mayor o menor que algún valor p.

binom.test(x, n, p) donde: x: número de éxitos n: número de intentos p: probabilidad de éxito en un ensayo dado Los siguientes ejemplos ilustran cómo usar esta función en R para realizar pruebas binomiales.

Ejemplo 1: prueba binomial de dos colas

Deseamos determinar si un dado cae o no en el número "3" durante 1/6 de las tiradas, por lo que tiramos el dado 24 veces y cae en "3" un total de 9 veces. Al realizar una prueba binomial para determinar si el dado realmente cae en "3" durante 1/6 de las tiradas.

```
binom.test(9, 24, 1/6)
```

El valor p de la prueba es 0,01176. Como esto es menos de 0.05, podemos rechazar la hipótesis nula y concluir que hay evidencia para decir que el dado no cae en el número "3" durante 1/6 de los lanzamientos.

Ejemplo 2: prueba binomial de cola izquierda

Desea determinar si es menos probable que una moneda caiga en cara en comparación con cruz, por lo que lanza la moneda 30 veces y descubre que cae en cara solo 11 veces. Realice una prueba binomial para determinar si es menos probable que la moneda caiga en cara en comparación con cruz.

```
binom.test(11, 30, 0.5, alternative="less")
```

El valor p de la prueba es 0,1002. Dado que no es inferior a 0,05, no podemos rechazar la hipótesis nula. No tenemos evidencia suficiente para decir que es menos probable que la moneda caiga en cara en comparación con cruz.

Ejemplo 3: prueba binomial de cola derecha

Una tienda fabrica widgets con un 80% de efectividad. Implementan un nuevo sistema que esperan que mejore la tasa de efectividad. Seleccionan aleatoriamente 50 widgets de una producción reciente y descubren que 46 de ellos son efectivos. Realice una prueba binomial para determinar si el nuevo sistema conduce a una mayor eficacia.

```
binom.test(46, 50, 0.8, alternative="greater")
```

El valor p de la prueba es 0,0185. Dado que es inferior a 0,05, rechazamos la hipótesis nula. Tenemos suficiente evidencia para decir que el nuevo sistema produce widgets efectivos a una tasa superior al 80%.

Pearson

```
my_data <- mtcars
head(my_data, 6)
library("ggpubr")
ggscatter(my_data, x = "mpg", y = "wt",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Miles/(US) gallon", ylab = "Weight (1000 lbs)")
```

Prueba preliminar para comprobar los supuestos de la prueba ¿La covariación es lineal? Sí, de la gráfica de arriba, la relación es lineal. En la situación en la que los diagramas de dispersión muestran patrones curvos, estamos tratando con una asociación no lineal entre las dos variables.

¿Los datos de cada una de las 2 variables (x, y) siguen una distribución normal? Utiliza la prueba de normalidad de Shapiro-Wilk y mira el gráfico de normalidad

```
# Shapiro-Wilk para mpg
shapiro.test(my_data$mpg) # => p = 0.1229
shapiro.test(my_data$wt) # => p = 0.09

#visual
# mpg
ggqqplot(my_data$mpg, ylab = "MPG")

# wt
ggqqplot(my_data$wt, ylab = "WT")
```

Ahora si, la prueba de Pearson

```
res <- cor.test(my_data$wt, my_data$mpg,
                    method = "pearson")
res
```

Spearman

La estadística rho de Spearman también se usa para estimar una medida de asociación basada en rangos. Esta prueba se puede utilizar si los datos no provienen de una distribución normal bivariada.

```
res2 <-cor.test(my_data$wt, my_data$mpg,  method = "spearman")
res2
```

Prueba de independencia, Chi-cuadrada

Prueba si existe una relación entre dos variables categóricas. Las hipótesis nula y alternativa son:

H0: las variables son independientes, no hay relación entre las dos variables categóricas. Conocer el valor de una variable no ayuda a predecir el valor de la otra variable H1: las variables son dependientes, existe una relación entre las dos variables categóricas. Conocer el valor de una variable ayuda a predecir el valor de la otra variable

La prueba de independencia Chi-cuadrado funciona comparando las frecuencias observadas con las frecuencias esperadas si no hubiera relación entre las dos variables categóricas (es decir, las frecuencias esperadas si la hipótesis nula fuera cierta).

Como solo hay una variable categórica en la base de datos Iris y la prueba de independencia de Chi-cuadrado requiere dos variables categóricas, agregamos la variable 'tamaño' que corresponde a 'pequeña' si la longitud del pétalo es más pequeña que la mediana de todas las flores, 'grande' en caso contrario:

```
dat <- iris

dat$size <- ifelse(dat$Sepal.Length < median(dat$Sepal.Length),
  "small", "big"
)

#creamos una tabla de contingencia
table(dat$Species, dat$size)
```

La tabla de contingencia da el número observado de casos en cada subgrupo. Por ejemplo, solo hay una flor setosa grande, mientras que hay 49 flores setosa pequeñas en el conjunto de datos.

También es una buena práctica dibujar un diagrama de barras para representar visualmente los datos:

```
ggplot(dat) +
  aes(x = Species, fill = size) +
  geom_bar()
```

Aquí puedes enchular tu grafica

Para este ejemplo, vamos a probar si existe una relación entre las variables Especie y tamaño. Para ello se utiliza la función chisq.test():

```
test <- chisq.test(table(dat$Species, dat$size))
test
```

Si aparece una advertencia como "La aproximación de chi-cuadrado puede ser incorrecta", significa que las frecuencias esperadas más pequeñas son inferiores a 5. Para evitar este problema, puedes:

- reunir algunos niveles (especialmente aquellos con un pequeño número de observaciones) para aumentar el número de observaciones en los subgrupos, o
- usar la prueba exacta de Fisher La prueba exacta de Fisher no requiere la suposición de un mínimo de 5 conteos esperados en la tabla de contingencia. Se puede aplicar en R gracias a la función fisher.test(). Esta prueba es similar a la prueba de Chi-cuadrado en términos de hipótesis e interpretación de los resultados.

Hablando de suposiciones, la prueba de independencia Chi-cuadrado requiere que las observaciones sean independientes. Por lo general, esto no se prueba formalmente, sino que se verifica en función del diseño del experimento y del buen control de las condiciones experimentales.

Si tienes observaciones dependientes (muestras pareadas), se deben usar las pruebas Q de McNemar o Cochran en su lugar. La prueba de McNemar se usa cuando queremos saber si hay un cambio significativo en dos muestras pareadas (típicamente en un estudio con una medida antes y después sobre el mismo tema) cuando las variables tienen solo dos categorías. La prueba Q de Cochran es una extensión de la prueba de McNemar cuando tenemos más de dos medidas relacionadas.