Practicos de Bioestadística 2

Derek Corcoran

2019-03-11

Contents

Requerimientos

Para comenzar el trabajo se necesita la última versión de R y RStudio (R Core Team, 2018).También se requiere de los paquetes *pacman*, *rmarkdown*, *tidyverse* y *tinytex*. Si no se ha usado R o RStudio anteriormente, el siguiente video muestra cómo instalar ambos programas y los paquetes necesarios para este curso en el siguiente link.

El código para la instalación de esos paquetes es el siguiente:

```
install.packages("pacman", "rmarkdown", "tidyverse", "tinytex")
```

En caso de necesitar ayuda para la instalación, contactarse con el instructor del curso.

0.1 Antes de comenzar

Si nunca se ha trabajado con `R` antes de este curso, una buena herramienta es provista por el paquete Swirl (Kross et al., 2017). Si deseas estar más preparado para el curso, realiza los primeros 7 módulos del programa *R Programming: The basics of programming in R* que incluye:

- Basic Building Blocks
- Workspace and Files
- Sequences of Numbers
- Vectors
- Missing Values
- Subsetting Vectors
- Matrices and Data Frames

El siguiente link muestra un video explicativo de cómo usar el paquete swirl Video

0.2 Descripción del práctico

Los prácticos de este curso se enfocan en aprender a realizar de manera práctica los conceptos enseñados en el cuso, pero además, usando herramientas interactivas y/o programáticas, el profundizar el entendimiento de ciertos conceptos teóricos y filosóficos del curso.

0.3 Objetivos del práctico

1. Aprender el uso de R como ambiente estadístico de limpieza, exploración, visualización de datos.

2. Conocer y aplicar de manera aplicada los conceptos enseñados en el curso de Bioestadística 2.
3. Aprender buenas prácticas de recolección y estandarización de bases de datos, con la finalidad de optimizar el análisis de datos y la revisión de éstas por pares.
4. Realizar análisis críticos de la naturaleza de los datos al realizar análisis exploratorios, que permitirán determinar la mejor forma de comprobar hipótesis asociadas a estas bases de datos.

0.4 Contenidos

- Capítulo 1 *Análisis exploratorio y el primer ANOVA*: En este capítulo se aprenderá a cómo explorar, resumir y visualizar una base de datos utilizando el paquete tidyverse (Wickham, 2017), además se realizarán un análisis básico de ANOVA
- Capítulo 2 *Supuestos de ANOVA y mínimos cuadrados*
- Capítulo 3 *Análisis de poder y primera tarea*
- Capítulo **??** *Referencias*
- Capítulo 4 *T de student*

0.5 Metodología

Clases prácticas donde cada estudiante trabajará con datos entregados para desarrollar análisis de datos. Además, se deberán generar informes, en base al trabajo con sus datos.

0.6 Evaluación

- Evaluación 1: Informe exploratorio de base de datos 25%
- Evaluación 2: Presentación 25%
- Evaluación 3: Informe final 50%

0.7 Presentación de introducción

Para la introducción de los prácticos seguiremos un a presentación que se encuentra en este link

0.8 Libros de consulta

Los principios de este curso están explicados en los siguientes libros gratuitos.

- Gandrud, Christopher. Reproducible Research with R and R Studio. CRC Press, 2013. Available for free in the following link
- Stodden, Victoria, Friedrich Leisch, and Roger D. Peng, eds. Implementing reproducible research. CRC Press, 2014. Available for free in the following link

Chapter 1

Exploración de datos y tu primer ANOVA

1.1 Actividad 1 Educación en Chile

En esta actividad exploraremos los resultados de la PSU en Chile para el año 2017. Pueden encontrar la base de datos original en Data Chile.

Trataremos de determinar, usando el puntaje de la PSU como medida, si existen brechas en la educación chilena por tipo de institución. Para ello, primero trabajaremos realizando análisis exploratorios en base a gráficos y tablas resumen usando funciones del paquete *tidyverse* (Wickham, 2017) en R.

La base de datos *EducacionChile.csv* se encuentra disponible en webcursos o en https://es.datachile.io/geo/chile#education.

1.1.1 Tablas resumen de los datos:

Lo primero que deben hacer es generar una tabla resumen usando el *tidyverse* usando las funciones *group_by* para agrupar por variables y *summarize* para resumir los datos, dentro de summarize podemos usar variables como:

- **mean()** promedio
- **sd()** desviación estándar
- **n()** número de muestras

a modo de ejemplo vemos la tabla 1.1 mostrando la media y número de muestras con la base de datos iris:

```
data("iris")
Table <- group_by(iris, Species) %>% summarize(Promedio = mean(Petal.Length), N = n())
```

```
knitr::kable(Table)
```

Basado en el resumen ¿Qué podemos decir de estos datos de educación en Chile?

Table 1.1: Resumen con la media y número de muestras del largo de pétalo de las flores de tres especies del género Iris

Species	Promedio	N
setosa	1.462	50
versicolor	4.260	50
virginica	5.552	50

1.1.2 Visualización de datos con ggplot2 (tidyverse)

El paquete *ggplot2* (Wickham, 2016) es una poderosa herramienta para graficar datos. Si desean ahondar en el uso de este paquete, pueden ver el siguiente link http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/. En este caso, aprenderemos a graficar *boxplots* y *jitterplots*, dos opciones para visualizar una variable categórica versus una cuantitativa.

1.1.2.1 Uso del ggplot2

Su función principal es *ggplot*, luego de cada función usaremos el símbolo + como usábamos el pipeline (%>%).

Primero usamos la función ggplot para determinar la base de datos y variables, acá las variables siempre van dentro de la función aes

```
ggplot(MiBaseDeDatos, aes(x = VariableX, y = VariableY))
```

Luego agregamos el tipo de gráfico que queremos para nuestra figura usando el + como pipeline

```
ggplot(MiBaseDeDatos, aes(x = VariableX, y = VariableY)) + geom_boxplot()
```

1.1.2.2 Ejemplo usando la base de datos iris

1.1.2.2.1 Boxplot

El siguiente código muestra como graficar un boxplot para la base de datos iris, la cual esta en R. En este caso graficaremos el largo del pétalo para cada especie (Figura 1.1).

```
data("iris")
ggplot(iris, aes(x = Species, y = Petal.Length)) + geom_boxplot()
```

En los Box Plots tenemos 4 visualizaciones:

- Mediana (linea gruesa)
- Caja (Cuantiles 25% y 75%)
- Bigotes (intervalo de confianza del 95%)
- Puntos Outlayers

Realice un boxplot de los datos de la educación de Chile, ¿Qué nos dice esto de los datos?

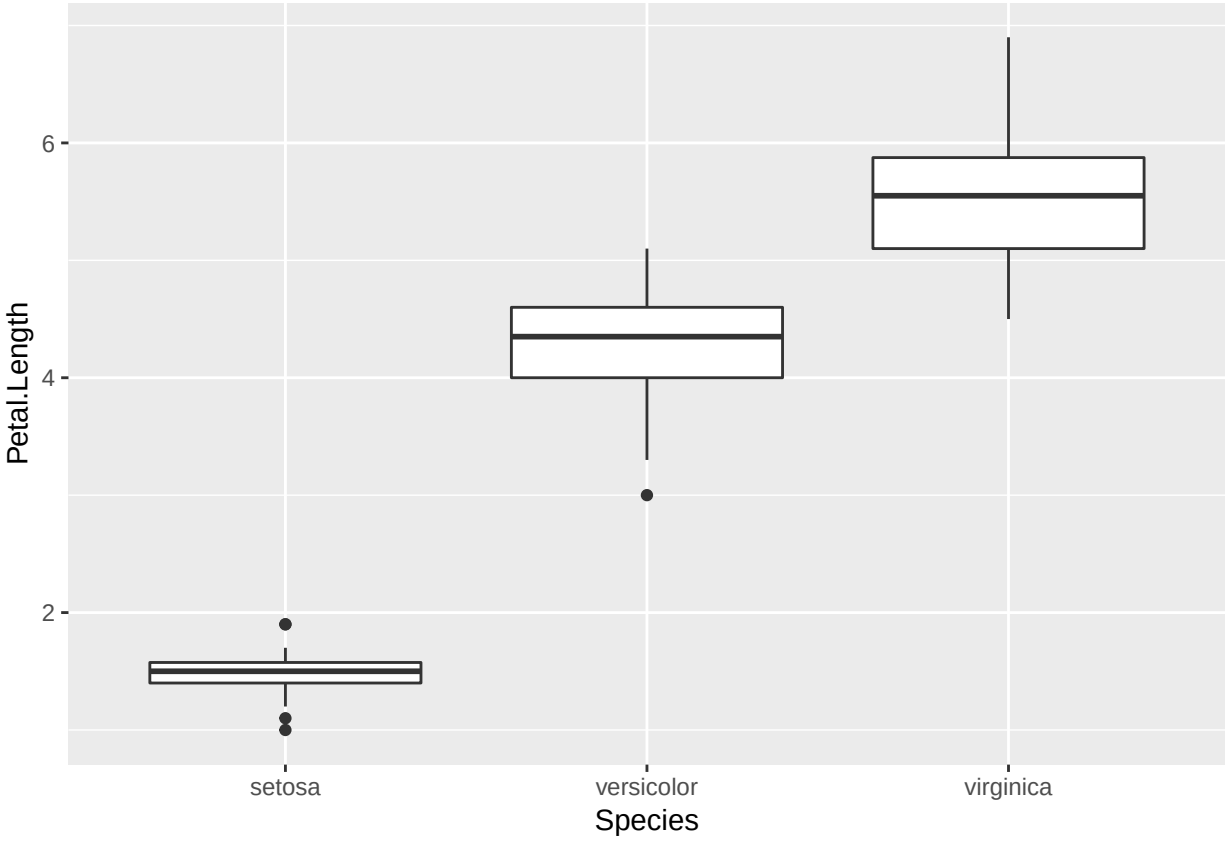

Figure 1.1: Box plot del largo de petalo de tres especies del género Iris

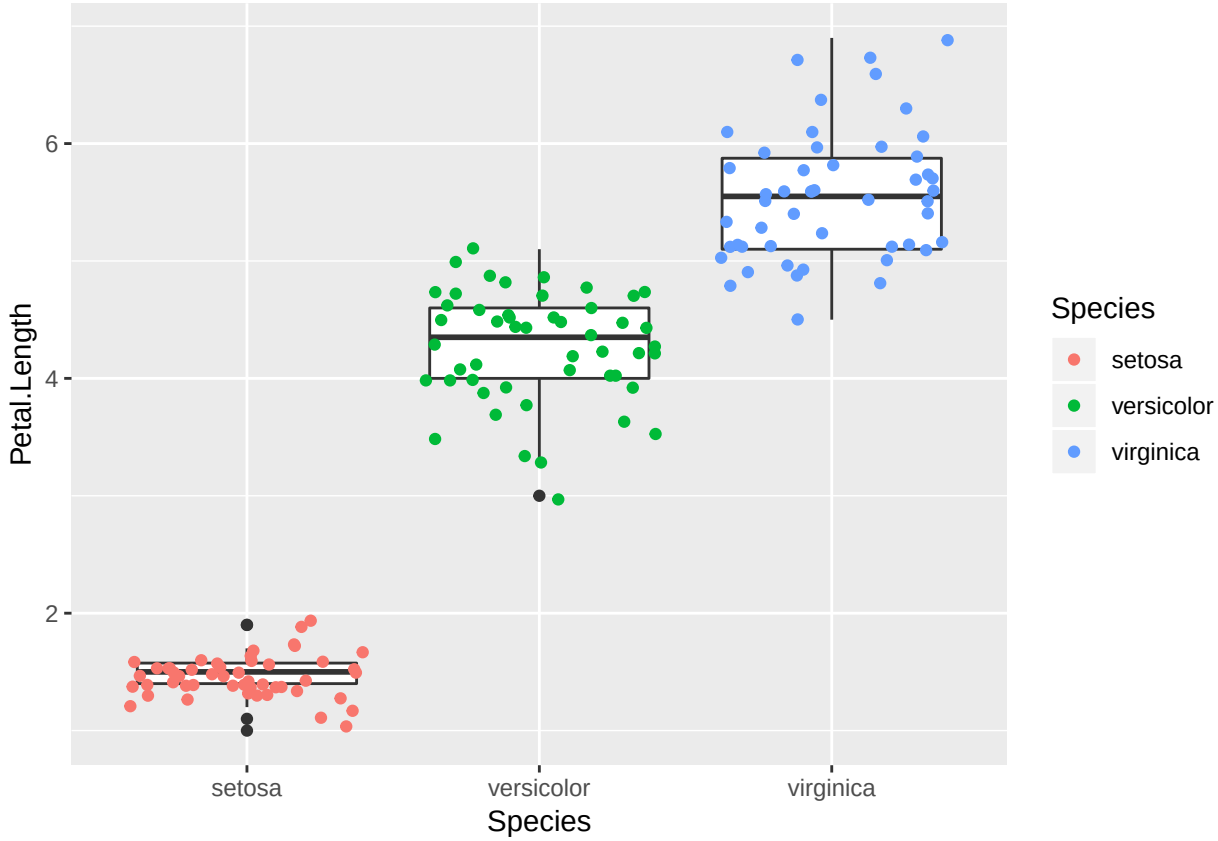

Figure 1.2: Box plot y jitter plot juntos para el largo de petalo de tres especies del género Iris

1.1.2.2.2 Jitter plot

El jitter plot suma un punto por cada observación, lo cual nos permite entender un poco más la naturaleza de los datos. En general se le agrega a un box plot para tener mayor claridad en los datos (Figura 1.2).

```
data("iris")
ggplot(iris, aes(x = Species, y = Petal.Length)) + geom_boxplot() + geom_jitter(aes(color = Species))
```

1.2 Actividad 2 Captación de CO2 en plantas

Utilizaremos base de datos CO_2 (Potvin et al., 1990) enviada al curso. Esta base de datos, también presente en R, tiene las siguientes variables

- **Plant**: Identidad de cada planta
- **Type**: Variedad de la planta (subespecie Quebec o Mississippi)
- **Treatment**: Tratamiento de la planta, algunas fueron enfriadas la noche anterior (Chilled)
- **conc**: Concentración ambiental de CO_2
- **Uptake**: Captación de CO_2 para cada planta en cada día

¿Hay diferencias entre la captación de CO_2 en plantas tratadas y no tratadas?

- Genere tablas resumenes que le permitan explorar esta pregunta

  - ¿Existen variables que puedan confundir el resultado? ¿como trataría los datos para lidiar con esto?

- Genere gráficos exploratorios para contestar esta pregunta

1.3 Actividad 3 Mi primer ANOVA

En *R* todos los modelos tienen la siguiente estructura **Funcion(y ~ x1 + x2 + ... + xn, data = MisDatos)**, donde la **Funcion** dice el modelo que queremos realizar (por ejemplo ANOVA, regresión lineal, modelos mixtos, etc.), **y** es la variable que queremos explicar, **x1** a **xn** son las variables explicativas, ~ es un símbolo que debe ser leído como explicado por y finalmente **data** es la base de datos que queremos utilizar, en un ANOVA (análisis de varianza), la función en cuestión es aov.

En el siguiente código vemos si el largo del pétalo de las flores del género *Iris*, pueden ser explicados por la especie a la que estas plantas pertenecen, por lo que generamos un modelo llamado *Primer.Anova* con la función **aov**.

```
Primer.Anova <- aov(Petal.Length ~ Species, data = iris)
```

Para acceder a la tabla de resultados utilizamos la función **summary**

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si establecemos el valor de alfa en 0.05 y al ver en la tabla que el valor de p es menor a alfa, rechazamos la hipótesis nula de que las medias son iguales, y decidimos que la media del largo de pétalo es distinta entre las especies.

1.3.1 Ejercicio

Determine si para la base de datos **CO2** la captación de CO_2 es distinto entre plantas con tratamiento de enfriamiento y sin enfriamiento.

1.3.2 Simulador de ANOVA

Web Page Blocked!

You have tried to access a web page which is in violation of your internet usage policy.

URL: http://admin.derek-corcoran-barrios.com/shiny/rstudio/sample-apps/Shiny2/?showcase=0
Category: Unrated
Client IP: 10.50.121.136
Server IP: 18.218.65.225
User name:
Group name:

To have the rating of this web page re-evaluated please click here.

Proceed Go Back

Chapter 2

Supuestos de ANOVA y mínimos cuadrados

2.1 Objetivos de este práctico

- Entender los supuestos de un ANOVA de una vía (independencia, aleatoriedad, homocedasticidad y normalidad)
- Entender el concepto de mínimos cuadrados
- Saber cuando realizar un ANOVA e interpretar sus resultados

2.2 Actividad 1 Sueño en mamíferos

En esta actividad intentaremos ver si hay diferencias en horas de sueño en mamíferos por Orden o dieta. Los datos fueron extraídos del trabajo de Savage and West (2007) y están incorporados en la base de datos de *ggplot2* con el nombre de *msleep*, pero estarán en webcursos en formato csv de todas formas. Para la guía los ejemplos se generarán en base a la base de datos *InsectSprays* que está en *R* y que fue extraída de Beall (1942), en la cual se testean la efectividad de insecticidas en Spray en la abundancia de insectos en plantaciones. Y en la base de datos *iris* que ya fue entregada, en la que se miden distintas características florales de especies del genero *Iris* (Anderson, 1935).

2.2.1 Homogeneidad de varianza

2.2.1.1 Inspección visual

Lo primero que intentaremos explorar de forma visual y a partir de tests si es que hay homogeneidad de varianza, para esto usaremos boxplots, y jitter plots (Figura 2.1), lo cual ya hemos hecho anteriormente:

```
ggplot(InsectSprays, aes(x = spray, y = count)) + geom_boxplot() + geom_jitter(aes(color = spray))
```

Para explorar visualmente si existe homogeneidad de varianza, se compraran las cajas y bigotes de los boxplots y se espera que tengan (Mas o menos distintos tamaños).

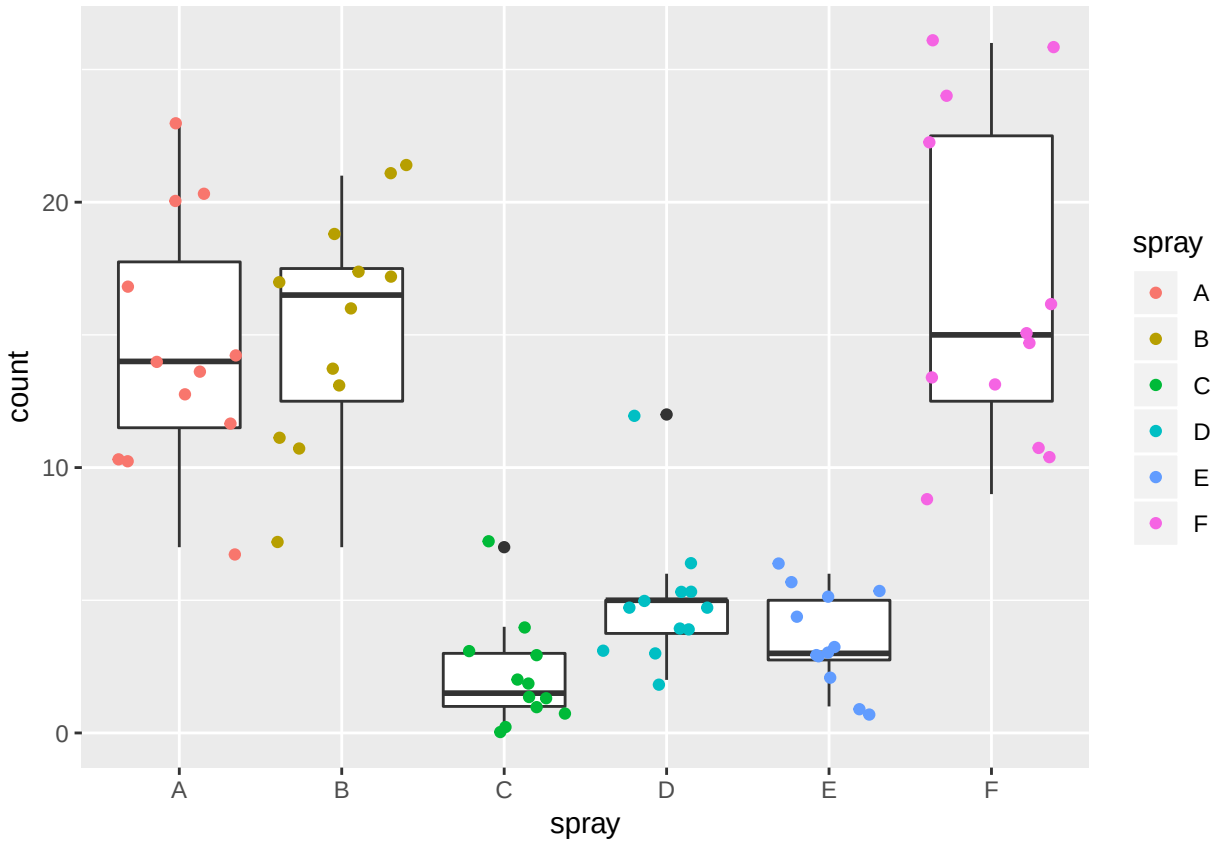

Figure 2.1: Cuenta de insectos según tipo de insecticida

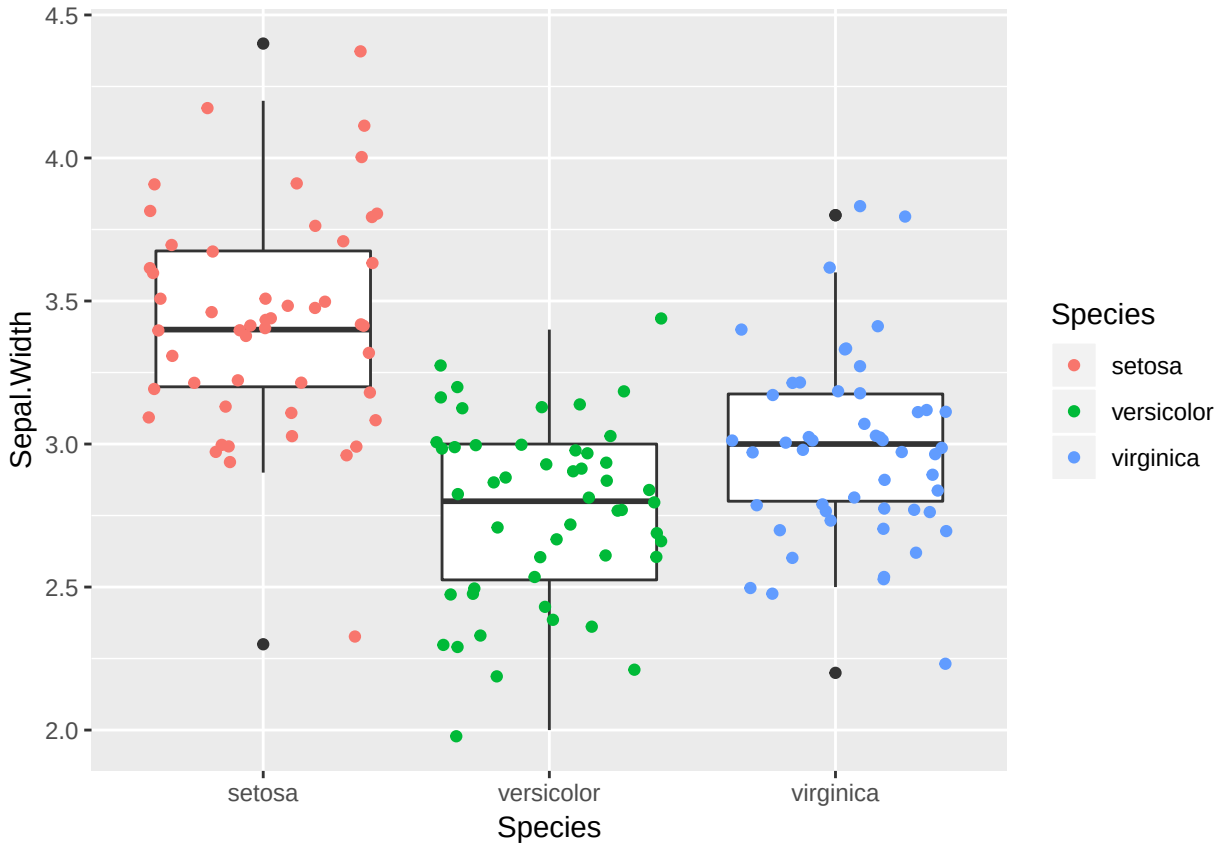

Figure 2.2: Ancho de sépalo según especie del género Iris

2.2.1.2 Test de Bartlett

Para realizar un test de homogeneidad de varianza se realiza el test de bartlett (Bartlett, 1937), en este se usa nuestra conocida formula $y \sim x$, esto es, y explicado por x junto a la función *bartlett.test*. Para nuestro caso usaríamos:

```
##
##  Bartlett test of homogeneity of variances
##
## data:  count by spray
## Bartlett's K-squared = 25.96, df = 5, p-value = 9.085e-05
```

Como en este caso, no el valor de p es menor a 0.05, decimos que no hay homogeneidad de varianza, por lo que no podemos hacer el test.

2.2.2 Normalidad de los residuales

En el caso de la base de datos *iris*, demostraremos inmediatamente que si hay homogeneidad de varianza en el ancho del sépalo (Figura 2.2):

```
##
##  Bartlett test of homogeneity of variances
```

Table 2.1: primeras 6 observaciones del dataframe resultante de augment

Sepal.Width	Species	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.std.resid
3.5	setosa	3.428	0.048	0.072	0.02	0.341	0.000	0.214
3.0	setosa	3.428	0.048	-0.428	0.02	0.339	0.011	-1.273
3.2	setosa	3.428	0.048	-0.228	0.02	0.340	0.003	-0.678
3.1	setosa	3.428	0.048	-0.328	0.02	0.340	0.006	-0.975
3.6	setosa	3.428	0.048	0.172	0.02	0.341	0.002	0.511
3.9	setosa	3.428	0.048	0.472	0.02	0.339	0.013	1.404

```
##
## data:  Sepal.Width by Species
## Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

Debido a ello, podemos testar si los residuales tienen una distribución normalidad de los residuales, para esto lo primero que debemos hacer es un ANOVA, como fue explicado en el práctico anterior y guardar este objeto con un nombre:

2.2.2.1 Extracción de los residuales del modelo

Para extraer los residuales, podemos hacerlo de dos formas, si solo queremos un vector de sus valores, podemos extraerlo desde el modelo mismo utilizando *$residuals*. Si queremos guardarlo en un dataframe mas completo podemos utilizar la función *augment* del paquete *broom*.

La segunda opción nos entregará más información que podremos utilizar más tarde, pero ambas sirven para testear normalidad, la siguiente tabla muestra las primeras 6 observaciones generadas por la función *augment*, donde *resid*, son los residuales (Ver tabla 2.1.

2.2.2.2 Inspección visual de los residuales

Existen dos formas de visualizar los residuales para determinar si la distribución de estos es o no es normal, histogramas y el qqplot.

2.2.2.2.1 Histograma

Los histogramas nos darán una representación visual para tratar de entender si la distribución es normal, para esto, solo necesitamos usar el comando *hist*, seguido del vector de los residuales, este es el comando para hacer el histograma (Figura 2.3) con cualquiera de las dos bases de datos, el resultado debiera ser el mismo:

```
hist(Residuales)
hist(Resultados$.resid)
```

2.2.2.2.2 QQplot

El qq plot es otra forma visual de establecer si los residuales son o no son normales, para esto, lo esperado es que la gráfica resultante sea una diagonal lo mas recta posible, para esto usaremos la función *qqnorm*, con nuestros residuales, de nuevo, podemos usar cualquiera de las dos versiones de nuestros datos:

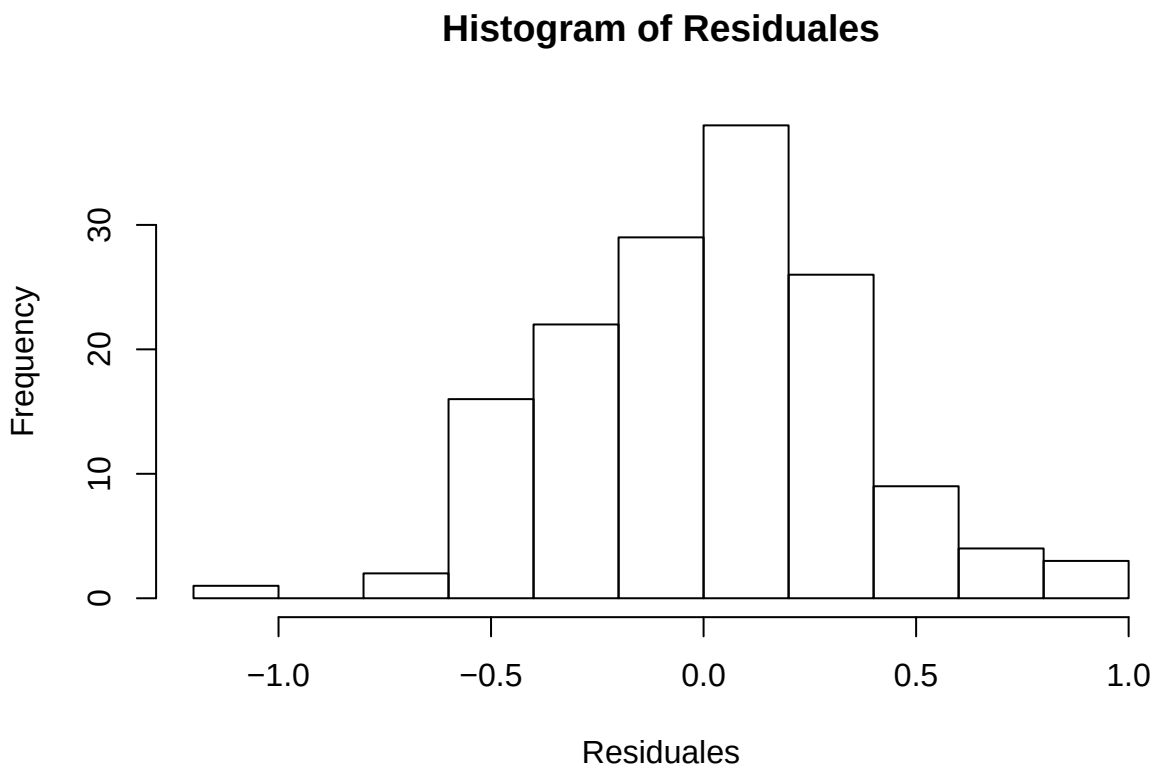

Figure 2.3: Histograma de los resiudales del modelo ANOVA

```
qqnorm(Residuales)
qqnorm(Resultados$.resid)
```

2.2.2.3 Test de Shapiro para determinar normalidad

La forma más sencilla de determinar normalidad es usando el test de Shapiro-Wilk de normalidad (Royston, 1995). Al igual que el test de Bartlett, si el valor de p es menor a 0.05, determinamos que la distribución de los datos no son normales, la función en *R* para este test es *shapiro.test*, y al igual que en los casos anteriores de *hist* y *qqpot*, solo necesitamos de usar un vector de residuales para ver el resultado del test. En nuestro caso:

```
shapiro.test(Residuales)
shapiro.test(Resultados$.resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Residuales
## W = 0.98948, p-value = 0.323
```

Ya que el valor de p es menor a 0.05, podemos decir que la distribución de nuestros residuales es normal, y por lo tanto el test cumple con los supuestos, y esto hace que sea valido el ANOVA, por lo que podemos ver nuestros resultados. La homogeneidad de Varianza es mas importante que la normalidad de residuales para estos casos, para ejemplos de lo que se debe hacer si se viola la normalidad ver Lix et al. (1996)

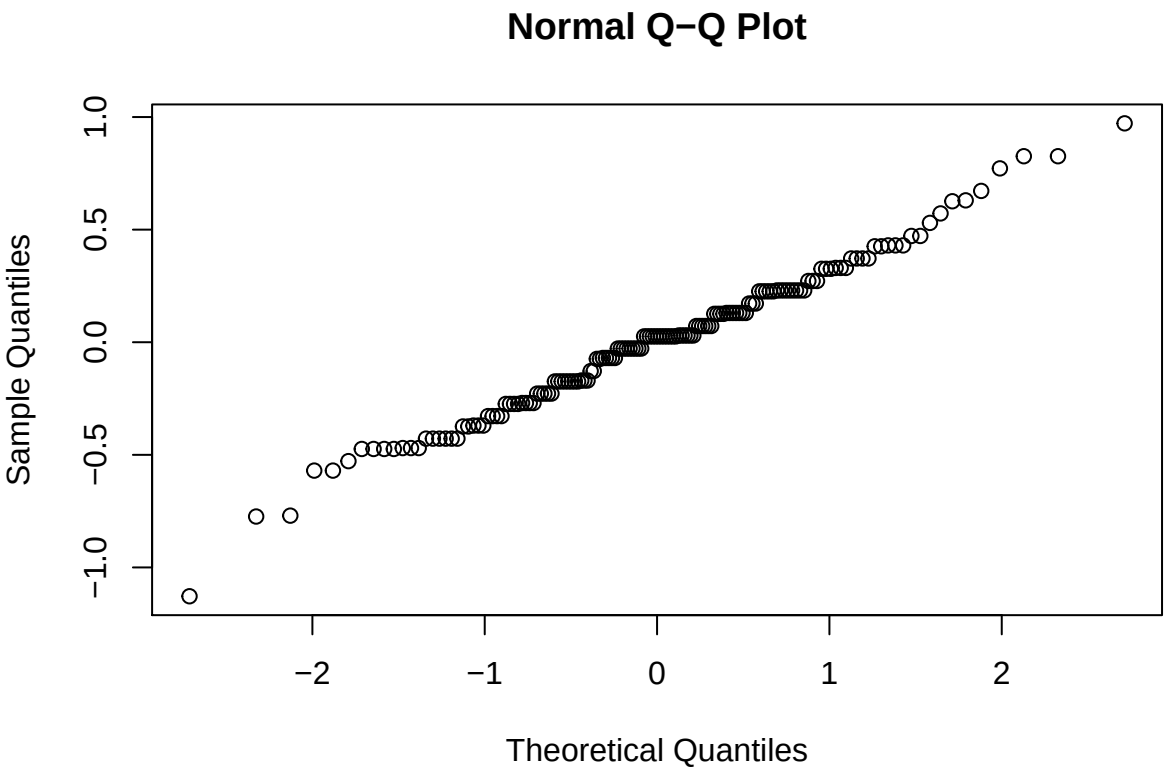

Figure 2.4: qqplot de los resiudales del modelo ANOVA

2.3 Actividad 2 Suma de cuadrados

Tanto los ANOVAS como las regresiones lineales se basan en minimizar la suma de cuadrados, es la suma de los cuadrados de los errores o residuales.

2.3.1 ¿Que es el error? ¿Por qué al cuadrado??

En la figura y en la formula vemos ejemplificado que es el error, también conocido como residual, este es simplemente el valor observado

$$Observado - Predicho$$

El objetivo de todo modelo es el de minimizar estos errores, al ajustar el mejor modelo posible.

Los errores siempre se calculan al cuadrado, discutiremos por que en clase

$$\sum_{i=1}^{n}(Observado - Predicho)^2$$

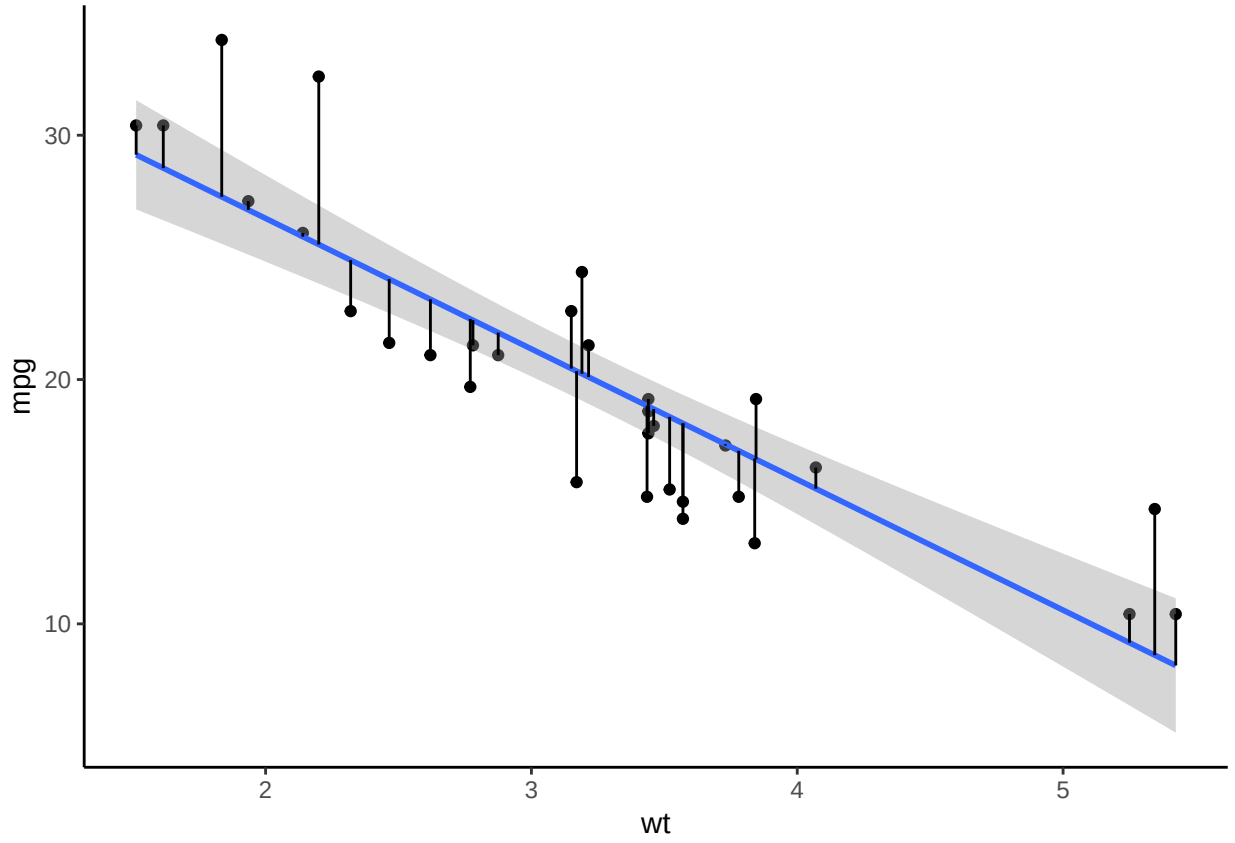

Figure 2.5: Errores de una regresión lineal ejemplificados con la linea entre el valor predicho y el observado

Web Page Blocked!

You have tried to access a web page which is in violation of your internet usage policy.

URL: http://admin.derek-corcoran-barrios.com/shiny/rstudio/sample-apps/Shiny1/?showcase=0
Category: Unrated
Client IP: 10.50.121.136
Server IP: 18.218.65.225
User name:
Group name:

To have the rating of this web page re-evaluated please click here.

Proceed Go Back

2.4 Referencias

Chapter 3

Análisis de poder y primera tarea

3.1 Obejtivos del práctico

- Entender cálculos de poder en base a matriz de confusión
- Primera tarea de práctico

3.2 Matriz de confusión

La matriz de confusión es una herramienta de toma de decisiones, en el caso especial de la toma de decisiones tenemos la siguiente matriz de confusión (Tabla 3.1)

Esto puede ser fácilmente ejemplificado con el problema de una alarma de humo (tabla3.2), en este caso cuando la alarma suena y no hay fuego y suena la alarma tenemos un error de tipo 1, en cambio si hay fuego y la alarma no suena tenemos un error de tipo 2

3.2.1 Poder y matriz de confusión

- Probabilidad de que suene la alarma cuando no hay fuego
  - α usualmente 5%
  - una de cada 20 alarmas es falsa
  - ¿Cuál es el α de una alarma de auto?
- Probabilidad de que no suene la alarma cuando hay fuego
  - β si es 10% uno de cada 10 fuegos no es detectado
  - poder es $1 - \beta$ confianza de que fuegos son detectados

Table 3.1: Tabla de confusión de errores

	Hipótesis nula cierta	Hipótesis alternativa cierta
Acepto hipótesis nula	No hay error	Error tipo 2
Acepto hipótesis alternativa	Error tipo 1	No hay error

Table 3.2: Matriz de confusión de una alarma de incendio

	No hay fuego	Hay fuego
No suena alarma	No hay error	Error tipo 2
Suena alarma	Error tipo 1	No hay error

3.3 Calculo de poder en R

Para hacer cálculos de poder en ANOVAS de una y dos vías en *R*, utilizamos el paquete *pwr2* (Lu et al., 2017). En este paquete podemos utilizar la función *pwr.1way* para determinar el poder de un ANOVA de una vía, los argumentos de esta función son:

- K: El número de grupos a testear
- n: Número de individuos por grupo
- *Alpha*: Nivel de significancia
- *Delta*: Valor mínimo a detectar
- *Sigma*: Desviación estándar de la muestra

Para cálculos precisos de n necesarios para muestras usar la siguiente app

Web Page Blocked!

You have tried to access a web page which is in violation of your internet usage policy.

URL: http://admin.derek-corcoran-barrios.com/shiny/rstudio/sample-apps/Shiny3/?showcase=0
Category: Unrated
Client IP: 10.50.121.136
Server IP: 18.218.65.225
User name:
Group name:

To have the rating of this web page re-evaluated please click here.

Proceed Go Back

3.4 Tarea

3.4.1 El problema

Una compañía que genera pesticidas descarga parte de sus desechos a un río. La ONG **RioSano**, dice que ha notado una alza en la mortalidad de los patos cortacorriente (*Merganneta armata*) del río.

Ante esto la empresa contrata un científico, el cual hace una estimación de la mortalidad de patos en 10 zonas del río en que descargan sus desechos, y lo compara con otros dos ríos no contaminados. Este científico dice que no hay diferencias significativas en la mortalidad de los patos de los ríos con desechos y sin desechos con una confianza del 95%. Para esto muestra como evidencia la figura 1 y tabla 3 e incluso hace públicos sus datos en el archivo *MuestraPatos.csv*.

La ONG *RioSano* lo contrata para determinar la validez del estudio y si es necesario generar un estudio extra. Ante esto:

1. Genere una matriz de confusión del problema y explique en este contexto que significaría el alfa y beta para este problema, y cual consideraría más relevante.

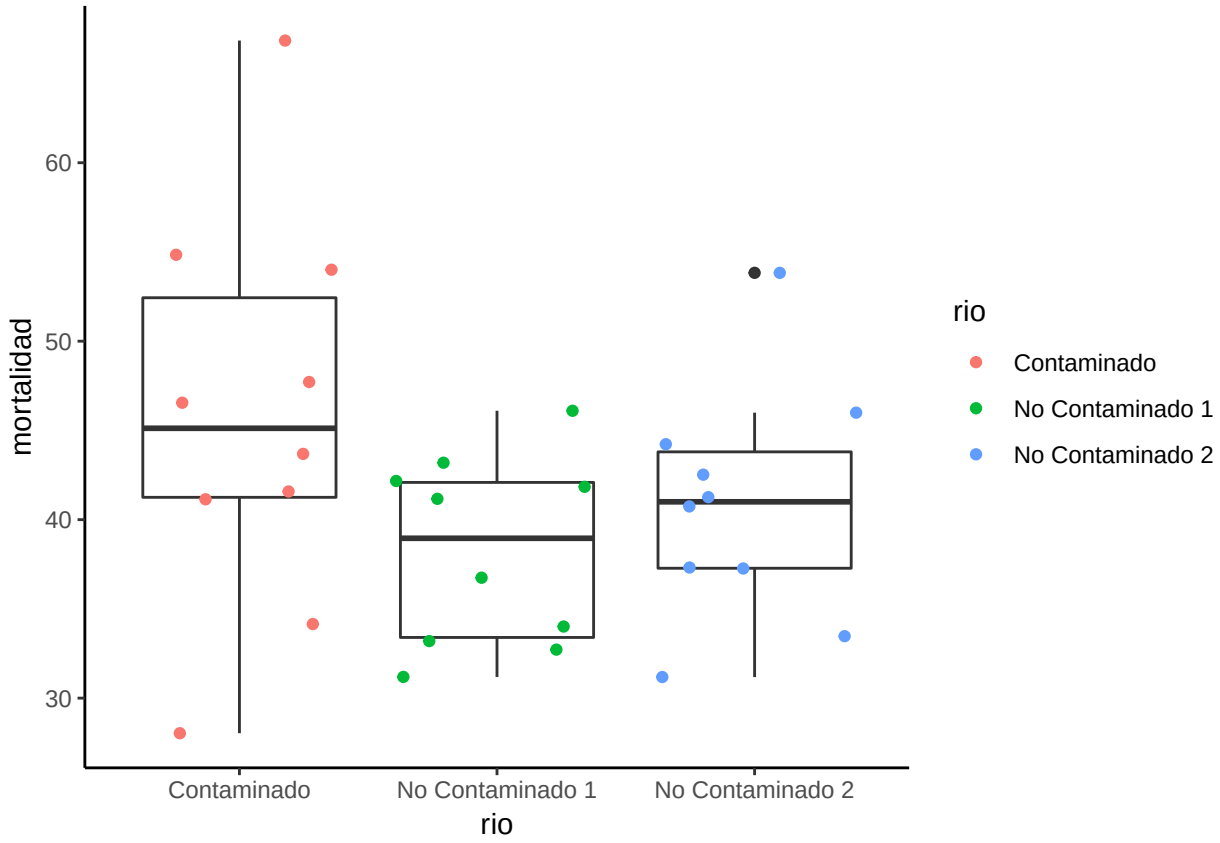

Figure 3.1: Mortalidades calculadas en 10 zonas de tres ríos

Table 3.3: Tabla de ANOVA de una vía de la mortalidad de patos de los tres ríos

term	df	sumsq	meansq	statistic	p.value
rio	2	301.2531	150.62656	2.359899	0.1136188
Residuals	27	1723.3436	63.82754	NA	NA

2. Diseñe el estudio que le gustaría hacer, determinando cuantas áreas debe muestrear por río, estime un delta mínimo que le gustaría determinar y el beta con el que se siente seguro y determine el n mínimo necesario para ese estudio. Justifique su respuesta

3. Dado este n mínimo realice lo siguiente

    - Realice un muestreo de n muestras por tipo de río del archivo *Patos.csv*
    - Genere gráficos y tablas exploratorias de los datos de su muestreo y describalas
    - Revise los supuestos del ANOVA para su base de datos tanto gráficamente como con tests y determine si se puede realizar el anova
    - Diga si según su diseño hay diferencias significativas en la mortalidad de patos entre los ríos

4. Cada zona a muestrear requiere de un monitoreo exhaustivo, que tiene un costo de 500.000 pesos (esto es 1.500.000 de pesos si consideramos los 3 ríos). La ONG *RioSano* consiguió 20.000.000 de pesos para este estudio. Dadas esas limitaciones, genere un balance de α, β y n dada esa limitación para hacer el mejor estudio posible dadas las consecuencias, justifique su respuesta.

Genere un informe para la ONG *RioSano* incorporando estos 5 puntos e incluya una introducción, metodología, resultados, discusión-conclusión y bibliografía, envíe el script de como generó los resultados

Chapter 4

Prueba t de Student

Puedes encontrar una versión interactiva de esta guía aquí.

La prueba t de student fue desarrollada por Gosset cuando trabajaba para la cervecería Guinness (Student, 1908). Esta prueba permite comparar las medias de una muestra con la media teórica de una población, o comparar dos poblaciones. Una de las características de la prueba de student, es que permite la alternativa de ver si dos medias son diferentes o, si uno busca más confianza determinar si una media es mayor, o menor que otra. Para la prueba t de Student, se determina un valor de t, usando la siguiente formula (ecuación (4.1)):

$$t = \frac{(\bar{x} - \mu)/(\frac{\sigma}{\sqrt{n}})}{s} \tag{4.1}$$

El estadístico t posee un valor de p asociado dependiendo de los grados de libertad de la prueba.

4.0.1 Pruebas de una muestra

Las pruebas de una muestra nos permiten poner a prueba si la media de una población son distintas a una media teórica. Como ejemplo veremos el caso de las erupciones del géiser *Old Faithful*, localizado en el Parque Nacional Yellowstone. Un guarda-parque del lugar dice que este géiser erupta cada 1 hora. Por suerte *R* posee una base de datos de Azzalini and Bowman (1990) llamada *faithfull*, la cual utilizaremos para determinar si esto es cierto o no usando la función `t.test`. Esta base de datos tiene dos columnas *eruptions*, que muestra la duración en minutos de cada erupción y *waiting* que presenta la espera en minutos entre erupciones.

Cuando usamos esta función con una muestra necesitamos llenar 2 argumentos:

- **x:** Un vector con los valores numéricos de a poner a prueba
- **mu:** La media teórica a poner a prueba
- **alternative:** Puede ser "two.sided", "less" o "greater", dependiendo de si uno quiere probar que la muestra posee una media distinta, menor o mayor que la media teórica.

En este caso haríamos lo siguiente

```
data("faithful")
t.test(x = faithful$waiting, mu = 60, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  faithful$waiting
## t = 13.22, df = 271, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 60
## 95 percent confidence interval:
##  69.27418 72.51994
## sample estimates:
## mean of x
##  70.89706
```

En este caso el valor de p nos dice que la media es diferente a 60.

4.0.1.1 Ejercicio 1

La base de datos *airquality* (incorporada como ejemplo en **R**), muestra entre otras variables las partículas de ozono en Nueva York, cada día de Mayo a Septiembre de 1973 entre las 13:00 y las 15:00 (Chambers et al., 1983). Supongamos que ustedes están a cargo de una agencia ambiental, y están estudiando en que meses deben reducir la actividad vehicular de Nueva York. Para esto planean disminuir a la mitad los pasajes del metro de Nueva York todos los meses que en promedio tengan sobre 55 ppb. Para esto deben comprobar estadisticamente que el mes en que harán esto tiene promedios sobre 55.

4.0.2 Pruebas de dos muestras

Las pruebas de dos muestras nos permiten ver si hay diferencias significativas entre las medias de dos muestras. En la base de datos `mtcars`, hay una columna que determina si los vehículos son de cambios manuales o automáticos. En este caso 0 significa automático y 1 significa manual. En la figura 4.1 podemos ver una inspección gráfica de las posibles diferencias.

Para hacer la comparación debemos agregar el argumento `var.equal` el cual en este caso asumiremos que es verdad, ya que en la próxima sección veremos los supuestos de la prueba t y las consecuencias de las violaciones de estos supuestos. En este caso podemos usar el símbolo `~` a ser leído como explicado por para la prueba t de dos muestras.

```
t.test(mpg ~ am, data = mtcars, var.equal =TRUE)
```

```
##
##  Two Sample t-test
##
## data:  mpg by am
## t = -4.1061, df = 30, p-value = 0.000285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.84837  -3.64151
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

En este caso se determinaría que los vehículos manuales (am = 1), son más eficientes que sus contra-partes automáticas.

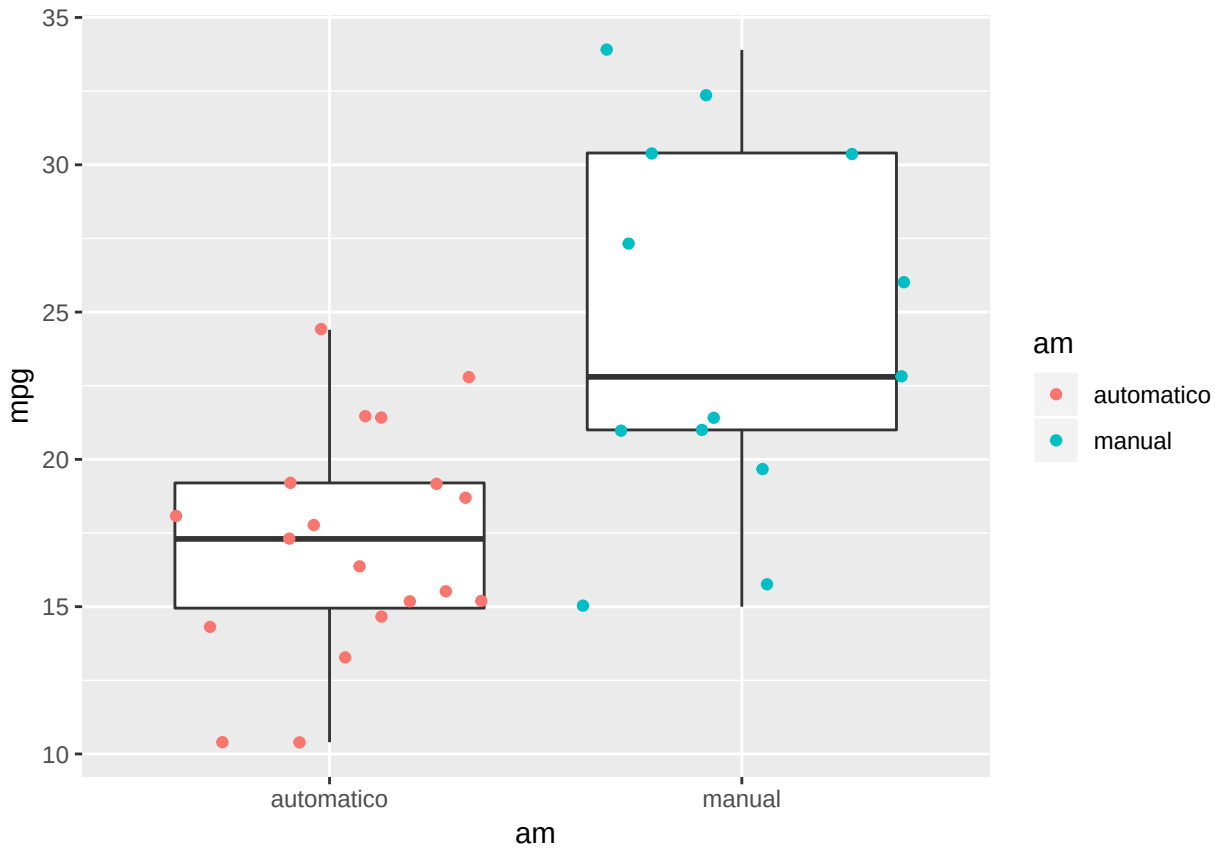

Figure 4.1: Comparación de eficiencia entre vehiculos automaticos y manuales

4.0.2.1 Ejercicio 2

Para el siguiente ejercicio usaremos la base de datos `BeerDark` disponible en webcursos o en el siguiente link. Esta base de datos posee 7 columnas, pero usaremos solo 4 de ellas:

- **Estilo:** Separa las cervezas entre Porters y Stouts
- **Grado_Alcoholico:** El grado alcohólico de las cervezas
- **Amargor:** Valor IBU (International Bittering Units), a mayor valor más amarga la cerveza
- **Color:** A mayor valor más oscura la cerveza.

Determinar si las cervezas Porter y Stouts son distintas en grado alcohólico, amargor y/o color.

4.1 Supuestos de la prueba de t y alternativas

Los supuestos de la t de student son las siguientes (Boneau, 1960)

- Independencia de las observaciones
- Distribución normal de los datos en cada grupo
- Homogeneidad de varianza

4.1.1 Prueba de una muestra

Como siempre la independencia de las muestras es algo que solo puede determinarse en base a el diseño del muestreo, y por otro lado, al haber solo una muestra, la homogeneidad de varianza no es un problema, en este caso solo podemos ver si la distribución es normal. Volviendo a nuestro ejemplo de una muestra, con la base de datos `faithfull`, veamos en base a un histograma (figura 4.2), qqplot (figura 4.3) y test de shapiro, si los datos son normales o no:

```
hist(faithful$waiting, xlab = "Minutos de espera entre erupciones")
```

```
qqnorm(faithful$waiting)
```

```
shapiro.test(faithful$waiting)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  faithful$waiting
## W = 0.92215, p-value = 1.015e-10
```

Como vemos en la figura 2, los datos no se ven normales, incluso se ven bimodales, lo cual significa que tiene 2 picos, en este caso uno al rededor de los 52 minutos y otro al rededor de los 85 minutos de espera (recordemos que la función `hist`, automáticamente usa el algoritmo de Sturges (1926), para determinar como dividir los datos y obtener el mejor histograma). Nuestras sospechas de no normalidad son confirmadas al ver el qqplot, que no sigue para nada la diagonal, y es reafirmado por el test de shapiro, cuyo valor mucho menor a 0.05, nos dice que la distribución no es normal. Dado esto, debemos apelar a un test de distribución libre como el de *Mann-Whitney*, la cual se realiza con la función `wilcox.test`, de la misma forma que es utilizada la función `t.test`, por lo tanto para nuestro ejemplo usamos:

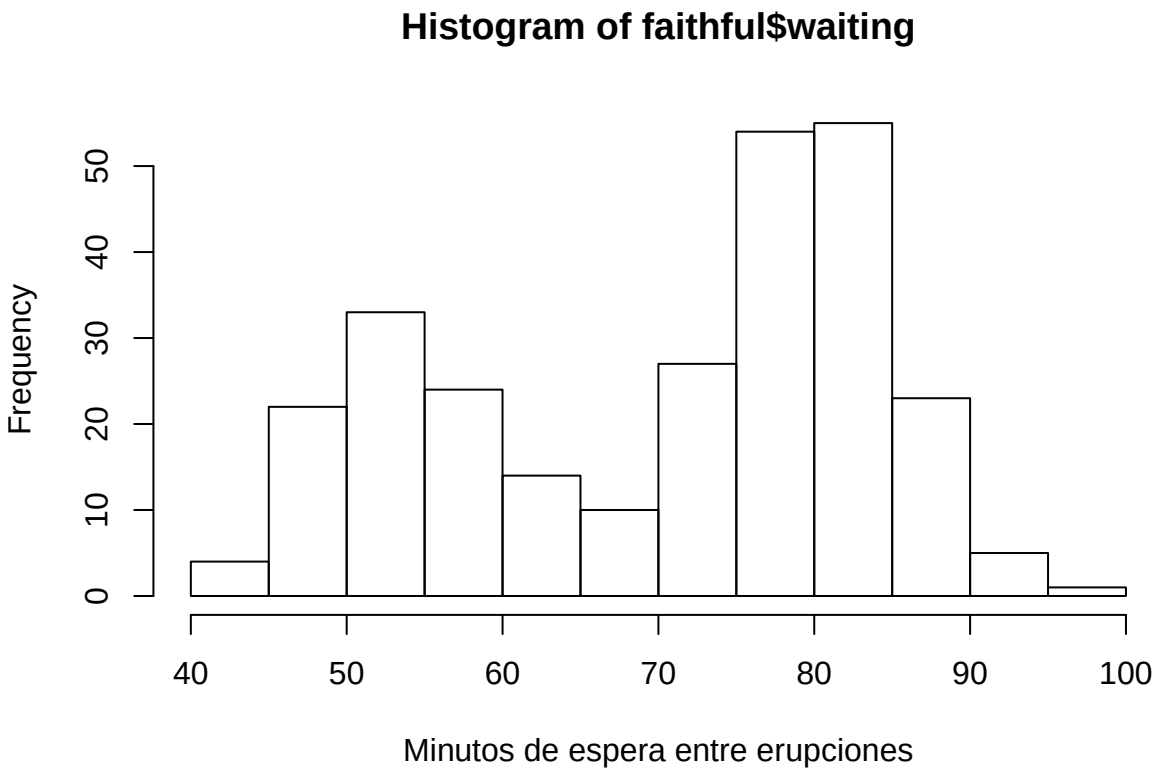

Figure 4.2: Histograma de los minutos de espera de el géiser Old Fiathful

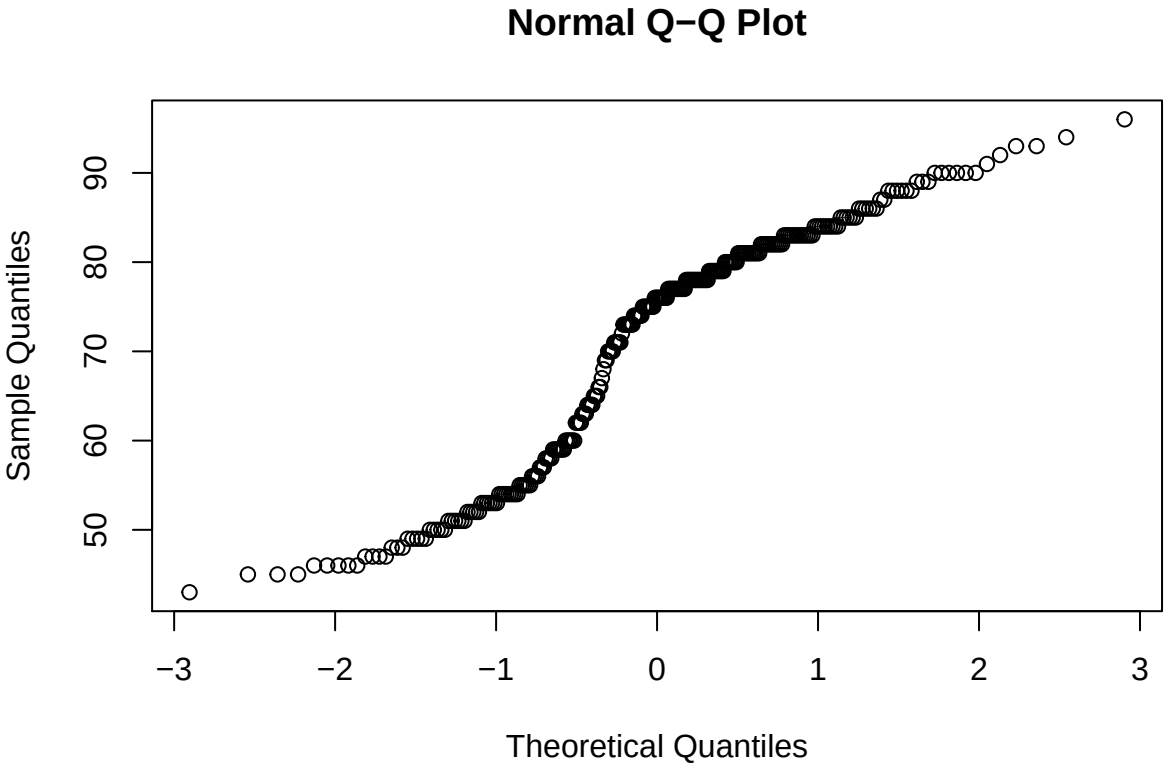

Figure 4.3: QQplot de los minutos de espera de el géiser Old Fiathful

```
data("faithful")
wilcox.test(x = faithful$waiting, mu = 60, alternative = "two.sided")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  faithful$waiting
## V = 31048, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 60
```

Que en este caso nos lleva a la misma conclusión que nuestro ejemplo anterior.

4.1.2 Prueba de dos muestras

Para una prueba de dos muestras, podemos testear tanto la homogeneidad de varianza como la normalidad, para ver las dos cosas al mismo tiempo podemos usar un gráfico de violín (figura 4.4). En este caso, las distribuciones no se ven muy diferentes a la normalidad, pero las varianzas se ven un tanto distintas, podemos seguir explorando esto visualmente usando la función `hist` previamente generando dos data frames, uno para autos automático y otro para manuales.

```
data("mtcars")
mt <- mtcars
mt$am <- ifelse(mtcars$am  == 0, "automatico", "manual")
mt <- as.data.frame(mt)
```

```
ggplot(mt, aes(x = am, y = mpg)) + geom_violin()
```

En este caso, las distribuciones no se ven muy diferentes a la normalidad, pero las varianzas se ven un tanto distintas, podemos seguir explorando esto separando los datos en vehículos automáticos y manuales para hacer histogramas, en este caso es importante que los ejes sean iguales, para eso en el histograma usaremos los parámetros ylim y xlim.

```
hist(manuales$mpg, xlim = c(10,35), ylim = c(0,5))
```

```
hist(autos$mpg, xlim = c(10,35), ylim = c(0,5))
```

Como vemos, los vehículos manuales no parecen tener distribución normal como se ve en la figura 4.5, esto podemos comprobarlo con el qqlot de los mismos datos (figura 4.7)

```
qqnorm(manuales$mpg)
```

4.1.2.1 Ejercicio 3

Como siempre la independencia de las muestras es algo que solo puede determinarse en base a el diseño del muestreo, y por otro lado, al haber solo una muestra, la homogeneidad de varianza no es un problema, en este caso solo podemos ver si la distribución es normal. Volviendo a nuestro ejercicio de una muestra, con la base de datos `airquality`, evalúe basado en histograma, qqplot y test de shapiro si se debe revaluar la hipótesis para los meses de julio y agosto

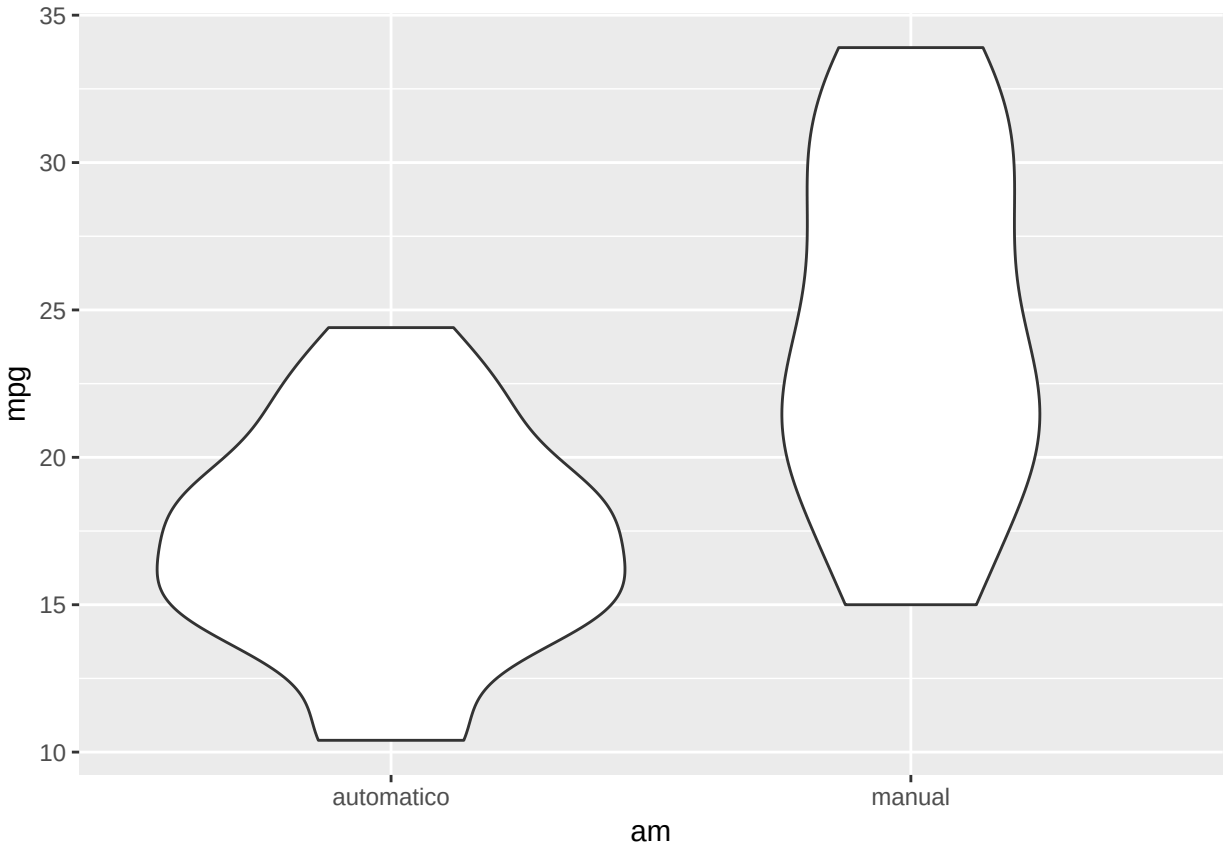

Figure 4.4: Comparación de distribuciones y varianzas de los vehiculos automáticos

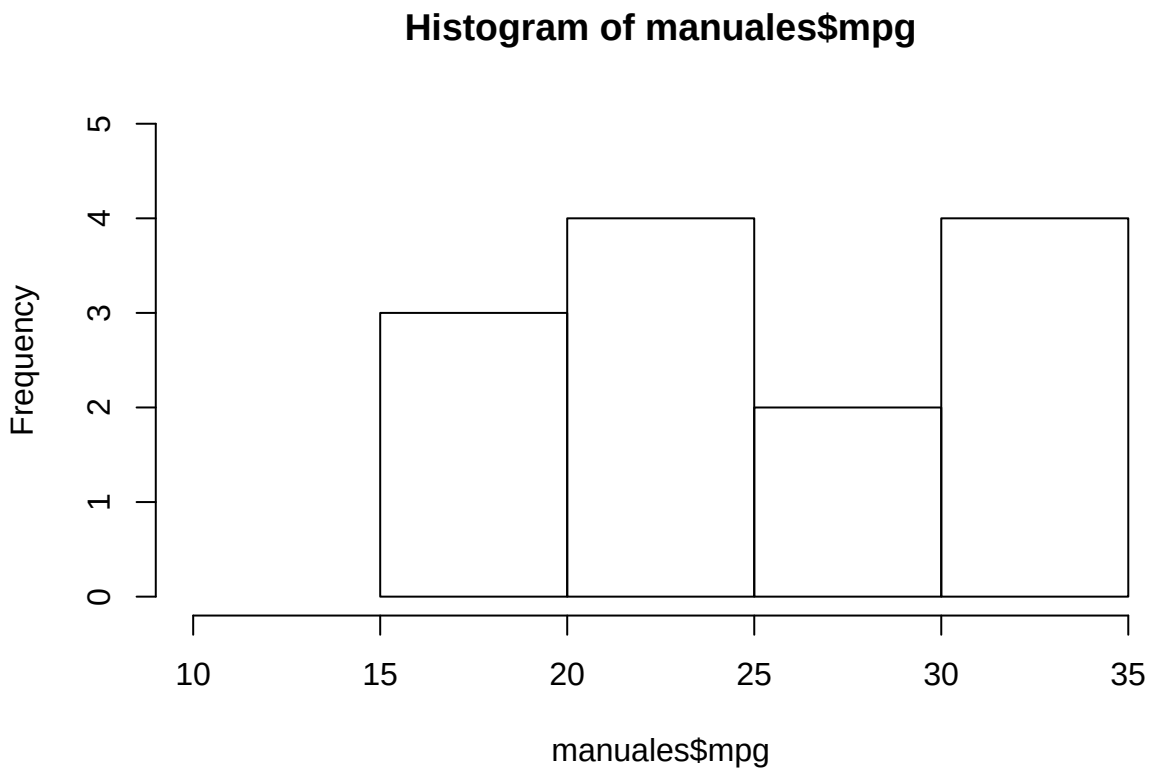

Figure 4.5: Histograma de vehiculos manuales

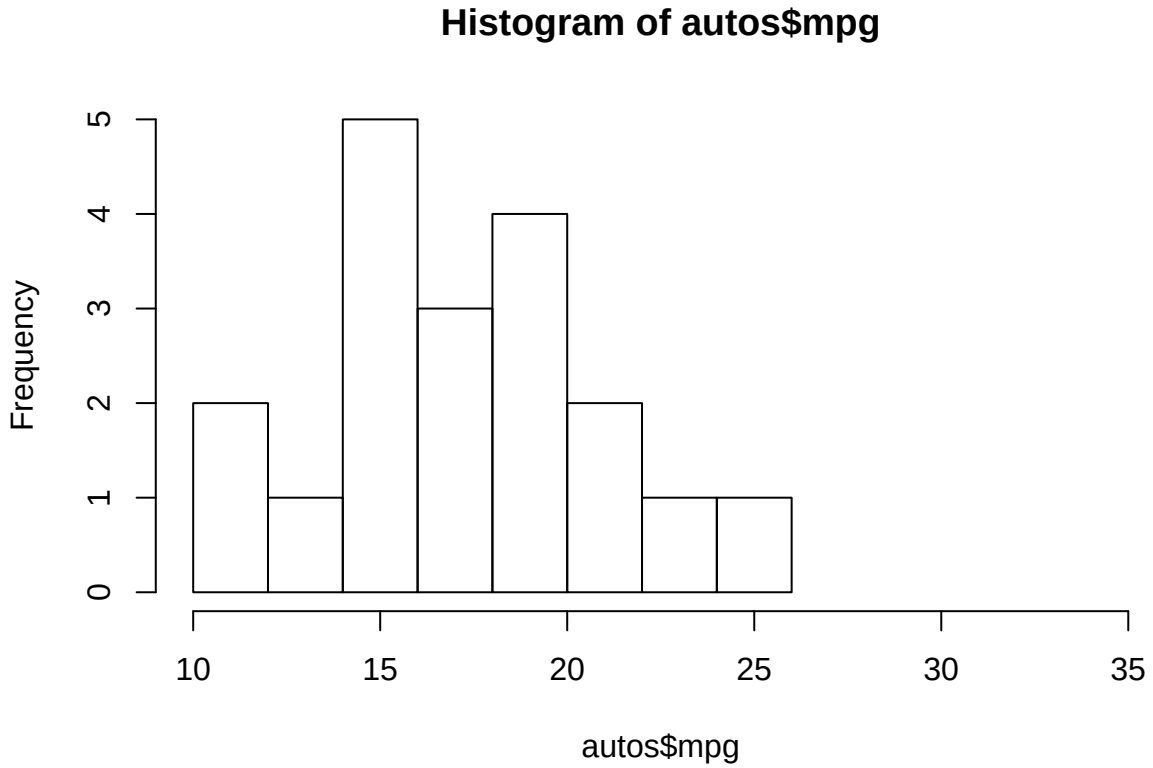

Figure 4.6: Histograma de vehiculos automáticos

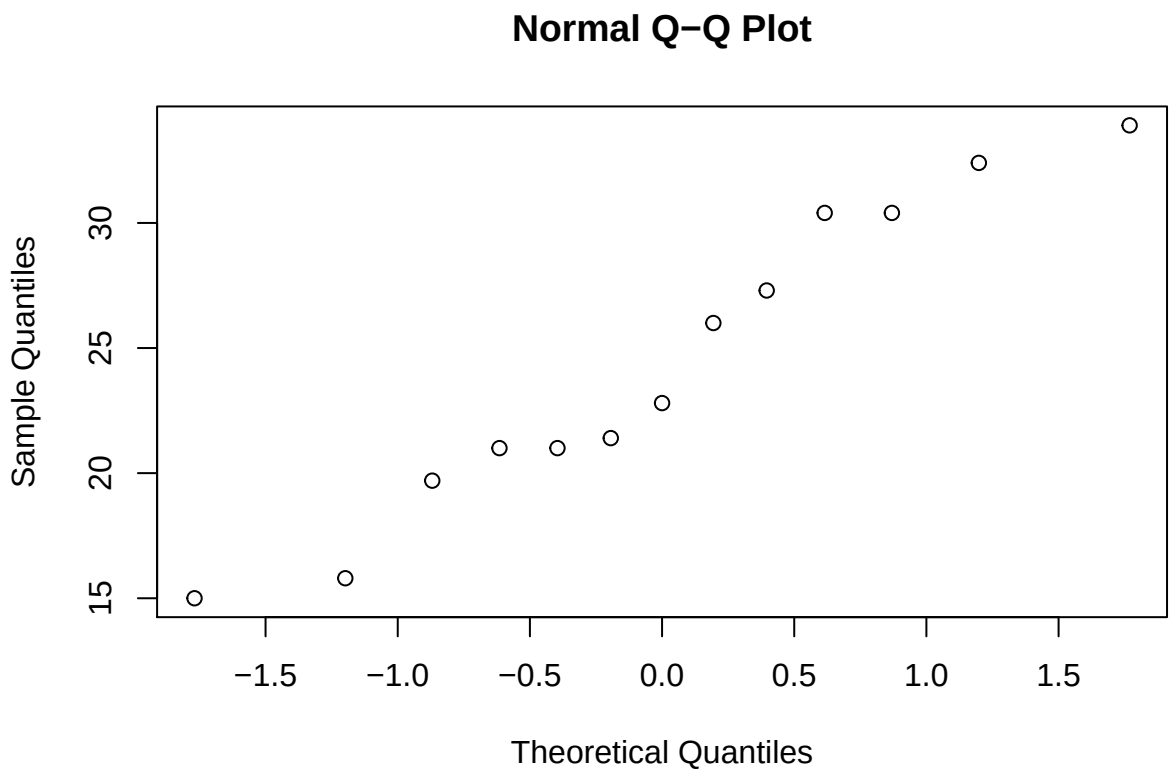

Figure 4.7: QQplot de eficiencia de vehiculos con cambios manuales

Para una prueba de dos muestras, podemos testear tanto la homogeneidad de varianza como la normalidad, para ver las dos cosas al mismo tiempo podemos usar un gráfico de violín `geom_violin` en *ggplot2*, lo cual puede seguir siendo explorando esto visualmente usando la función `hist` generando dos data frames, uno por cada clase de datos.

Evalúe si es necesario revaluar la hipótesis de que el amargor es distinto entre ambos estilos de cerveza

4.2 Bibliografía

Bibliography

Anderson, E. (1935). The irises of the gaspe peninsula. *Bulletin of the American Iris society*, 59:2–5.

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the old faithful geyser. *Applied Statistics*, pages 357–365.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A*, 160(901):268–282.

Beall, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. *Biometrika*, 32(3/4):243–262.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological bulletin*, 57(1):49.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). Graphical methods for data analysis. 1983. *Wadsworth, Belmont, CA*, 35.

Kross, S., Carchedi, N., Bauer, B., and Grdina, G. (2017). *swirl: Learn R, in R*. R package version 2.4.3.

Lix, L. M., Keselman, J. C., and Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance f test. *Review of educational research*, 66(4):579–619.

Lu, P., Liu, J., and Koestler, D. (2017). *pwr2: Power and Sample Size Analysis for One-way and Two-way ANOVA Models*. R package version 1.0.

Potvin, C., Lechowicz, M. J., and Tardif, S. (1990). The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures. *Ecology*, 71(4):1389–1400.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Royston, P. (1995). Remark as r94: A remark on algorithm as 181: The w-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):547–551.

Savage, V. M. and West, G. B. (2007). A quantitative, theoretical framework for understanding mammalian sleep. *Proceedings of the National Academy of Sciences*, 104(3):1051–1056.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.