Estadística Avanzada para Ciencias Naturales

Dr. Luciano Selzer

2018-03-13

Contents

Chapter 1

Reglamento

Para aprobar la asignatura el alumno deberá tener obtener una nota de cursada mínina de 60%, aprobar los parciales. La nota se calcula ponderando los siguientes items.

Actividad	Ponderación
Asistencia a clases, participación y evaluación de pares	5%
Entrega de Ejercicios	10%
Cuestionario de Comprensión	10%
Laboratorios en R	15%
Parcial 1	30%
Parcial 2	30%
Total	100%

1.1 Asistencia a clases, participación y evaluación de pares

Se espera que los alumnos vayan a clases y participen activamente de la misma, con preguntas y respuestas. La asistencia y participación es una parte pequeña pero no insignificante de la nota de cursada. Si bien puede pedirse a algún alumno en particular que responda una pregunta o resuelva un ejercicio, se espera que sean participativos sin tener que ser llamados.

A través del cuatrimestre también se pedirá a los alumnos que completen algunas evaluaciones de pares. Estas serán usadas para asegurar que todos los miembros del equipo contribuyan al éxito del grupo y poder resolver los posibles problemas de manera temprana.

1.2 Entrega de Ejercicios

Al principio de cada unidad se harán un pequeño número de ejercicios de forma manual.

El objetivo de los problemas es ayudar a desarrollar una comprensión profunda de los temas y preparar a los alumnos para los parciales y el proyecto. Se evaluará la precisión y la completitud. Para recibir la nota deben mostrar todo el trabajo.

Se espera que los alumnos colaboren, y se alienta, pero cada uno debe entregar su propio trabajo. Si se detectan copias ambos recibirán 0 para ese grupo de problemas.

La nota menor no se tomará en cuenta.

1.3 Laboratorios

Cada laboratorio se completará de forma individual. Se puede realizar consultas entre alumnos y se alienta la colaboración. Pero la entrega de informes se debe realizar de forma individual. Es importante que el informe sea completamente reproducible. Se puede asegurar de ello usando el botón *knit*. Deben entregar el archivo *.Rmd* y el *.html*.

La nota menor no se tomará en cuenta.

1.4 Cuestionario de Comprensión

Al final de la unidad se harán cuestionarios para evaluar la comprensión que tiene el alumno sobre el tema. El mismo se hará en la plataforma de Moodle, y tendrán una semana para realizarla. Luego de ese tiempo tendrán 0 puntos. Se harán 10 preguntas de elección múltiple.

La menor puntuación no será tenida en cuenta.

1.5 Parciales

La asignatura contará con 2 parciales prácticos (se considera parcial aprobado con el 60% bien desarrollado y si no se cometen errores conceptuales básicos en lo referente a los temas propios de la asignatura), con sus respectivos recuperatorios; como requerimiento para aprobar la cursada. Además de asistir al 80% de las clases prácticas de acuerdo con lo que estipula la Resolución N° 350/14. La asignatura tiene *promoción*. Para promocionar es necesario aprobar con más del 80% los parciales sin ir a recuperatorio, y tener una nota de cursada mayor al 80%. Y realizar un trabajo final que consiste en poner en práctica los conocimientos adquiridos.

Examen Final: El énfasis estará en los conocimientos teóricos y su interpretación con la aplicación. Se tendrá en cuenta la síntesis que alumno realice con los nuevos conceptos adquiridos. Para ello, el examen final consistirá un trabajo final donde los alumnos deberán analizar e interpretar datos provistos por la asignatura y un examen escrito donde deberán saber qué técnicas usar para resolver diferentes problemas. Los exámenes se rendirán en las fechas previstas en el Calendario Académico de la Facultad

1.6 Cronograma

Semana	Fecha	Día	Tema
1	13/3/2018	Martes	Visualización
1	15/3/2018	Jueves	Análisis Reproducible - Manejo de Datos
2	20/3/2018	Martes	Planificación de experimentos. ANOVA de un factor- Supuestos del ANOVA.
2	22/3/2018	Jueves	ANOVA: Contrastes- Modelo aleatorio
3	27/3/2018	Martes	Métodos no paramétricos para comparar varias muestras
3	29/3/2018	Jueves	Jueves Santo
4	3/4/2018	Martes	Anova de dos factores.
4	5/4/2018	Jueves	Modelo aleatorio. Modelo Mixto
5	10/4/2018	Martes	Diseños experimentales
5	12/4/2018	Jueves	Pruebas no paramétricas: Análisis, para dos muestras, varias muestras y para varias muestras relacionadas
6	17/4/2018	Martes	Repaso Regresión lineal simple

Semana	Fecha	Día	Tema
6	19/4/2018	Jueves	Regresión lineal múltiple
7	24/4/2018	Martes	Representación Matricial ANCOVA
7	26/4/2018	Jueves	Modelos Lineales generalizados
8	1/5/2018	Martes	Día del Trabajo
8	3/5/2018	Jueves	Modelos Lineales generalizados
9	8/5/2018	Martes	Parcial
9	10/5/2018	Jueves	Modelos Lineales generalizados
10	15/5/2018	Martes	Análisis Multivariado: Medidas de distancias y similitudes
10	17/5/2018	Jueves	Recuperatorio
11	22/5/2018	Martes	Análisis Multivariado: Clusters
11	24/5/2018	Jueves	Análisis Multivariado: Análisis de Componentes Principales
12	29/5/2018	Martes	Análisis Multivariado: Análisis de Coordenadas Principales
12	31/5/2018	Jueves	Análisis Multivariado: Análisis de Redundancia y Análisis de correspondencias canónicas
13	5/6/2018	Martes	Análisis Multivariado: Análisis discriminante.
13	7/6/2018	Jueves	Análisis Multivariado: Análisis discriminante.
14	12/6/2018	Martes	Introducción a Inferencia Multimodelo
14	14/6/2018	Jueves	Introducción a GAM
15	19/6/2018	Martes	Introducción a GAM
15	21/6/2018	Jueves	Parcial
16	26/6/2018	Martes	Introducción a Estadística Bayesiana
16	28/6/2018	Jueves	Recuperatorio

Chapter 2

Introducción a *R* y RStudio

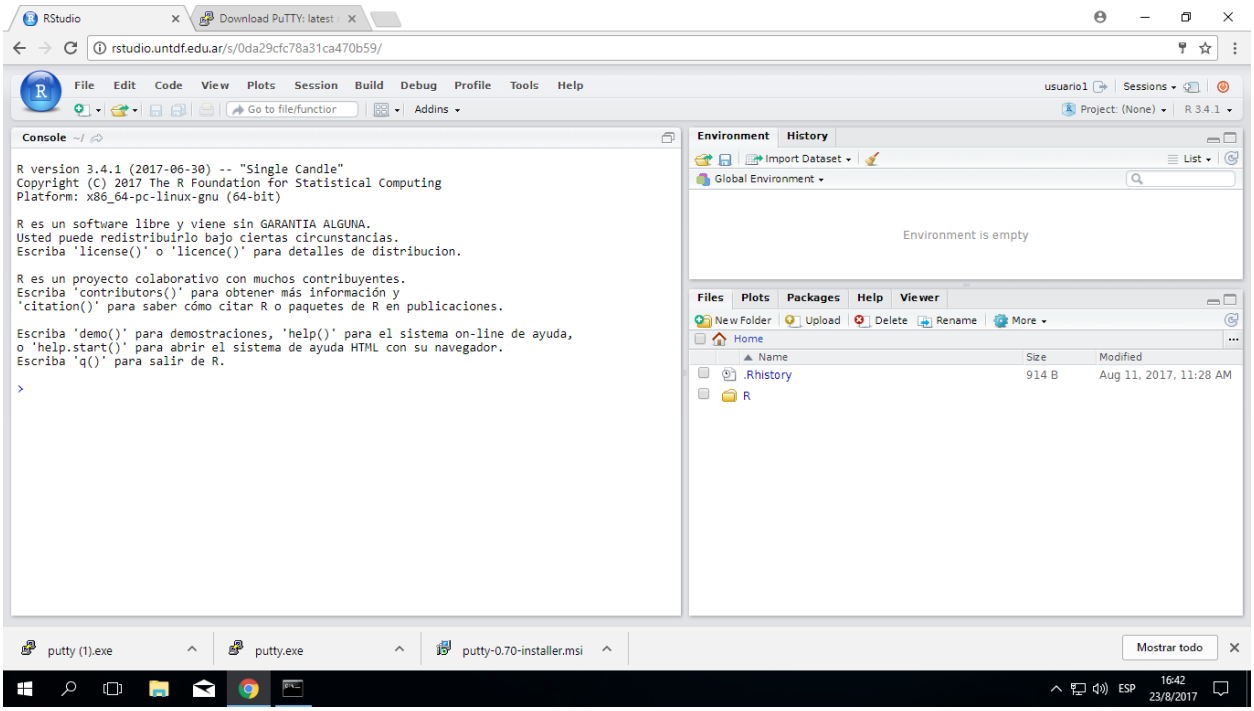
En la materia de Estadística Avanzada para Ciencias Naturales vamos a usar métodos estadísticos complejos. Por lo que, en muchos casos, es muy poco práctico realizar los cálculos a mano. Por eso, es necesario usar software estadístico específico para poder trabajar eficientemente. Existen múltiples programas, entre los más conocidos podes mencionar Statistica, Stata, SPSS, SAS, o Infostat, desarrollado en Argentina. Todos ellos son comerciales, y por lo tanto hay que pagar por las licencias de uso. Por otro lado, *R* es gratis y es usado ampliamente en el mundo. Otra de las ventajas, es la gran comunidad de usuarios y desarrolladores que se ha formado. Lo que hace que esté siendo constantemente actualizado y que las últimas técnicas estadísticas estén, muchas veces, implementadas directamente en R. Sin embargo, una de las dificultades es que hay que escribir comandos para hacer que funcione. Lo que es una desventaja al aprender pero luego se convierte en una ventaja, ya que permite automatizar tareas tediosas y además realizar análisis **reproducibles** de datos.

Otro problema de *R* es su interfaz muy poco amigable. Por eso se han desarrollado otras interfaces (llamadas entornos de desarrollo integrado o IDE por sus siglas en inglés) que hacen más fácil trabajar con este programa. Hay varias: RStudio, Tinn-R, RKward, etc. Nosotros vamos a usar RStudio por ser la más trabajo tiene encima y está más pulida.

Se puede bajar e instalar ambos programas en cualquier computadora. Hay que instalar *R* descargando desde https://cran.r-project.org/ y seguir las instrucciones del instalador. Para bajar RStudio hay que ir a https://www.rstudio.com/products/rstudio/download/#download y también seguir las instrucciones de instalación.

Para las clases tenemos instalado estos programas en un servidor de la Universidad y se puede acceder desde cualquier red (LAN o WiFi) de la sede de Yrigoyen entrando a https://rstudio.untdf.edu.ar. Verán una pantalla de login.

Una vez que entren verán una pantalla así.



2.1 RStudio

La interfaz de RStudio está dividida en varios paneles, y cada uno tiene varias pestañas. Arriba a la derecha está el *espacio de trabajo* (*Environment*), es donde van a aparecer los objetos que creen a medida que trabajan en R. La otra pestaña es el *historial* (*History*), donde quedan guardados todos los comandos que hayan ejecutado. Abajo de estos dos hay un panel con varias pestañas. *Archivos* (*Files*), muestra los archivos. *Gráficos* (*Plots*) es donde van a aparecer los gráficos que vayamos haciendo. *Paquetes* (*Packages*) muestra las librerías que tenemos y sus paquetes instalados y con un tilde los cargados (más adelante vamos a ver que son los paquetes). La *ayuda* (*Help*) es donde vamos a poder la ayuda de funciones de **R**. Y además está las pestaña del *Visor* (*Viewer*) que nos muestra una vista de los documentos que creemos.

Por el lado izquierdo está la *consola*, es donde pasa toda la acción. Todo lo que hagamos va a ser escrito como un orden o comando ahí y luego vamos a ver el resultado ahí o si es un gráfico en el panel de gráficos. Cada vez que iniciemos RStudio va a mostrar la consola con un mensaje que indica la versión de *R* y otros detalles. Debajo de ese mensaje está el *prompt*. Aquí es donde *R* espera que se ingresen los comandos. Y para interactuar con *R* hay que decirle que tiene que hacer. Los comandos y su sintaxis han evolucionado a lo largo de décadas y ahora proveen a los usuarios una manera natural de acceder y procesar datos, aplicar procedimientos estadísticos, etc.

Se puede usar *R* como una calculadora. Podemos poner una cuenta a realizar en el prompt y *R* nos devolverá el resultado. Por ejemplo, podemos poner:

```
2 + 2
```

```
## [1] 4
```

Prueben escribirlo en su consola justo después del “>”.

También es posible guardar los resultados en un objeto:

```
x <- 2 + 2
```

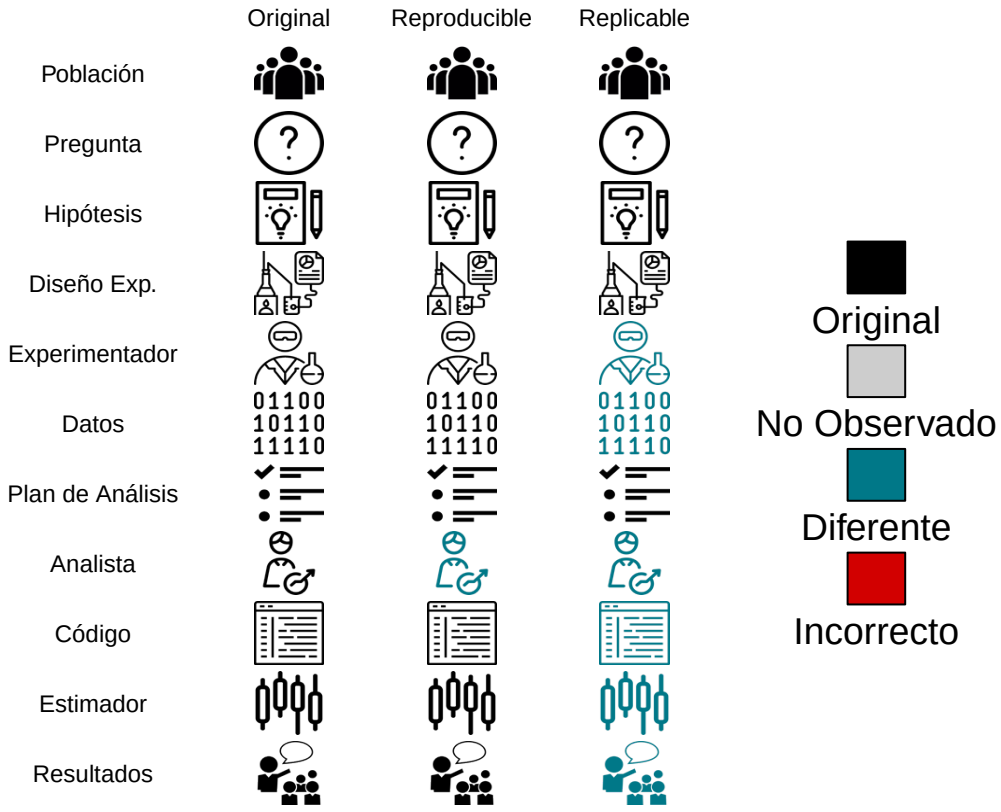


Figure 2.1: Diferencias entre reproducible y replicable.

Prueben hacerlo en su consola. En este caso, parece que no pasó nada. No apareció el resultado. Pero si observan en panel del espacio de trabajo verán que hay un nuevo objeto llamado `x`.

Prueben que sucede si escriben `x` en la consola.

2.2 Análisis Reproducible

Una ventaja que tiene *R* respecto a otros programas estadísticos es que permite reproducir el análisis de los datos. Reproducible significa que a partir de los mismos datos otro analista va a llegar a los mismos resultados. En cambio, replicable es que otro experimentador al repetir el experimento va a llegar a resultados diferentes, que pueden ser o no similares (Figura 2.1)

Los programas de interfaz gráfica, que no usan código o no lo proveen, complican la reproducibilidad de los análisis. Esto se debe a que es más complicado de comunicar como se realizó el análisis. La ventaja del código es que queda todo explícito en él.

Hay varias formas de trabajar con R. Una es de forma interactiva en la consola. Como cuando pusieron `2 + 2`. Esto es muy útil cuando estamos probando si algo funciona. Pero no guardamos el código de esta forma. Aunque, en verdad queda guardado en el orden en que lo ejecutamos en el historial no es útil porque se va sobreescribiendo y va quedando lo que funcionó y lo no que no. Otra forma de usar *R* es utilizando *scripts* (archivos con extensión *.R*). Es indispensable para crear nuevas funciones pero para analizar datos tiene sus desventajas. Ya que, si bien el código se puede comentar anteponiendo # a la línea que queremos comentar, es límitado el formato que podemos usar en el comentario. Además, tendremos que volver a correr el script para ver los resultados si no los guardamos explicitamente en un documento, hoja de cálculo, o imagen. Por último, tenemos los documentos de programación letrada. La programación letrada consiste en

mezclar código con texto plano, como en un procesador de texto.

En R, hay varias aproximaciones a esto. La que más éxito ha tenido es *knitr*. *To knit* es tejer en inglés, lo que hace es "tejer" el documento final con el resultado del código, es decir los análisis que hagamos, y texto explicando que hicimos, porque, y como. Es decir, podemos escribir un paper o informe completo. Para darle formato al texto se usa *markdown* que permite usar marcas livianas para poner *cursivas*, **negritas** o ~~tachado~~.

2.2.1 Rmarkdown

Hay muchas opciones para formatear el texto. La idea detrás de markdown es que se pueda escribir en un procesador de texto sencillo y las marcas sean fáciles de poner y no interrumpan la lectura. Algunos ejemplos:

2.2.1.1 Énfasis

```
*cursiva*   **negrita**

_cursiva_   __negrita__
```

2.2.1.2 Títulos

```
# Título 1

## Título 2

### Título 3
```

2.2.1.3 Listas

```
Lista Desordenada
* Item 1
* Item 2
    + Item 2a
    + Item 2b
Lista Ordenada
1. Item 1
2. Item 2
3. Item 3
    + Item 3a
    + Item 3b
```

2.2.1.4 Saltos de línea manuales

Termina una línea con dos o más espacios:

Las rosas son rojas,
las violetas son azules.

2.2.1.5 Vínculos

Usa una dirección http simple o agrega un vínculo a una frase:

```
http://example.com
```

```
[frase vínculada](http://example.com)
```

2.2.1.6 Imágenes

Imágenes en la web o en el mismo directorio de trabajo:

```
![alt text](http://example.com/logo.png)
```

```
![alt text](figures/img.png)
```

Ejercicio 2.1 (Probando markdown)**.** Descarguen un archivo de RMarkdown usando este código en la consola:

```
download.file("url", "ejercicio-1.Rmd")
```

Una vez descargado, abranlo desde el panel *Files*.

Los prácticos en general se harán en un archivo similar a este. En la parte superior encontran el encabezado entre guiones. Ahí deberán poner sus nombres y el nombre del grupo.

Ejercicio 2.2 (Personalizando)**.** Cambien en encabezado y pongan sus nombres, el nombre del grupo y la fecha de hoy.

Abajo encontraran espacio para ir contestando la preguntas. Una consideración que deben tomar en cuenta es que todo el texto que escriban va a ser considerado como un *único* párrafo a menos que este separado por **una línea en blanco**.

Por ejemplo, prueben escribir esto en el documento (pueden copiarlo):

> Mucho antes de que el lector haya llegado a esta parte de mi obra se le habrán ocurrido una multitud de dificultades. Algunas son tan graves, que aun hoy dia apenas puedo reflexionar sobre ellas sin vacilar algo; pero, según mi leal saber y entender, la mayor parte son solo aparentes, y las que son reales no son, creo yo, funestas para mi teoria.
>
> Estas dificultades y objeciones pueden clasificarse en los siguientes grupos:
>
> 1° Si las especies han descendido de otras especies por suaves gradaciones, ¿por qué no encontramos en todas partes innumerables formas de transicion? ¿Por qué no está toda la naturaleza confusa, en lugar de estar las especies bien definidas según las vemos?
>
> 2° ¿Es posible que un animal que tiene, por ejemplo, la conformación y costumbres de un murciélago pueda haber sido formado por modificación de otro animal de costumbres y estructura muy diferentes? ¿Podemos creer que la selecci6n natural pueda producir, de una parte, un órgano insignificante, tal como la cola de la jirafa, que sirve de mosqueador, y, de otra, un órgano tan maravilloso como el ojo?
>
> 3° ¿Pueden los instintos adquirirse y modificarse por selección natural? ¿Qué diremos del instinto que lleva a la abeja a hacer celdas y que prácticamente se ha anticipado a los descubrimientos de profundos matemáticos?
>
> 4° ¿Cómo podemos explicar que cuando se cruzan las especies son esteriles o producen descendencia esteril, mientras que cuando se cruzan las variedades su fecundidad es sin igual?

Luego, hagan clic en el botón *Knit* que tienen arriba, al lado de un ovillo y una aguja de tejer. El desafío es mantener el formato con cada párrafo separado.

La misma consideración se debe tener en cuenta para otros formatos como títulos o listas. Un consejo para ver como va quedando el documento, es tejerlo seguido. Así podremos ver cualquier problema pronto.

2.3 Integrando código

En los documentos de RMarkdown se puede integrar bloques de código (de *R* y otros lenguajes). Para insertar un bloque pueden hacer clic en el botón de "Insert/R" que hay arriba o por el atajo del teclado "Ctrl+Alt+I".

````
```{r}
Acá va el código

```
````

Es importante mantener las comillas invertidas tal cual están ya que con ellas se define donde empieza y termina el bloque. Entre las llaves se incluye como se va a ejecutar el código (con *R* en este caso). También se puede poner nombre al bloque, cosa que es muy recomendable porque sino van a estar nombrados como chunk-#, donde # son números consecutivos. Ahora, imaginen que el chunk-34 de 60 falla. Va a ser un poco tedioso buscarlo, con nombre será más sencillo saber donde estar el fallo. Además se pueden poner otras opciones, como ocultar el código, cambiar el tamaño de figuras, etc. Luego de las llaves, en *la línea siguiente*, deben introducir el código que quieran ejecutar, siempre teniendo en cuenta de dejar la línea con las comillas invertidas tal cual está. También es buena idea dejar una línea en blanco luego.

Ejercicio 2.3. Incluyan un bloque de código y pongan un nombre descriptivo. Luego, escriban una operación matemática simple. Finalmente, tejan el documento.

Chapter 3

Visualización de Datos

Para hacer el tutorial ingresen este código en la consola:
`download.file("git.io/visualizacion.Rmd", file = "visualizacion.Rmd")`
A continuación abran el archivo `visualizacion.Rmd` y hagan clic en el botón `Run document` arriba.

3.1 Introducción

Una de las formas más útiles de visualizar la información es mediante gráficos (aunque si son pocos datos es preferible una tabla). De hecho, el primer paso antes de analizar los datos debe ser hacer un gráfico de los valores que tienen. Un gráfico de dispersión si es bidimensional o un histograma si solo tiene una dimensión.

Aunque hay varios sistemas gráficos (`base`, `lattice`, `ggobi`, `plotly`) vamos a usar `ggplot2` por su facilidad de uso y potencia para hacer gráficos complejos a partir de componentes simples.

Este paquete sigue una idea que se llama gramática de gráficos propuesta por Wilkinson en donde los gráficos pueden dividirse en cuatro partes:

- Los **datos** y como se **mapea** (`aes`) esos datos a las diferentes atributos estéticos. Es decir que columna corresponde al eje x, al eje y, forma, color, etc.
- Las formas geométricas (`geom`) que representan como se ven los datos. Como puntos, líneas, barras de error, etc.
- Transformaciones estadísticas de los datos (`stats`) resumen los datos de forma útil. Por ejemplo, para agregar la media por grupo o una línea de regresión sin haberlos calculado antes.
- Escalas a las que se mapean los datos (`scale`). Estas pueden ser escalas de color, forma, etc.
- Un sistema de coordenadas (`coord`), que describe como se proyectan estos datos. Por defecto se usa el sistema cartesiano. Pero hay otros disponibles como el polar.
- Un sistema paneles (`facet`) que describe como dividir los datos en distintos paneles.

Adicionalmente a la gramática, se agrega un sistema de temas que permite modificar la totalidad de elementos que hacen al gráfico como fuentes, líneas de ejes, etc.

3.2 El conjunto de datos `mpg`

En los datos que vienen con el paquete `ggplot2` está `mpg`. Contiene los datos de rendimiento, cilindrada y otros más de algunos modelos de autos. Los datos están una estructura rectangular llamada *data frame*, cada

columna es una variable y cada fila una observación recolectadas por la Agencia de Protección ambiental de EE.UU.

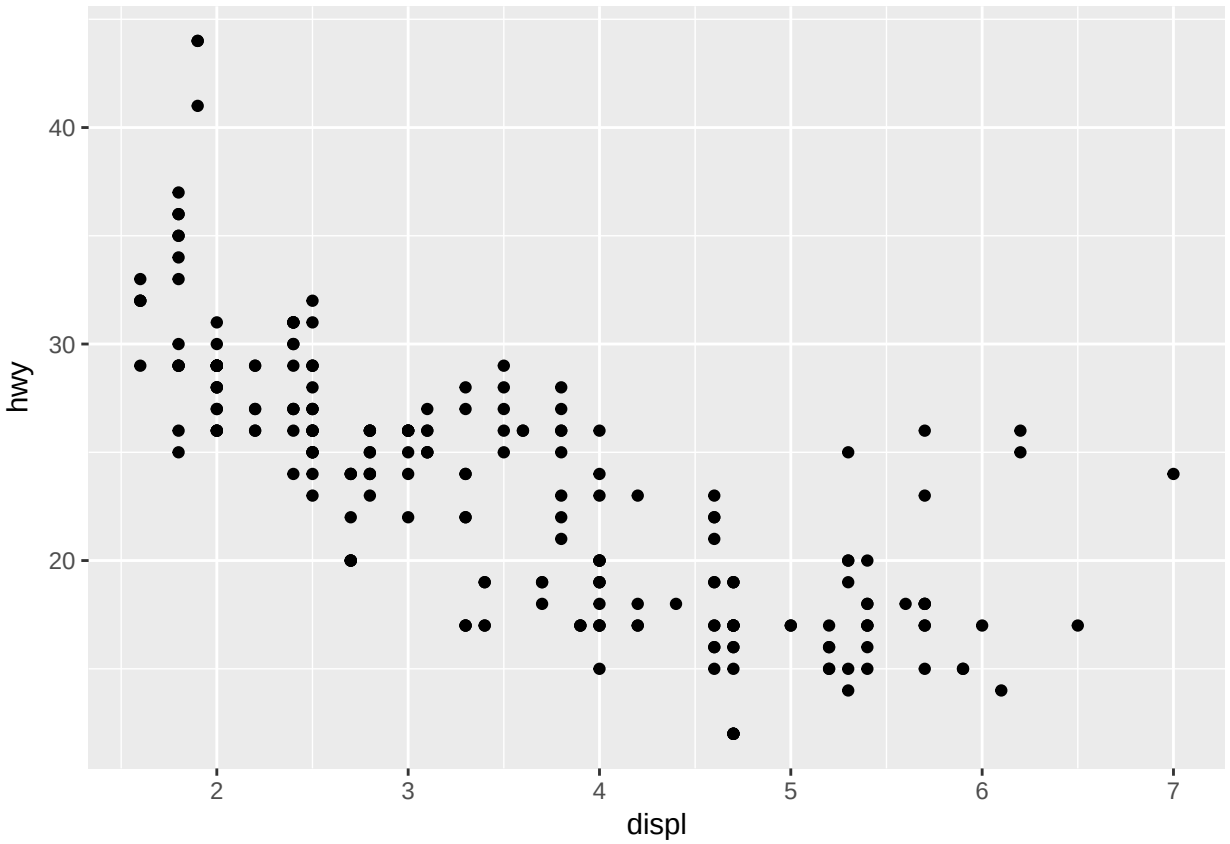
```
library(tidyverse)
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr>
##  1 audi         a4        1.80  1999     4 auto(l~ f        18    29 p
##  2 audi         a4        1.80  1999     4 manual~ f        21    29 p
##  3 audi         a4        2.00  2008     4 manual~ f        20    31 p
##  4 audi         a4        2.00  2008     4 auto(a~ f        21    30 p
##  5 audi         a4        2.80  1999     6 auto(l~ f        16    26 p
##  6 audi         a4        2.80  1999     6 manual~ f        18    26 p
##  7 audi         a4        3.10  2008     6 auto(a~ f        18    27 p
##  8 audi         a4 quat~  1.80  1999     4 manual~ 4        18    26 p
##  9 audi         a4 quat~  1.80  1999     4 auto(l~ 4        16    25 p
## 10 audi         a4 quat~  2.00  2008     4 manual~ 4        20    28 p
## # ... with 224 more rows, and 1 more variable: class <chr>
```

3.3 Gráficos con ggplot

Podemos hacer un gráfico de la siguiente forma:

```
library(ggplot2)
ggplot(data = mpg) +
 geom_point(mapping = aes(x = displ, y = hwy))
```



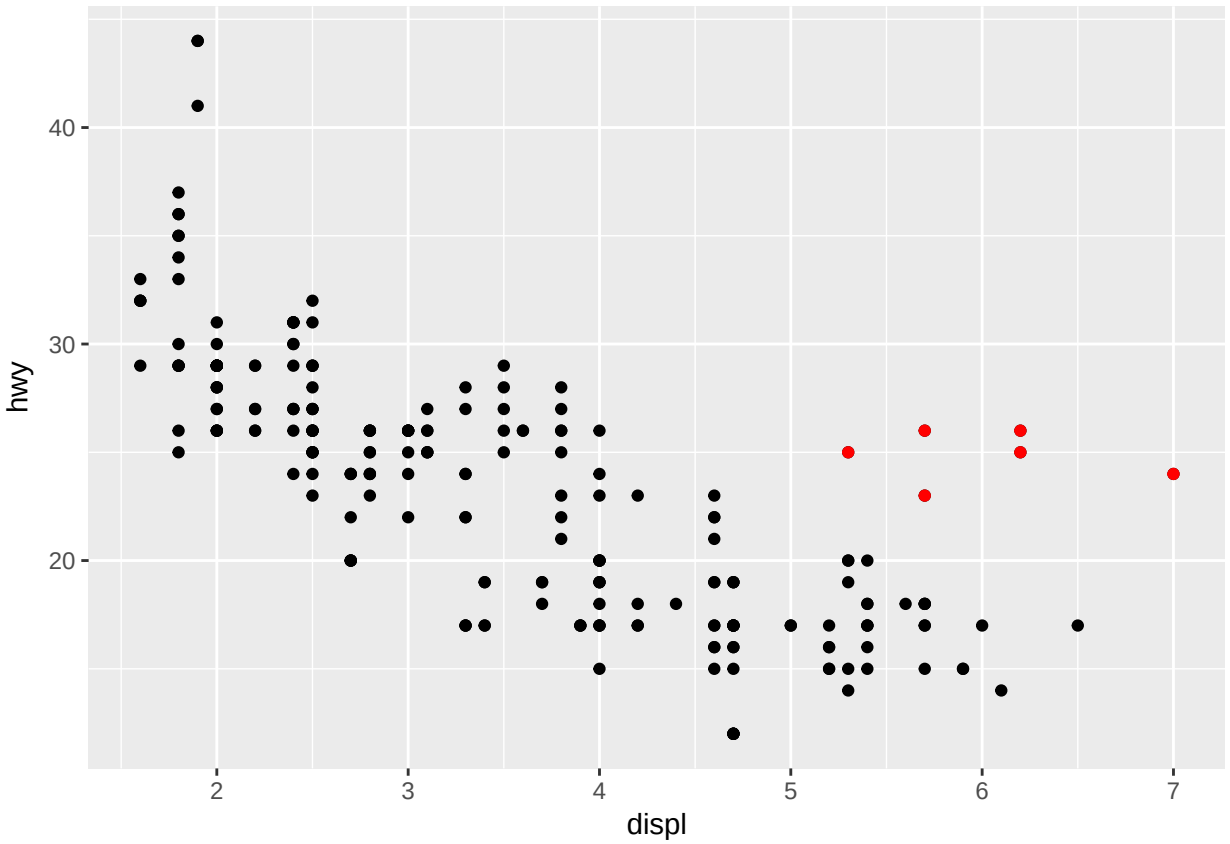
Prueben escribiendo el código en la consola.

En el gráfico vemos que hay una tendencia a disminuir el rendimiento a medida que aumenta la cilindrada.

En la llamada de Hay distintas partes. La llamada a `ggplot` donde especificamos el nombre de los datos que vamos a usar. Es la que inicializa el gráfico. Pero aquí no especifica nada de como graficarlo. Sin embargo, es necesario empezar siempre por esta función y luego ir agregando capas. Luego agregamos una capa de puntos. Ambos están unidos por un `+`. Cada vez que deseemos agregar una capa, lo haremos con ese símbolo `+`. Por otro lado, especificamos el mapeo de las columnas de los datos a las ordenadas y abscisas dentro del argumento `mapping`. Hemos graficado el tamaño del motor (en litros), `displ` en las ordenadas y el rendimiento en millas por galón en las abscisas. Siempre que quereamos mapear una columna a alguna parte del gráfico lo hemos de hacer dentro la función `aes()`, de *aestetics* que significa ésteticas en inglés.

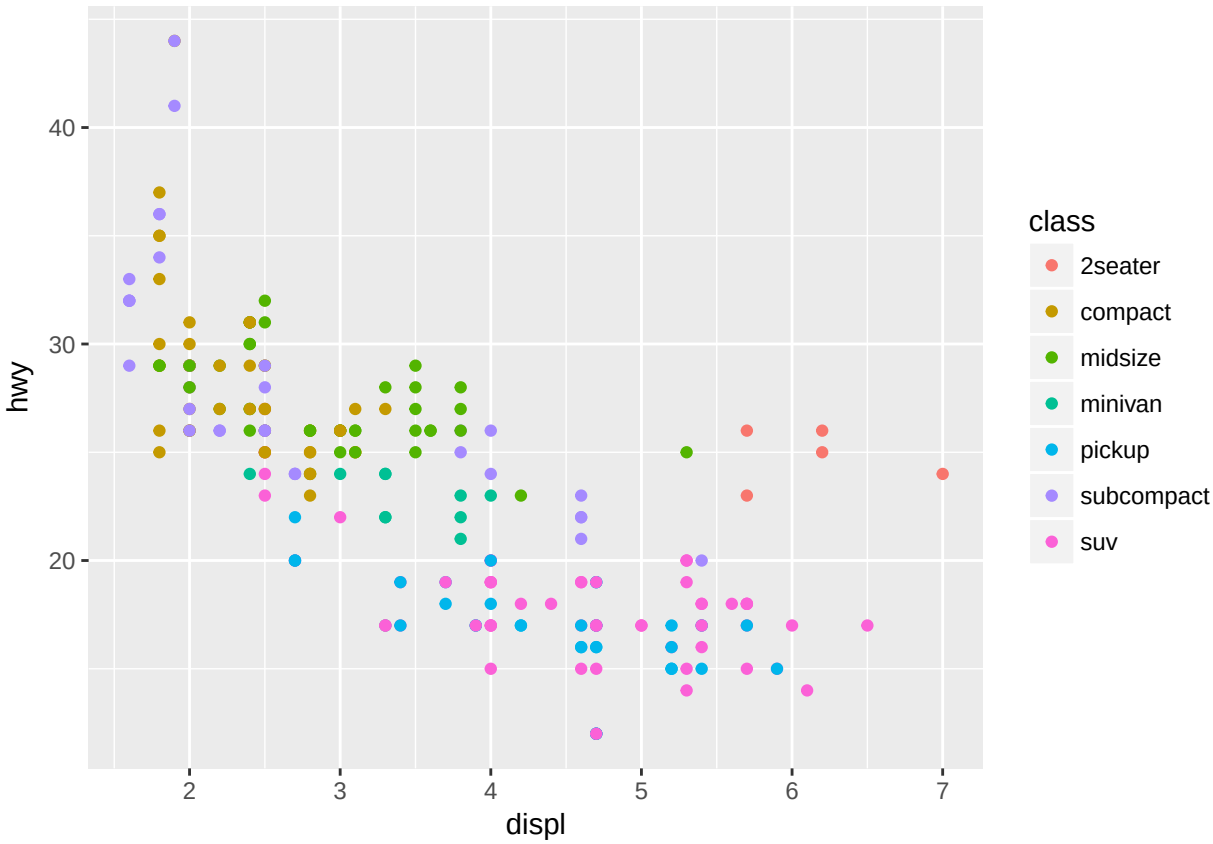
3.4 Mapeando

En el gráfico anterior vemos que hay unos puntos que no siguen la tendencia general. Aquí están resaltados con rojo.



Para saber más el porque de estos puntos podríamos agregar más información al gráfico como por ejemplo el tipo de auto. Una opción es agregar colores.

```
ggplot(data = mpg) +
 geom_point(mapping = aes(x = displ, y = hwy, color = class))
```



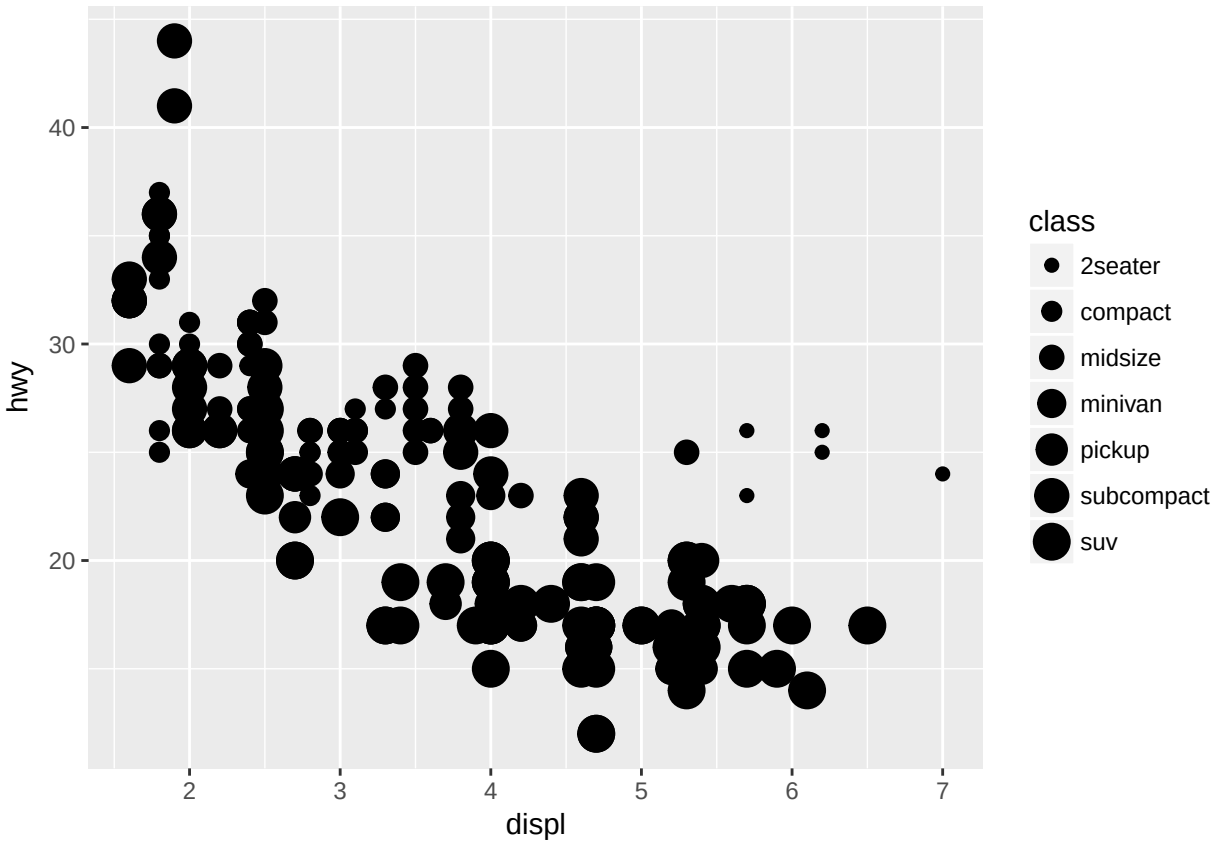
Ahora podemos ver, que en general los autos con cilindrada grande son camionetas (*pickup*) o suv. Y que los que tienen cilindrada grande pero rendimiento mayor son autos deportivos.

Agregamos el color mapeando `class` a la éstetica de color. `ggplot` le asigna automáticamente un color a cada nivel de `class`. Y también genera la leyenda apropiada.

También podemos mapear el tamaño del punto a `class`. En este caso recibiremos un `warning` porque no tiene mucho sentido mapear el tamaño con una variable discreta desordenada. Es decir, que no hay una correspondencia entre el tamaño del punto y la clase.

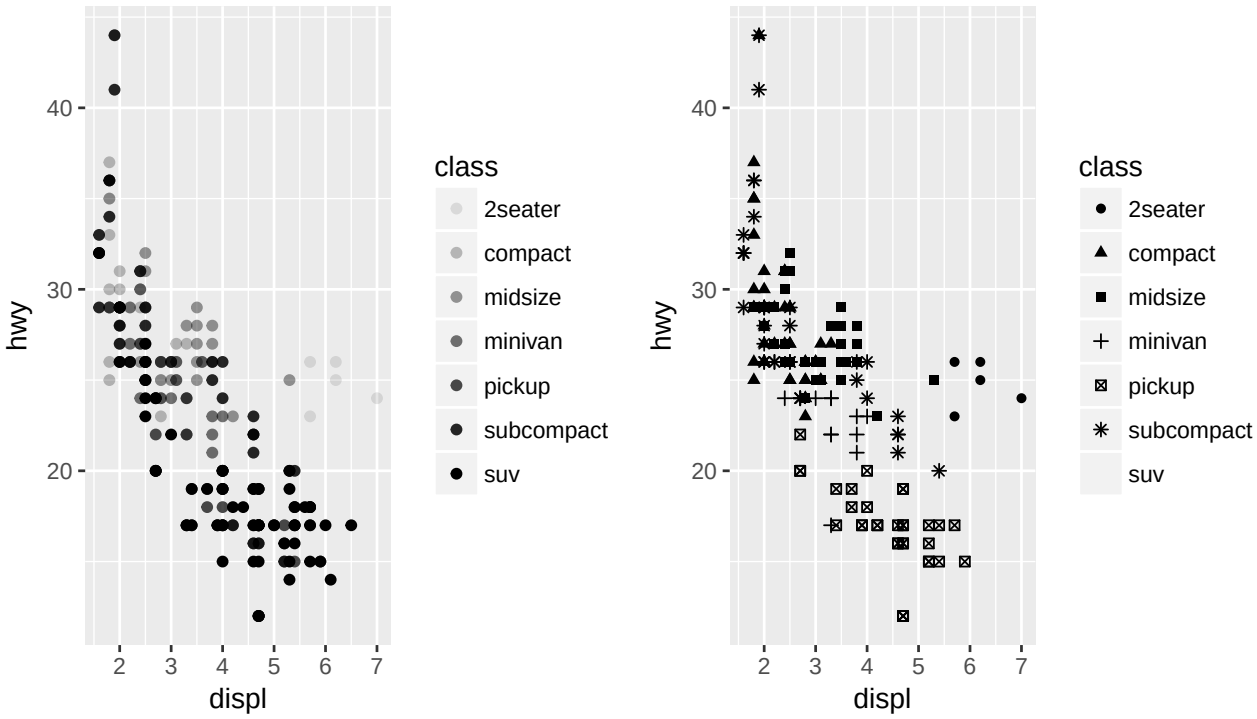
```
ggplot(data = mpg) +
 geom_point(mapping = aes(x = displ, y = hwy, size = class))
```

```
## Warning: Using size for a discrete variable is not advised.
```



También podriamos mapear la clase a la transparencia de los puntos (`alpha`) o a la forma (`shape`)

```
# Izquierda
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, alpha = class))

# Derecha
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, shape = class))
```



Si reproducen el código en sus computadoras verán que que ambos dan advertencias. Así como no tienen mucho sentido mapear el tamaño a algo sin orden intríseco, tampoco lo tiene mapear la transparencia. Por otro lado, noten que en el gráfico de la derecha ¡faltan los puntos de *suv*! Esto es porque ggplot solo asigna automaticamente hasta 6 símbolos diferentes para los puntos. Si queremos más hay que hacerlo de forma manual.

En general, uno mapea una variable a alguna característica del gráfico asociandola dentro de `aes()`. `ggplot` se encarga de los detalles de pasar esa asociación a las distintas capas, de generar los niveles apropiados y de hacer la leyenda. De hecho, podemos ver que x e y también son características del gráfico pero en vez de mostrar una leyenda genera las marcas en los ejes.

También es posible configurar alguna estética a un valor específico. Como por ejemplo hacer que todos los puntos sean rojos

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "red")
```

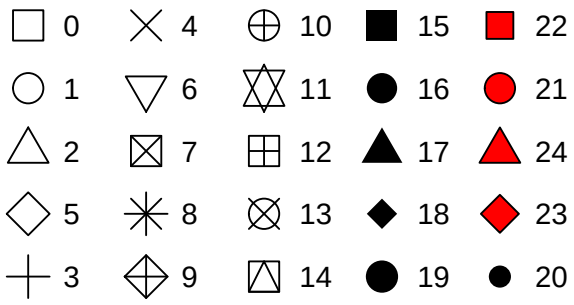
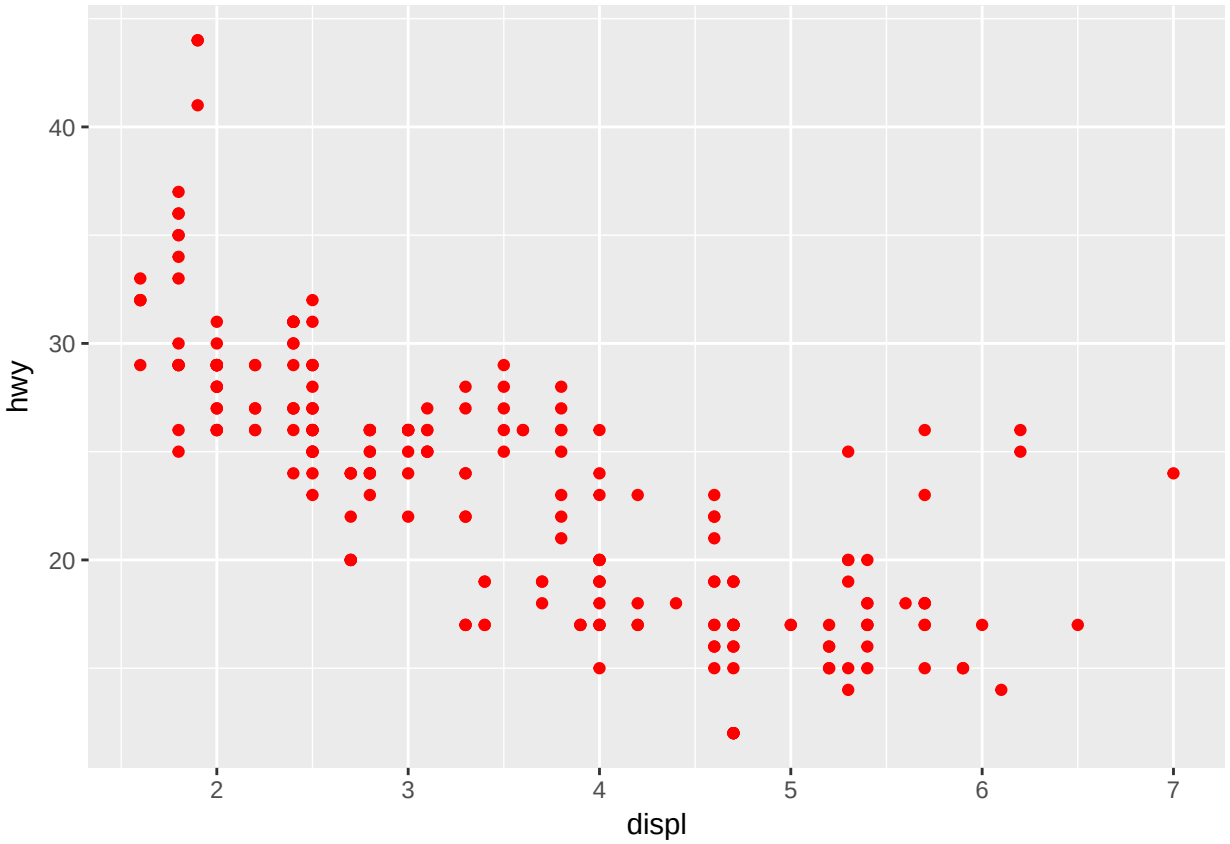


Figure 3.1: Valores númericos y la forma asociada a cada uno. En R hay 25 formas diferentes. Algunas parecen repetirse pero no es así. Por ejemplo, las formas 0, 15 y 22 son todos cuadrados. Pero las formas del 0-15 tienen el color definido por el borde, usan 'color' para cambiar el color. Del 15 a 18 son formas rellenas que usan 'fill' para cambiar el color del relleno. Y de la forma 21 a 23 son formas con relleno y borde que usan ambas 'fill' y 'color'.



Acá el color no muestra ninguna información extra. También es posible cambiar el:

- el color por el nombre que tenga sentido o en hexadecimal.
- el tamaño de los puntos en mm.
- la forma del punto según los valores que se muestran acá abajo

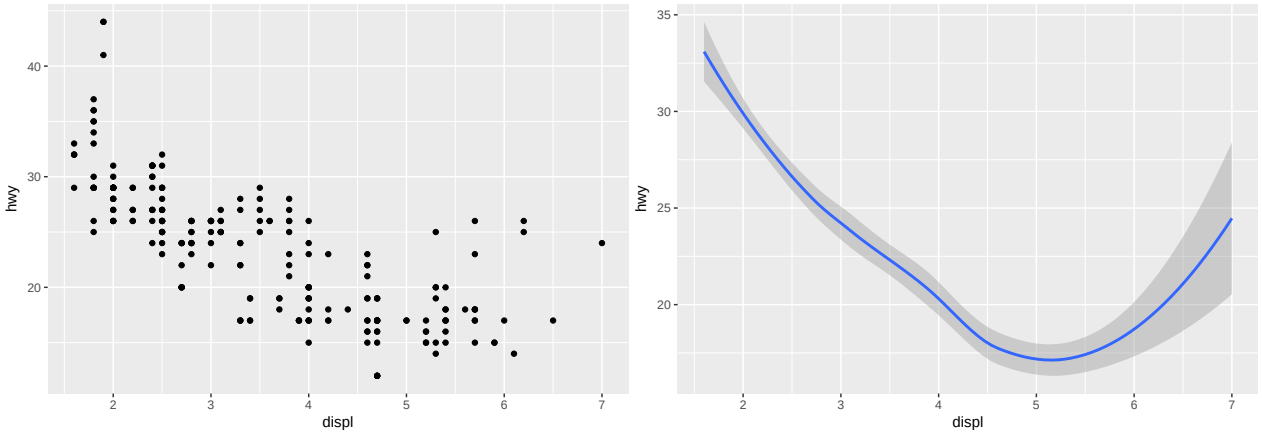
Valores númericos y la forma asociada a cada uno. En R hay 25 formas diferentes. Algunas parecen repetirse pero no es así. Por ejemplo, las formas 0, 15 y 22 son todos cuadrados. Pero las formas del 0-15 tienen el color definido por el borde, usan `color` para cambiar el color. Del 15 a 18 son formas rellenas que usan `fill` para cambiar el color del relleno. Y de la forma 21 a 23 son formas con relleno y borde que usan ambas `fill` y `color`.

Table 3.1: Gráficos comunes con ggplot

Gráfico	geom
Barras	geom_col geom_bar
Puntos	geom_point
Cajas y Barras	geom_boxplot
Histograma	geom_histogram
Lineas	geom_line
Barras de Error	geom_errorbar

3.5 Formas geometricas

¿En que se parecen los gráficos de abajo?



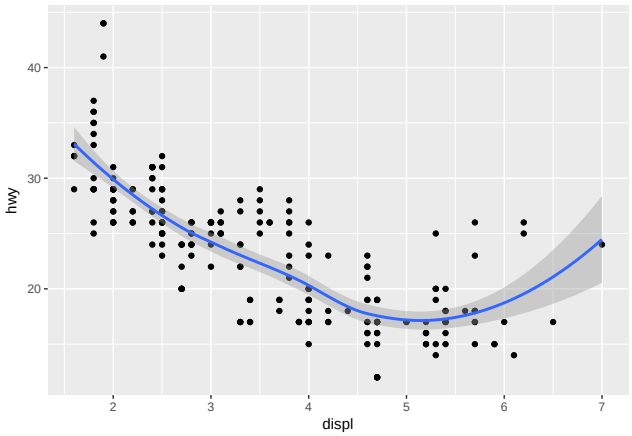
Ambos tienen las mismas variables, pero están representados por distintas formas. En el idioma de ggplot cada forma es un **geom**. Y cada geom es una forma geométrica de representar los datos. Hay muchos **geom** (más de 30 en el paquete y muchos más en extensiones) y todos empiezan con `geom_`), por ejemplo:

Pueden ver más en la ayuda de ggplot en R (usando la pestaña de ayuda o usando `help(nombre_de_función)` en la consola) o en la documentación online que tiene la ventaja de tener graficados los ejemplos http: //ggplot2.tidyverse.org/reference/ .

Para hacer los gráficos de arriba usamos:

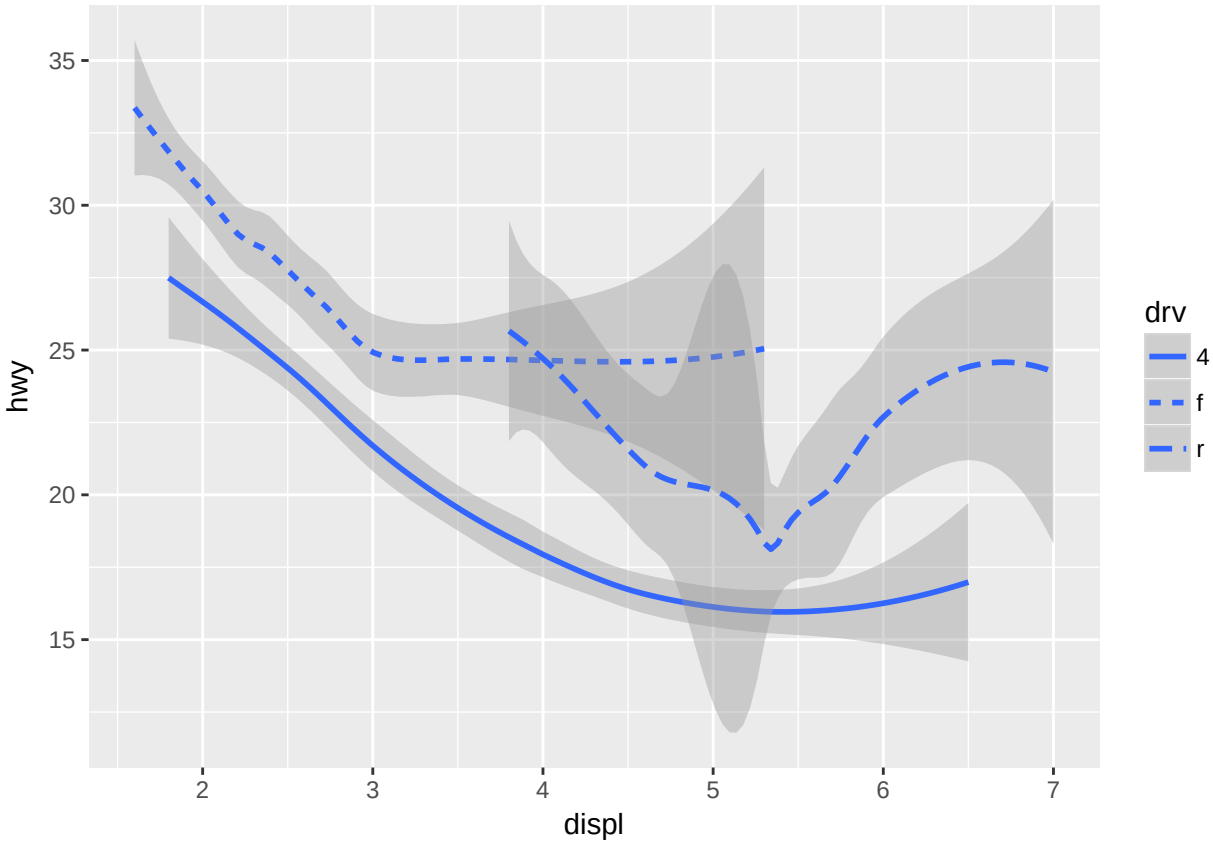
```
# Izquierda
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))

# Derecha
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

Todos los *geoms* van luego de `ggplot` y se unen con un `+`. En ggplot cada forma geométrica es una capa y pueden combinarse varias en un mismo gráfico.



Además, todos los `geoms` tienen un argumento `mapping` para la estética. Claro que no todos aceptan los mismas argumentos. No tienen sentido ponerle relleno a una línea o cambiar el tipo de línea a un punto. Pero si se puede cambiar el tipo de línea de 'geom_smooth:

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv))
```



Acá `geom_smooth` separa tres líneas según el valor de `drv`, que es la tracción. Una línea lisa para las que son 4x4 (4), rayas cortas para tracción delantera (f) y rayas largas para tracción trasera (r).

Podemos ver más claramente porque tiene esta forma geom_smooth graficando los puntos de cada grupo:

Muchos `geoms` que resumen la información de varios datos con una sola forma tienen un argumento de éstetica llamado `group` que agrupa la observaciones que son iguales en una variable.

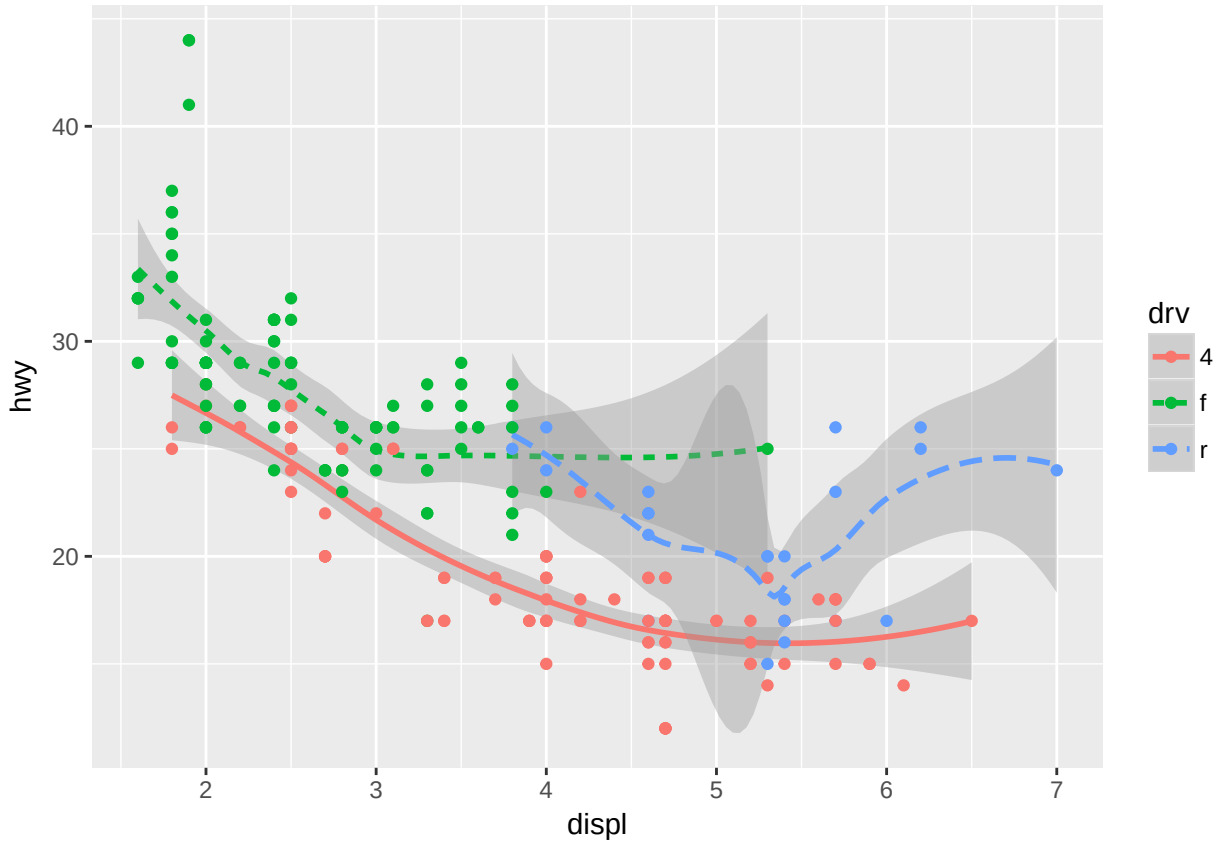
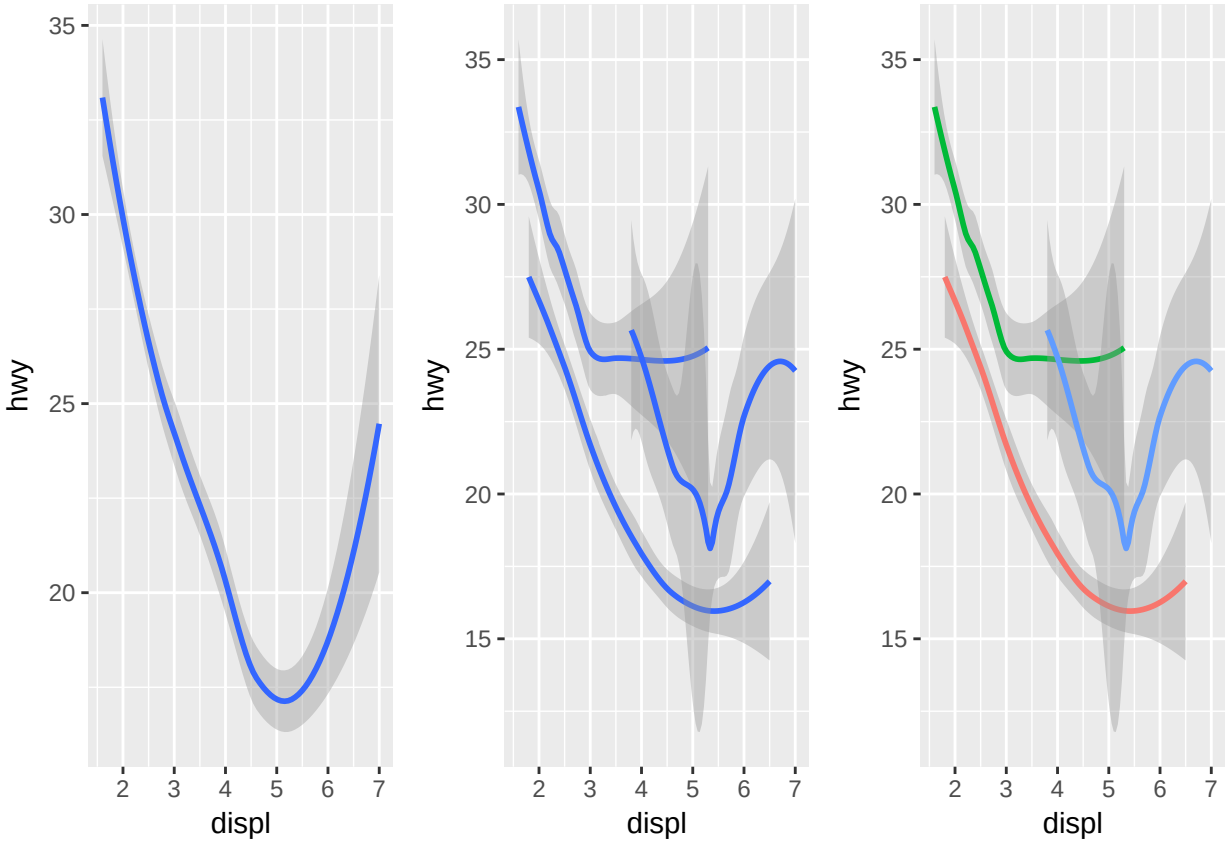


Figure 3.2: Varios geoms pueden usarse en un mismo gráfico.

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy))

ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy, group = drv))

ggplot(data = mpg) +
  geom_smooth(
    mapping = aes(x = displ, y = hwy, color = drv),
    show.legend = FALSE
  )
```



Arriba vimos que podiamos usar dos geoms en un mismo gráfico (y podríamos agregar más si quisieramos). Pero al hacerlo hemos duplicado el `mapping` en los dos `geoms`:

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv, color = drv)) +
  geom_point(mapping = aes(x = displ, y = hwy, color = drv))
```

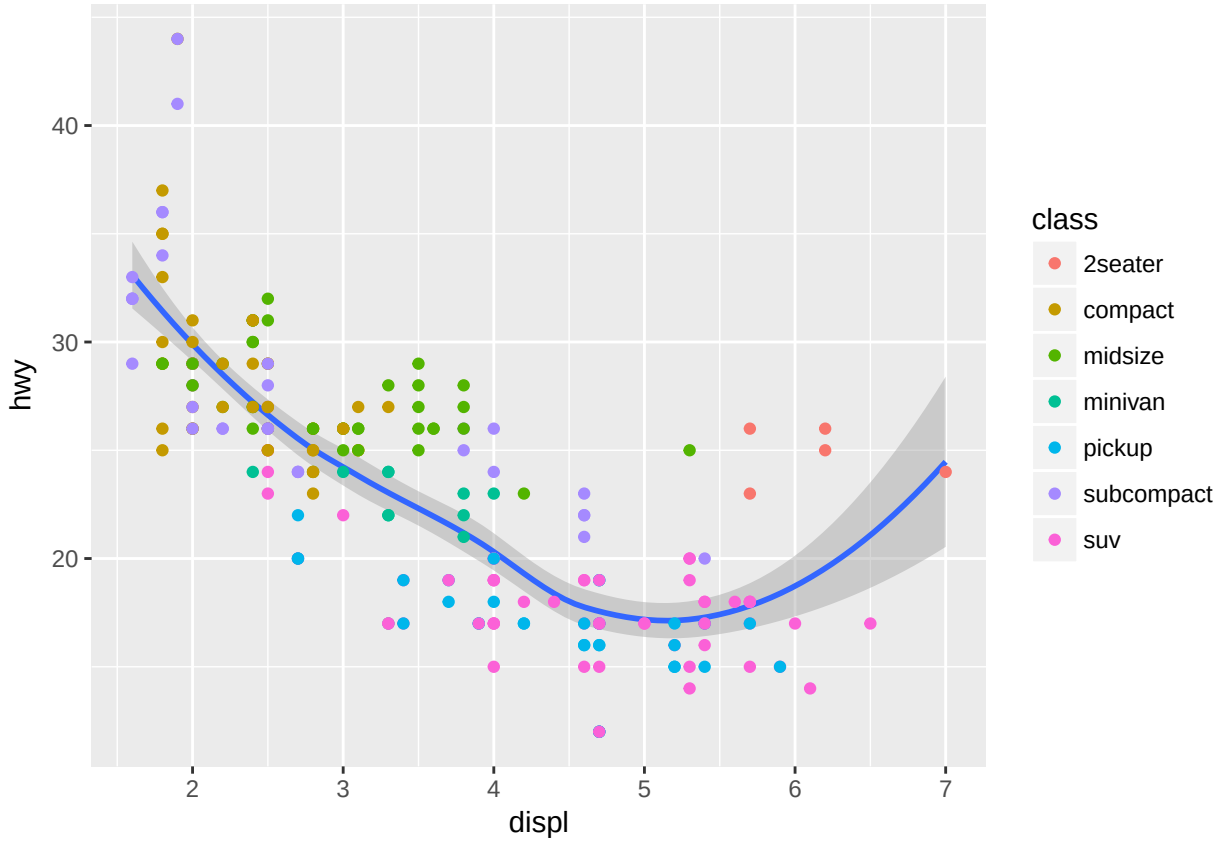
Podemos evitarlo si ponemos el `mapping` dentro de la llamada de ggplot:

```
ggplot(data = mpg, mapping=  aes(x = displ, y = hwy, linetype = drv, color = drv)) +
  geom_smooth() +
  geom_point()
```

Esto funciona porque las capas heredan el `mapping` de `ggplot`. Por eso, va a funcionar en todas las capas que pongamos, de forma *global*. A veces, esto introduce ciertos errores cuando usamos varias fuentes de datos y las variables no están en presentes en todos los conjuntos. Es posible cambiar el `mapping` de una

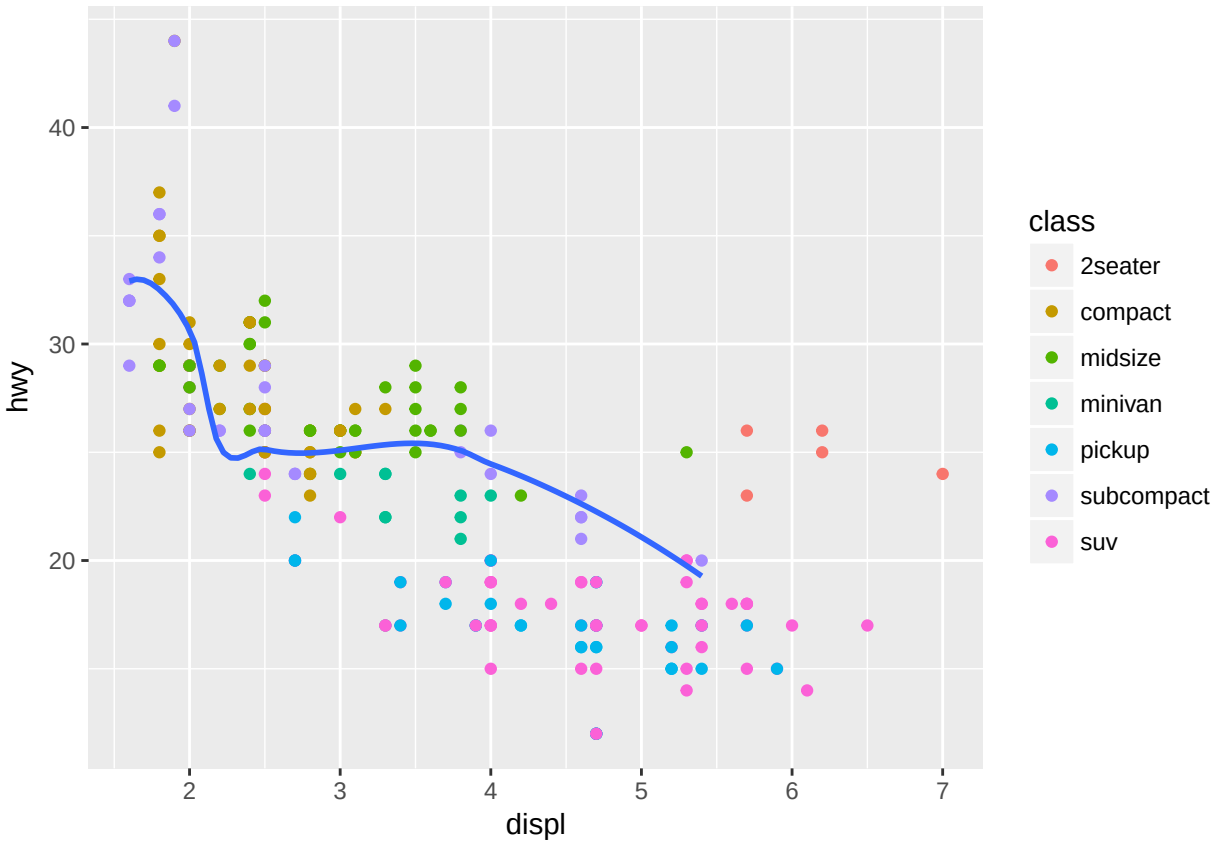
capa específica y va a ser utilizada *solo* en esa capa; es decir, de forma de forma *local*.

```
ggplot(data = mpg, mapping=  aes(x = displ, y = hwy)) +
  geom_smooth() +
  geom_point(mapping = aes(color = class))
```



O también podemos definir otro conjunto de datos para el `geom`:

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping = aes(color = class)) +
  geom_smooth(data = filter(mpg, class == "subcompact"), se = FALSE)
```

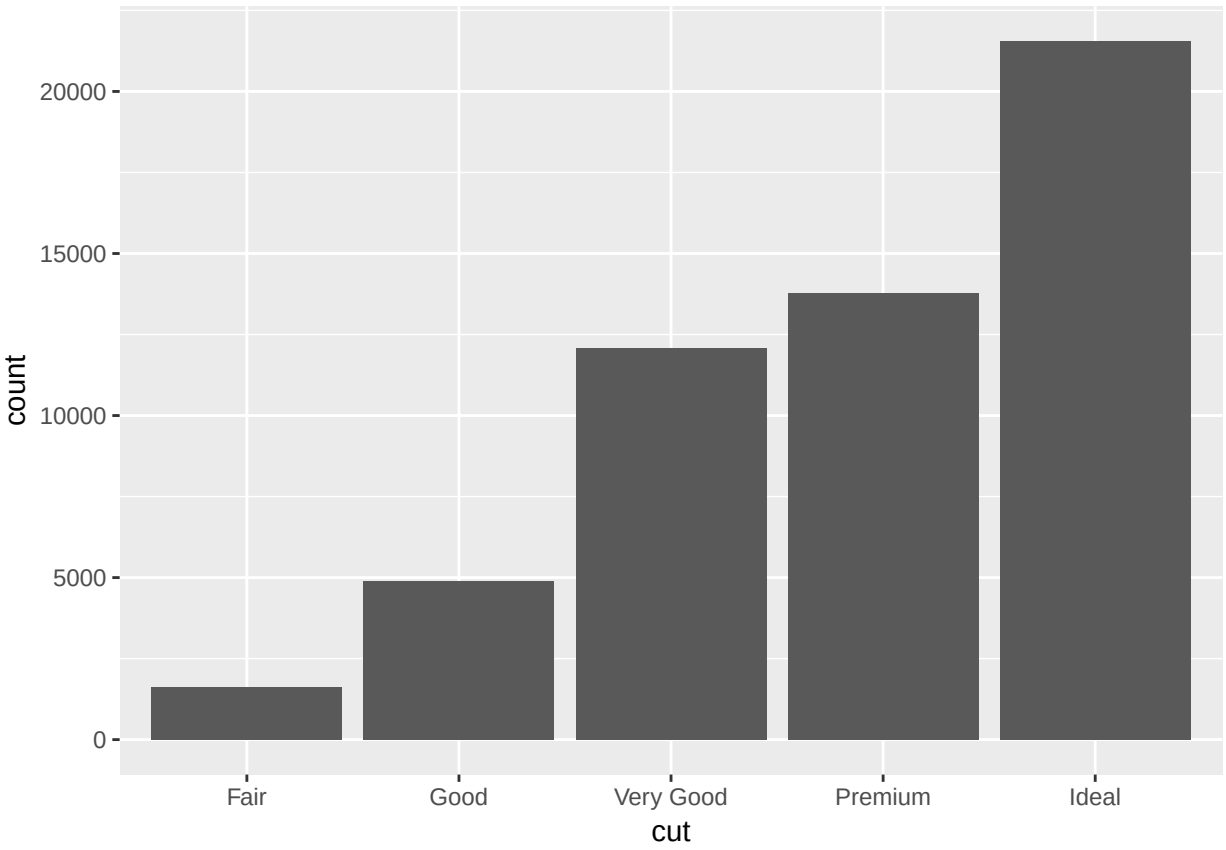


Todavía no vieron `filter`, pero ya lo verán más adelante.

3.6 Transformaciones Estadísticas

Pensemos en los gráficos de barras. En `ggplot2` se hacen con `geom_bar`. A primera vista los gráficos de barras parecen simples. En este caso estamos graficando datos de diamantes, del conjunto de datos `diamonds` que contiene cerca 54000 datos. En el gráfico vemos que hay muchos más diamantes de cortes buenos que regulares.

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))
```



En el eje x está puesto el corte, y en el eje y está puesto la cuenta (frecuencia) de cada uno. Pero si vemos el conjunto de datos veremos que esta última ¡no está! Entonces ¿De dónde salió? Algunos *geoms* grafican los datos puros pero otros aplican transformaciones estadísticas a los datos y crean nuevas variables:

- Los gráficos de barras, histograms y polígonos de frecuencia cuentan y juntan los datos.
- `geom_smooth` usa modelos para mostrar las tendencias de los datos.
- `geom_boxplot` crea un sumario de estadísticos robustos para mostrar los datos.

El algoritmo usado para calcular las nuevas variables se llama `stat`. Abajo vemos como funciona `stat_count`

Empieza con los datos

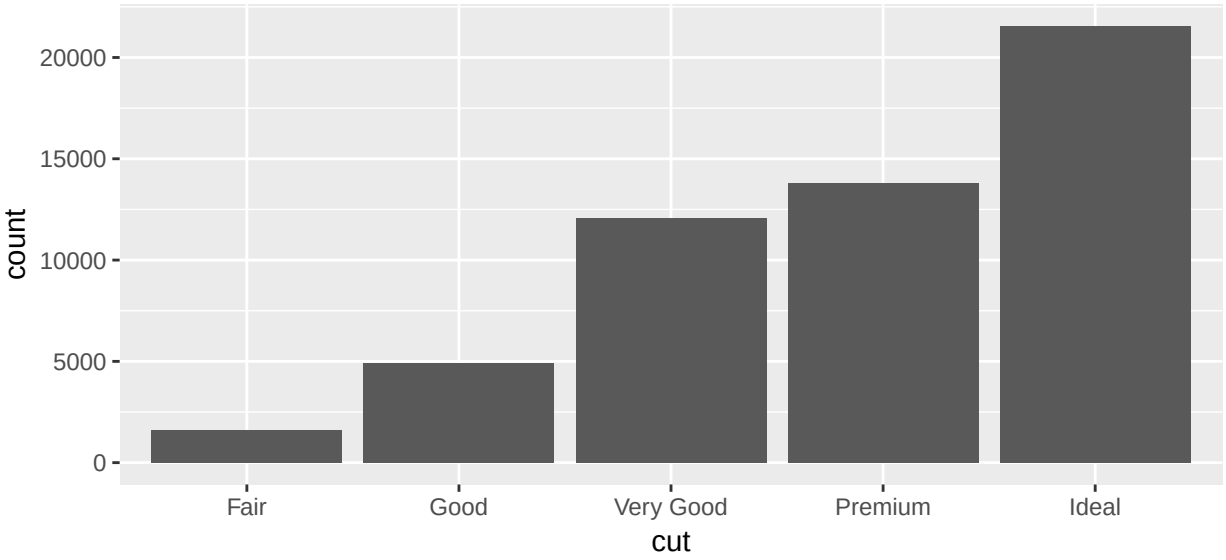
```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1 0.230 Ideal     E     SI2      61.5   55.   326  3.95  3.98  2.43
## 2 0.210 Premium   E     SI1      59.8   61.   326  3.89  3.84  2.31
## 3 0.230 Good      E     VS1      56.9   65.   327  4.05  4.07  2.31
## 4 0.290 Premium   I     VS2      62.4   58.   334  4.20  4.23  2.63
## 5 0.310 Good      J     SI2      63.3   58.   335  4.34  4.35  2.75
## 6 0.240 Very Good J     VVS2     62.8   57.   336  3.94  3.96  2.48
```

`geom_bar` calcula las nuevas variables usando el `stat count` que devuelve un nuevo data.frame

```
## # A tibble: 5 x 3
##   cut       count  prop
##   <ord>     <int> <dbl>
## 1 Fair       1610    1.
## 2 Good       4906    1.
## 3 Very Good 12082    1.
```

```
## 4 Premium   13791    1.
## 5 Ideal     21551    1.
```

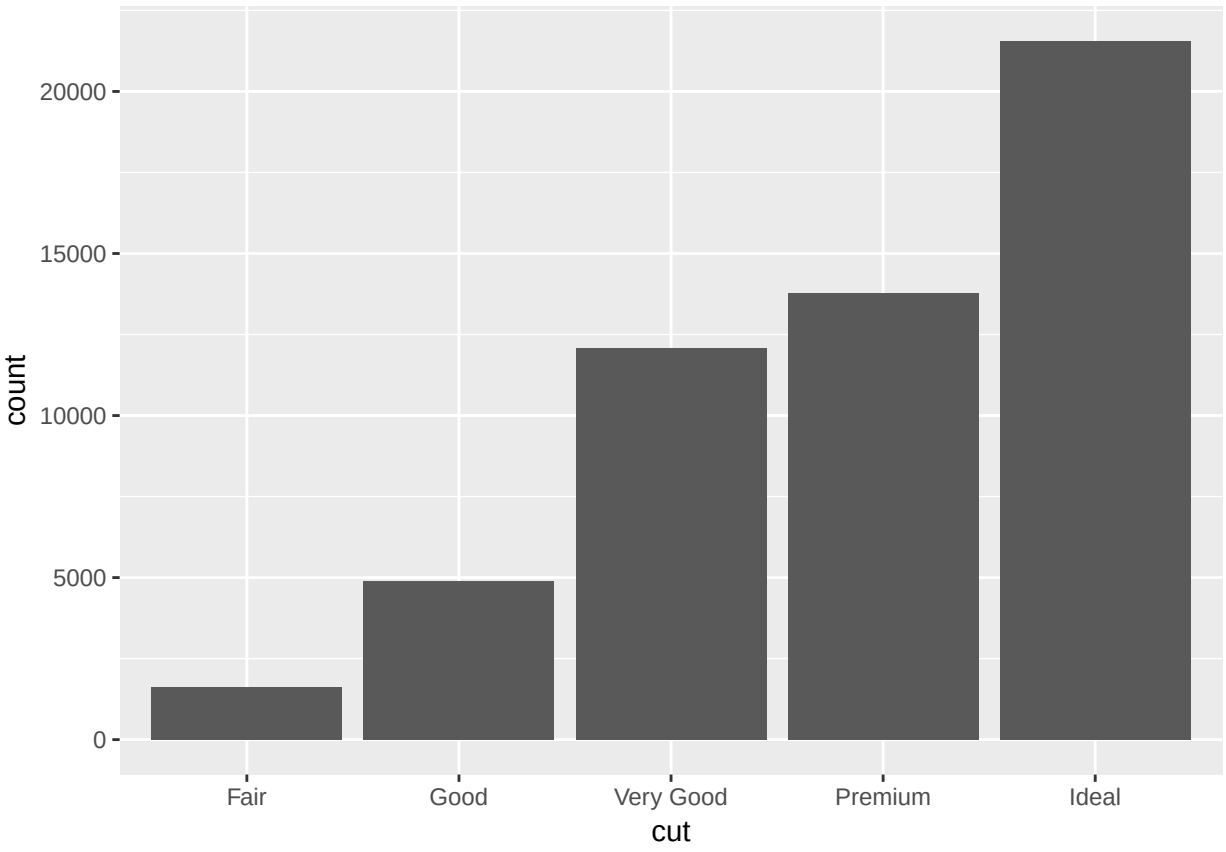
`geom_bar` luego usa esos datos para graficar:



Podés saber que `stat` usa cada `geom` usando la ayuda. Por ejemplo, `?geom_bar` usa por defecto `stat_count` y `stat_count` usa por defecto `geom_bar` para mostrar sus resultados y ambos están descriptos en la misma página de ayuda. Podemos ver que es calculado en la sección *Computed Variables*.

Es posible intercambiar `geom` por su `stat`. Por ejemplo:

```
ggplot(data = diamonds) +
  stat_count(mapping = aes(x = cut))
```

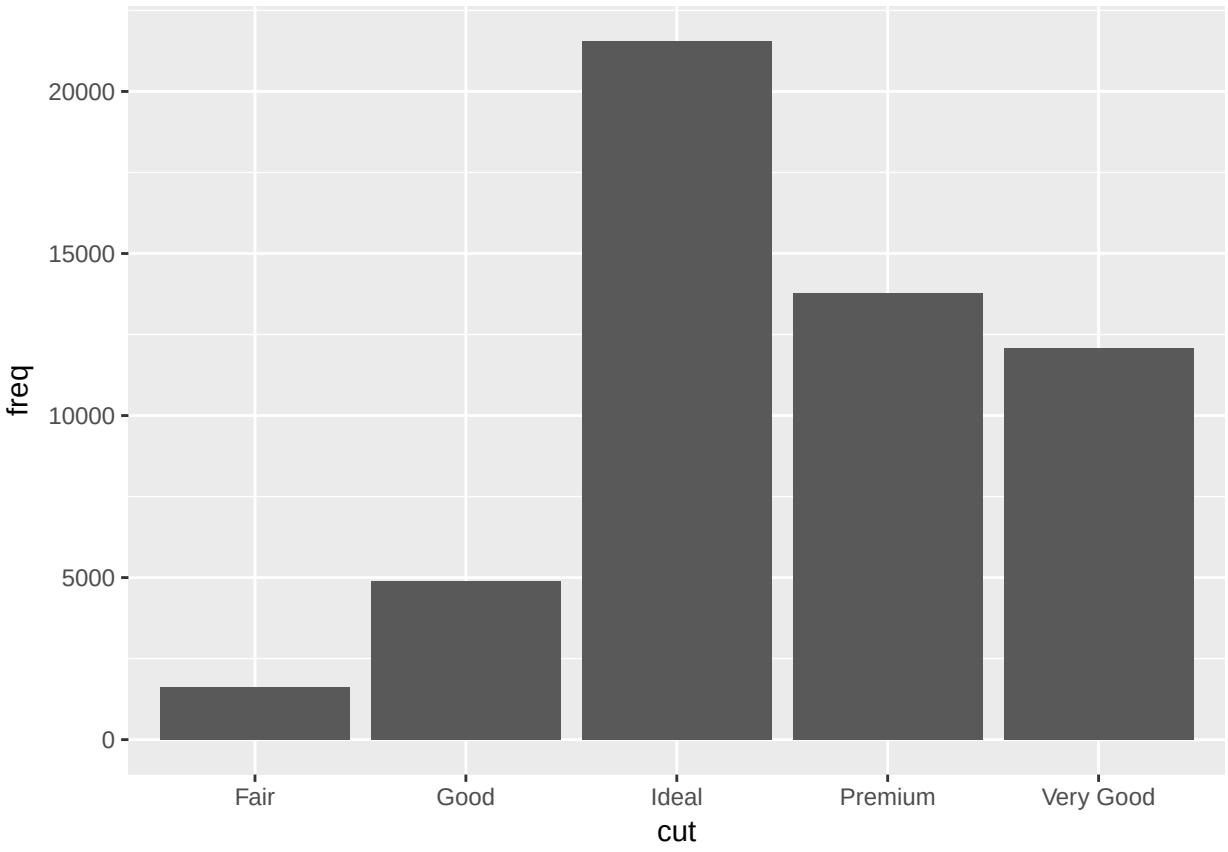


Esto funciona porque cada *stat* tiene un *geom* por defecto y cada *geom* tiene un *stat*. Lo que significa que podes usar cada *geom* sin preocuparte por las transformaciones subyacentes.

Hay veces que querrás cambiar los valores por defecto:

1. Cuando tengas las variables precomputadas y desees graficarlas. En el código de abajo cambio el *stat* de `geom_bar` por `stat_identity` (identidad). Esto me permite graficar la altura de la variable *y* a algún valor del conjunto de datos.

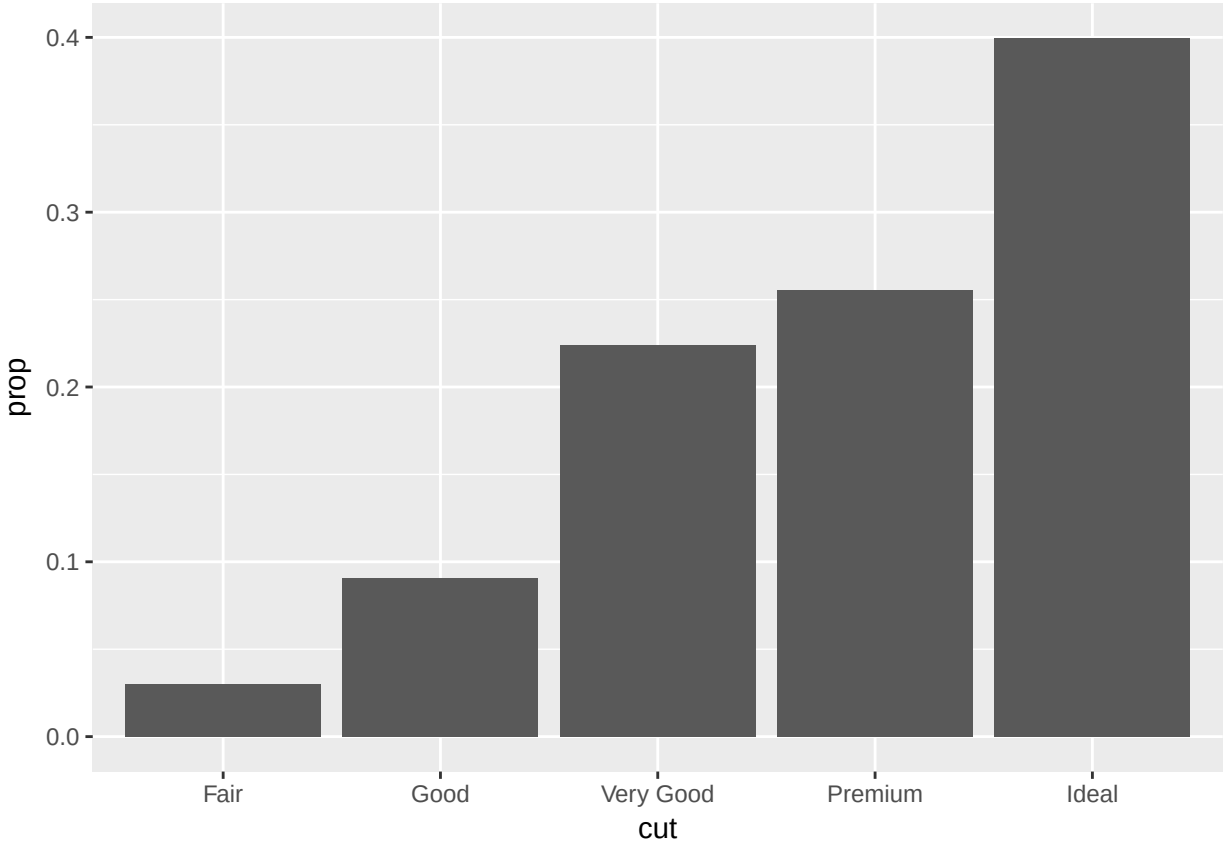
```
demo <- tribble(
  ~cut,          ~freq,
  "Fair",        1610,
  "Good",        4906,
  "Very Good",   12082,
  "Premium",     13791,
  "Ideal",       21551
)

ggplot(data = demo) +
  geom_bar(mapping = aes(x = cut, y = freq), stat = "identity")
```



No te preocupes si no entiendes que hace `tribble` o `<-`. Todavía no lo hemos visto pero quizá puedas entender que hacen por su contexto

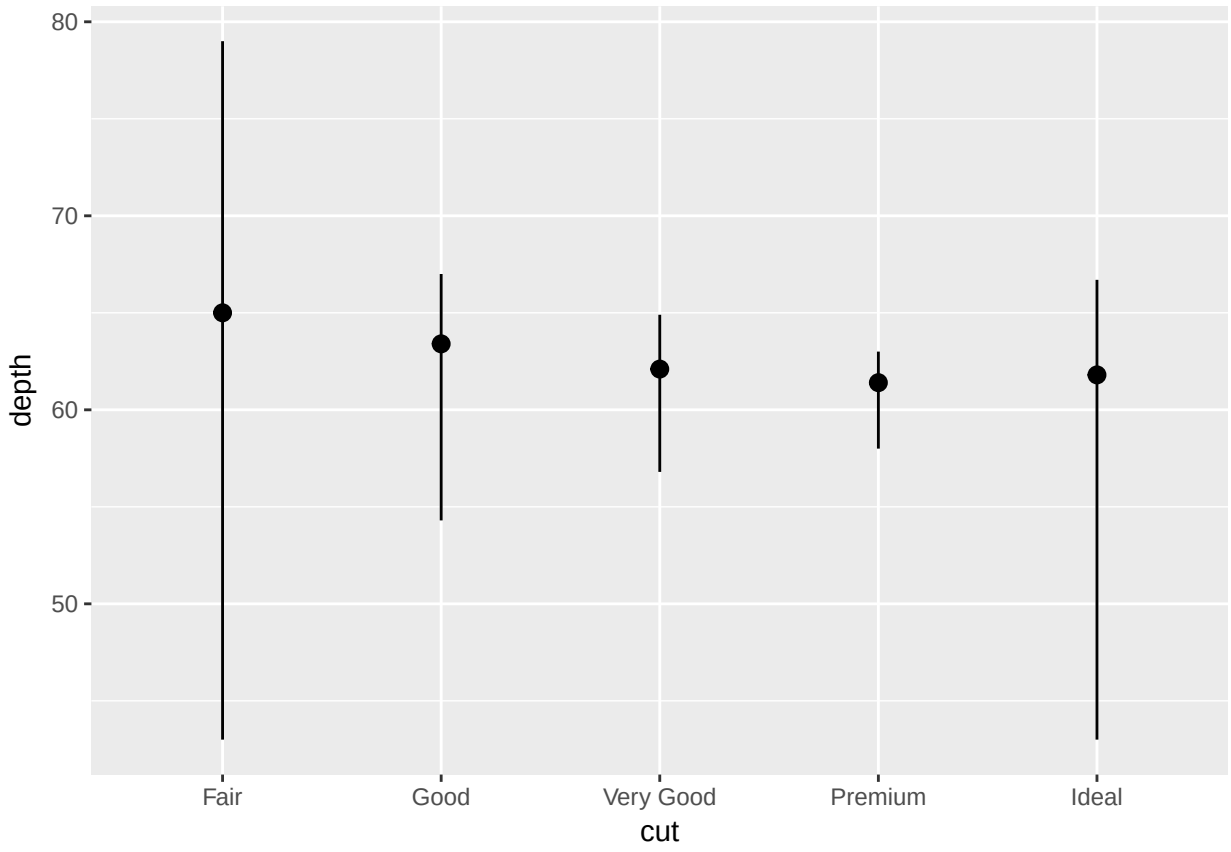
2. Muchos *stats* computan varias variables y quizás quieras mostrar otra. Abajo, en lugar de graficar la frecuencia o cuenta, grafico la proporción o frecuencia relativa.

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, y = ..prop.., group = 1))
```



3. Quizás quieras llamar la atención sobre ciertos medidas de resumen que has calculado. Puedes hacer esto con `stat_summary`.

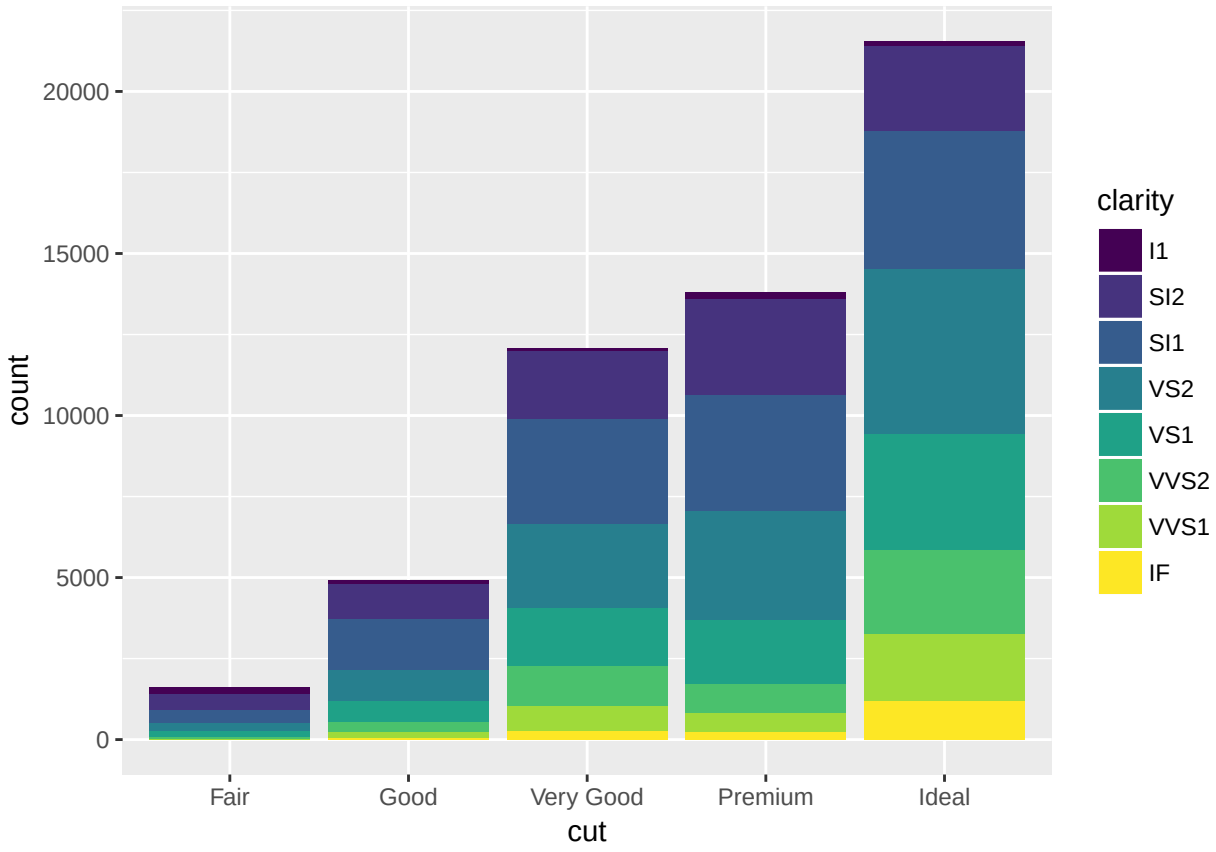
```
ggplot(data = diamonds) +
  stat_summary(
    mapping = aes(x = cut, y = depth),
    fun.y = median,
    fun.ymax = max,
    fun.ymin = min

  )
```



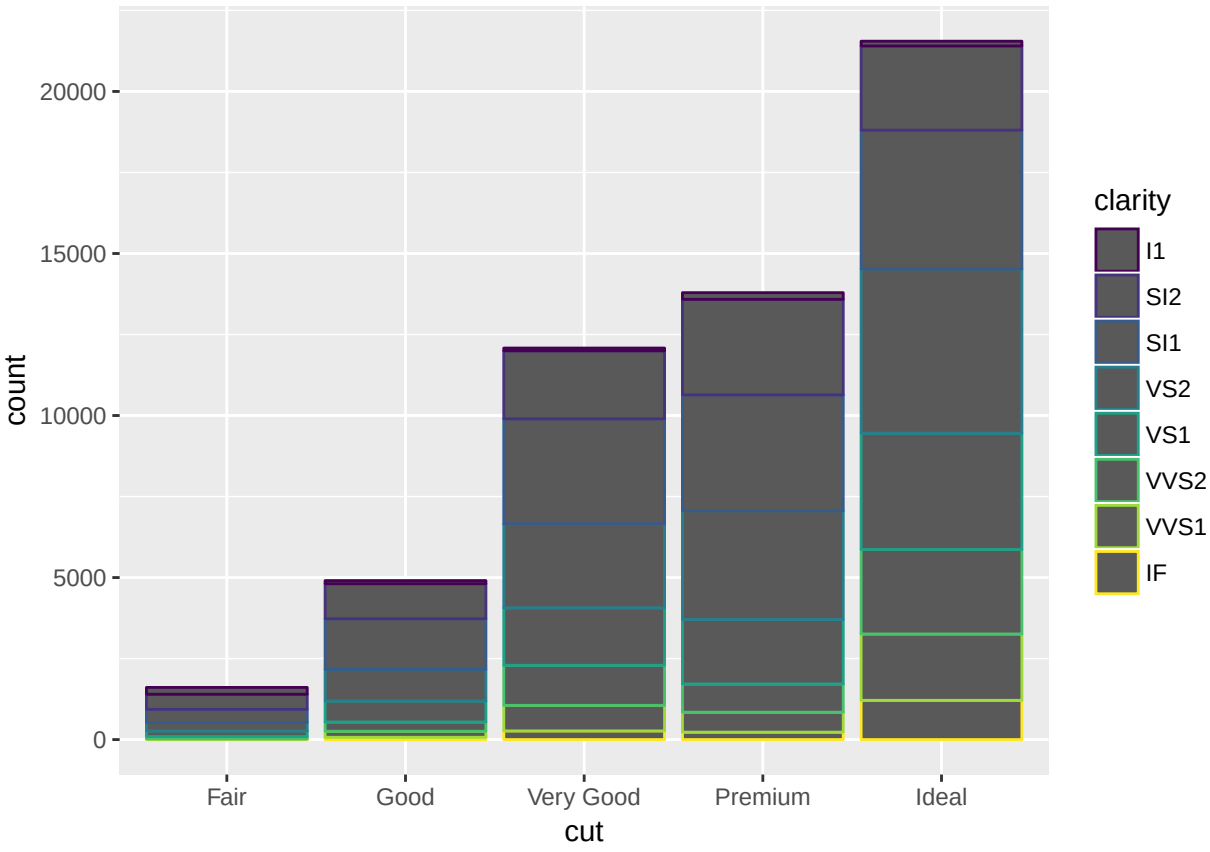
3.7 Ajuste de Posiciones

El gráfico anterior revela algo interesante de los gráficos de barras. Tienen relleno (*fill*) y tienen color.

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = clarity))
```



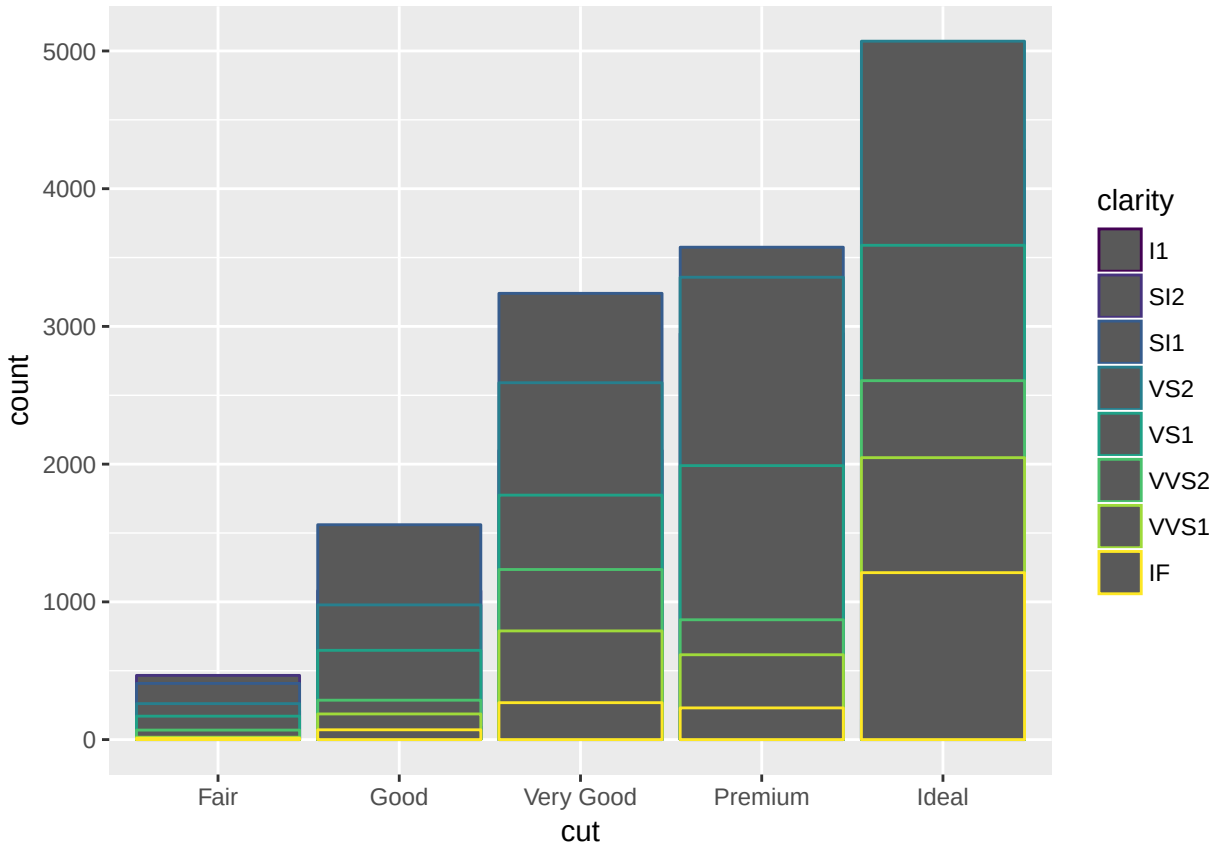

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, color = clarity))
```



Es más claro trabajar con el relleno porque es más visible en el gráfico. Pero vemos que las distintas barras están apiladas, lo que dificulta la comparación. Se debe a que la posición de barras es apilada (`stack`) por defecto. Podemos cambiarla modificando el argumento `position` de `geom_bar()`. Entre las otras posiciones que podemos elegir están identidad (`identity`), esquivar (`dodge`) y relleno (`fill`) .

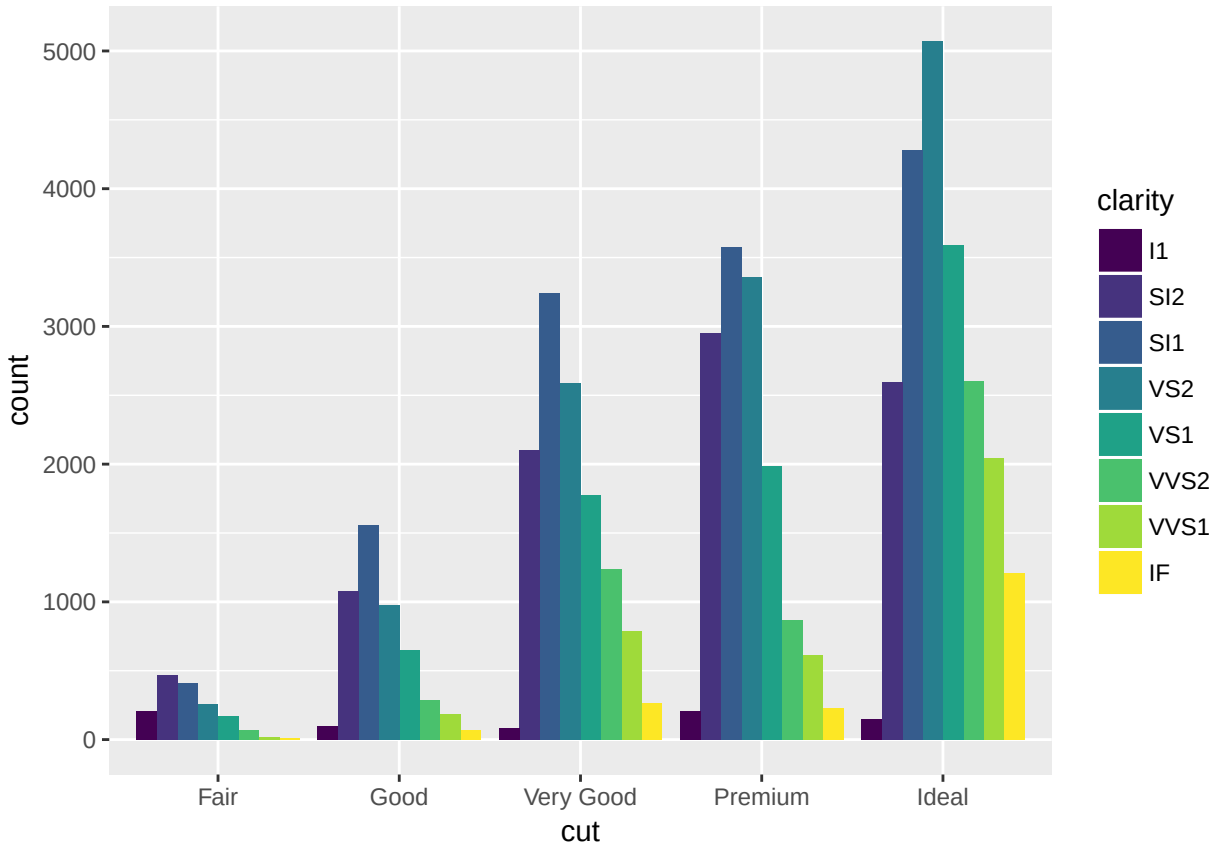
- Identidad hace que las barras (o otro *geoms*) caigan unas encimas de otras. No es muy útil porque las barras se superponen y hace muy difícil la interpretación. Se puede mejorar agregando transparencia y usando color y no relleno, pero es complicado de interpetar si la barras se superponen o están apiladas.

```
ggplot(data = diamonds) +
   geom_bar(mapping = aes(x = cut, color = clarity), position = "identity")
```



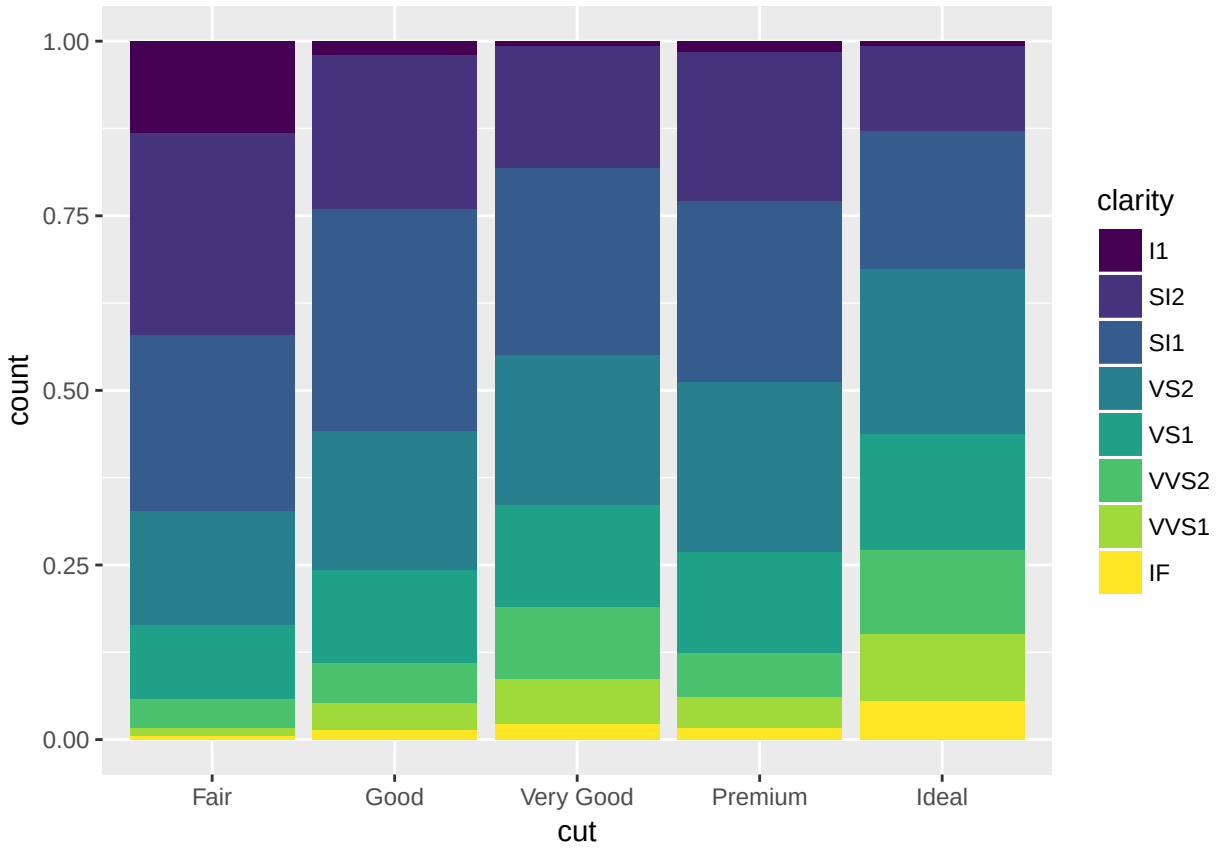
- Esquivar es quizás la más útil junto con relleno. Hace que las barras estén una al lado de la otra. Lo que hace que sea sencillo comparar la altura de estas.

```
ggplot(data = diamonds) +
   geom_bar(mapping = aes(x = cut, fill = clarity), position = "dodge")
```



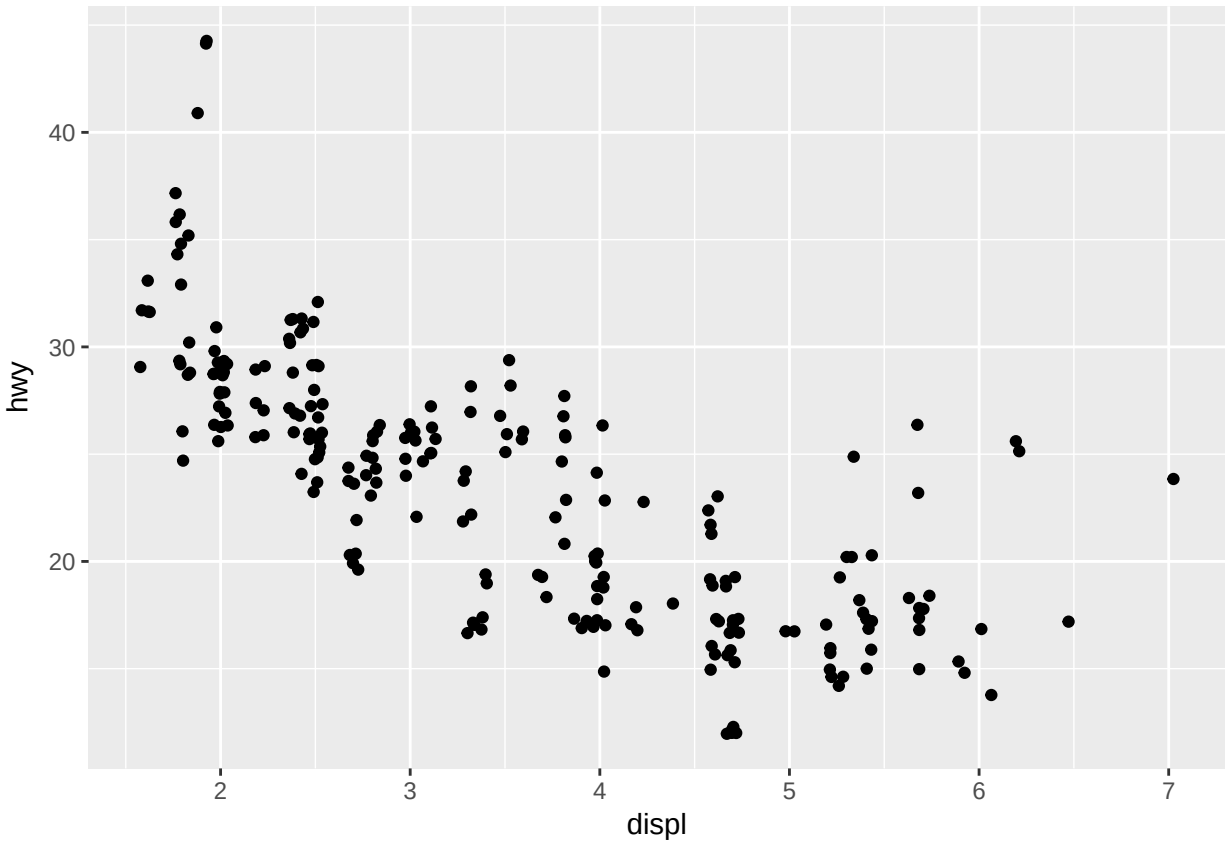
- Recién vimos que que con esquivar podemos poner las barras una al lado de otra. Pero, más allá de comparar la cantidad de diamantes en cada uno, ya que la cantidad es muy diferente en cada corte resulta más útil comparar las proporciones de cada una de las claridades. Relleno funciona de forma similar al apilado, pero estadariza cada columna a longitud uno. Entonces se ve las proporciones o frecuencias relativas de cada uno de los niveles de `clarity`.

```
ggplot(data = diamonds) +
   geom_bar(mapping = aes(x = cut, fill = clarity), position = "fill")
```



Hay otros ajustes de posiciones que no son útiles para los gráficos de barras pero son muy útiles para los gráficos de puntos. En el gráfico de dispersión entre `displ` y `hwy` hay menos puntos que el total 31 vs 234. Muchos puntos se superponen, por eso vemos menos. Podemos evitarlo añadiendo un poco de movimiento aleatorio a cada punto.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), position = "jitter")
```



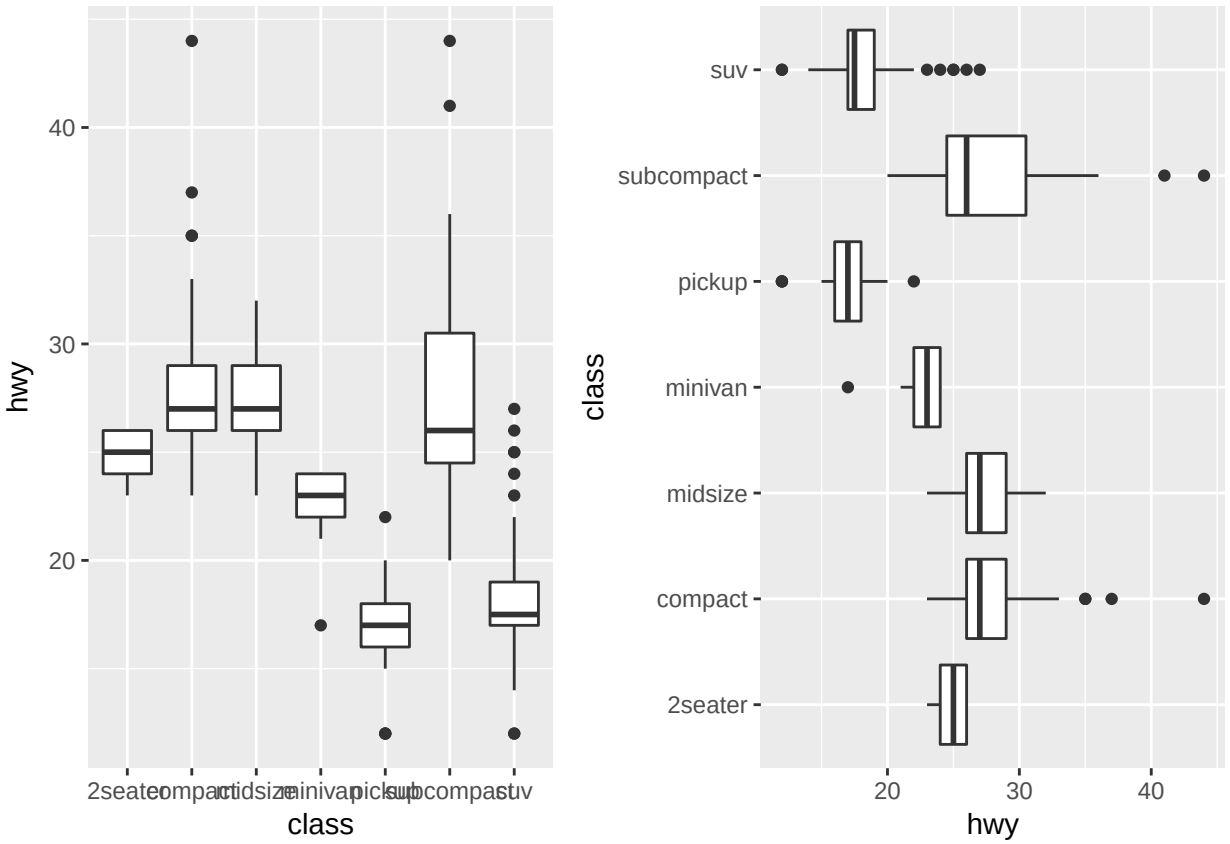
Si bien el gráfico no va a ser exacto, muestra más información que en el caso donde se superponen los puntos.

Podes obtener más información el ayuda de cada uno: `?position_dodge`, `?position_identity`, `?position_fill`, `?position_stack`, `?position_jitter`

3.8 Sistemas de Coordenadas

Hasta ahora estuvimos graficando en un sistema de coordenadas cartesianas. Pero es posible cambiarlo, por ejemplo intercambiando el eje x e y.

```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
  geom_boxplot()
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
  geom_boxplot() +
  coord_flip()
```



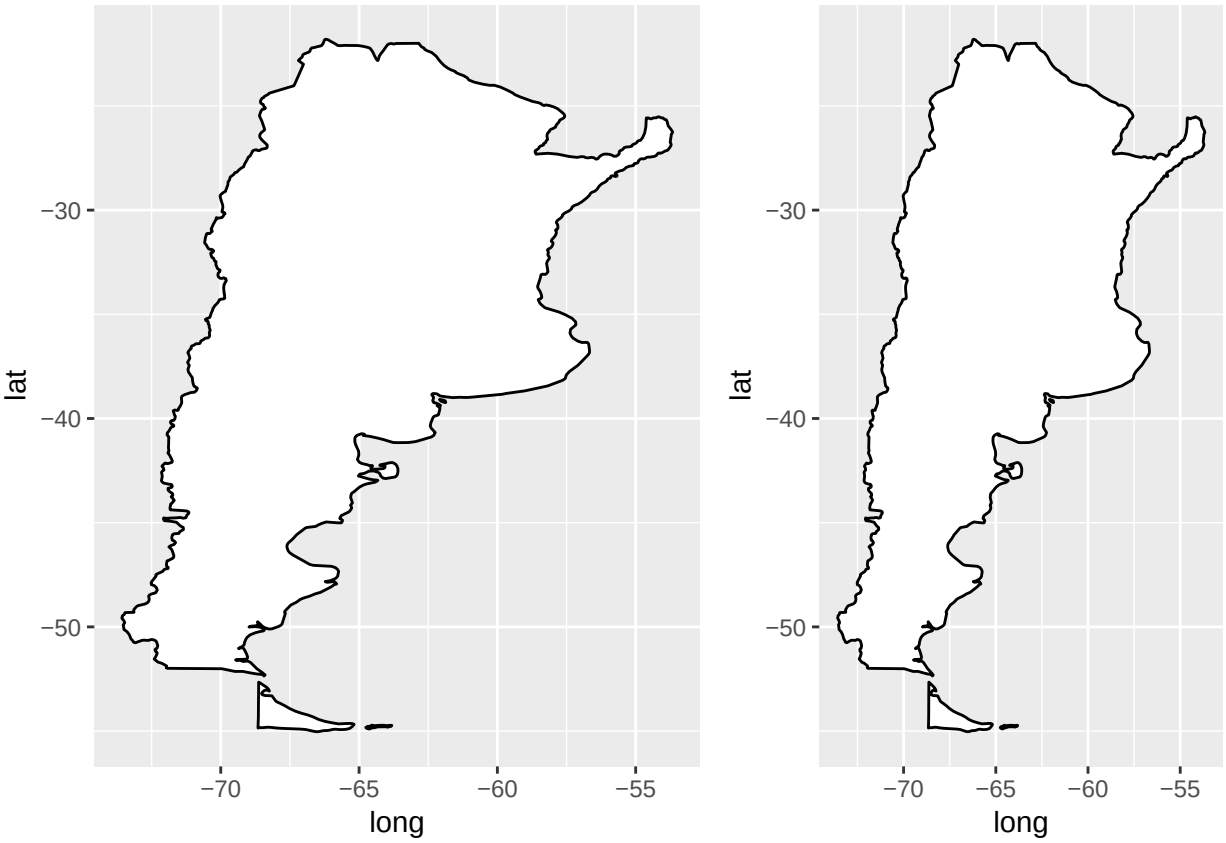
En el primer caso las etiquetas del eje x se superponen, pero en el segundo es fácil verlas. No es la única forma de solucionar este problema. También es posible cambiar el ángulo de las etiquetas para que no se superpongan.

Otras veces es mejor reemplazar las coordenadas cartesianas por coordenadas geográficas.

```
arg <- map_data("world", region = "Argentina")

ggplot(data = arg, mapping = aes(x = long, y = lat, group = group)) +
  geom_polygon(fill = "white", color = "black")

ggplot(data = arg, mapping = aes(x = long, y = lat, group = group)) +
  geom_polygon(fill = "white", color = "black") +
  coord_quickmap()
```

```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map

## <ggproto object: Class CoordQuickmap, CoordCartesian, Coord, gg>
##     aspect: function
##     default: FALSE
##     distance: function
##     expand: TRUE
##     is_linear: function
##     labels: function
```

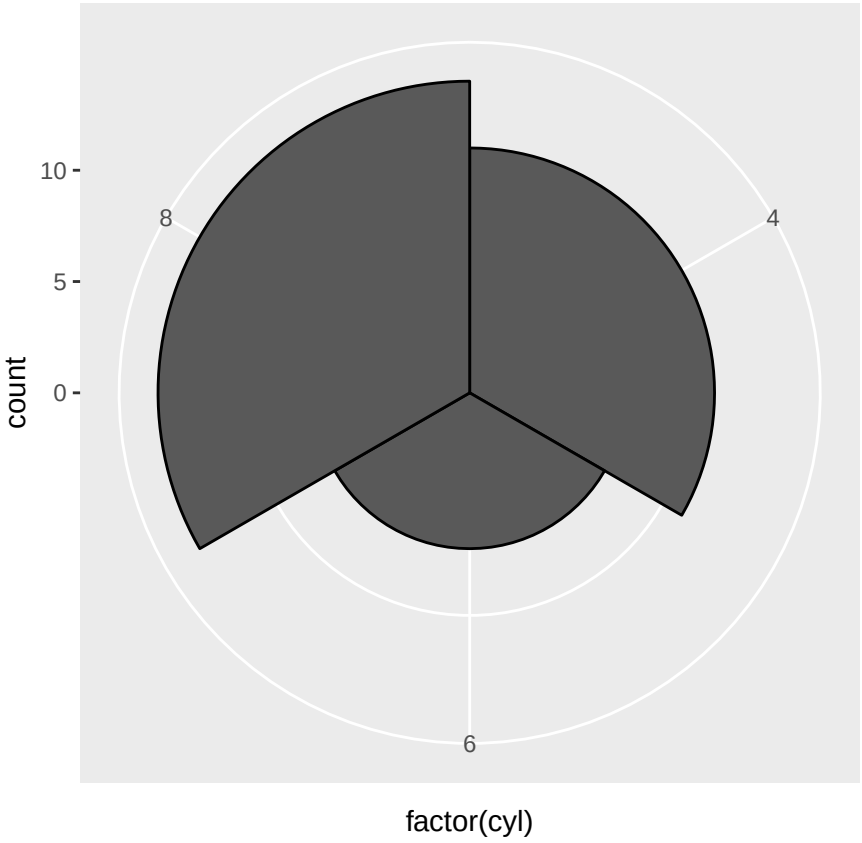
```
##     limits: list
##     modify_scales: function
##     range: function
##     render_axis_h: function
##     render_axis_v: function
##     render_bg: function
##     render_fg: function
##     setup_data: function
##     setup_layout: function
##     setup_panel_params: function
##     setup_params: function
##     transform: function
##     super:  <ggproto object: Class CoordQuickmap, CoordCartesian, Coord, gg>
```



Esto evita que el mapa se deforme, ya que los grados de longitud no miden lo mismo en todas las latitudes. Si van a hacer muchos mapas les recomiendo que vean la extesión `ggmap` que tiene muchas utilidades para hacer mejores mapas.

También existen la coordenadas polares. Un gráfico de torta, que les recomiendo que no lo usen por los problemas de percepción que tiene, es un gráfico de barras apiladas en coordernadas polares.

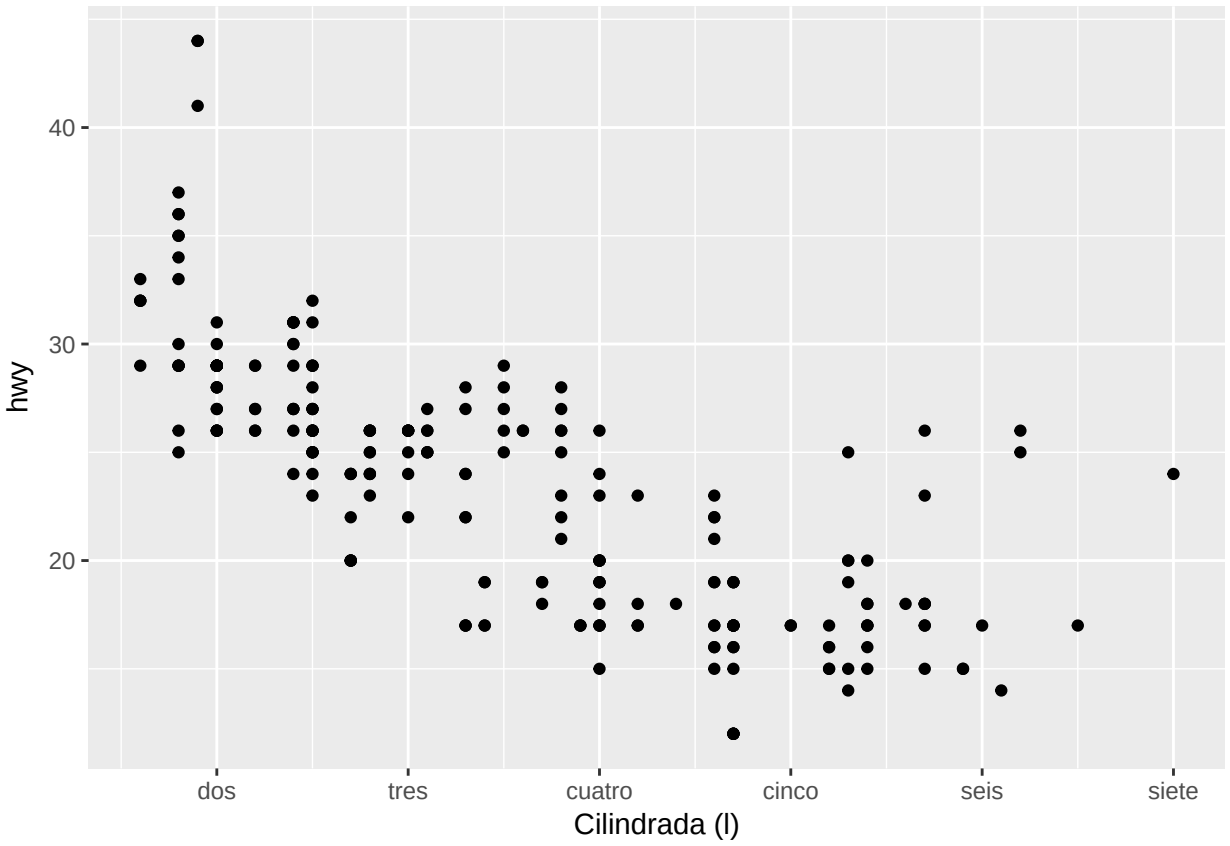
```
cxc <- ggplot(mtcars, aes(x = factor(cyl))) +
  geom_bar(width = 1, colour = "black")
cxc + coord_polar()
```



3.9 Personalizando el gráfico

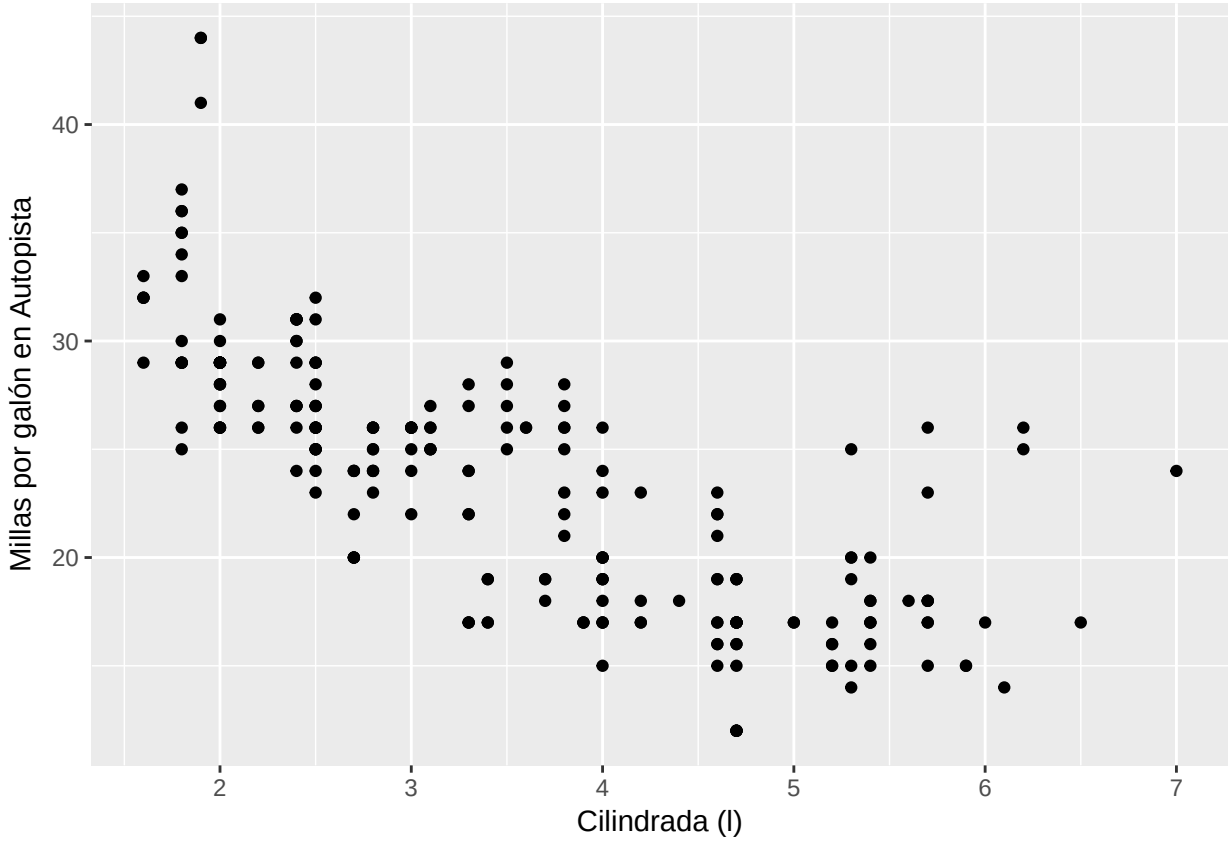
Hay varias maneras de personalizar los gráficos. Por un lado, las estéticas pueden ser personalizadas cambiando las distintas `scales` (escalas). Para cambiar el eje x se usa `scale_x_*` donde * es el tipo de dato que tiene el eje: si es númerico se usa `continuous` y si es categórico se usa `discrete`. Se pueden cambiar muchas cosas: el título del eje (`name`), el lugar de las marcas (`breaks`), las etiquetas de las marcas (`labels`), y muchas más opciones.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))+
  scale_x_continuous(name = "Cilindrada (l)", breaks = 1:7,
                     labels = c("uno", "dos", "tres", "cuatro", "cinco",
                                "seis", "siete"))
```



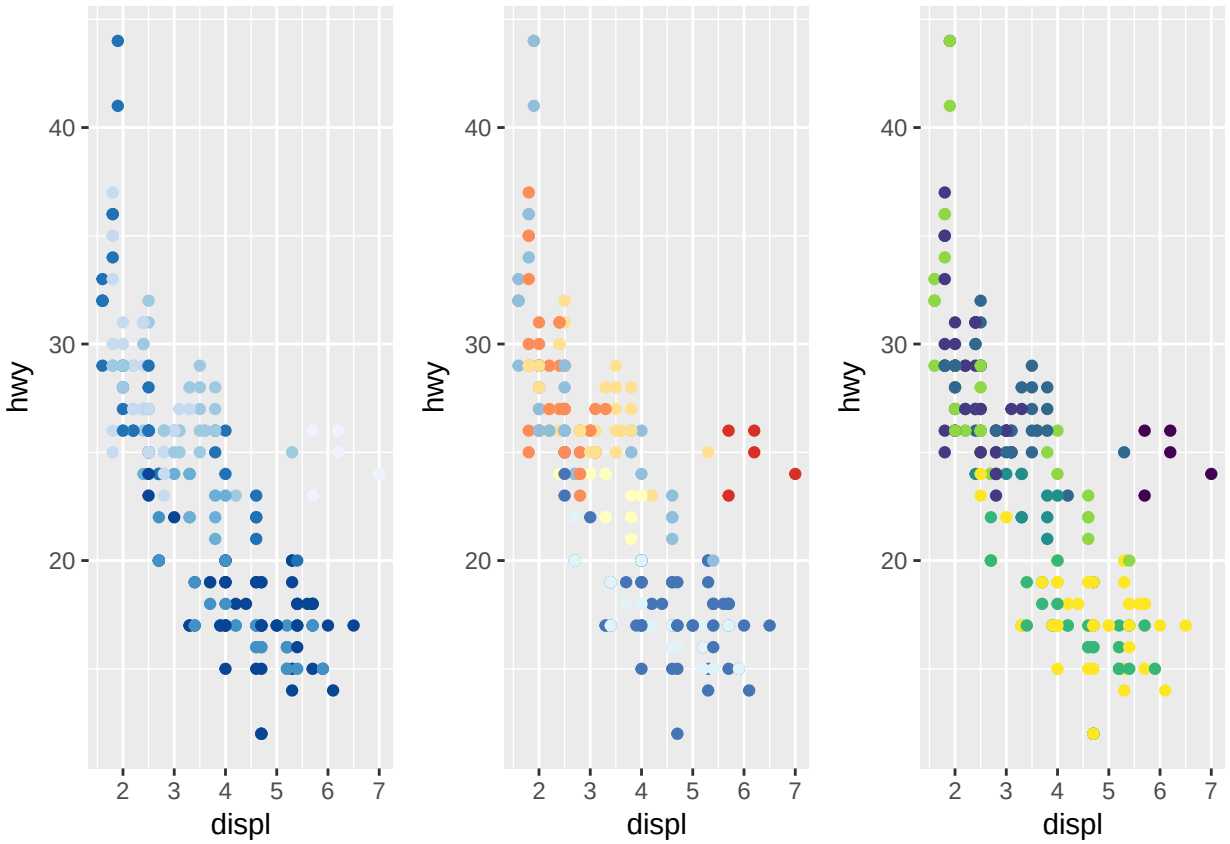
Un atajo para modificar los nombres de los ejes es usar la función `labs()`, pero solo se pueden modificar los nombres de los ejes y nada más.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))+
  labs(x = "Cilindrada (l)", y = "Millas por galón en Autopista")
```



También se puede modificar los colores que se asignan.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, colour = class))+
  scale_color_brewer("Clase")

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, colour = class))+
  scale_color_brewer("Clase", palette = "RdYlBu")

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, colour = class))+
  scale_color_viridis_d("Clase")
```

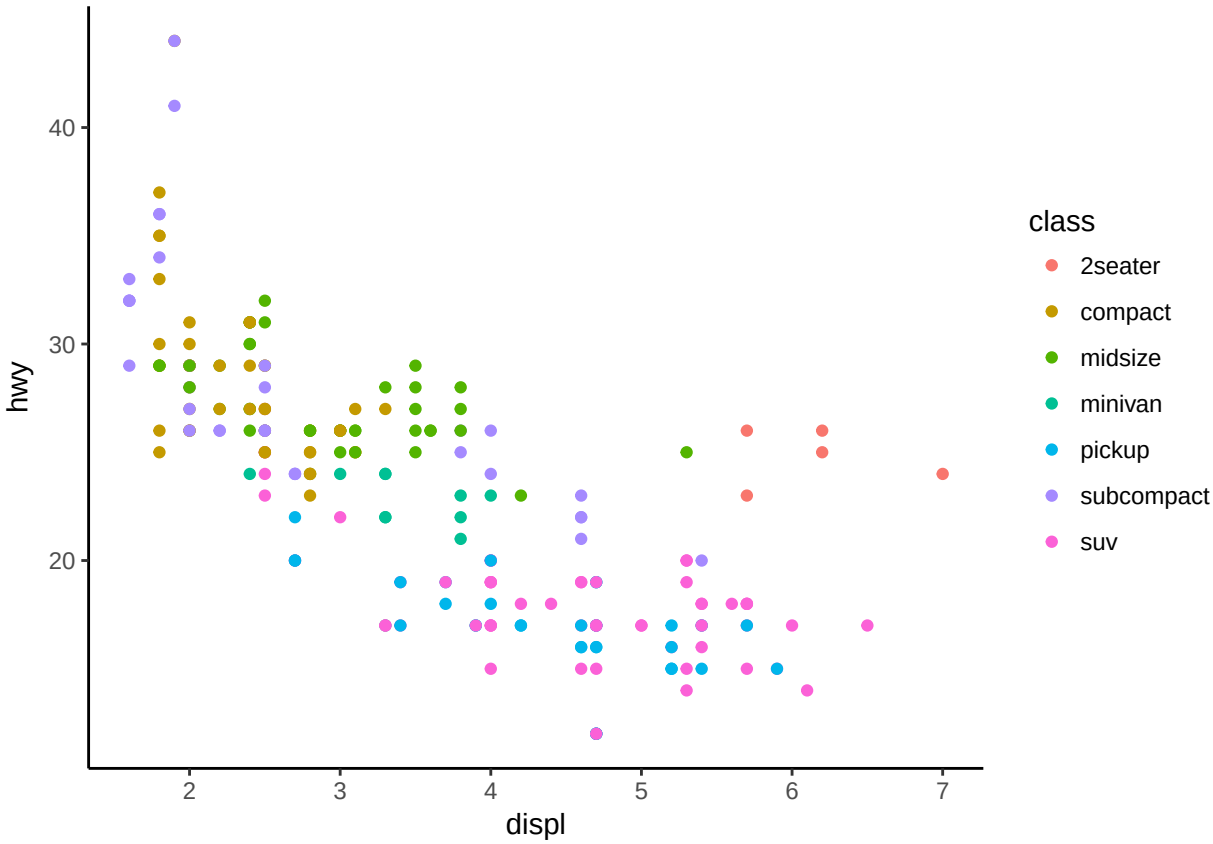


Hay muchas más opciones disponibles, ya que como dice el dicho: "Para gustos, los colores". Si quieren conocerlas te recomiendo que lean la ayuda de cada una o visiten el sitio de ggplot2.

Vemos que hay patrón común con las escalas, todas empiezan por `scale`, luego sigue por lo que se quiere modificar: el eje, `x` o `y`; el color, `color`; relleno, `fill`; la forma, `shape`; el tipo de línea `linetype`, etc. Cada uno de las estéticas tiene su escale. Luego, salvo alguna excepción, siguen por el tipo de dato o en el caso de los colores por el método de creación del color. Vale la pena agregar que cada escale tiene su versión manual para un control total de la apariencia.

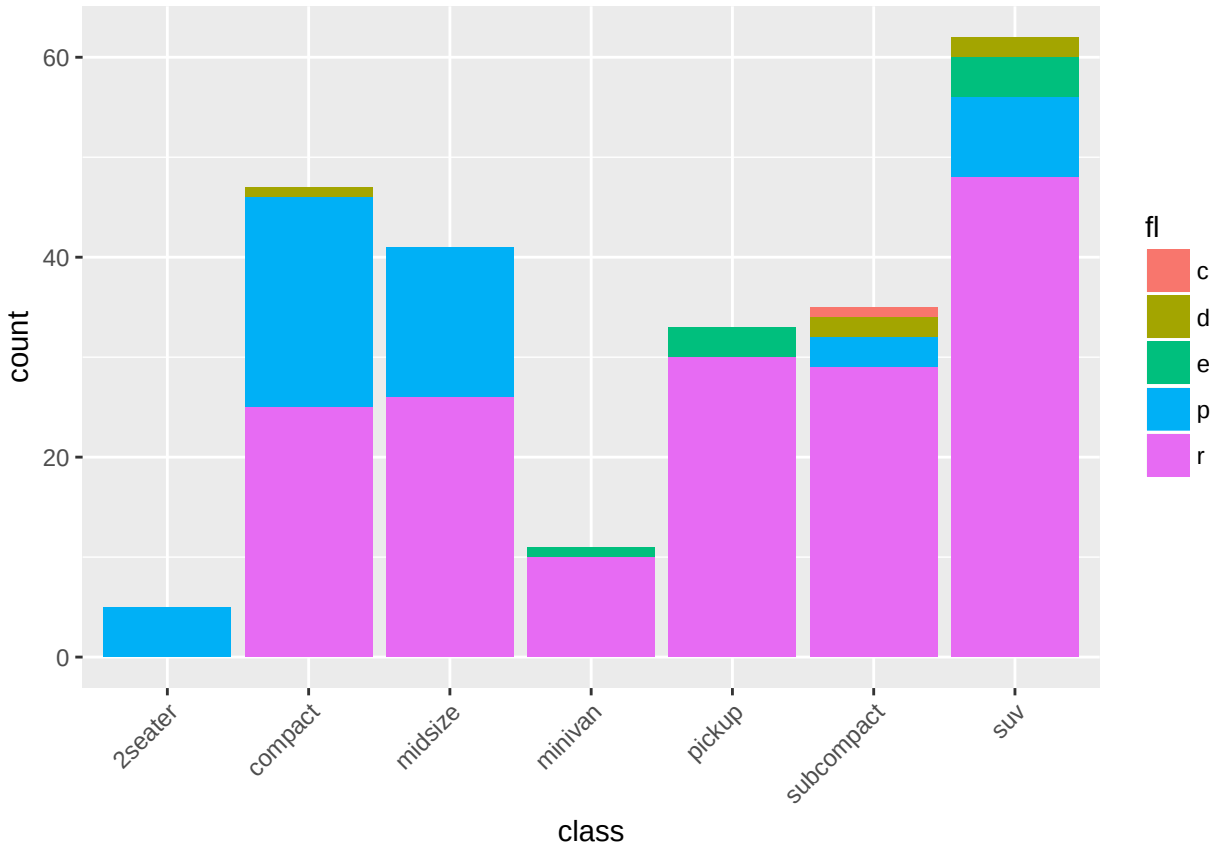
Por otro lado están los elementos del gráfico que modifican la apariencia general del gráfico. El tipo y tamaño de letra, el color del fondo, el grosor de la líneas de los ejes, la dirección de marcas, la dirección del texto, y ¡todo lo demás!. Todo esto está unido a lo que es el tema (`theme`) del gráfico. Se pueden guardar las modificaciones para usarla facilmente y ya vienen algunas opciones en `ggplot` y hay más en el paquete `ggthemes` y otros.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, colour = class)) +
  theme_classic()
```



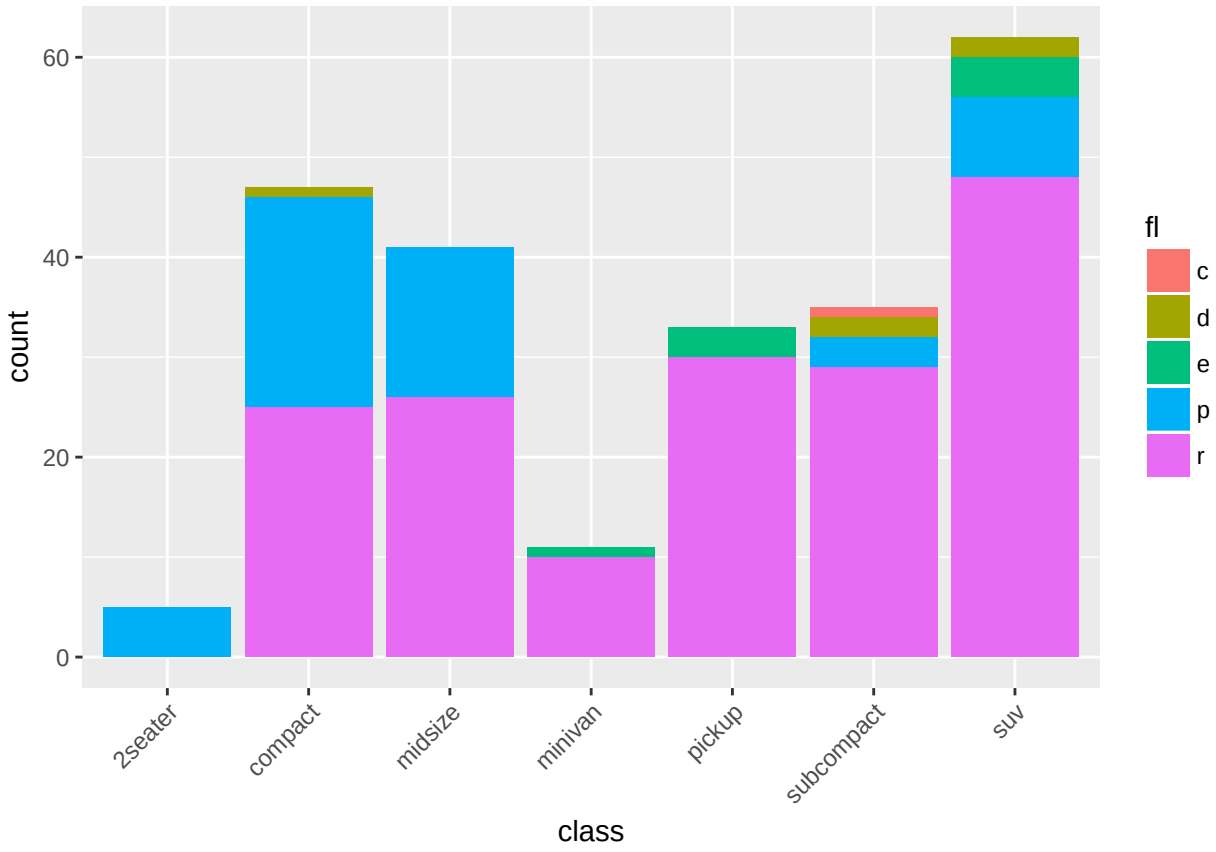
Para modificar algún elemento en particular usamos la función `theme()` al final del gráfico. Dentro de la llamada a `theme` modificamos el argumento que queremos cambiar usando la función `element_*()`.

```
ggplot(data = mpg) +
  geom_bar(mapping = aes(x = class, fill = fl)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```



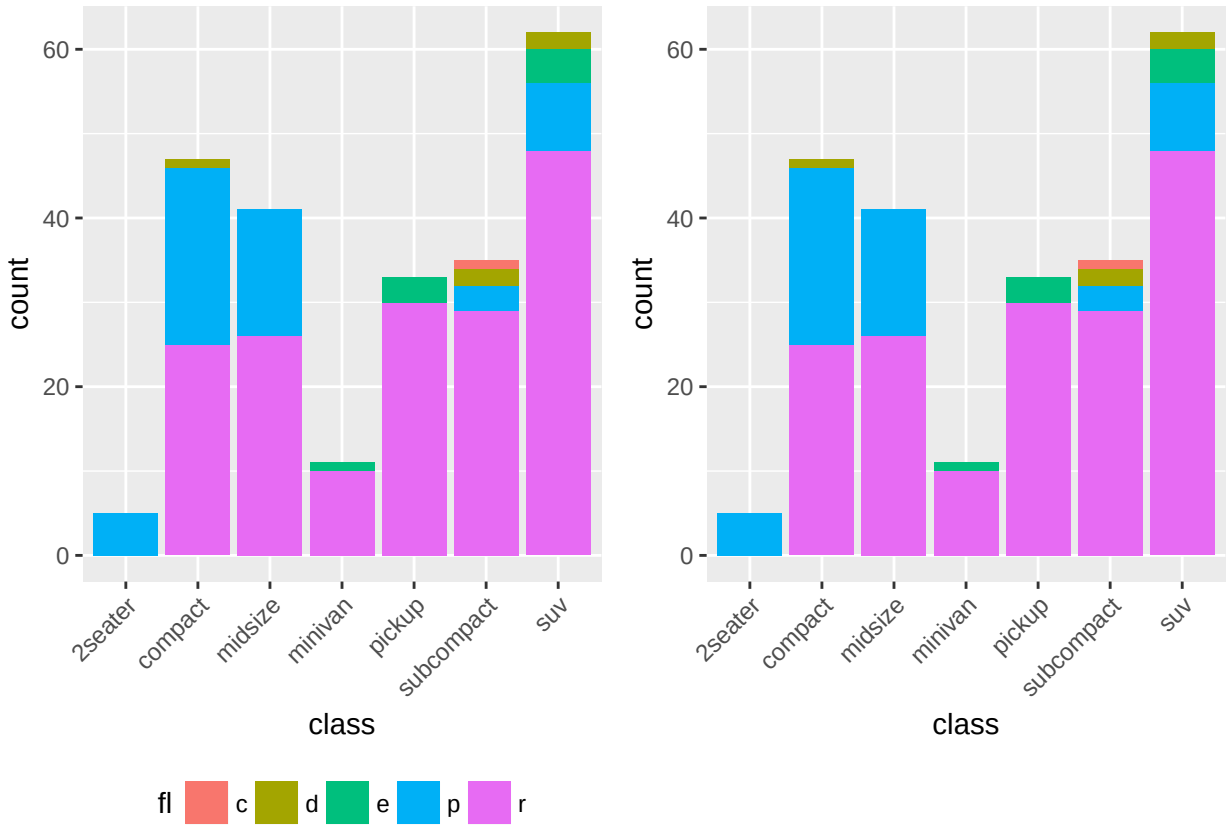
Sí, es bastante complicado. Pero por suerte se puede guardar y reutilizar.

```
x_45 <- theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

ggplot(data = mpg) +
  geom_bar(mapping = aes(x = class, fill = fl)) +
  x_45
```



También es posible cambiar la posición de la leyenda o eliminarla completamente.

```
ggplot(data = mpg) +
  geom_bar(mapping = aes(x = class, fill = fl)) +
  theme(legend.position = "bottom")

ggplot(data = mpg) +
  geom_bar(mapping = aes(x = class, fill = fl)) +
  theme(legend.position = "none")
```


60
40
20
0
count
2seater
compact
midsize
minivan
pickup
subcompact
suv
class
fl
c
d
e
p
r

Chapter 4

Manejo de datos

> Antes de comenzar bajen el archivo donde realizarán su informe reproducible. En la consola copien este código:
> `download.file(url = "git.io/informe-manejo", destfile = "informe-manejo-de-datos.Rmd")`
> Pueden abrirlo desde la pestaña de archivos, a la derecha. Cambien el nombre por el suyo en el encabezado y mientras leen este capítulo respondan las preguntas.

Una parte muy importante del análisis de datos, es el manejo de ellos. Como seleccionar columnas, filtrar datos, y realizar operaciones sobre ellos. Vamos a usar el paquete `dplyr` y `tidyr` para el manejo. Los paquetes extienden la funcionalidad de *R* agregando nuevas funciones.

```
library("dplyr")

nombres <- readRDS("data/nombres-1980-1999.RDS")
```

Revisemos el código. Con `library("dplyr")` cargamos el paquete dplyr. Luego, leemos el archivo que contiene los datos y le asignamos el nombre `nombres`. Si no le asignasemos ningún nombre, se leerían los datos, imprimiendose en la consola y luego se borrarían de la memoría.

Para revisar su contenido podemos escribir el nombre del objeto o usar la función `glimpse`

```
glimpse(nombres)
```

Ejercicio 4.1. Escriban el nombre del objeto o usen la función `glimpse` para ver que tiene dentro el objeto `nombres`.

1. ¿Cuantas columnas tiene y como se llaman?
2. ¿Que tipo de dato tiene cada columna?

En *R* existen diversos tipos de dato, en estos datos solo hay 2: entero (`integer`) y carácter (`character`). El primero son números enteros y el segundo es texto. Con el primero se puede hacer operaciones matemáticas y con el segundo otro tipo de operaciones, pero no matemáticas. Es importante comprobar que los tipos de datos se correspondan con lo que esperamos. Si no los resultados pueden no ser los correctos o dar errores. Por ejemplo, el tipo de dato númerico puede ser leído como `chr` y no podremos calcular la media.

Table 4.1: Operadores Lógicos en R.

Operador	Descripción
<	menor que
<=	menor o igual que
>	mayor que
>=	mayor o igual que
==	exactamente igual a
!=	no igual a
!x	no x
x | y	x *O* y
x & y	x *E* y
isTRUE(x)	comprueba si x es verdadero

4.1 Seleccionando datos

Muchas veces solo nos interesa un subconjunto de datos. Una forma de seleccionar datos es usando la función `filter()`.

```
nombres %>%
  filter(nombre == "Luciano") %>%
  filter(anio == 1984)
```

Acá empezamos a ver varias cosas nuevas. Primero tenemos el símbolo `%>%` conocido en inglés como *pipe*, la traducción más correcta al español es tubo. Lo que hace este símbolo es enviar la salida de la operación a la izquierda a la función de la derecha. Prueben poner cada comando en orden y ver cual es la salida. Esto es:

```
nombres
```

Luego,

```
nombres %>%
  filter(nombre == "Luciano")
```

La función `filter()` filtra un conjunto de datos según los valores de la columna/s que seleccionemos cuyos valores sean igual a `Luciano` en este caso. Y luego filtramos la columna `anio` solo los años que sean iguales a 1984.

Ejercicio 4.2. Prueben cambiar el nombre por el suyo y el año por su año de nacimiento.

El operador que usamos para la igualdad es `==`. Este operador, de igualdad, es parte de la familia de operadores lógicos, o booleanos en terminología de ciencias de la información. Son lógicos porque van a comparar valores y dar como resultado **verdadero** (`TRUE`) o **falso** (`FALSE`). En la Tabla 4.1 podemos ver la lista de operadores lógicos.

Los primeros cinco son bastante sencillos y los han estado usando desde la primaria. Así que vamos a explicar en más profundidad los otros. El símbolo `!=` va a devolver `TRUE` cuando los valores sean diferentes al que pusimos. Por ejemplo:

```
# x una secuencia de 1 a 10
x <- 1:10
# Todos los valores distintos a 5
x != 5
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Ejercicio 4.3 (Nombres comunes)**.** ¿Cómo filtrarían los nombres raros excluyéndolos del conjunto de datos?

Guarden el resultado como nombres_comunes y calculen el total por nombre.

Nota: Por coherencia, definamos nombres raros como los que son menos de 100.

Otro operador muy útil es el de negación `!` que invierte las comparaciones, convierte los falsos en verdaderos y los verdaderos en falsos. Siguiendo nuestro ejemplo:

```
!x != 5
```

```
##  [1] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
```

Es un ejemplo trivial, que podría haber sido resuelto más sencillamente usando `==`. Pero es muy útil cuando queremos seleccionar todos los datos que no cumplan un conjuto de condiciones. Lo que nos lleva al operador `|` (O) y el operador `&` (Y). El primero va a devolver verdadero cuando *al menos uno* de los valores sea verdadero. Por ejemplo:

```
TRUE | TRUE
```

```
## [1] TRUE
```

```
TRUE | FALSE
```

```
## [1] TRUE
```

```
FALSE | TRUE
```

```
## [1] TRUE
```

```
FALSE | FALSE
```

```
## [1] FALSE
```

Por otro lado, el operador lógico Y `&` solo devuelve verdadero cuando *ambos valores* son verdaderos.

```
TRUE & TRUE
```

```
## [1] TRUE
```

```
TRUE & FALSE
```

```
## [1] FALSE
```

```
FALSE & TRUE
```

```
## [1] FALSE
```

```
FALSE & FALSE
```

```
## [1] FALSE
```

Los operadores se evalúan en el orden que aparecen a menos que haya paréntesis, entonces se evalúa primero dentro del paréntesis y luego fuera.

Ejercicio 4.4. ¿Qué resultado darán las siguientes evaluaciones? Piensen que resultado tendría que dar y luego comprueben lo que piensan con lo que les devuelve R.

1. TRUE | FALSE | TRUE
2. TRUE | FALSE & TRUE
3. TRUE | (FALSE & TRUE)
4. TRUE != FALSE & TRUE
5. !(TRUE | FALSE) & TRUE

Estos dos últimos operadores son muy importantes porque nos permiten comprobar distintas condiciones. Por ejemplo, no hay un operador para seleccionar todos los valores entre a y b (siendo a y b dos números cualesquiera). Podemos hacerlo combinando por un lado, x $>$ a y x $<$ b ¿Y cómo debemos combinar estas

dos comparaciones? ¿Usando el operador `&` o el `|`? Queremos los valores que cumplen con ambas condiciones, que sean mayores que a y menores que b, por lo tanto debemos usar el operador `&`.

```
# Si a = 3 y b = 6
( x > 3 ) & ( x < 6)
```

```
##  [1] FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
```

Estos valores corresponden a la posición de los valores que cumplen o no con la condición. Usando corchetes `[]` podemos seleccionar solo los verdaderos

```
x[( x > 3 ) & ( x < 6)]
```

```
## [1] 4 5
```

La función `filter()` hace algo similar para conjuntos de datos (`data.frames` o `tibbles`).

Ejercicio 4.5. Anteriormente usamos dos operaciones de `filter()` para seleccionar el nombre y el año. Pero es posible usar solo una con los operadores lógicos que vimos. Intenten hacerlo.

Finalmente está `isTRUE()` que devuelve `TRUE` cuando el objeto es `TRUE` lo que suena bastante obvio. Pero es parte de una familia que permite comprobar si un objeto es del tipo esperado. Por ejemplo: `is.numeric()` comprueba que el objeto es un vector con algún tipo de número.

Otra forma de seleccionar datos es por posición. Es decir, seleccionar los primeras diez filas:

```
nombres_comunes %>%
  slice(1:10)
```

```
## # A tibble: 10 x 3
##    nombre          anio cantidad
##    <chr>          <int>    <int>
##  1 Aaron           2012      152
##  2 Aaron           2013      167
##  3 Aaron           2014      200
##  4 Aaron Benjamin  2012      108
##  5 Aaron Benjamin  2013      120
##  6 Aaron Benjamin  2014      125
##  7 Abel            1982      102
##  8 Abel            1989      103
##  9 Abel            1990      102
## 10 Abigail         1991      132
```

O seleccionar las primeras 10 filas que corresponden números primos:

```
nombres_comunes %>%
  slice(c(2, 3, 5, 7, 11, 13, 17, 19, 23, 29))
```

```
## # A tibble: 10 x 3
##    nombre          anio cantidad
##    <chr>          <int>    <int>
##  1 Aaron           2013      167
##  2 Aaron           2014      200
##  3 Aaron Benjamin  2013      120
##  4 Abel            1982      102
##  5 Abigail         1992      120
##  6 Abigail         1994      198
##  7 Abigail         1998      210
##  8 Abigail         2010      171
##  9 Abigail         2014      278
```

```
## 10 Abril           1999      813
```

También es posible eliminar las filas según posición:

```
nombres_comunes %>%
  slice(-(1:10)) #Tengan en cuenta los parentesis extra
```

```
## # A tibble: 33,665 x 3
##    nombre   anio cantidad
##    <chr>   <int>    <int>
##  1 Abigail  1992      120
##  2 Abigail  1993      165
##  3 Abigail  1994      198
##  4 Abigail  1995      182
##  5 Abigail  1996      159
##  6 Abigail  1997      167
##  7 Abigail  1998      210
##  8 Abigail  1999      174
##  9 Abigail  2010      171
## 10 Abigail  2011      152
## # ... with 33,655 more rows
```

Ejercicio 4.6. 1. ¿Qué sucede si olvidan los paréntesis en el código de arriba? 2. Seleccionen las últimas 10 filas.

4.2 Seleccionando columnas

La función para seleccionar columnas es `seletc()`. Hay muchas formas de seleccionar columnas. La más obvia es por nombre de la columna:

```
nombres_comunes %>%
  select(nombre, cantidad)
```

```
## # A tibble: 33,675 x 2
##    nombre         cantidad
##    <chr>             <int>
##  1 Aaron               152
##  2 Aaron               167
##  3 Aaron               200
##  4 Aaron Benjamin      108
##  5 Aaron Benjamin      120
##  6 Aaron Benjamin      125
##  7 Abel                102
##  8 Abel                103
##  9 Abel                102
## 10 Abigail             132
## # ... with 33,665 more rows
```

También es posible seleccionar varias columnas usando secuencias:

```
nombres_comunes %>%
  select(nombre:cantidad)
```

```
## # A tibble: 33,675 x 3
##    nombre          anio cantidad
##    <chr>          <int>    <int>
```

```
##  1 Aaron           2012      152
##  2 Aaron           2013      167
##  3 Aaron           2014      200
##  4 Aaron Benjamin  2012      108
##  5 Aaron Benjamin  2013      120
##  6 Aaron Benjamin  2014      125
##  7 Abel            1982      102
##  8 Abel            1989      103
##  9 Abel            1990      102
## 10 Abigail         1991      132
## # ... with 33,665 more rows
```

De la misma forma se puede eliminar columnas usando el signo `-`.

```
nombres_comunes %>%
  select(-anio)
```

```
## # A tibble: 33,675 x 2
##    nombre         cantidad
##    <chr>             <int>
##  1 Aaron               152
##  2 Aaron               167
##  3 Aaron               200
##  4 Aaron Benjamin      108
##  5 Aaron Benjamin      120
##  6 Aaron Benjamin      125
##  7 Abel                102
##  8 Abel                103
##  9 Abel                102
## 10 Abigail             132
## # ... with 33,665 more rows
```

Se pueden renombrar columnas

```
nombres_comunes %>%
  select(año = anio)
```

```
## # A tibble: 33,675 x 1
##      año
##    <int>
##  1  2012
##  2  2013
##  3  2014
##  4  2012
##  5  2013
##  6  2014
##  7  1982
##  8  1989
##  9  1990
## 10  1991
## # ... with 33,665 more rows
```

Pero se eliminan las no seleccionadas. Se puede renombrar sin tener que seleccionar el resto usando la función `rename`

```
nombres_comunes %>%
  select(año = anio)
```

```
## # A tibble: 33,675 x 1
##      año
##    <int>
##  1  2012
##  2  2013
##  3  2014
##  4  2012
##  5  2013
##  6  2014
##  7  1982
##  8  1989
##  9  1990
## 10  1991
## # ... with 33,665 more rows
```

Hay muchas más formas de seleccionar columnas, pueden referirse a la ayuda `?select`, `?select_at` y también a este excelente [tutorial][https://suzan.rbind.io/2018/01/dplyr-tutorial-1/] (en inglés).

4.3 Agregando columnas

Otra operación muy común es agregar nuevas columnas o variables. Por ejemplo al transformar los datos es siempre **mala idea** sobreescribir los datos originales.

Para esta operación sirve la función `mutate()`. Dado un *data frame* computa una valor para cada fila. Por ejemplo:

```
nombres %>%
  mutate(log_cantidad = log10(cantidad))
```

```
## # A tibble: 3,749,133 x 4
##    nombre         anio cantidad log_cantidad
##    <chr>         <int>    <int>        <dbl>
##  1 A Aron Misael  2012        2        0.301
##  2 A Mi           1984        1        0.
##  3 A N A          2012        1        0.
##  4 A Reum         1983        5        0.699
##  5 A Reum         1987        7        0.845
##  6 A Sang         1994        4        0.602
##  7 Aaaraon        2013        1        0.
##  8 Aadil          1992        1        0.
##  9 Aage Andres    1990        1        0.
## 10 Aage Carlos    1985        1        0.
## # ... with 3,749,123 more rows
```

Cualquier operación que funcione con vectores funciona con `mutate()`. También funciona se pueden modificar columnas si el nombre que asignamos ya está usado dentro de nuestro *data frame*.

4.4 Operaciones por grupos

Muchas veces van a necesitar calcular por grupos: la suma, media, varianza, etc. Por ejemplo, calcular el número total de personas con cada nombre. Podrían hacerlo de esta forma:

```
nombres_comunes %>%
  filter(nombre == "Luciano") %>%
  summarise(total = sum(cantidad))
```

```
## # A tibble: 1 x 1
##   total
##   <int>
## 1 13244
```

Y repetirlo cambiando el nombre para cada uno de los nombres. Por su puesto, esta forma de hacer las cosas es muy incómoda y propensa a errores. Hay una forma más fácil y es usando `group_by()`. Un ejemplo:

```
# No intenten hacerlo en sus computadoras
# Los datos tienen más de 3 millones de registros y va a tomar un tiempo
nombres_comunes %>%
  group_by(nombre) %>%
  summarise(total = sum(cantidad))
```

```
## # A tibble: 4,272 x 2
##    nombre           total
##    <chr>            <int>
##  1 Aaron              519
##  2 Aaron Benjamin     353
##  3 Abel               307
##  4 Abigail           2656
##  5 Abril             4163
##  6 Adolfo             688
##  7 Adrian            4186
##  8 Adrian Alberto    2009
##  9 Adrian Alejandro  3460
## 10 Adrian Eduardo     106
## # ... with 4,262 more rows
```

Como pueden ver estos datos distan bastante de estar limpios ya que hay muchos errores de entrada de datos, como nombres todo en mayúsculas, versiones del mismo nombre con tilde y sin tilde, etc. Para evitar todo ese "ruido", podríamos filtrar los nombres raros que son mayoría de las entradas.

Además de usar la función `summarise()` se puede usar la función `mutate` que va a hacer que queden la misma cantidad de casos. Por ejemplo, calcular el número acumulado de personas con el mismo nombre en a través de los años:

```
nombres_comunes %>%
  group_by(nombre) %>%
  arrange(anio, .by_group = TRUE) %>%
  mutate(acumulado = cumsum(cantidad))
```

```
## # A tibble: 33,675 x 4
## # Groups:   nombre [4,272]
##    nombre          anio cantidad acumulado
##    <chr>          <int>    <int>     <int>
##  1 Aaron           2012      152       152
##  2 Aaron           2013      167       319
##  3 Aaron           2014      200       519
##  4 Aaron Benjamin  2012      108       108
##  5 Aaron Benjamin  2013      120       228
##  6 Aaron Benjamin  2014      125       353
##  7 Abel            1982      102       102
```

```
##  8 Abel            1989      103       205
##  9 Abel            1990      102       307
## 10 Abigail         1991      132       132
## # ... with 33,665 more rows
```

Acá hay una función nueva, `arrange()`. Lo que hace. Esta función ordena de manera creciente (0-9, a-z) un conjunto de datos. Por ejemplo:

```
nombres_comunes %>%
  arrange(cantidad)
```

```
## # A tibble: 33,675 x 3
##    nombre              anio cantidad
##    <chr>              <int>    <int>
##  1 Adolfo              1981      101
##  2 Adrian Alberto      1990      101
##  3 Adrian Maximiliano  1988      101
##  4 Agustin Adrian      1998      101
##  5 Agustina Alejandra  1997      101
##  6 Ailin               1993      101
##  7 Alan Benjamin       2014      101
##  8 Alan Gabriel        1989      101
##  9 Alan Matias         1989      101
## 10 Alberto Martin      1987      101
## # ... with 33,665 more rows
```

Si queremos que sea decreciente (9-0, z-a), hay que agregar la función `desc()`. Por ejemplo:

```
nombres_comunes %>%
  arrange(desc(cantidad))
```

```
## # A tibble: 33,675 x 3
##    nombre       anio cantidad
##    <chr>       <int>    <int>
##  1 Maria Belen  1993     6946
##  2 Maria Belen  1992     6249
##  3 Maria Belen  1994     6098
##  4 Juan Pablo   1982     5561
##  5 Maria Belen  1991     5307
##  6 Maria Laura  1980     5144
##  7 Maria Belen  1995     5094
##  8 Benjamin     2013     4964
##  9 Maria Laura  1981     4747
## 10 Benjamin     2012     4726
## # ... with 33,665 more rows
```

También se pueden poner varios criterios para que ordene según ellos. Por ejemplo, por cantidad y luego por orden alfabético.

```
nombres_comunes %>%
  arrange(cantidad, nombre)
```

```
## # A tibble: 33,675 x 3
##    nombre              anio cantidad
##    <chr>              <int>    <int>
##  1 Adolfo              1981      101
##  2 Adrian Alberto      1990      101
##  3 Adrian Maximiliano  1988      101
```

```
##  4 Agustin Adrian      1998      101
##  5 Agustina Alejandra  1997      101
##  6 Ailin               1993      101
##  7 Alan Benjamin       2014      101
##  8 Alan Gabriel        1989      101
##  9 Alan Matias         1989      101
## 10 Alberto Martin      1987      101
## # ... with 33,665 more rows
```

Ejercicio 4.7 (Orden de totales)**.** Ordenen el resultado del total de nombres que calcularon en el ejercicio 4.3

4.5 Formato Ancho y Formato Largo

Los datos en general vienen en uno de dos formatos:

- Formato Ancho: cada fila se corresponde a varias observaciones, parte de la información está en el nombre de las columnas.
- Formato Largo: cada fila corresponde a una única observación.

Table 4.2: Ejemplo de formato ancho. Cada fila corresponde a un individuo denotado por el nombre y cada columna corresponde al puntaje obtenido en una prueba bajo distintos tiempos.

Name	50	100	150	200	250	300	350
Carla	1.2	1.8	2.2	2.3	3	2.5	1.8
Mace	1.5	1.1	1.9	2	3.6	3	2.5
Lea	1.7	1.6	2.3	2.7	2.6	2.2	2.6
Karen	1.3	1.7	1.9	2.2	3.2	1.5	1.9

La tabla anterior es un ejemplo de formato ancho. Es cómoda para leer para nosotros pero no es cómoda para trabajar para las computadoras. Por ejemplo, ¿Cómo hacen para indicar que columna es la variable independiente y cual es la de respuesta? No se puede porque la variable indepediente es ¡el nombre de la columna!

La forma para poder graficarlo, o analizarlo facilmente es poner esto datos en formato largo. Así quedará una columna para el nombre, otra con los tiempos y otra con el puntaje.

Table 4.3: Ejemplo de formato largo. Cada fila corresponde a un observación denotado por el nombre, el tiempo y el puntaje.

Name	time	score
Carla	50	1.2
Mace	50	1.5
Lea	50	1.7
Karen	50	1.3
Carla	100	1.8
Mace	100	1.1
Lea	100	1.6
Karen	100	1.7
Carla	150	2.2
Mace	150	1.9

Name	time	score
Lea	150	2.3
Karen	150	1.9
Carla	200	2.3
Mace	200	2
Lea	200	2.7
Karen	200	2.2
Carla	250	3
Mace	250	3.6
Lea	250	2.6
Karen	250	3.2
Carla	300	2.5
Mace	300	3
Lea	300	2.2
Karen	300	1.5
Carla	350	1.8
Mace	350	2.5
Lea	350	2.6
Karen	350	1.9

De esta forma, será fácil indicar que columna corresponde al eje de las ordenadas, cual al de las abscisas para hacer un gráfico o cual es la variable independiente y cual la dependiente en una regresión.

Hay una función que nos permite llevar los datos en formato ancho a formato largo: `gather()` (*recoger*). Tiene tres argumntos: `data` el objeto, `key` el nombre de la nueva columna que contendrá la identificación del dato, es decir los viejos nombres de columna, y `value` la nueva columna que contendrá los valores. Por ejemplo:

```
gather(data = race, key = time, value = score)
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
##    time score
## 1  Name Carla
## 2  Name  Mace
## 3  Name   Lea
## 4  Name Karen
## 5    50   1.2
## 6    50   1.5
## 7    50   1.7
## 8    50   1.3
## 9   100   1.8
## 10  100   1.1
## 11  100   1.6
## 12  100   1.7
## 13  150   2.2
## 14  150   1.9
## 15  150   2.3
## 16  150   1.9
## 17  200   2.3
## 18  200     2
## 19  200   2.7
## 20  200   2.2
```

```
## 21  250     3
## 22  250   3.6
## 23  250   2.6
## 24  250   3.2
## 25  300   2.5
## 26  300     3
## 27  300   2.2
## 28  300   1.5
## 29  350   1.8
## 30  350   2.5
## 31  350   2.6
## 32  350   1.9
```

¡Pero que ha pasado aquí! No es el mismo resultado que en la Tabla 4.3. Por defecto, la función `gather()` recoge todas las columnas del objeto. Para evitar que lo haga con todas hay que indicar con un signo menos.

```
race_largo <- gather(data = race, key = time, value = score, -Name)
race_largo
```

```
##     Name time score
## 1  Carla   50   1.2
## 2   Mace   50   1.5
## 3    Lea   50   1.7
## 4  Karen   50   1.3
## 5  Carla  100   1.8
## 6   Mace  100   1.1
## 7    Lea  100   1.6
## 8  Karen  100   1.7
## 9  Carla  150   2.2
## 10  Mace  150   1.9
## 11   Lea  150   2.3
## 12 Karen  150   1.9
## 13 Carla  200   2.3
## 14  Mace  200   2.0
## 15   Lea  200   2.7
## 16 Karen  200   2.2
## 17 Carla  250   3.0
## 18  Mace  250   3.6
## 19   Lea  250   2.6
## 20 Karen  250   3.2
## 21 Carla  300   2.5
## 22  Mace  300   3.0
## 23   Lea  300   2.2
## 24 Karen  300   1.5
## 25 Carla  350   1.8
## 26  Mace  350   2.5
## 27   Lea  350   2.6
## 28 Karen  350   1.9
```

También funciona indicar cuales son las columnas que debe recoger escribiendo su nombre.

```
gather(data = race, key = time, value = score, `50`, `100`, `150`, `200`, `250`,
       `300`, `350`)
```

```
##     Name time score
## 1  Carla   50   1.2
```

```
## 2   Mace   50   1.5
## 3    Lea   50   1.7
## 4  Karen   50   1.3
## 5  Carla  100   1.8
## 6   Mace  100   1.1
## 7    Lea  100   1.6
## 8  Karen  100   1.7
## 9  Carla  150   2.2
## 10  Mace  150   1.9
## 11   Lea  150   2.3
## 12 Karen  150   1.9
## 13 Carla  200   2.3
## 14  Mace  200   2.0
## 15   Lea  200   2.7
## 16 Karen  200   2.2
## 17 Carla  250   3.0
## 18  Mace  250   3.6
## 19   Lea  250   2.6
## 20 Karen  250   3.2
## 21 Carla  300   2.5
## 22  Mace  300   3.0
## 23   Lea  300   2.2
## 24 Karen  300   1.5
## 25 Carla  350   1.8
## 26  Mace  350   2.5
## 27   Lea  350   2.6
## 28 Karen  350   1.9
```

Claro que en este caso es mucho más largo hacerlo de esta forma y además hay que delimitar cada número con acentos graves “‘” porque no es un nombre válido en *R*. Los nombres válidos son aquellos que empiezan con letras o puntos y no contienen signos de operaciones matemáticas ni espacios dentro.

Para poner los datos en formato ancho está la función `spread()` (*expandir*). Los argumentos son los mismos que tiene `gather()`.

```
spread(data = race_largo, key = time, value = score)
```

```
##    Name 100 150 200 250 300 350  50
## 1 Carla 1.8 2.2 2.3 3.0 2.5 1.8 1.2
## 2 Karen 1.7 1.9 2.2 3.2 1.5 1.9 1.3
## 3   Lea 1.6 2.3 2.7 2.6 2.2 2.6 1.7
## 4  Mace 1.1 1.9 2.0 3.6 3.0 2.5 1.5
```

4.6 Por su cuenta

Lean los datos de

```
load(url("git.io/calidad-del-aire-2017.RData"))
```

Son datos de calidad del aire de la ciudad de Buenos Aires

1. ¿Qué columnas hay y cuantos datos encuentran?
2. ¿En que tipo de formato está? ¿Largo o ancho?
3. Cambien los datos a formato largo. *Pista*: Usen `gather` para llegar a `FECHA`, `HORA`, `Columna`, `Valor`, luego Usen la función `separate` para separar el lugar del tipo de variable. Finalmente, con spread

llevenlo a un formato más adecuado para trabajar.
4. Calculen el promedio por lugar para las distintas variables. *Pista*: el formato más largo que crearon como paso intermedio arriba hace que el trabajo sea más corto.
5. Ordenen los lugares por la contaminación con material particulado.
6. Grafiquen cada uno de los datos de cada lugar, por fecha. Usen facetas para cada variable y un color distinto para cada lugar.
7. Seleccionen los datos de parque Centenario.
8. Grafiquen los datos de parque Centenario.

Chapter 5

ANOVA

5.1 Algunos conceptos importantes

Factor: Un factor es una variable independiente a ser estudiada en una investigación. Ejemplo: Temperatura, Dieta

Nivel: El nivel de un factor es una forma particular de ese factor. Ejemplo: Temperatura: 0ºC, 10ºC y 20ºC. Dieta: Con aditivos proteicos y Sin aditivos proteicos

Estudios Uní y Multifactoriales: Estudios de un factor, únicamente un factor es de interés. En estudios Multifactoriales, dos o más factores son investigados simultáneamente. Ejemplo: Un factor: cantidades de suplemento de proteínas de una clase determinada. Más de un factor: cantidades y clases de suplementos de proteínas.

Factores Experimentales y de Clasificación: En cualquier investigación basada sobre datos observacionales, los factores bajo estudio son factores de clasificación. Un factor de clasificación corresponde a la característica de las unidades bajo estudio y no las que están bajo control del investigador, no pueden ser manipuladas experimentalmente. Por otro lado, un factor experimental es aquel donde los niveles del factor son asignados al azar a las unidades experimentales.

Factores cualitativos y cuantitativos: Un factor cualitativo es aquel donde los niveles difieren con respecto a un atributo cualitativo. Por otro lado, un factor cuantitativo es aquel que es descrito por una cantidad numérica sobre una escala.

Tratamientos: Es el procedimiento cuyo efecto se mide y se compara con otros tratamientos. En estudios unifactoriales un tratamiento corresponde a un nivel de un factor. En estudios Multifactoriales, un tratamiento corresponde a una combinación de niveles de factores.

5.2 Diseño de Estudios de ANOVA

Elección del tratamiento: La elección de los tratamientos a ser incluidos en una investigación es básicamente una decisión del investigador. En una investigación científica, los tratamientos incluidos deberían poder suministrar conocimientos sobre el mecanismo subyacente al fenómeno bajo estudio.

Definición del tratamiento: Al seleccionar un conjunto de tratamientos, es importante definir cada tratamiento cuidadosamente y considerarlo con respecto a cada uno de los demás tratamientos para asegurarse, en lo posible, que el conjunto dé respuestas eficientes relacionadas con los objetivos del experimento.

Tratamiento Control o Testigo: Un tratamiento control consiste en la aplicación de procedimientos idénticos a las unidades experimentales que aquellos usados con los otros tratamientos, excepto por los efectos bajo investigación.

Un tratamiento control es requerido cuando la efectividad general de los tratamientos bajo estudio no es conocida, o cuando la efectividad general de los tratamientos es conocida pero no es consistente bajo todas las condiciones.

Unidad experimental o unidad básica de estudio: Es la unidad de material a la cual se aplica un tratamiento. Es la mínima unidad de muestreo. No siempre coincide con la unidad de muestreo. Ejemplo: se aplica un tratamiento en una maceta y se muestrean tres hojas de cada maceta.

Observación individual: Son las mediciones que se hacen en cada una de las unidades experimentales.

Muestra: Es el conjunto de observaciones individuales, se expresa en términos de observaciones individuales y no de unidades experimentales, es la única información que uno posee.

5.3 Planificación De Experimentos

Las inferencias que pueden hacerse, a partir de los resultados de un experimento, dependen de la forma en que fue hecho el experimento. Es una buena práctica hacer un proyecto de los propósitos de cualquier experimento. Este proyecto constará de tres partes:

Enumeración de las finalidades: debe incluir una determinación del campo sobre el cual se harán las generalizaciones, o, en otras palabras, la población respecto de la cual se espera hacer inferencias.

descripción del experimento: Se ha usado el término tratamiento para denominar los diferentes procesos cuyos efectos van a ser medidos y comparados. En la selección de los tratamientos es importante definir claramente cada uno de ellos y entender el papel que jugará para alcanzar los objetivos del experimento.

bosquejo del método de análisis de los resultados: Las características del experimento que deben ser tenidas en cuenta en la enumeración de finalidades son: el número de repeticiones, los tipos de material experimental que se van a usar, las mediciones que se van a hacer. Finalmente, el bosquejo debería describir, con algún detalle, el método propuesto para sacar conclusiones de los resultados.

5.4 Usos Del ANOVA

Los estudios de un solo factor son utilizados para comparar efectos de diferentes niveles de un factor, para determinar el “mejor” nivel del factor y la semejanza. En estudios multifactoriales, el ANOVA es empleado para determinar si los diferentes factores interactúan, que factores son claves; cuales combinaciones de factores son las “mejores”, etc.

5.5 MODELO I DE ANOVA. NIVELES DEL FACTOR FIJOS

5.5.1 Distinción Entre Modelos I Y II de ANOVA

El modelo I de ANOVA se aplica en casos tales como una comparación de un número determinado de tratamientos, y donde las conclusiones se restringen a aquellos niveles del factor incluidos en el estudio. También se conoce como modelo de efectos fijos. El modelo II de ANOVA se aplica a un tipo diferente de situación, donde las conclusiones se extenderán a una población de niveles del factor del cual los niveles bajo estudio son una muestra. Es decir que se trata de un modelo de efectos aleatorios.

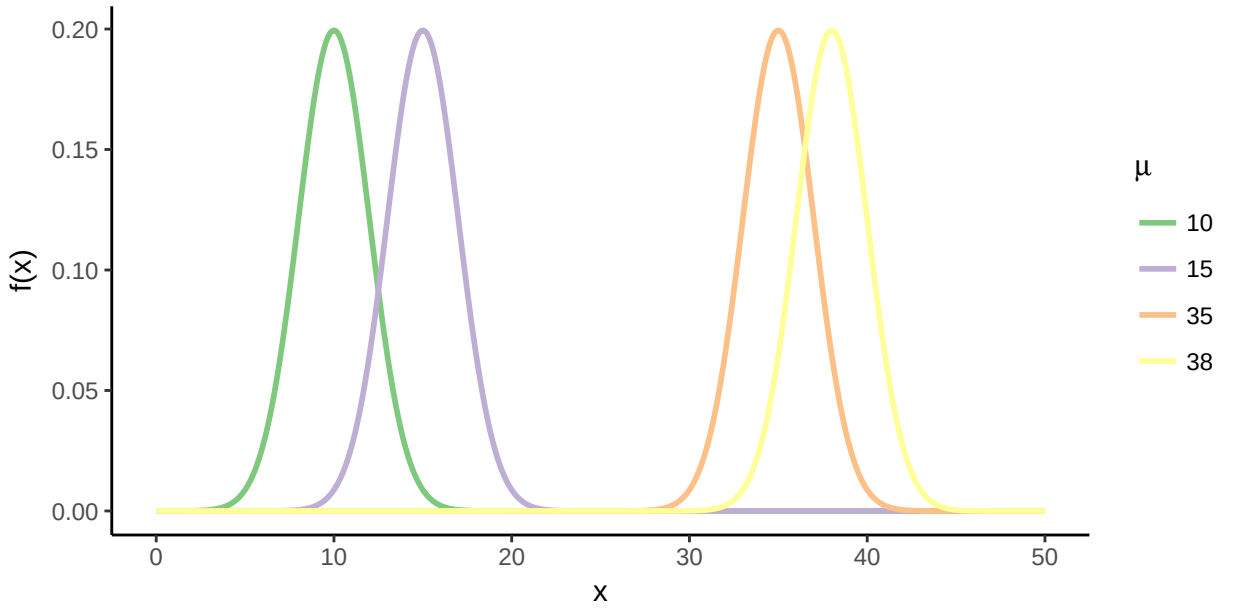


Figure 5.1: Densidad de distribuciones de cuatro distribuciones normales con igual varianza y distinta media

5.5.2 Ideas Básicas

Los elementos básicos del modelo I de ANOVA para un estudio de un factor son muy simples. Correspondiendo a cada nivel del factor, hay una distribución de probabilidades de respuestas. El modelo I de ANOVA supone:

1. Cada una de las distribuciones en probabilidades es **normal**

2. Cada distribución en probabilidad tiene la **misma varianza** (desviación estándar).

3. Las observaciones para cada nivel del factor son observaciones **aleatorias** de la correspondiente distribución y son **independientes** de las observaciones de cualquier otro nivel del factor.

La Figura 5.1 ilustra estas condiciones: la normalidad de la distribución en probabilidades y la variabilidad constante. Las distribuciones en probabilidad difieren sólo con respecto a sus medias. El análisis de los datos de las muestras de las distribuciones en probabilidades de los niveles de los factores se desarrolla usualmente en dos pasos:

Determinar si las medias de los niveles de los factores son las mismas.

Si las medias de los niveles del factor no son las mismas, examinar como difieren y cuales son las consecuencias de las diferencias.

5.6 Comprobación de los Supuestos

Los modelos de ANOVA son razonablemente robustos, aunque se produzcan ciertos alejamientos del supuesto de normalidad.

5.6.1 Prueba para igualdad de varianzas

5.6.1.1 Prueba de Bartlett

Las hipótesis son

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_i^2 = \ldots = \sigma_I^2$$
$$H_a : \text{no todos los } \sigma_i^2 \text{ son iguales}$$

Sean $S_1^2, \ldots, S_I^2$indican las varianzas muestrales de I poblaciones normales, y Gl_i indica los grados de libertad asociados con la varianza muestral S_i^2.

Bartlett ha demostrado que una función de $[\ln(C\text{MD}) - ln(MGD)]$, ($M$GD: media geométrica pesada; CMD: cuadrado medio dentro) para grandes tamaños muestrales, sigue aproximadamente la distribución χ^2 con $(I-1)$ grados de libertad cuando las varianzas poblacionales son iguales. La prueba estadística es:

$$B = \frac{GL_t}{C}\left[\ln(CM_D) - ln(M\text{MGD})\right]$$

, donde

$$C = 1 + \frac{1}{3(I-1)}\left[\left(\sum_{i=1}^{I}\frac{1}{GL_i}\right) - \frac{1}{GL_T}\right]$$

El término C es siempre mayor que 1.

La prueba estadística se reduce a:

$$B = \frac{1}{C}\left[(GL_t)\ln(CM_D) - \sum_{i=1}^{I}(GL_i)\, lnS_i^2\right]$$

se calcula el estadístico B. La regla de decisión es:

Si $B < \chi^2_{(1-\alpha;I-1)}$, no se rechaza H_0

Si $B > \chi^2_{(1-\alpha;I-1)}$, se rechaza H_0

Esta aproximación se considera apropiada cuando los grados de libertad son mayores o iguales que cuatro.

Cuando la prueba se usa para un modelo de ANOVA de un factor se tiene:

$GL_i = n_i - 1$ y $GL_T = \sum_{i=1}^{I}(n_i - 1) = N - I$

La prueba de Bartlett es bastante sensible a la falta de normalidad. Si las varianzas muestrales son menores que la unidad, sus logaritmos serán negativos. Por lo tanto, es conveniente utilizar un código multiplicativo para hacer las varianzas mayores que la unidad. Este código no afecta en modo alguno a la prueba estadística.

5.6.1.2 Prueba de Levene Modificada

Las hipótesis son

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_i^2 = \ldots = \sigma_I^2$$
$$H_a : \text{no todos los } \sigma_i^2 \text{ son iguales}$$

Primero se calcula la desviación absoluta de las Y_ij observaciones de sus respectivas medianas del nivel del factor $\tilde{Y}_i$

$$d_{ij} = \left|Y_{ij} - \tilde{Y}_i\right|$$

Entonces la prueba de Levene determina si los valores esperados de las desviaciones absolutas son iguales. Si las varianzas son iguales entonces los valores esperados de las desviaciones absolutas también serán iguales. La prueba de Levene usa el estadístico F^*

$$F_L^* = \frac{CM_{ET}}{CM_D}$$

donde

$$CM_{ET} = \frac{\sum n_i \left(\overline{d}_{i\bullet} - \overline{d}_{\bullet\bullet}\right)^2}{I-1}$$

$$CM_D = \frac{\sum\sum \left(\overline{d}_{ij} - \overline{d}_{i\bullet}\right)^2}{N-1}$$

$$\overline{d}_{i\bullet} = \sum_j \frac{d_{ij}}{n_i}$$

$$\overline{d}_{..} = \frac{\sum\sum d_{ij}}{N}$$

Si las varianzas son iguales y los tamaños muestrales no son extremadamente pequeños, $\mathbf{F_L^*}$ sigue aproximadamente una distribución F con $(I-1)$ y $(N-I)$ grados de libertad.

5.6.2 Prueba de Kolmogorov - Smirnov (modificación de Lilliefors) para estudiar Normalidad

Dada una muestra aleatoria, se calcula su media y su varianza muestral, luego se calculan los datos normalizados Z_i. Se ordenan los datos de menor a mayor, se calculan las frecuencias acumuladas observadas, las esperadas para los Z_i, y luego se calculan las diferencias, en valor absoluto entre las frecuencias acumuladas observadas y las esperadas. Se define $D_{max} = max\left|F_i - \hat{F}_i\right|$ este estadístico se compara con el valor de tablas d_{max} al nivel de significación α.

5.6.3 Residuos

El residuo ε_{ij} es definido como la diferencia entre el valor observado y el ajustado:

$$\varepsilon_{ij} = y_{ij} - \overline{y}_{i\bullet}$$

Así, un residuo representa la desviación de una observación individual de la respectiva media estimada del nivel del factor. A veces es útil trabajar con los residuos estandarizados, que se expresan como:

$$\varepsilon^{\otimes}_{ij} = \frac{\varepsilon_{ij} - \overline{\varepsilon}}{\sqrt{CM_D}}$$

Los residuos "*semistudentizados*", los residuos "*studentizados*", y los residuos "*studentizados borrados*" son a menudo útiles para diagnosticar los alejamientos del modelo de ANOVA.

Los residuos "*semistudentizados*" se calculan como:

$$\varepsilon^*_{ij} = \frac{\varepsilon_{ij}}{\sqrt{CM_D}}$$

Los residuos "*studentizados*" se calculan como:

$$r_{ij} = \frac{\varepsilon_{ij}}{S\left(\varepsilon_{ij}\right)}$$

donde

$$S\left(\varepsilon_{ij}\right) = \sqrt{\frac{CM_D\left(n_i - 1\right)}{n_i}}$$

Finalmente, los residuos "*studentizados borrados*" se calculan

$$t_{ij} = \varepsilon_{ij}\left[\frac{N - I - 1}{SC_D\left(1 - \frac{1}{n_i}\right)\varepsilon^2_{ij}}\right]^{\frac{1}{2}}$$

5.6.4 Gráficos de Residuos

Estos gráficos son muy importantes para el diagnóstico de problemas con el modelo. Incluye:

1. *Residuos vs. las medias de tratamientos*: Dado que las valores ajustados de cada nivel de factor se corresponde a la media, todos los valores de los residuales de ese nivel se alinearán en sobre esa media (Figura 5.2-a). Si no hay problemas con el modelo, entonces los residuales deberían tener la misma dispersión.

2.Residuos vs. el tiempo u otra secuencia: Si los datos fueron tomados de forma aleatoria no debería verse un patrón definido (Figura 5.2-b).

3. Gráficos de puntos de los residuos: Este gráfico es similar al primero. Denuevo, en todos lo niveles del factor los residuales deberían tener la misma dispersión alrededor del cero(Figura 5.2-c). Además, dado que están graficados los residuales estandarizados, cualquier residual mayor 3 debería ser investigado por ser muy extremo.

4. Gráficos de probabilidad normal de los residuos: También llamado gráfico cuantil-cuantil o *qqplot* (Figura 5.2-d). Aquí se grafican los cuantiles de una normal teórica vs los cuantiles muestrales. Idealmente, deberían seguir una línea recta de pendiente 1 y ordenada 0.

5.6.4.1 Diagnóstico de los alejamientos de los supuestos del Modelo de ANOVA

Heterogeneidad de Varianzas: El Modelo de ANOVA requiere que los términos del error tengan varianzas constantes para todos los niveles del factor. Cuando los tamaños de las muestras son iguales o no difieren mucho, esta suposición puede ser estudiada usando los residuos, los residuos "studentizados" o los residuos "semistudentizados". Gráficos de los residuos vs. las medias de los niveles del factor o los gráficos de puntos de los residuos son útiles. Cuando los tamaños de las muestras difieren mucho, los residuos "studentizados" deberían ser usados en estos gráficos. La constancia de la varianza del error se ve en estos gráficos pues los puntos tienen aproximadamente la misma dispersión alrededor del cero para cada nivel.

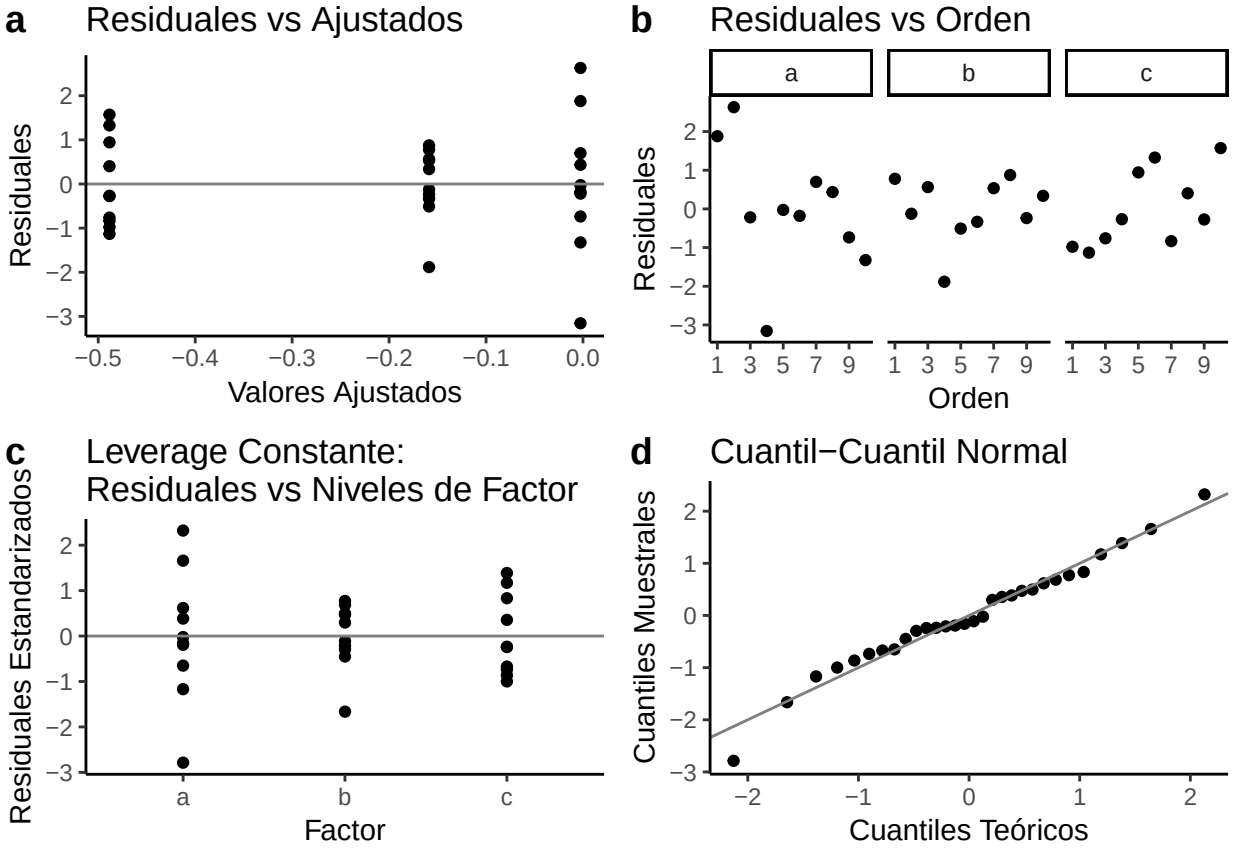


Figure 5.2: Gráficos de residuales para modelos de ANOVA. a - residuales vs valores predichos. b - residuales vs orden de toma de datos. c - residuales vs niveles del factor. d - gráfico de probabilidad normal.

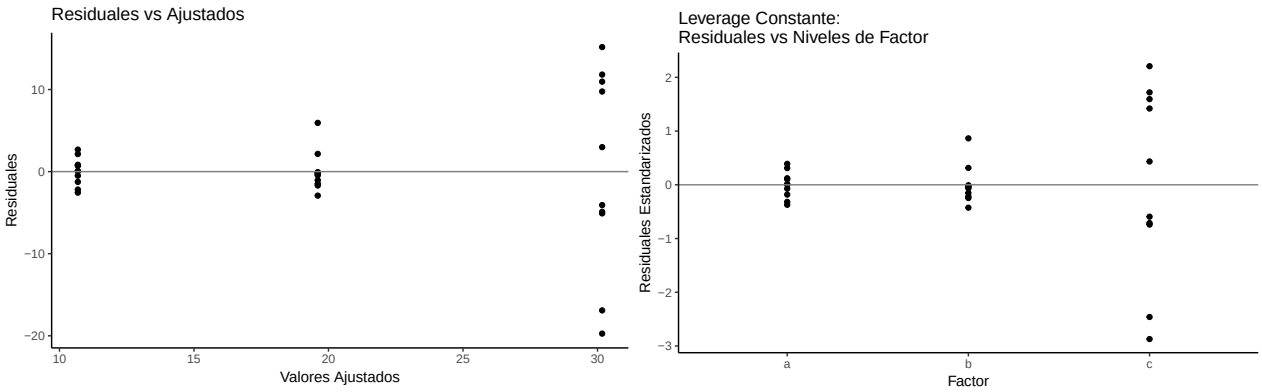


Figure 5.3: Residuales vs Valores Ajustados o predichos. Este gráfico muestra que los residuales uno de los niveles muestra mayor dispersión que el resto de los datos.

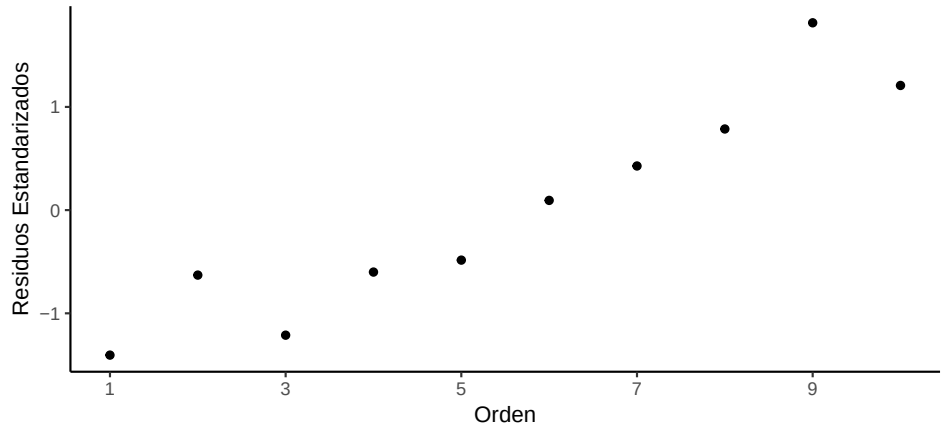


Figure 5.4: Residuales vs Orden. Los residuales muestran falta de independencia al haber una correlación entre ellos.

La Figura 5.3 muestra un caso en el que las varianzas de los errores no son constantes. En este caso los términos del error del nivel c del factor tienen una varianza mayor que los otros dos niveles del factor.

Cuando los tamaños de las muestras, para los diferentes niveles del factor son grandes, los histogramas de los residuos para cada tratamiento, son una manera efectiva de examinar la constancia de la varianza de los términos del error.

Falta de independencia en los términos del error: En todos aquellos casos en que los datos son obtenidos en una secuencia de tiempo, un gráfico de secuencia de residuos es aconsejable para examinar si los términos del error están correlacionados. La Figura 5.4 muestra un caso en el cual los residuos aparecen altamente correlacionados. Esto puede pasar porque el operario tiende a sobreestimar a medida que pasa el tiempo o también porque los equipos se descalibran.

La siguiente Figura 5.5 muestra un caso donde la varianza decrece con el tiempo.

Cuando los datos son ordenados en alguna a otra secuencia lógica, tal como una secuencia geográfica, también debe verificarse si existe correlación entre los términos del error de acuerdo a este orden.

Otros usos del análisis de residuos: Este tipo de análisis se puede usar para detectar "outliers". También es útil para determinar si modelo de ANOVA de un factor es el adecuado; pues puede determinar la omisión de alguna variable importante que explica las observaciones. También puede ser usado para determinar la falta de normalidad de los términos del error. Esto se realiza graficando los cuantiles de los residuales observados vs los esperados.

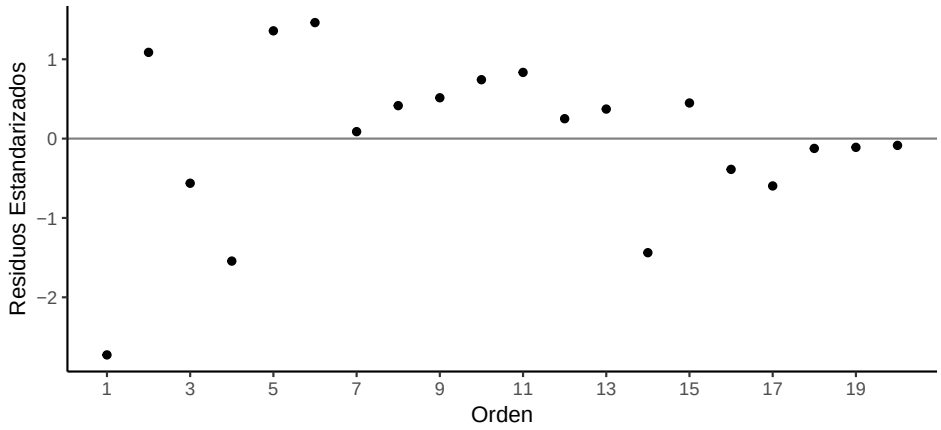


Figure 5.5: Residuales vs Orden. Los residuales muestran que la varianza decrece, ya que al principio son mayores y al final son menores

5.7 Transformaciones

5.7.1 Transformaciones para estabilizar las Varianzas

Varianza *proporcional a* μ_i : El estadístico muestral $\mathbf{S_i^2/\overline{Y}_i}$ tenderá a ser constante. Este tipo de situaciones a menudo se encuentra cuando la variable observada es un número entero. Para estos casos, una transformación raíz cuadrada es útil para estabilizar la varianza:

$Y^{'} = \sqrt{Y}$ o $Y^{'} = \sqrt{Y + \frac{1}{2}}$ o$Y^{'} = \sqrt{Y} + \sqrt{Y + 1}$

Desviación estándar proporcional a μ_i: $\mathbf{S_i/\overline{Y}_i}$tiende a ser contante para los diferentes niveles del factor. Una transformación útil para estabilizar la varianza es la transformación logarítmica:

$Y^{'} = \log Y\, o Y^{'} = \log(Y + 1)$

Desviación estándar proporcional a μ_i^2: $\mathbf{S_i^2/\overline{Y}_i^2}$En este caso tiende a ser constante. la transformación apropiada es la recíproca:

$$Y^{'} = \frac{1}{Y}$$

La variable dependiente es una proporción: Una transformación apropiada para este caso es la transformación angular o arcoseno:

$$Y^{'} = \text{arcsen}\sqrt{Y}$$

5.7.2 Transformaciones para corregir la falta de normalidad

La transformación que ayuda a corregir la heterogeneidad de varianzas usualmente también es efectiva para hacer que las distribuciones de los términos del error sean más normales.

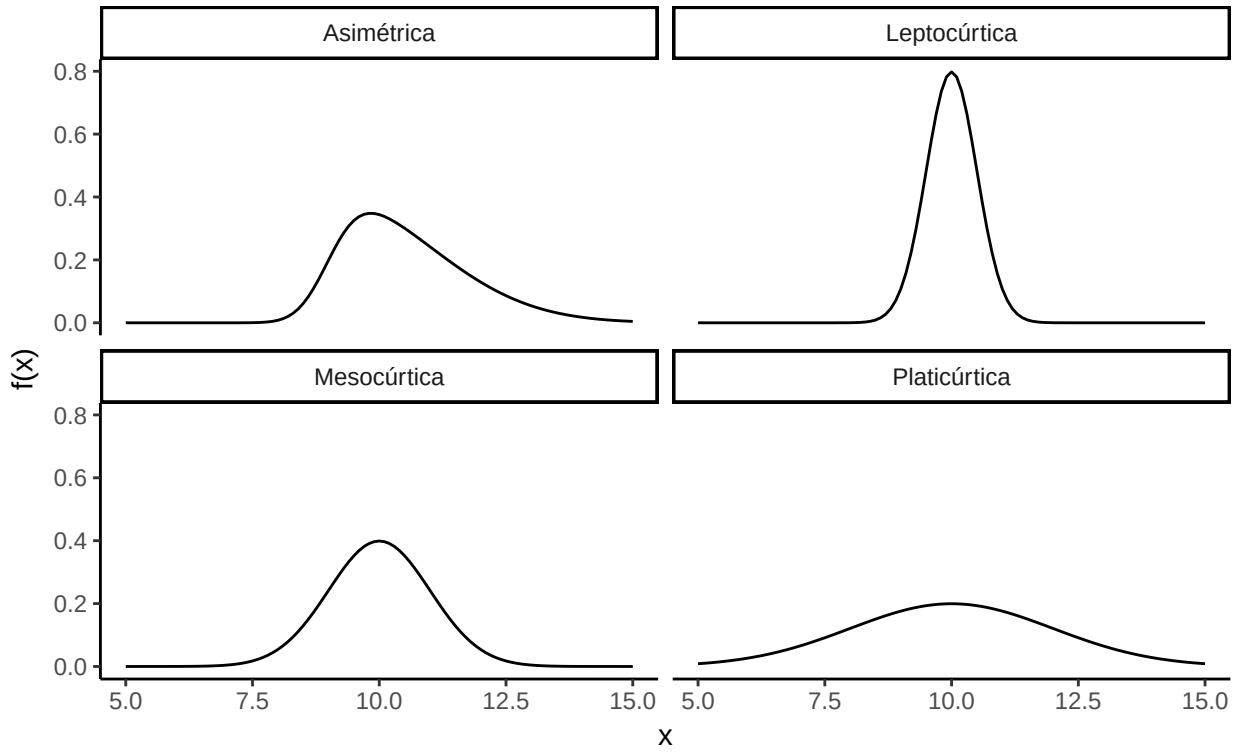


Figure 5.6: Funciones de densidad para curvas asimétrica, mesocúrtica, leptocúrtica, platicúrtica

5.7.3 Efectos Del Alejamiento De Los Supuestos Del Modelo

5.7.3.1 Normalidad

Para el modelo I de ANOVA, la falta de normalidad no es importante, en tanto ese alejamiento no sea extremo. La kurtosis es más importante que la asimetría en términos de efectos sobre las inferencias (Figura 5.6).

La prueba F es poco afectada por la falta de normalidad, ya sea en términos del nivel de significación o de la potencia de la prueba. Para el Modelo II de ANOVA, la falta de normalidad tiene serias implicaciones.

5.7.3.2 Heterogeneidad de varianzas

Para el modelo de efectos fijos, la prueba de F es ligeramente afectada si los tamaños muestrales son iguales o no difieren mucho. La prueba de F y los análisis relacionados son robustos frente a la heterogeneidad de varianzas cuando los tamaños muestrales son aproximadamente iguales.

Para el modelo de efectos aleatorios, la heterogeneidad de varianzas puede tener efectos pronunciados sobre las inferencias acerca de los componentes de la varianza, aun con tamaños muestrales iguales.

5.7.3.3 Independencia de los términos del error

La falta de independencia puede tener serios efectos sobre las inferencias en el análisis de la varianza, para el modelo de efectos fijos y para el de efectos aleatorios.

5.8 Formulación Del Modelo I De ANOVA.

Denotaremos por I el número de niveles del factor bajo estudio, y denotaremos cualquiera de estos niveles por el subíndice i ($i = 1, 2, \ldots, I$). El número de casos para el i-ésimo nivel del factor es simbolizado por n_i, y el número total de casos en el estudio es denotado por N, donde:

$$N = \sum_{i=1}^{I} n_i$$

Además, Y_{ij} denotará la j-ésima observación para el i-ésimo nivel del factor. Dado que el número de casos para el i-ésimo nivel del factor es denotado por n_i, tendremos $j = 1, 2, \ldots, n_i$.

El modelo I de ANOVA se puede plantear como sigue:

$$y_{ij} = u_i + \varepsilon_{ij}$$

donde:

- y_{ij} es el valor de la j-ésima observación para el i-ésimo nivel del factor o tratamiento.
- μ_i es un parámetro
- ε_{ij} son variables independientes $N(0, \sigma^2)$
- $i = 1, 2, \ldots, I; j = 1, 2, \ldots, n_i$

5.8.1 Características importantes del modelo

El valor observado de Y en el j-ésimo ensayo del i-ésimo nivel del factor o tratamiento es la suma de dos componentes: a) un término constante μ_i, y b) un término del error aleatorio ε_{ij}.

Dado que $E(\varepsilon_{ij}) = 0$, se sigue que:

$$E(Y_{ij}) = \mu_i$$

Dado que μ_i es una constante, se sigue que:

$$Var(Y_{ij}) = Var(\varepsilon_{ij}) = \sigma^2$$

Así como cada ε_{ij} esta normalmente distribuido, también lo está cada Y_{ij}.

Se asume que los términos del error son independientes

El modelo de ANOVA puede ser re-enunciado como:

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

5.8.2 Interpretación De Las Medias De Los Niveles Del Factor

Datos observacionales: la media del nivel del factor μ_i corresponde a las medias para las diferentes poblaciones del nivel del factor.

Datos Experimentales: la media del nivel del factor μ_i representa la media de la respuesta que debería obtenerse si el i-ésimo tratamiento fuera aplicado a todas las unidades en la población de las unidades experimentales sobre las cuales se harán las inferencias.

5.8.3 Ajustando El Modelo

Supongamos que tenemos I tratamientos o niveles de un factor y que aplicamos cada uno de ellos a un grupo de unidades experimentales.Los datos se podrían consignar de la siguiente forma:

Tratamientos	T_1	T_2	...	T_i	...	T_I	
	$y_{1\mathbf{1}}$	$y_{2\mathbf{1}}$	$\cdots$	y_{i1}	$\cdots$	y_{I1}	
	$y_{1\mathbf{2}}$	$y_{2\mathbf{2}}$	$\cdots$	y_{i2}	$\cdots$	y_{I2}	
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	...	$\vdots$	
	y_{1j}	y_{2j}	$\cdots$	$y_{i\mathbf{j}}$	$\cdots$	$y_{I\mathbf{j}}$	
$\sum_{j=\mathbf{1}}^{n_i} y_{i\mathbf{j}}$	$\sum_{j=\mathbf{1}}^{n_1} y_{1j}$	$\sum_{j=\mathbf{1}}^{n_2} y_{2j}$	$\cdots$	$\sum_{j=\mathbf{1}}^{n_i} y_{i\mathbf{j}}$	$\cdots$	$\sum_{j=\mathbf{1}}^{n_I} y_{I\mathbf{j}}$	$\sum_{j=\mathbf{1}}^{I} y_{i\mathbf{j}}$
	n_1	n_2	...	n_i	...	n_I	$\sum_{i=\mathbf{1}}^{I} n_i = N$
$\overline{y_{i\bullet}}$	$\overline{y_{1.}}$	$\overline{y_{2.}}$	$\cdots$	$\overline{y_{i\bullet}}$	$\cdots$	$\overline{y_{I.}}$	$\frac{\sum_{i\mathbf{j}} y_{i\mathbf{j}}}{N} = \overline{y_{\bullet\bullet}}$
	S_1^2	S_2^2	...	S_i^2	...	S_I^2	

donde

T_i es el tratamiento o nivel del factor i; con $i = 1, 2, \ldots, I$

y_{ij} es la observación sobre la unidad experimental j con el tratamiento i; $ j = 1,2, ..., n_{i}$.

N tamaño de la muestra

$\overline{y}_{i\bullet}$ es la media muestral de cada tratamiento.

n_i es el número de observaciones con el tratamiento i

$\overline{y}_{\bullet\bullet}$ es la media total, para todas las observaciones.

S_i^2 es la varianza muestral para el tratamiento i

5.8.4 Estimadores De Mínimos Cuadrados

De acuerdo al criterio de mínimos cuadrados la suma de los cuadrados de las desviaciones de las observaciones alrededor de sus valores esperados puede ser minimizada con respecto a los parámetros. Para un modelo de ANOVA, tenemos que:

$$E\left(Y_{ij}\right) = \mu_i$$

Así, la cantidad a ser minimizada es:

$$\sum_i \sum_j \left(y_{ij} - \mu_i\right)^2$$

Esta expresión se puede escribir como:

$$\sum_j \left(y_{1j} - \mu_1\right)^2 + \sum_j \left(y_{2j} - \mu_2\right)^2 + \ldots + \sum_j \left(y_{Ij} - \mu_I\right)^2$$

La media muestral minimiza una suma de desviaciones al cuadrado

$$\hat{\mu_i} = \overline{y}_{i\bullet} \tag{5.1}$$

5.8.4.1 Comentarios

1. Los estimadores de mínimos cuadrados (5.1) son también estimadores de máxima verosimilitud para el error normal (ε_{ij}) del modelo de ANOVA.
2. Para derivar el estimador de mínimo cuadrados de u_i , es necesario minimizar, con respecto a u_i, el i-ésimo componente de la suma de cuadrados en:

$$\sum_j (y_{ij} - \mu_i)^2$$

Diferenciando con respecto a μ_i, se obtiene:

$$\frac{\partial \sum_j (y_{ij} - \mu_i)^2}{\partial \mu_i} = \sum -2 (y_{ij} - \mu_i)$$

Esta derivada se iguala a cero y se reemplaza el parámetro μ_i por su estimador:

$$-2 \sum_{j=1}^{n_i} (y_{ij} - \mu_i) = 0$$

$$\sum_{j=1}^{n_i} y_{ij} = n_i \hat{\mu}_i$$

$$\hat{\mu}_i = \overline{Y}_{i\bullet}$$

5.9 Partición De La Suma De Cuadrados Total

La variabilidad total de las observaciones y_{ij}, sin usar la información sobre los niveles del factor, es medida en términos de la desviación de cada observación y_{ij} alrededor de la media total $\overline{y}_{\bullet\bullet}$:

$$y_{ij} - \overline{y}_{\bullet\bullet}$$

Cuando se utiliza la información sobre los niveles del factor, las desviaciones son aquellas de cada observación y_{ij} alrededor de su respectiva media estimada $\overline{y}_{\mathrm{i}\bullet}$:

$$y_{ij} - \overline{y}_{\mathrm{i}\bullet}$$

La diferencia entre la desviación total y la desviación anterior refleja la diferencia entre la media estimada del nivel del factor y la media total:

$$(y_{ij} - \overline{y}_{\bullet\bullet}) - (y_{ij} - \overline{y}_i) = \overline{y}_{\mathrm{i}\bullet} - \overline{y}_{\bullet\bullet}$$

Así, la desviación total $y_{ij} - \overline{y}_{\bullet\bullet}$ puede ser vista como la suma de dos componentes:

La desviación de la media estimada del nivel del factor alrededor de la media total.

La desviación de y_{ij} alrededor de la media de su nivel del factor. Esta desviación es simplemente el residuo ε_{ij} .

Elevando al cuadrado se obtiene:

$$\sum_i \sum_j (y_{ij} - \overline{y}_{\bullet\bullet})^2 = \sum_i n_i (\overline{y}_{\mathrm{i}\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum_i \sum_j (y_{ij} - \overline{y}_{\mathrm{i}\bullet})^2$$

El primer miembro de igualdad representa la variabilidad total de las y_{ij} observaciones y es denotado como la *suma de cuadrados total (SCT)*:

$$SC_T = \sum_i \sum_j (y_{ij} - \overline{y}_{\bullet\bullet})^2$$

El primer término del segundo miembro de la igualdad será indicado como SCE, la *suma de cuadrados entre* tratamientos:

$$SC_E = \sum_i n_i (y_{\mathrm{i}\bullet} - \overline{y}_{\bullet\bullet})^2$$

El segundo término se indica como SCD, la *suma de cuadrados dentro* de tratamientos o la *suma de cuadrados del error.*

$$SC_D = \sum_i \sum_j (y_{ij} - \overline{y}_{\mathrm{i}\bullet})^2 = \sum_i \sum_j \varepsilon_{ij}^2$$

Así, podemos escribir:

$$SCT \;=\; SCE \;+\; SCD$$

La suma total de los cuadrados para el modelo de análisis de la varianza se compone en consecuencia de dos partes.

5.9.1 Fórmulas computatorias

$$\begin{gathered} SC_T = \textstyle\sum_i \sum_j y_{ij}^2 - N\overline{y}_{\bullet\bullet}^2 \\ SC_E = \textstyle\sum_i n_i y_{\mathrm{i}\bullet}^2 - N\overline{y}_{\bullet\bullet}^2 \\ SC_D = \textstyle\sum_i \sum_j y_{ij}^2 - n_i \overline{y}_{\mathrm{i}\bullet}^2 = SC_T - SC_E \end{gathered}$$

5.10 Grados De Libertad

Correspondiendo a la descomposición de la suma de cuadrados total, se puede obtener los grados de libertad asociados.

La SCT tiene $(N - 1)$ grados de libertad asociados. Hay en conjunto N desviaciones $Y_{ij} - \overline{Y}_{\bullet\bullet}$, pero un grado de libertad se pierde debido a que las desviaciones no son independientes a causa de que la suma de ellas debe ser cero.

$\sum_i \sum_j (y_{ij} - \overline{y}_{\bullet\bullet}) = 0$

La SCE (entre tratamientos) tiene $(I - 1)$ grados de libertad asociados. Hay I desviaciones de las medias de los niveles de los factores $\overline{Y}_{\mathrm{i}\bullet} - \overline{Y}_{\bullet\bullet}$, pero un grado de libertad se pierde porque las desviaciones no son independientes a causa de que la suma pesada debe ser cero. $\sum_i n_i (y_{\mathrm{i}\bullet} - \overline{y}_{\bullet\bullet}) = 0$.

La SCD tiene $(N - I)$ grados de libertad asociados. Esto puede verse considerando el componente de la SCD para el i-ésimo nivel del factor:

$$\sum_j (y_{ij} - \overline{y}_{i\bullet})^2$$

La expresión es equivalente a la suma de cuadrados total considerando sólo el i-ésimo nivel del factor. Así, hay $n_i - 1$ grados de libertad asociados con esta suma de cuadrados. De esta forma la SCD es una suma de sumas de cuadrados, los grados de libertad asociados son la suma de los grados de libertad de sus términos:

$$(n_1 - 1) + (n_2 - 1) + \ldots + (n_I - 1) = N - I$$

Los grados de libertad, al igual que la suma de cuadrados, son aditivos.

5.11 Cuadrados Medios

Los cuadrados medios se obtienen dividiendo la suma de cuadrados por sus grados de libertad asociados. Se tiene:

$$CM_E = \frac{SC_E}{I-1}$$

$$CM_D = \frac{SC_D}{N-I}$$

CM_E, es el cuadrado medio entre los tratamientos.

CM_D, es el cuadrado medio dentro de los tratamientos o del error.

5.11.1 Esperanza de los Cuadrados Medios

Los valores esperados del CM_D y CM_E pueden ser vistos como:

$$E(CM_D) = \sigma^2$$

$$E(CM_E) = \sigma^2 + \frac{\sum n_i (\mu_i - \mu_\bullet)^2}{I-1} (\#eq : eCM_E) \tag{5.2}$$

donde

$$\mu_\bullet = \frac{\sum n_i \mu_i}{N}$$

1. El CM_D es un estimador insesgado de la varianza del error llamado ε_{ij}, tanto si las medias u_i son iguales como si no.
2. Cuando todas las medias μ_i de los niveles del factor son iguales y por lo tanto iguales a la media pesada $\mu_\bullet$, entonces $E(CM_E) = \sigma^2$ dado que el segundo término se vuelve cero. Cuando las medias de los niveles del factor no son iguales, el CM_E tiende en promedio a ser mayor que el CM_D, dado que el segundo término de la Ecuación @ref(eq:eCM_E) será positivo. Esto es intuitivamente razonable, como se ilustra en la figura para cuatro tratamientos. En la situación planteada se asume que todos los tamaños muestrales son iguales, o sea $n_i = n$. Cuando todos los μ_i son iguales, entonces todos lo

$\overline{Y}_{i\bullet}$ siguen la misma distribución en el muestreo, con una mediaμ_c y una varianza σ^2/n. Cuando las μ_i no son iguales, por otro lado, las $\overline{Y}_{\mathrm{i}\bullet}$ siguen diferentes distribuciones en el muestreo, cada una con la misma variabilidad σ^2/n pero centradas sobre medias diferentes μ_i (Figura 5.1).

En consecuencia, los $\overline{Y}_{\mathrm{i}\bullet}$ tenderán a diferir unos de otros tanto si los μ_i difieren como si son iguales, y en consecuencia la SCE tenderá a ser mayor cuando las medias de los niveles de los factores no son las mismas que cuando ellas son iguales. Esta propiedad de la SCE es utilizada en la construcción de la prueba estadística para determinar si las medias de los niveles del factor son iguales o no. Si la SCE y la SCD son de la misma magnitud, esto sugiere que las medias μi de los niveles del factor son iguales. Si la SCE es substancialmente mayor que la SCD, esto sugeriría que los μ_i no son iguales.

5.11.2 Comentarios

1. Para encontrar el valor esperado del CM_D, se ve que puede ser expresado como sigue:

$$\begin{aligned} CM_D &= \tfrac{1}{N-I}\sum_i\sum_j\left(Y_{ij}-\overline{Y}_{\mathrm{i}\bullet}\right)^2 \\ &= \tfrac{1}{N-I}\sum_i\left[(n_i-1)\,\tfrac{\sum_j\left(Y_{ij}-\overline{Y}_{\mathrm{i}\bullet}\right)^2}{(n_i-1)}\right] \end{aligned}$$

Indicamos la varianza muestral de las observaciones para el i-ésimo nivel del factor como s_i^2:

$$s_i^2 = \frac{\sum_j\left(Y_{ij}-\overline{Y}_{\mathrm{i}\bullet}\right)^2}{n_i-1}$$

Por lo tanto, el CM_D puede ser expresado de la siguiente forma:

$$CM_D = \frac{1}{N-I}\sum_i\left(n_i-1\right)s_i^2$$

Dado que la varianza muestral es un estimador insesgado de la varianza poblacional, la cual es σ^2 para todos los niveles del factor, se obtiene:

$$\begin{aligned} E\left(CM_D\right) &= \frac{1}{N-I}\sum_i\left(n_i-1\right)E\left(s_i^2\right) \\ &= \frac{1}{N-I}\sum_i\left(n_i-1\right)\sigma^2 \\ &= \sigma^2 \end{aligned}$$

2. Se puede derivar el valor esperado de la CM_E para el caso especial en que todos los tamaños muestrales ni son los mismos, o sea $n_i = n$. El resultado general para este caso especial:

$$E\left(CM_E\right) = \sigma^2 + \frac{n\sum\left(\mu_i-\mu_\bullet\right)^2}{I-1} \text{ cuando } n_i = n$$

De esta forma, cuando todos los tamaños muestrales de los niveles del factor son n, el CM_E se vuelve:

$$CM_E = \frac{n\sum\left(\overline{Y}_{\mathrm{i}\bullet}-\overline{Y}_{\bullet\bullet}\right)^2}{I-1} \text{ cuando } n_i = n$$

Para derivar el $E(CM_D)$, se considera el modelo:

$$Y_{ij} \;=\; u_i \;+\; \varepsilon_{ij}$$

Promediando el Y_{ij} para el i-ésimo nivel del factor, se obtiene:

$$\overline{Y}_{i\bullet} = \mu_i + \overline{\varepsilon}_{i\bullet}$$

donde $\overline{\varepsilon}_{i\bullet}$ es el promedio de los ε_{ij} para el i-ésimo nivel del factor:

$$\overline{\varepsilon}_{\mathrm{i}\bullet} = \frac{\sum_j \varepsilon_{ij}}{n}$$

Promediando los Y_{ij} sobre todos los niveles del factor, se obtiene:

$$\overline{Y}_{\bullet\bullet} = \mu_\bullet + \overline{\varepsilon}_{\bullet\bullet}$$

donde $\mu_\bullet$ para $n_i = n$:

$$\mu_\bullet = \frac{n\sum \mu_i}{\mathrm{nI}} = \frac{\sum \mu_i}{I} \text{ donde } n_i = n$$

y $\overline{\varepsilon}_{\bullet\bullet}$ es el promedio de todos los ε_{ij} :

$$\overline{\varepsilon}_{\bullet\bullet} = \frac{\sum\sum \varepsilon_{ij}}{\mathrm{nI}}$$

Cuando los tamaños muestrales son iguales, se tiene:

$$\overline{Y}_{\bullet\bullet} = \frac{\sum Y_{\mathrm{i}\bullet}}{I} \quad \overline{\varepsilon}_{\bullet\bullet} = \frac{\sum \varepsilon_{\mathrm{i}\bullet}}{I}$$

Operando se obtiene:

$$\overline{Y}_{\mathrm{i}\bullet} - \overline{Y}_{\bullet\bullet} = (\mu_i + \overline{\varepsilon}_{\mathrm{i}\bullet}) - (\mu_\bullet + \overline{\varepsilon}_{\bullet\bullet}) = (\mu_i - \mu_\bullet) + (\overline{\varepsilon}_{\mathrm{i}\bullet} - \overline{\varepsilon}_{\bullet\bullet})$$

Elevando al cuadrado y sumando sobre los niveles del factor, se obtiene:

$$\sum \left(\overline{Y}_{\mathrm{i}\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 = \sum \left(\mu_i - \mu_\bullet\right)^2 + \sum \left(\overline{\varepsilon}_{\mathrm{i}\bullet} - \overline{\varepsilon}_{\bullet\bullet}\right)^2 + 2\sum \left(\mu_i - \mu_{\bullet\bullet}\right)\left(\overline{\varepsilon}_{\mathrm{i}\bullet} - \overline{\varepsilon}_{\bullet\bullet}\right)$$

Se desea encontrar el $E\left\{\sum \left(\overline{Y}_{\mathrm{i}\bullet} - \overline{Y}_{\bullet\bullet}\right)^2\right\}$, y por lo tanto se necesita encontrar el valor esperado de cada uno de los términos de la derecha:

3. Dado que $\sum \left(\mu_i - \mu_\bullet\right)^2$ es una constante, su valor esperado es:

$$E\left\{\sum \left(\mu_i - \mu_\bullet\right)^2\right\} = \sum \left(\mu_i - \mu_\bullet\right)^2$$

4. Antes de encontrar el valor esperado del segundo término de la derecha, consideremos la expresión:

$$\frac{\sum \left(\overline{\varepsilon}_{\mathrm{i}\bullet} - \overline{\varepsilon}_{\bullet\bullet}\right)^2}{I-1}$$

Esto es una varianza muestral, dado que $\overline{\varepsilon}_{\bullet\bullet}$ es la media muestral de los I términos $\overline{\varepsilon}_{\mathrm{i}\bullet}$. Se sabe que la varianza muestral es un estimador insesgado de la varianza de la variable, en este caso de $\overline{\varepsilon}_{\mathrm{i}\bullet}$. Pero $\overline{\varepsilon}_{\mathrm{i}\bullet}$ es la media de n términos independientes del error ε_{ij}. Así:

$$\mathrm{Var}\left(\overline{\varepsilon}_{\mathrm{i}\bullet}\right) = \frac{\mathrm{Var}\left(\varepsilon_{ij}\right)}{n} = \frac{\sigma^2}{n}$$

Por lo tanto:

$$E\left\{\frac{\sum \left(\overline{\varepsilon}_{i\bullet} - \overline{\varepsilon}_{\bullet\bullet}\right)^2}{I-1}\right\} = \frac{\sigma^2}{n}$$

en consecuencia:

$$E\left\{\sum \left(\overline{\varepsilon}_{\mathrm{i}\bullet} - \overline{\varepsilon}_{\bullet\bullet}\right)^2\right\} = \frac{\left(I-1\right)\sigma^2}{n}$$

5. Dado que tanto $\overline{\varepsilon}_{\mathrm{i}\bullet}$ como $\overline{\varepsilon}_{\bullet\bullet}$ son medias de los ε_{ij} , los cuales tiene un valor esperado, se sigue que:

$$E\left(\overline{\varepsilon}_{\mathrm{i}\bullet}\right) = 0 \; E\left(\overline{\varepsilon}_{\bullet\bullet}\right) = 0$$

por tanto:

$$E\left\{2\sum \left(\mu_i - \mu_\bullet\right)\left(\overline{\varepsilon}_{i\bullet} - \overline{\varepsilon}_{\bullet\bullet}\right)\right\} = 2\sum \left(\mu_i - \mu_\bullet\right) E\left(\overline{\varepsilon}_{\mathrm{i}\bullet} - \overline{\varepsilon}_{\bullet\bullet}\right) = 0$$

Ya se ve que:

$$E\left\{\sum \left(\overline{Y}_{\mathrm{i}\bullet} - \overline{Y}_{\bullet\bullet}\right)^2\right\} = \sum \left(\mu_i - \mu_\bullet\right)^2 + \frac{\left(I-1\right)\sigma^2}{n}$$

Entonces:

$$\begin{aligned} E\left(CM_E\right) = E\left\{\frac{n\sum \left(\overline{Y}_{\mathrm{i}\bullet} - \overline{Y}_{\bullet\bullet}\right)^2}{I-1}\right\} &= \frac{n}{I-1}\left[\sum \left(\mu_i - \mu_\bullet\right)^2 + \frac{\left(I-1\right)\sigma^2}{n}\right] \\ &= \sigma^2 + \frac{n\sum \left(\mu_i - \mu_\bullet\right)^2}{I-1} \end{aligned}$$

TABLA DE ANÁLISIS DE LA VARIANZA

ANOVA de un factor

Fuente de variación	SC	GL	CM	E(CM)
Entre tratamientos	$\sum_i n_i \left(y_{\mathrm{i}\bullet} - \overline{y}_{\bullet\bullet}\right)^2$	$I - 1$	$CM_E = \frac{SC_E}{I-1}$	$\sigma^2 + \frac{1}{I-1}\sum n_i \left(\mu_i - \mu_\bullet\right)^2$

Fuente de variación	SC	GL	CM	E(CM)
Error (dentro de tratamientos)	$\sum_i \sum_j \left(y_{ij} - \overline{y}_{i\bullet}\right)^2$	$N - I$	$CM_D = \frac{SC_D}{N-I}$	σ^2
Total	$\sum_i \sum_j \left(y_{ij} - \overline{y}_{\bullet\bullet}\right)^2$	$N - 1$		

5.12 Prueba F para la Igualdad de las Medias de los Niveles del Factor

Las conclusiones alternativas a ser consideradas son:

$$H_0 : u_1 = u_2 = \ldots = u_I$$
$$H_a : \text{no todos los } \mu_i \text{ son iguales}$$

5.12.1 Prueba Estadística

La prueba estadística a ser usada para elegir entre las hipótesis planteadas, es:

$$F^* = \frac{CM_E}{CM_D}$$

La prueba apropiada es de una cola a la derecha.

5.12.2 Distribución de F*

Cuando todas las medias de los tratamientos son iguales, cada observación Y_{ij} tiene el mismo valor esperado. En vista de la aditividad de la suma de cuadrados y de los grados de libertad, del teorema de Cochran se sigue que:

Cuando H_0 se verifica, SCE/σ^2 y SCD/σ^2 son variables distribuidas como χ^2 independientes. Por lo tanto: cuando H_0 se verifica, F^* se distribuye como $F_{(I-\ 1)(\ N\ -\ I)}$.

Si H_a se verifica, esto es, si los μ_i no son todos iguales, F^* no sigue una distribución F. Es más, sigue una distribución compleja llamada distribución $F_{no\ central}$.

5.12.3 Regla De Decisión

Dado que se sabe que F^* se distribuye como $F_{(I-\ 1)(\ N\ -\ I)}$. cuando se verifica H_0 y que grandes valores de F^* llevan a concluir H_a , la regla de decisión para controlar el nivel de significación α es:

Si $F^* \leq F_{(1-\alpha;I-1;\ N-I)}$ no se rechaza H_0.

Si $F^* > F_{(1-\alpha;I-1;\ N-I)}$ se rechaza H_0.

donde $F^* \leq F_{(1-\alpha;I-1;\ N-I)}$ es el percentil del $(1-\alpha) \times 100$ de la distribución de F.

5.12.4 Comentario

Si hay sólo dos niveles del factor esto es $I = 2$, se ve fácilmente que la prueba empleando F^* es equivalente a la prueba de "t" a dos colas para dos poblaciones. La prueba de F tiene $(1,\ N-2)$ grados de libertad, y la prueba "t" tiene $(n_1 + n_2 - 2)$ o $(N-2)$ grados de libertad, así ambas pruebas conducen a regiones críticas equivalentes. Para comparar las medias de dos poblaciones, la prueba de "t" debe preferirse.

5.13 Formulación Alternativa Del Modelo I

MODELO I DE ANOVA - MODELO DE LOS EFECTOS DEL FACTOR

Con esta formulación las medias de los tratamientos son expresadas de un modo equivalente por medio de la identidad:

$$\mu_i \equiv \mu_\bullet + (\mu_i - \mu_\bullet)$$

donde $u_\bullet$ es una constante. Se denotará la diferencia:

$$(u_i - u_\bullet) = \alpha_i$$

esto implica que:

$u_i = u_\bullet + \alpha_i$

La diferencia $u_i = u_\bullet + \alpha_i$ es llamada el efecto del i-ésimo nivel del factor.

El modelo I de ANOVA puede ser expresado como sigue:

$$Y_{ij} = u_\bullet + \alpha_i + \varepsilon_{ij}$$

donde:

$u_\bullet$ es una componente constante común a todas las observaciones.

α_i es el efecto del i-ésimo nivel del factor (constante para cada nivel del factor)

ε_{ij} son variables independientes que se distribuyen $N(0,\ \sigma^2)$

$$i = 1,\ 2, \ldots,\ I;\ j = 1,\ 2, \ldots,\ n_i$$

El modelo de ANOVA es llamado el modelo de los efectos del factor pues se expresa en términos de los efectos del factor α_i en distinción del modelo de las medias de las celdas, el cual se expresa en términos de las medias de los tratamientos.

El modelo de los efectos del factor es un modelo lineal, como su modelo equivalente de las medias de las celdas.

5.13.1 Definición de $\mu_\bullet$

Medias no pesadas: A menudo, una definición de $\mu_\bullet$ como un promedio no pesado para todas las medias de los niveles del factor μ_i puede ser útil:

$$\mu_\bullet = \frac{\sum_{i=1}^{I} \mu_i}{I}$$

Esta definición implica que

$$\sum_{i=1}^{I} \alpha_i = 0$$

pues:

$$\sum \alpha_i = \sum (\mu_i - \mu_\bullet) = \sum \mu_i - I\mu_\bullet$$

y

$$\sum \mu_i = I\mu_\bullet$$

Así la definición de la constante general . implica una restricción sobre los μ_i, en este caso que su suma debe ser cero.

Medias pesadas: La constante . también puede definirse como un promedio pesado de las medias de los niveles del factor μ_i:

$$\mu_\bullet = \sum_{i=1}^{I} f_i \mu_i$$

donde los f_i son pesos definidos tales que $\sum f_i = 1$ La restricción sobre los α_i es entonces:

$$\sum_{i=1}^{I} f_i \alpha_i \ = 0$$

La elección de los pesos f_i puede depender de la significación de las medidas resultantes de los efectos de los niveles del factor. Por ejemplo, los pesos se pueden dar de acuerdo a: a) una medida conocida de importancia o b) de acuerdo al tamaño muestral.

Cuando los tamaños muestrales son iguales se usa una media no pesada.

5.14 Prueba Para La Igualdad De Las Medias De Los Niveles Del Factor

Dado que el modelo de efectos del factor es equivalente al modelo de las medias de las celdas, la prueba para igualdad de las medias de los niveles del factor es la misma prueba estadística F^*. La única diferencia está en el planteo de las hipótesis. Para el modelo de las medias de las celdas las hipótesis son:

$$H_0 : \ u_1 = u_2 = \ldots = u_I$$
$$H_a : \ \text{no todos los } u_i \text{ son iguales}$$

Para el modelo de los efectos del factor, estas mismas hipótesis en términos de los efectos del factor son:

$$H_0 : \ \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$$
$$H_a : \ \text{no todos los } \alpha_i \text{ son iguales}$$

5.15 Análisis De Los Efectos Del Nivel Del Factor

Si la prueba de F lleva a la conclusión de que las medias de los niveles del factor µi difieren, se sigue que hay una relación entre el factor y la variable dependiente. En este caso, un análisis cuidadoso de la naturaleza de los efectos de los niveles del factor es usualmente emprendido. Esto se hace de dos maneras:

1. Un análisis directo de los efectos de los niveles de interés del factor usando técnicas de estimación.
2. Pruebas estadísticas con respecto a los efectos de los niveles del factor de interés.

5.15.1 Gráficos de las estimaciones de las medias de los niveles del factor

Se dispone de dos tipos de gráficos (1) una línea, la que es apropiada tanto si los tamaños de las muestras n_i son iguales como si no; y (2) un gráfico de probabilidad normal, la que es apropiada si los tamaños de las muestras n_i no son iguales.

5.15.2 Estimación de los efectos de los niveles del factor

Las estimaciones de los efectos de los niveles del factor usualmente empleadas incluyen:

1. Estimación de la media de un nivel del factor
2. Estimación de la diferencia entre dos medias de dos niveles de un factor.
3. Estimación de un contraste entre las medias de los niveles del factor.
4. Estimación de una combinación lineal de las medias de los niveles del factor.

5.15.2.1 Estimación de la media del nivel del factor

Un estimador insesgado de la media del nivel del factor µi, fue obtenido como:

$$\hat{\mu}_i = \overline{Y}_{i\bullet}$$

Este estimador tiene media y varianza:

$$E\left(\overline{Y}_{i\bullet}\right) = \mu_i$$

$$\text{Var}\left(\overline{Y}_{i\bullet}\right) = \frac{\sigma^2}{n_i}$$

El último resultado se sigue pues $\overline{Y}_{i\cdot} = \mu_i + \overline{\varepsilon}_{i\bullet}$, la suma de una constante a una media de n_i términos independientes ε_{ij}, cada uno de los cuales tiene una varianza σ^2. En consecuencia $\overline{Y}_{i\bullet}$ está normalmente distribuido pues los términos del error ε_{ij} son variables aleatorias normales e independientes.

La varianza estimada de $\overline{Y}_{i\bullet}$ se simboliza $S^2(\overline{Y}_{i\bullet})$

$$S^2 = \frac{CM_D}{n_i}$$

Se puede demostrar que $\frac{\overline{Y}_{i\bullet} - \mu_i}{\sqrt{\frac{CM_D}{n_i}}}$, se distribuye como t_{N-I}.

Se sigue que los límites del intervalo de confianza del $(1-\alpha)$ para u_i son:

$$\overline{Y}_{i\bullet} \pm t_{\left(1-\frac{\alpha}{2};N-I\right)}\sqrt{\frac{CM_D}{n_i}}$$

5.15.2.2 Estimación de la diferencia entre dos medias de niveles del factor

Frecuentemente dos tratamientos o niveles de un factor son comparados por estimación de la diferencia D entre las dos medias de los niveles del factor, o sea, u_i y $u_{i'}$:

$$D = u_i - u_{i'}$$

Tal comparación entre dos medias de niveles del factor será llamada comparación de a pares. Un estimador puntual es:

$$\hat{D} = \overline{Y}_{\mathrm{i}\bullet} - \overline{Y}_{i'\bullet}$$

Este estimador puntual es insesgado:

$$E\left(\hat{D}\right) = \mu_i - \mu_{i'}$$

Dado que $\overline{Y}_{\mathrm{i}\bullet}$ y $\overline{Y}_{i'.}$ son independientes, la varianza es:

$$\mathrm{Var}(\hat{D}) = Var\left(\overline{Y}_{\mathrm{i}\bullet}\right) + Var\left(\overline{Y}_{i'.}\right) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$$

La varianza estimada está dada por:

$$S^2\left(\hat{D}\right) = CM_D\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$$

$\hat{D}$ está normalmente distribuido por ser una combinación lineal de variables independientes normales.

Se sigue, de estas características, que:

$\frac{\hat{D}-D}{S\left(\hat{D}\right)} \sim t_{N-I}$ para el modelo de ANOVA

De esta manera un intervalo de confianza de $(1-\alpha)$ para D está dado por:

$$\hat{D} \pm t_{\left(1-\frac{\alpha}{2};N-I\right)}\sqrt{CM_D\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}$$

5.15.2.3 Estimación de contrastes

Un contraste es una comparación que involucra a dos o más medias e incluye al caso anterior de comparación de la diferencia entre un par de medias. Un contraste se simboliza como f y es una combinación lineal de las medias de los niveles del factor µi donde los coeficientes ci suman cero:

$$f = \sum_{i=1}^{I} c_i \mu_i \text{ donde} \sum_{i=1}^{I} c_i = 0$$

Ejemplos: Supongamos que estamos estudiando 4 niveles de un factor

$f = \mu_1 - \mu_2$; aquí $c_1 = 1$; $c_2 = -1$; $c_3 = 0$ y $c_4 = 0$; y $\sum_{i=1}^{4} c_i = 0$.

$f = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$; aquí $c_1 = \frac{1}{2}$; $c_2 = \frac{1}{2}$; $c_3 = -\frac{1}{2}$ y $c_4 = -\frac{1}{2}$; y $\sum_{i=1}^{4} c_i = 0$.

$f = \frac{\mu_1+\mu_3}{2} - \frac{\mu_2+\mu_4}{2}$; aquí $c_1 = \frac{1}{2}$; c2 = -1/2; $c_3 = \frac{1}{2}$ y $c_4 = -\frac{1}{2}$; y $\sum_{i=1}^{4} c_i = 0$.

Un estimador insesgado de un contraste f es:

$$\hat{f} = \sum_{i=1}^{I} c_i \overline{Y}_{i\bullet}$$

Dado que los $\overline{Y}_{i\bullet}$ son independientes, la varianza de $\hat{f}$ es:

$$\text{Var}\left(\hat{f}\right) = \sum_{i=1}^{I} c_i^2 \text{Var}\left(\overline{Y}_i\right) = \sum_{i=1}^{I} c_i^2 \left(\frac{\sigma^2}{n_i}\right) = \sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{n_i}$$

Un estimador insesgado de esta varianza es:

$$S^2\left(\hat{f}\right) = CM_D \sum_{i=1}^{I} \frac{c_i^2}{n_i}$$

$\hat{f}$ está normalmente distribuida pues es una combinación lineal de variables independientes normalmente distribuidas. Se puede demostrar que:

$\frac{\hat{f}-f}{\sqrt{CM_D \sum_{i=1}^{I} \frac{c_i^2}{n_i}}} \sim t_{(N-I)}$ para el modelo de ANOVA

5.15.3 Comparaciones múltiples

1. Planeados o A Priori: Se proponen antes de ver los resultados del experimento, pueden ser significativos aunque el ANOVA no dé significativo.
2. No Planeados o A Posteriori: Se plantean a la vista de los resultados, se hacen sólo si el ANOVA da significativo.

5.15.3.1 Planeados

1) *LSD*: el criterio de la prueba para examinar si existen diferencias significativas entre medias se llama diferencia mínima significativa y se simboliza LSD:

$$\begin{aligned}\text{LSD}_\alpha &= t_{(\alpha(2);N-I)}\sqrt{CM_D\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}\\ &= t_{(\alpha(2);N-I)}\sqrt{CM_D\left(\frac{2}{n}\right)} \text{ para } n_i = n\ \forall\ i\end{aligned}$$

Se calcula $\left|\overline{Y}_{i\bullet} - \overline{Y}_{i'.}\right|$ y este valor se compara con LSD, en caso de que el primero sea mayor las diferencias son significativas.

2) *Método de Bonferroni o método "t"*: Este método es aplicable ya sea que los tamaños muestrales sean iguales o no; o si se hacen comparaciones de a pares o contrastes.

- Un contraste

$$f = \sum_{i=1}^{I} c_i \mu_{i\bullet}$$

estimo por medio de

$$\hat{f} = \sum_{i=1}^{I} c_i \overline{y}_{i\bullet}$$

$$H_{0f}:\ f = 0$$

$$\begin{aligned}\text{Var}\left(\hat{f}\right) &= \text{Var}\left(\sum c_i \overline{y}_{i\bullet}\right)\\ &= \sum c_i^2 \text{Var}\left(\overline{y}_{i\bullet}\right)\\ &= \sum c_i^2 \frac{\hat{\sigma^2}}{n}\\ &= \hat{\sigma^2}\sum_i \frac{c_i^2}{n_i}\\ &= CM_D \sum_i \frac{c_i^2}{n_i}\\ \Rightarrow \text{ES}\left(\hat{f}\right) &= \sqrt{CM_D \sum_i \frac{c_i^2}{n_i}}\end{aligned}$$

- Intervalo de confianza

$$\hat{f} \pm t_{\alpha(2);N-I}\sqrt{CM_D \sum_i \frac{c_i^2}{n_i}}$$
$$\text{Si } 0 \in [\ ] \Rightarrow \textit{no rechazo } H_0$$
$$\varepsilon = \frac{\hat{f}}{\sqrt{CM_D \sum_i \frac{c_i^2}{n_i}}}$$
$$\text{VC} = t_{\alpha(2);N-I}$$
$$\text{Si } \varepsilon > \text{VC} \Rightarrow \textit{rechazo } H_0$$

- m contrastes $f_1,\ f_2,\ \ldots,\ f_m$

Se fija el número de contrastes a realizarse, junto con el experimento, m contrastes. Se toma como nivel para cada contraste $\frac{\alpha}{m} = \alpha_i \Rightarrow \sum_i \alpha_i = \alpha$.

- Intervalo de confianza

$$\hat{f}_i \pm t_{\frac{\alpha}{m}(2);N-I}\sqrt{CM_D \sum_i \frac{c_i^2}{n_i}}$$

5.15.3.2 No Planeados:

1) *Método de Scheffé*: Este método da, para cada contraste, intervalos de confianza de la forma:

$$\hat{f} \pm \text{VC ES}\left(\hat{f}\right)$$

donde

$$\hat{f} = \sum_i c_i \overline{y}_{i\bullet}$$
$$\text{VC} = \sqrt{(I-1)\, F_{I-1;N-I;\alpha}} = S$$
$$\text{ES}\left(\hat{f}\right) = \sqrt{CM_D \sum_i \frac{c_i^2}{n_i}}$$

Si usamos el estadístico

$\varepsilon = \frac{\hat{f}}{\text{ES}\left(\hat{f}\right)}$ rechazo H_0 sí $\varepsilon > S$

2) *Método de Tukey*

Utiliza el método de rangos *studentizado*. Supongamos que se tiene r observaciones independientes $Y_1,\ \ldots,\ Y_r$ de una distribución normal con media u y varianza σ^2. Llamamos w al rango de estas observaciones; así:

$$w = max(Y_i) - min(Y_i)$$

Supongamos que se tiene una estimación S^2 de la varianza σ^2 la cual está basada sobre grados de libertad. El cociente w/s es llamado *rango studentizado* y se denota:

$$q\,(r, \nu) = \frac{w}{S}$$

$$\varepsilon = \frac{\overline{y}_{\text{i}\bullet\text{max}} - \overline{y}_{\text{i}\bullet\text{min}}}{S_{\overline{y}_{\bullet\bullet}}} \sim q_{I;N-I}$$
$$S_{\overline{y}_{\bullet\bullet}} = \sqrt{\frac{CM_D}{n}}$$

Sólo se puede usar si los $n = n_i \ \forall$ i. Si n $\neq$ n_i se usa $n = min(n_i \ , \ n_j)$

Si el $\varepsilon > q_{I;N-I;\alpha}$, rechazo H_0.

5.15.3.3 Ni planeados ni no planeados

Contrastes Ortogonales: son contrastes tales que:

$$\begin{aligned} f &= \sum c_i \mu_i \\ \sum c_i &= 0 \\ \sum c_i^j c_i^{j'} &= 0 \ \forall j \neq j' \end{aligned} \tag{5.3}$$

donde j y j' son contrastes diferentes.

Para aplicar estos contrastes se supone que los $n_i = n \ \forall i$; y que el número de contrastes ortogonales es el mismo que los grados de libertad entre, es decir el número de tratamientos menos uno.

Ejemplo:

Supongamos que tenemos cuatro niveles de un factor

T1	T2	T3	T4

Entonces sólo se pueden hacer 3 contrastes. Sean, por ejemplo:

$$\begin{aligned} f_1 &= \mu_1 - \frac{\mu_2}{3} - \frac{\mu_3}{3} - \frac{\mu_4}{3} \\ f_2 &= \mu_3 - \mu_4 \\ f_3 &= \mu_2 - \frac{\mu_3}{2} - \frac{\mu_4}{2} \\ c_1 &= \begin{pmatrix} 1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{pmatrix} \\ c_2 &= \begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix} \\ c_3 &= \begin{pmatrix} 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \end{aligned}$$

Se puede comprobar que cada uno es un contraste porque $\sum c_i = 0$ y si hacemos la multiplicación de a pares se comprueba su ortogonalidad:

$$\begin{aligned} \sum c_i^1 c_i^2 = 1 \quad &\times 0 \quad + - \frac{1}{3} \quad \times 0 \quad + - \frac{1}{3} \quad \times 1 \quad + - \frac{1}{3} \quad \times -1 \quad = 0 \\ \sum c_i^1 c_i^3 = 1 \quad &\times 0 \quad + - \frac{1}{3} \quad \times 1 \quad + - \frac{1}{3} \quad \times - \frac{1}{2} \quad + - \frac{1}{3} \quad \times - \frac{1}{2} \quad = 0 \\ \sum c_i^2 c_i^3 = 0 \quad &\times 0 \quad + 0 \quad \times 1 \quad + 1 \quad \times - \frac{1}{2} \quad + - 1 \quad \times - \frac{1}{2} \quad = 0 \end{aligned}$$

Esto es equivalente a tener una matriz ortogonal de coeficientes. Para estudiar la significación de los contrastes se debe encontrar un estadístico y compararlo con algún valor crítico. La idea es descomponer la SCE en SC independientes cada un grado de libertad. En el ejemplo la SCE se descompone en:

$$SCE \ = \ SC_{1, \ 2, \ 3, \ 4} \ + \ SC_{3, \ 4} \ + \ SC_{2,3,4}$$

El procedimiento para calcular cada una de las sumas de cuadrados es:

$$SC_{f_i} = \frac{\hat{f}_i^2}{\frac{\sum c_i^2}{n}} = \frac{\hat{f}_i^2 n}{\sum c_i^2}$$
$$\text{donde } \hat{f}_i = \sum c_i \overline{Y}_{\text{i}\bullet}$$

El cociente

$$\frac{SC_{f_i}}{CM_D} \sim F_{1,N-I;\alpha}$$

Ejemplo 1.- Comparaciones

Se midió el contenido de nitrógeno tres suelos.

```
##
## Study: nitro_aov ~ "trt"
##
## LSD t Test for nitro
##
## Mean Square Error:  274.3452
##
## trt,  means and individual ( 95 %) CI
##
##     nitro      std r      LCL      UCL Min Max
## A 275.250 20.04816 8 263.0717 287.4283 242 300
## B 293.625 13.79376 8 281.4467 305.8033 282 320
## C 288.625 15.19340 8 276.4467 300.8033 264 314
##
## Alpha: 0.05 ; DF Error: 21
## Critical Value of t: 2.079614
##
## least Significant Difference: 17.22271
##
## Treatments with the same letter are not significantly different.
##
##     nitro groups
## B 293.625      a
## C 288.625     ab
## A 275.250      b
```

Existen diferencias entre los sitios A y B en lo que a contenido de nitrógeno se refiere.

Ejemplo 2.- Comparaciones a posteriori

Test de Sheffé

```
##
## Study: ratas_aov ~ "especie"
##
## Scheffe Test for presion
##
## Mean Square Error  : 9.25
##
## especie,  means
##
##   presion      std  r Min Max
```

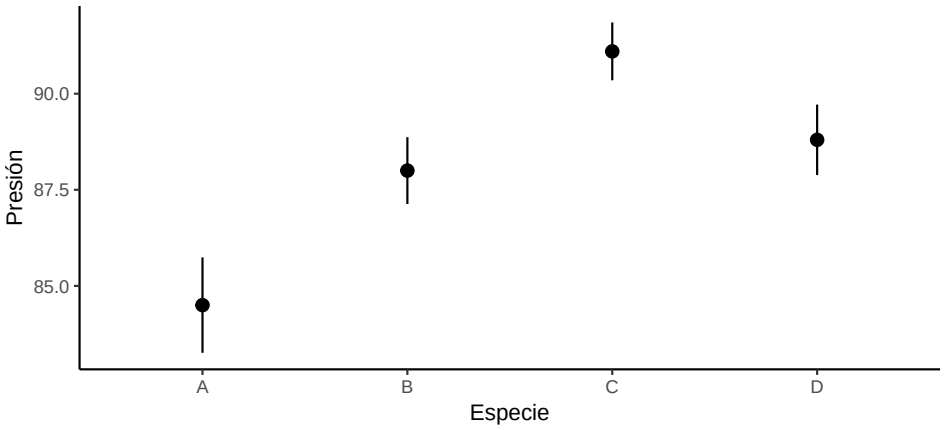


Figure 5.7: Presion de cuatro especies de ratas. media +- error estándar, n = 10.

```
## A    84.5 3.922867 10  79  92
## B    88.0 2.748737 10  84  92
## C    91.1 2.378141 10  87  95
## D    88.8 2.898275 10  85  93
##
## Alpha: 0.05 ; DF Error: 36
## Critical Value of F: 2.866266
##
## Minimum Significant Difference: 3.988455
##
## Means with the same letter are not significantly different.
##
##   presion groups
## C    91.1      a
## D    88.8      a
## B    88.0     ab
## A    84.5      b
```

Test de Tukey

```
##
## Study: ratas_aov ~ "especie"
##
## HSD Test for presion
##
## Mean Square Error:  9.25
##
## especie,  means
##
##   presion      std  r Min Max
## A    84.5 3.922867 10  79  92
## B    88.0 2.748737 10  84  92
## C    91.1 2.378141 10  87  95
## D    88.8 2.898275 10  85  93
##
## Alpha: 0.05 ; DF Error: 36
## Critical Value of Studentized Range: 3.808798
##
```

```
## Minimun Significant Difference: 3.663185
##
## Treatments with the same letter are not significantly different.
##
##   presion groups
## C    91.1      a
## D    88.8      a
## B    88.0     ab
## A    84.5      b
```

Polinomios ortogonales

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
especie	3	224.6	74.87	8.094	0.0003005
especie: A vs C-D	1	198	198	21.41	4.672e-05
especie: A vs B	1	0.1333	0.1333	0.01441	0.9051
especie: C vs D	1	26.45	26.45	2.859	0.09948
Residuals	36	333	9.25	NA	NA

Tabla: Modelo de Análisis de la Varianza Existen diferencias entre A vs C –D y no existen entre C y D.

Ejercicio: Comprobar si los contrastes son ortogonales.

Prueba de homogeneidad de varianzas

	Df	F value	Pr(>F)
group	3	0.7086	0.5532
	36	NA	NA

Tabla: Prueba de Levene para homogeneidad de varianzas (centro = mediana).

	Sum Sq	Df	F value	Pr(>F)
especie	224.6	3	8.094	0.0003005
Residuals	333	36	NA	NA

Tabla: ANOVA (Tipo II)

Medias marginales estimadas

especie	lsmean	SE	df	lower.CL	upper.CL
A	84.5	0.9618	36	82.55	86.45
B	88	0.9618	36	86.05	89.95
C	91.1	0.9618	36	89.15	93.05
D	88.8	0.9618	36	86.85	90.75

Pruebas post hoc

Prueba	contrast	estimate	SE	df	t.ratio	p.value
Tukey	A - B	-3.5	1.4	36	-2.57	0.06553

Prueba	contrast	estimate	SE	df	t.ratio	p.value
Tukey	**A - C**	**-6.6**	**1.4**	**36**	**-4.85**	**0.00013**
Tukey	**A - D**	**-4.3**	**1.4**	**36**	**-3.16**	**0.01606**
Tukey	B - C	-3.1	1.4	36	-2.28	0.12197
Tukey	B - D	-0.8	1.4	36	-0.59	0.93501
Tukey	C - D	2.3	1.4	36	1.69	0.34314
Bonferroni	A - B	-3.5	1.4	36	-2.57	0.08604
Bonferroni	**A - C**	**-6.6**	**1.4**	**36**	**-4.85**	**0.00014**
Bonferroni	**A - D**	**-4.3**	**1.4**	**36**	**-3.16**	**0.01908**
Bonferroni	B - C	-3.1	1.4	36	-2.28	0.17213
Bonferroni	B - D	-0.8	1.4	36	-0.59	1
Bonferroni	C - D	2.3	1.4	36	1.69	0.59688
LSD	**A - B**	**-3.5**	**1.4**	**36**	**-2.57**	**0.01434**
LSD	**A - C**	**-6.6**	**1.4**	**36**	**-4.85**	**2e-05**
LSD	**A - D**	**-4.3**	**1.4**	**36**	**-3.16**	**0.00318**
LSD	**B - C**	**-3.1**	**1.4**	**36**	**-2.28**	**0.02869**
LSD	B - D	-0.8	1.4	36	-0.59	0.56009
LSD	C - D	2.3	1.4	36	1.69	0.09948
Scheffe	A - B	-3.5	NA	NA	NA	0.1041
Scheffe	**A - C**	**-6.6**	NA	NA	NA	**4e-04**
Scheffe	**A - D**	**-4.3**	NA	NA	NA	**0.0301**
Scheffe	B - C	-3.1	NA	NA	NA	0.1779
Scheffe	B - D	-0.8	NA	NA	NA	0.9506
Scheffe	C - D	2.3	NA	NA	NA	0.4253

5.16 Planificación Del Tamaño Muestral

Diseño De Estudios De ANOVA

La planificación de los tamaños muestrales es una parte integral del diseño en un estudio de ANOVA. Se asumirá que todos los niveles del factor tienen el mismo tamaño muestral

5.16.1 Potencia De La Prueba F

La potencia de la prueba F es la probabilidad de rechazar H_0 cuando H_0 es falsa, o también se puede pensar como la probabilidad de no rechazar H_a cuando H_a es cierta. Específicamente la potencia está dada por la siguiente expresión:

$$P = P\left(F^* > F_{(\alpha;I-1,N-I)} \,|\phi\ \right)$$

donde ϕ es un parámetro de no-centralidad, que es una medida de cuan distintas son las μ_i:

$$\phi = \frac{1}{\sigma}\sqrt{\frac{\sum n_i \left(\mu_i - \mu_\bullet\right)^2}{I}}$$

y

$$\mu_\bullet = \frac{\sum n_i \mu_i}{N}$$

Cuando todos los tamaños muestrales son iguales, el parámetro es:

$$\phi = \frac{1}{\sigma}\sqrt{\frac{n\sum\left(\mu_i - \mu_\bullet\right)^2}{I}} \text{con } n = n_i$$

donde:

$$\mu = \frac{\sum \mu_i}{I}$$

Para determinar la potencia, se necesita utilizar la distribución F no-centrada, dado que ésta es la distribución muestral de F^* cuando H_a es cierta. Los cálculos son bastantes complejos, pero se han preparado gráficos que permiten determinar la potencia relativamente fácil. Estos son los gráficos de Pearson-Hartley de la potencia de la prueba F. La curva a utilizar depende del número de niveles del factor, del tamaño muestral y del nivel de significación empleado en la regla de decisión. Estos gráficos se usan de la siguiente forma:

Cada página se refiere a diferentes ν_1, los grados de libertad del numerador de F^*. Para el modelo de ANOVA $\nu_1 = I1$.

Dos niveles de significación, indicados por α, son usados en los gráficos, $\alpha = 0.05$ y $\alpha = 0.01$. Hay dos escalas de X, dependiendo de cual es el nivel de significación empleado. De esta forma, el grupo de curvas de la izquierda corresponde a $\alpha = 0.05$ y el de la derecha a $\alpha = 0.01$.

Hay curvas separadas para diferentes valores de ν_2, los grados de libertad del denominador de F^*. Para el modelo de ANOVA $\nu_2 = NI$. Las curvas son indicadas de acuerdo al valor de ν_2, en la parte superior del gráfico. Dado que sólo son usados en la tabla valores seleccionados de ν_2, es necesario interpolar para valores intermedios de ν_2

La escala de X representa a ϕ, el parámetro no-central.

La escala de Y da la potencia $1 - \beta$, donde β es la probabilidad de cometer el error de tipo II.

Ejemplo 3.- Consideremos el caso donde $\nu_1 = 2$, $\nu_2 = 10$, $\phi = 10$ y $\alpha = 0.05$. Si buscamos en la tabla la potencia es $1 - \beta = 0.983$ aproximadamente.

Una forma alternativa para determinar la potencia es especificar la mínima diferencia que se desea detectar entre las medias de las dos poblaciones más diferentes. Designaremos a esta diferencia mínima detectable δ, calculamos entonces:

$$\phi = \sqrt{\frac{n\delta^2}{2IS^2}}$$

Ejemplo 4.- Supongamos que especificamos que trabajaremos con $n = 10$, y que deseamos detectar, entre cuatro tratamientos, diferencias entre las medias de al menos 4 unidades. De un estudio piloto se sabe que $S^2 = 7.5888$. Trabajamos con $\alpha = 0.05$.

$I = 4 \nu_1 = 3\ n = 10\ \nu_2 = 36\ \delta = 4.0\ S^2 = 7.5888$

$$\phi = \sqrt{\frac{n\delta^2}{2IS^2}}$$
$$\phi = \sqrt{\frac{10\left(4.0\right)^2}{2\left(4\right)\left(7.5888\right)}}$$
$$\phi = \sqrt{2.6355}$$
$$\phi = 1.62$$

En la tabla obtenemos una potencia igual a 0.72; lo que implica una probabilidad de cometer el error de tipo II del 28%.

Si bien es deseable estimar la potencia antes de realizar el ANOVA, es útil, también, preguntarse con que potencia se ha realizado un ANOVA. Esto es especialmente interesante si la H_0 no se ha rechazado, pues entonces es deseable saber cuan bien la prueba detecta las diferencias entre las medias de la población.

Calculamos ϕ, de la siguiente forma:

$$\phi = \sqrt{\frac{(I-1)\,(CM_E - S^2)}{IS^2}}$$

Ejemplo 5.- Los datos de la tabla corresponden a una muestra recogida de tres poblaciones de aves geográficamente aisladas. Se midió la longitud del pico con una precisión de un décimo de mm; obteniéndose los siguientes datos:

Población		
A	B	C
4.2	3.8	3.0
3.3	4.1	3.5
2.8	5.0	4.5
4.3	4.6	4.4
3.7	5.1	
4.5		
3.6		

$H_0:\ u_A = u_B = u_C$
$H_a:$ No todas las medias de las poblaciones son iguales.

Fte. de Variación	SC	GL	CM	F	P	VC
Entre Dentro	1.7977143 5.0322857	2 13	0.8988571 0.3870989	2.322	0.137322	3.806
Total	6.83	15				

No rechazamos H_0 con $p > 0.05$

$$\phi = \sqrt{\frac{(I-1)\,(CM_E - S^2)}{IS^2}} = \sqrt{\frac{(3-1)\,(0.898871 - 0.387098)}{3\,(0.3870989)}} = 0.9388053$$

Consultando la tabla la Potencia es 0.25; por lo cual la probabilidad de cometer error de tipo II es aproximadamente de 0.75.

- Se puede observar que grandes valores de ϕ están asociados con grandes potencias, de las ecuaciones vistas anteriormente se ve que ϕ se incrementa con:
- incremento del tamaño muestral;
- incremento entre las diferencias de las medias de las poblaciones (medida ya sea por CM_E , o por $\sum (\mu_i - \mu)^2$ o por la mínima diferencia detectable);
- un bajo número de niveles del factor o de tratamientos;
- una disminución de la variabilidad dentro de las poblaciones, σ^2, estimada por S^2 o el CM_D.

Ejemplo 6.- Veamos qué pasa con el experimento anterior al aumentar el tamaño de la muestra.

POBLACIÓN		
A	B	C
3.9	4.6	3.7
3.5	4.1	4.2
4.1	4.5	3.6
4.4	4.4	4.0
4.4	3.7	3.3
4.6	4.6	3.5
3.3	3.9	4.0
3.9	4.6	4.4
4.4	4.5	3.5
3.6	3.7	4.1
3.7	4.1	3.9
3.4	4.2	4.3

Fte. de Variación	SC	GL	CM	F	p	VC
Entre	0.9433318	2	0.4716659	3.179	0.05458498	3.285
Dentro	4.89465511	33	0.14832288			
Total	5.8379869	35				

$$\phi \approx 1.21$$

La potencia es entonces 0.4

Para el uso de las tablas vistas anteriormente se hace necesario la realización de un experimento. Pero existen tablas que proporcionan los tamaños muestrales adecuados directamente. Este método es aplicable cuando todos los niveles del factor tienen el mismo tamaño muestral, esto es $n = n_i$.

La planificación del tamaño de la muestra usando estas tablas se hace en términos del parámetro de no-centralidad, para tamaños muestrales iguales. Sin embargo, en lugar de requerir una especificación directa de los niveles de u_i para los cuales es importante controlar la probabilidad de cometer el error de tipo II; esta tabla sólo requiere una especificación del rango mínimo de las medias de los niveles del factor para los cuales es importante detectar diferencias entre los u_i, con alta probabilidad. Este rango mínimo se indica Δ:

$$\Delta = max\,(u_i)\,min\,(u_i)$$

Las siguientes especificaciones son necesarias para hacer uso de la tabla:

1. El nivel de significación α
2. La magnitud del rango mínimo Δ de los u_i, la cual es importante detectar con alta probabilidad. La magnitud de σ, la desviación estándar de Y, debe también ser especificada para entrar en la tabla en términos del cociente: $\frac{\Delta}{\sigma}$
3. El nivel de β. Entrar en la tabla en términos de $1 - \beta$.

Cuando se usa la tabla están disponibles cuatro niveles de α $(0.2; 0.1; 0.05\ y\ 0.01)$. También hay cuatro niveles de a través de la potencia. La tabla provee tamaños muestrales para estudios de $I = 2, \ldots, 10$ niveles del factor o tratamientos.

Ejemplo 7.- 1) Supongamos que se quiere con un rango mínimo $\Delta = 3$, para comparar cuatro tratamientos. Se sabe por estudios anteriores que σ es aproximadamente igual a 2. Los niveles para controlar los errores son:

$$\alpha = 0.05 \ \beta = 0.10 \ o \ P = 1 - \beta = 0.90$$

Entramos a la tabla para $\frac{\Delta}{\sigma} = \frac{3}{2} = 1.5$; $\alpha = 0.05$; $1 - \beta = 0.9$ e $I = 4$. Encontramos que $n = 14$.

Especificación de $\frac{\Delta}{\sigma}$ directa: El rango mínimo también se puede especificar en términos de unidades de desviación estándar.

$$\frac{\Delta}{\sigma} = \frac{k\sigma}{\sigma} = k$$

En nuestro ejemplo supongamos que el rango de las medias es k = 2 o más. Supongamos que las otras especificaciones son:

$$\alpha = 0.01 \ \beta = 0.05 \ o \ 1 - \beta = 0.95$$

En la tabla encontramos que $n = 9$.

2) En el ejemplo se hace necesario incrementar el tamaño de la muestra. Para ello nos preguntamos cuál es el tamaño de muestra necesario para, trabajando con $\alpha = 0.05$, tener una potencia de 0.80 para detectar diferencias tan pequeñas como 0.7. Suponemos que $S^2 = 0.3870989$ es una buena estimación de σ^2.

Entramos a la tabla para $\frac{\Delta}{\sigma} = \frac{0.6}{\sqrt{0.387089}} \approx 1$; $\alpha = 0.05$; $1 - \beta = 0.8$ e $I = 3$

Encontramos que $n = 21$.

5.17 Modelo II De ANOVA: Niveles Del Factor Aleatorios

Existen situaciones en las cuales los niveles del factor o los tratamientos empleados no tienen un interés en sí mismos, pero constituyen una muestra de la población. El Modelo II de ANOVA está diseñado para este tipo de situaciones.

5.17.1 Modelo Aleatorio de Medias de Celdas.

El modelo II de ANOVA para un factor es:

$$Y_{ij} \ = \ u_i \ + \ \varepsilon_{ij}$$

donde

u_i son variables independientes $\sim N\left(\mu_\bullet, \sigma_\mu^2\right)$

εij son variables independientes $\sim N\left(0, \sigma^2\right)$

u_i y ε_{ij} son variables aleatorias independientes

$$i = 1, \ 2, \ldots, \ I; \ j = 1, \ 2, \ \ldots, \ n_i$$

5.17.2 Características importantes del Modelo

El valor esperado de una observación Y_{ij} es:

$$E(Y_{ij}) = u_\bullet$$

esto se debe a que:

$$\begin{aligned} E\left(Y_{ij}\right) &= E\left(u_\bullet\right) + E\left(\varepsilon_{ij}\right) \\ &= u_\bullet + 0 \\ &= u_\bullet \end{aligned}$$

La varianza de Y_{ij}, que se indica σ^2_Y, es:

$$\text{Var}\left(Y_{ij}\right) = \sigma^2_Y = \sigma^2_\mu + \sigma^2$$

A causa de que la varianza de Y en este modelo es la suma de dos componentes, este modelo se llama, algunas veces, un modelo de componentes de la varianza.

Los Y_{ij} están normalmente distribuidos pues son una combinación lineal de variables independientes, u_i y ε_{ij}, distribuidas normalmente

Las Y_{ij} para el modelo aleatorio son sólo independientes si pertenecen a diferentes tratamientos o niveles del factor. Se puede demostrar que la covarianza para cualesquiera dos observaciones Y_{ij} e $Y_{ij'}$, para el mismo nivel i con un modelo II es:

$$Cov(Y_{ij},\ Y_{ij'}) = \sigma^2_Y \ \forall\ j \neq j$$

El modelo II supone que la covarianza entre cualesquiera dos observaciones para el mismo nivel del factor es constante para todos los niveles del factor.

Una vez que los niveles del factor han sido seleccionados, el modelo II asume que dos observaciones cualesquiera para el mismo nivel del factor son independientes pues la media del nivel del factor µi es entonces fijada y las dos observaciones difieren sólo por los términos del error ε_{ij}.

5.17.3 Cuestiones de Interés

Cuando el modelo aleatorio es apropiado, uno no está particularmente interesado en inferencias sobre un u_i particular incluido en el estudio, ya sea si es grande o pequeño, pero sí en inferencias acerca de la población completa de mu_i. Específicamente, el interés a menudo se centra sobre la media de los mu_i, u, y en la variabilidad de los mu_i medida por σ^2_μ.

Dado que σ^2_μ es una medida directa de la variabilidad de los mu_i, el efecto de esa variabilidad, a menudo, es medido por el cociente:

$$\frac{\sigma^2_\mu}{\sigma^2_\mu + \sigma^2}$$

1. El cociente toma valores entre 0 ($\sigma^2 = \infty$) y 1 ($\sigma^2 = 0$).
2. El denominador es σ^2_Y.

En vista de las propiedades 1 y 2, el cociente mide la proporción de la variabilidad total de los Y_{ij} que se debe a la variabilidad en los $\mu_{i\bullet}$.

5.17.4 Prueba para $\sigma_\mu^2 = 0$

Consideremos como decidir entre

$$H_0 : \ \sigma_\mu^2 = 0$$

$$H_a : \ \sigma_\mu^2 > 0$$

H_0 implica que todos los mu_i son iguales, esto es, $mu_i = u_\bullet$. H_a implica que los mu_i difieren.

La diferencia entre los dos modelos aparece en los valores esperados de los cuadrados medios. Se puede demostrar de misma forma que lo hemos hecho para el modelo I, que:

$$E(CM_D) = \sigma^2$$

$$E\left(CM_E\right) = \sigma^2 + \ n\sigma_\mu^2$$

donde

$$n = \frac{1}{I-1}\left[\left(\sum n_i\right) - \frac{\sum n_i^2}{\sum n_i}\right]$$

Sí todos los $n_i = n$, entonces $n = n$

Es claro que si $\sigma_\mu^2 = 0$, el CM_D y el CM_E tienen el mismo valor esperado σ^2. Por otro lado $E\left(CM_E\right) > E(CM_D)$ dado que $n > 0$ siempre. En consecuencia, grandes valores de la prueba estadística:

$$F^* = \frac{CM_E}{CM_D}$$

nos llevará a rechazar H_0. Dado que F^* sigue la distribución F cuando H_0 es verdadera, la regla de decisión es la misma que para el modelo I:

Si $F^* \leq F_{(1-\alpha;I-1;\ N-I)}$ no se rechaza H_0.

Si $F^* > F^* \leq F_{(1-\alpha;I-1;\ N-I)}$ se rechaza H_0.

Ejemplo 8.- Un laboratorio emplea una cierta técnica para determinar el contenido de fósforo en el forraje del ganado bovino. La cuestión planteada es "*si las determinaciones de fósforo dependen de las técnicas empleadas para el análisis*". Para contestar esta pregunta se seleccionaron al azar cuatro técnicas con cinco observaciones para la misma tanda de forraje, obteniéndose los siguientes resultados:

Técnica 1	Técnica 2	Técnica 3	Técnicas 4
34	37	34	36
36	36	37	34
34	35	35	37
35	37	37	34
34	37	36	35

H_0: La determinación del contenido de fósforo no difiere entre las técnicas.

H_a : La determinación del contenido de fósforo difiere entre técnicas.

Fte. de Variación	SC	GL	CM	F	p	VC	General	Ejemplo
Entre	9	3	3	2.4	0.10589	3.23886	$\sigma^2 + n\dot{t}\sigma^2_\mu$	$\sigma^2 + 5\sigma^2_\mu$
Dentro	20	16	1.25				σ^2	σ^2
Total	29	19						

No se rechaza H_0

Niveles del Factor	n_i	Media muestral	Varianza muestral
1	5	34.6	0.8
2	5	36.4	0.8
3	5	35.8	1.7
4	5	35.2	1.7

5.17.5 Estimación De $\mu_\bullet$

Se sabe que:

$$E(Y_{ij}) = u_\bullet$$

Así, un estimador insesgado de $\mu_\bullet$ es:

$$\hat{\mu}_i = \overline{Y}_{\bullet\bullet}$$

Se puede demostrar que la varianza de este estimador es:

$$S^2\left(\overline{Y}_{\bullet\bullet}\right) = \frac{\sigma^2_\mu}{I} + \frac{\sigma^2}{N} = \frac{n\sigma^2_\mu + \sigma^2}{N}$$

Recordar que $N = I\ n$.

Se ve que la varianza está formada por dos componentes.

Un estimador insesgado de esta varianza es:

$$S^2\left(\overline{Y}_{\bullet\bullet}\right) = \frac{CM_E}{N}$$

es un estimador insesgado pues, cuando ni = n:

$$E\left(CM_E\right) = n\sigma^2_\mu + \sigma^2$$

Se puede demostrar que:

$\frac{\overline{Y}_{\bullet\bullet} - \mu_\bullet}{S\left(\overline{Y}_{\bullet\bullet}\right)} \sim t_{(I-1)}$, cuando $n_i = n$.

Así, de la forma usual se obtienen los límites del intervalo de confianza para μ • :

$$\overline{Y}_{\bullet\bullet} \pm t_{I-1;\alpha(2)} S\left(\overline{Y}_{\bullet\bullet}\right)$$

Ejemplo 9.- En el estudio de contenido de fósforo del forraje del ganado bovino. Se tiene:

$$\overline{Y}_{\bullet\bullet} = 35.5 \; CM_E = 3 \; N = 20$$

Necesitamos $t_{3;\ 0.05(2)} = 3.182$ y $S^2\left(\overline{Y}_{\bullet\bullet}\right) = \frac{3}{20} = 0.15$, entonces $S\left(\overline{Y}_{\bullet\bullet}\right) = 0.38729833$; el intervalo de confianza del 95% es:

$$34.27 \leq \; u_{\bullet} \leq 36.73$$

5.17.6 Estimación De $\sigma_{\mu}^2/\left(\sigma_{\mu}^2 + \sigma^2\right)$

El cociente $\sigma_{\mu}^2/\left(\sigma_{\mu}^2 + \sigma^2\right)$ revela el alcance del efecto de la varianza entre los mu_i. Para desarrollar un intervalo de confianza para este cociente, se supone que todos los tamaños muestrales de los niveles del factor son iguales.

Comenzaremos obteniendo un intervalo de confianza para el cociente $\frac{\sigma_{\mu}^2}{\sigma^2}$. El CM_E y el CM_D son variables aleatorias independientes para el modelo II de ANOVA, lo mismo que para el modelo I. Cuando $n_i = n$, se puede demostrar que:

$$\frac{CM_E}{n\sigma_{\mu}^2 + \sigma^2} + \frac{CM_D}{\sigma^2} \sim F_{I-1,N-I}$$

Así, se puede escribir la probabilidad:

$$P\left(F_{\left(1-\frac{\alpha}{2}\right);I-1,N-I} \leq \frac{CM_E}{n\sigma_{\mu}^2 + \sigma^2} + \frac{CM_D}{\sigma^2} \leq F_{\left(\frac{\alpha}{2}\right);I-1,N-I}\right) = 1 - \alpha$$

Reordenando las desigualdades, se obtienen los siguientes límites S e I para $\frac{\sigma_{\mu}^2}{\sigma^2}$

$$I = \frac{1}{n}\left[\frac{CM_E}{CM_D}\left(\frac{1}{F_{\frac{\alpha}{2};I-1,N-I}}\right) - 1\right]$$
$$S = \frac{1}{n}\left[\frac{CM_E}{CM_D}\left(\frac{1}{F_{1-\frac{\alpha}{2};I-1,N-I}}\right) - 1\right]$$

donde I es el límite inferior y S el superior.

Los límites I^* y S^* para $\frac{\sigma_{\mu}^2}{\sigma_{\mu}^2+\sigma^2}$ pueden ser obtenidos como sigue:

$$I^* = \frac{I}{1+I} S^* = \frac{S}{1+S}$$

Ejemplo 9 cont.- En nuestro ejemplo

$$CM_E = 3 \; CM_D = 1.25 \; n = 5 \; I = 4 \; N = 20$$

Para construir el intervalo de confianza del 95% se necesita:

$$F_{0.975;\ 3,\ 19} = 0.071 \; F_{0.025;\ 3,\ 19} = 3.093$$

De esta manera los límites del 95% para $\frac{\sigma_{\mu}^2}{\sigma^2}$ son:

$$I = \frac{1}{5}\left[\frac{3}{1.25}\left(\frac{1}{3.093}\right) - 1\right] = -0.077 S = \frac{1}{5}\left[\frac{3}{1.25}\left(\frac{1}{0.071}\right) - 1\right] = 6.561$$

Cuando el límite inferior del intervalo de confianza para $\frac{\sigma_\mu^2}{\sigma^2}$ es negativo, la práctica usual es considerarlo como 0. Entonces el intervalo de confianza es:

$$0 \leq \frac{\sigma_\mu^2}{\sigma^2} \leq 6.561$$

Finalmente, los límites de confianza para $\frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma^2}$ son:

$$0 \ \leq \ \frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma^2} \leq \ 0.87$$

Concluimos que la variabilidad de la media de las determinaciones de fósforo se encuentra entre 0 y 87% de la varianza total.

Estimación de 2 y σ_μ^2

Un estimador insesgado para σ^2 es:

$$\hat{\sigma}^2 = CM_D$$

Y el intervalo de confianza se obtiene como:

$$\frac{(N-I)\,S^2}{\chi^2_{0.025;N-I}} \leq \sigma^2 \leq \frac{(N-I)\,S^2}{\chi^2_{0.975,N-I}}$$

También se puede obtener un estimador insesgado de σ_μ^2:

$$E(CM_D) = \sigma^2$$

$$E(CM_E) = \sigma^2 + n\sigma_\mu^2$$

Se sigue que:

$$\hat{\sigma}_\mu^2 = \frac{CM_E - CM_D}{n}$$

Ejemplo 9 cont.-

$$CM_D = 1.25 \ \chi^2_{0.975,16} = 6.908 \ \chi^2_{0.025;16} = 28.845$$

El intervalo de confianza es:

$$0.693 = \frac{16\,(1.25)}{28.845} \leq \sigma^2 \leq \frac{16\,(1.25)}{6.908} = 2.895$$

Una estimación insesgada de σ_μ^2 es:

$$\hat{\sigma}_\mu^2 = \frac{3 - 1.25}{5} = 0.35$$

5.17.7 Modelo De Efectos Aleatorios

El modelo se puede expresar como:

$$Y_{ij} = u_{\bullet} + \alpha_i + \varepsilon_{ij}$$

donde

$\mu_{\bullet}$ es una componente constante común a todas las observaciones

α_i son variables aleatorias independientes $\sim N(0, \sigma_{\mu}^2)$

ε_{ij} son variables aleatorias independientes $\sim N(0, \sigma^2)$

α_i y ε_{ij} son independientes

$$i = 1, 2, \ldots, I; j = 1, 2, \ldots, n_i$$

Chapter 6

Problemas ANOVA Simple

Para analizar datos con ANOVA en R hay que conocer unas pocas funciones como mínimo.

- `aov()` ajusta el modelo de ANOVA especificado a los datos.
- `summary()` muestra un resumen del resultado, junto con la típica tabla de ANOVA.
- `autoplot()` realiza automaticamente los gráficos de diagnóstico más comunes.

La función `aov()` tiene dos argumentos principales. En primer lugar, la formula que define el modelo. Las formulas estadísticas tienen dos partes, una izquierda y una derecha y se separan con el signo `~`. La parte izquierda define los términos dependientes, que será una variable el caso de estadística univariada o o varias en estadística multivariada. La parte derecha define los términos explicatorios o independientes. Por ejemplo, `y ~ x` indica que `y` depende de la variable `x`.

La otra parte importante es el argumento `data` que indica donde se encuentran esas variables. Si les aparece un error del tipo *object 'y' not found* es muy probable que hayan especificado mal este argumento o se lo hayan olvidado.

Un ejemplo concreto, con los datos de contenido de nitrógeno en tres suelos. La variable dependiente es el nitrógeno, y la explicatoria es el tipo de suelo. Estas corresponden a las columnas `nitro` y `trt` respectivamente.

```
library(tidyverse)
# Cargar datos
nitro <- read_csv("data/nitrogeno.csv")
nitro
```

```
## # A tibble: 8 x 3
##       A     B     C
##   <int> <int> <int>
## 1   270   309   281
## 2   255   295   264
## 3   278   320   291
## 4   294   283   285
## 5   292   285   314
## 6   300   288   298
## 7   242   282   298
## 8   271   287   278
```

```
# Poner los datos en formato largo
# Una columna para la variable dependiente
# Una columna para la variable explicatoria
nitro <- nitro %>%
```

```
  gather(trt, nitro)
nitro
```

```
## # A tibble: 24 x 2
##    trt   nitro
##    <chr> <int>
##  1 A       270
##  2 A       255
##  3 A       278
##  4 A       294
##  5 A       292
##  6 A       300
##  7 A       242
##  8 A       271
##  9 B       309
## 10 B       295
## # ... with 14 more rows
```

```
nitro_aov <- aov(nitro ~ trt, data = nitro)
nitro_aov
```

```
## Call:
##    aov(formula = nitro ~ trt, data = nitro)
##
## Terms:
##                       trt Residuals
## Sum of Squares  1444.083  5761.250
## Deg. of Freedom        2        21
##
## Residual standard error: 16.56337
## Estimated effects may be unbalanced
```

Por si solo, la impresión de los resultados no da demasiada información. En primer lugar, el la llamada que usamos para calcular el ANOVA. En segundo lugar, cuales son los términos del modelo, junto con sus repectivas suma de cuadados (*Sum of Squares*) y grados de libertad (*Deg. of Freedom*). Y finalmente, el error estándar residual o sea $\sqrt{\hat{\sigma}^2}$.

Para obtener la tabla de ANOVA es necesario usar la función `summary`

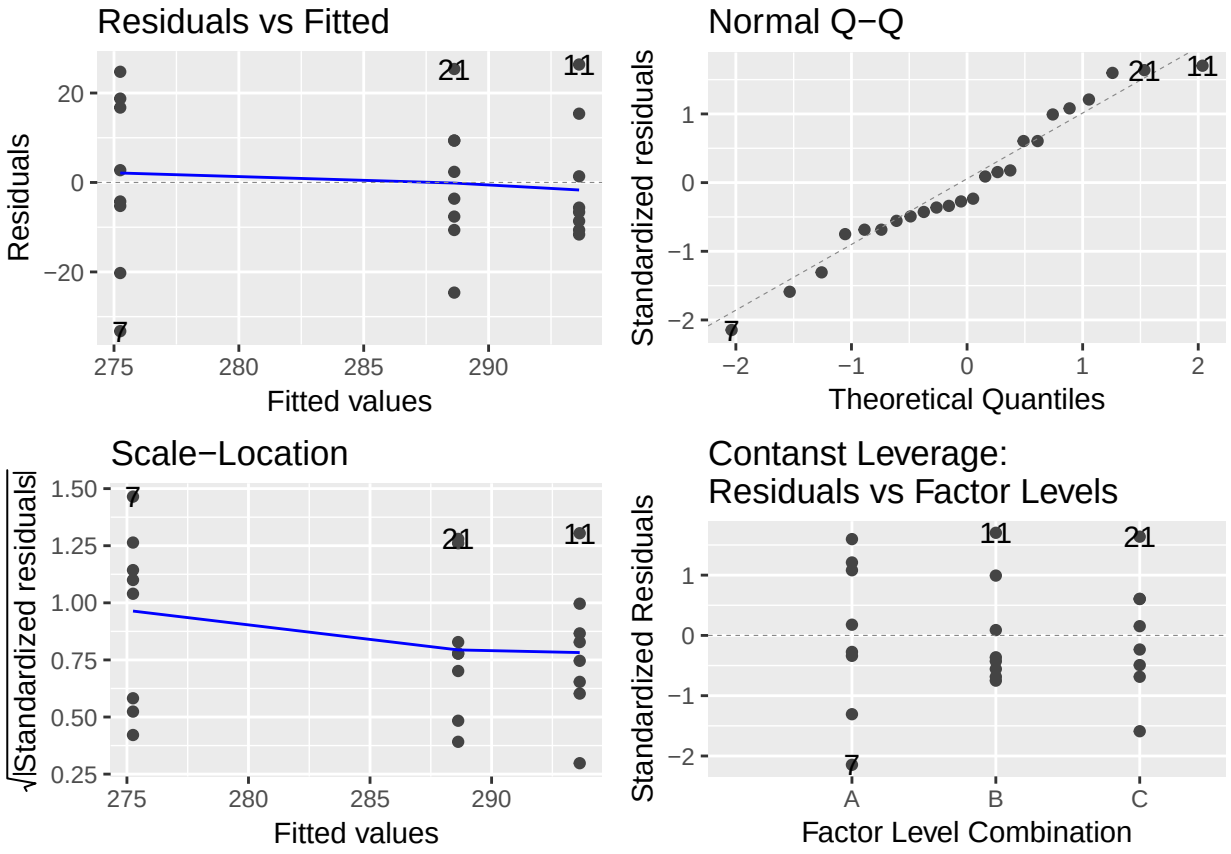
```
summary(nitro_aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## trt          2   1444   722.0   2.632 0.0955 .
## Residuals   21   5761   274.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como vemos, esto nos devuelve la típica tabla de ANOVA.

Además, para saber si el ajuste ha sido adecuado podemos ver los gráficos de residuales

```
library(ggfortify)
autoplot(nitro_aov)
```



Por lo que vemos en el los gráficos no hay motivo para preocuparse por la falta de cumplimiento de los supuestos. Pero para estar seguros podemos usar las pruebas que vimos en la teoría: la prueba de bartlett y la prueba de levene.

```
bartlett.test(nitro ~ trt, data = nitro)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  nitro by trt
## Bartlett's K-squared = 1.0316, df = 2, p-value = 0.597
```

```
library(car)
leveneTest(nitro ~ trt, data = nitro)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.7634 0.4786
##       21
```

Como ven ambas funciones trabajan de manera similar a `aov()`. Ambas aceptan formulas y necesitan del argumento `data`.

Ejercicio 6.1. ¿Qué concluirían si el test de Bartlett rechaza la hipótesis nula y el de Levene no lo hace?

6.0.1 Recordatorio

Recuerden que pueden calcular las medias, desvios, etc. por grupo usando la función `group_by()` y a continuación la función `summarise()`. Por ejemplo:

```
nitro %>%
  group_by(trt) %>%
  summarise(
    media = mean(nitro),
    desvio_estandar = sd(nitro),
    n = n()
  )
```

```
## # A tibble: 3 x 4
##   trt   media desvio_estandar     n
##   <chr> <dbl>           <dbl> <int>
## 1 A      275.            20.0     8
## 2 B      294.            13.8     8
## 3 C      289.            15.2     8
```

Con esta misma secuencia se puede calcular el test de Lilliefors para normalidad. Aunque la secuencia algo más compleja porque el objeto que devuelve este test no es un único número sino que son varios.

```
nitro %>%
  group_by(trt) %>%
  summarise(normalidad = list(lillie.test(nitro))) %>%
  mutate(statistic = map_dbl(normalidad, "statistic"),
         p.value = map_dbl(normalidad, "p.value"))
```

```
## # A tibble: 3 x 4
##   trt   normalidad  statistic p.value
##   <chr> <list>          <dbl>   <dbl>
## 1 A     <S3: htest>     0.173  0.682
## 2 B     <S3: htest>     0.283  0.0580
## 3 C     <S3: htest>     0.144  0.897
```

Voy a explicar que es lo que hice. Los pasos hasta `summarise()` son similares a los que se usan para calcular la media, desvío estándar, etc. La función `lillie.test` devuelve varios números distintos y ¡`summarise` quiere solo un número! Para resolver este incoveniente hay que envolver todos estos números en otra objeto. Imaginen que cada número está dentro de una caja, si metemos todos estas cajas dentro de un cajón entonces tenemos solo un cajón por cada uno de nuestros niveles. Al cajón de está analogía se lo conoce como lista en R (`list`) y puede contener cajas de de cualquier tipo ¡Incluso mezcladas!. Ahora, la lista es cómoda para trabajar como estructura intermedia en nuestros calculos, pero no es cómoda para ver los resultados. De cada uno esos cajones queremos extraer la caja con los números que nos interesan, el valor del estadístico y la probabilidad ($P(X > x)$). Para eso use la función `mutate` que agrega o cambia el valor de una columna y la función `map`. Es un poco compleja de explicar ahora como funciona esta última función, pero por ahora solo necesitan saber que la estamos usando para extraer la cajas de los cajones. Entonces, para ponerlo en castellano lo que estoy haciendo se puede traducir como: con los datos de `nitro` agrupalos por la variable `trt`; luego resumilos en una nueva variable `normalidad`, que es el resultado de probar la normalidad con la función `lillie.test()` de la variable `nitro`; luego, a ese resultado, agregar la variable `statistic` que es igual a extraer la caja `statistic` del cajón `normalidad` y la variable `p.value` que es igual a extraer la caja `p.value` del cajón `normalidad`.

Ejercicio 6.2. 1. ¿Cuales son las hipótesis que se prueban en la prueba de Lilliefors?

2. ¿Se rechaza alguna de ellas?

6.1 Problemas

Primero bajen la planilla para completar los problemas:

```
download.file("git.io/informe-anova.Rmd", "informe-anova.Rmd")
```

1. Se lleva a cabo una experiencia para poner a prueba el efecto de 6 fertilizantes sobre el crecimiento de la soja, obteniéndose la siguiente tabla de ANOVA:

Fte.de.Variación	SC	GL	F	p
Entre	754.25	5	2.23	0.0644
Dentro	3646.08	54		

- Calcular la potencia. Enuncie sus conclusiones.
- ¿Cual es el n necesario para tener una potencia de 90%?

Para calcular la potencia en R se puede usar la función `power.anova.test()` del paquete `pwr`. Esta función puede calcular tanto la potencia de una prueba como el n necesario para alcanzar cierta potencia, dependiendo de que dato falte completar. Para eso necesita:

- El número de grupos
- El n de cada grupo (esto implica que solo da resultados aproximados con datos desbalanceados)
- la varianza entre grupos.
- la varianza dentro de los grupos.

A partir de la tabla de ANOVA es posible derivar todos los datos.

```
# Número de grupos. Recordar que GL entre = K - 1
g <- 5 + 1
# Número de réplicas por grupo N-I/I
n <- (54 + g) / g

# bv <- (as.numeric(as.character(soja$SC[1])) +
#          as.numeric(as.character(soja$SC[2]))) / (n*g-1)/n

# Varianza entre grupo CM entre /n
bv <- 754.25 / 5 / n

# Varianza dentro de grupos CM dentro
wv <- 3648.08/ 54



power.anova.test(groups = g, n = n, between.var = bv, within.var = wv)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##           groups = 6
##                n = 10
##     between.var = 15.085
##      within.var = 67.55704
##       sig.level = 0.05
##           power = 0.6821423
##
## NOTE: n is number in each group
```

2. Fueron ensayados dos cebos distintos para estimar si existía una diferencia significativa en función de su consumo por ratones silvestres. Los datos de la tabla se obtuvieron en cinco sitios diferentes por cebo y están expresados como porcentaje de consumo:

```
cebos
```

```
##    A  B
## 1 10 15
## 2 15 20
## 3 12 16
## 4 20 25
## 5 14 20
```

- Analizar la significación de estas observaciones.
- ¿Qué transformación es conveniente utlizar teniendo en cuenta el tipo de dato?
- La potencia de la prueba.

3. Durante el estudio del control del fotoperíodo de la reproducción del alga roja Porphira, se llevó a cabo un experimento para examinar el efecto de la interrupción de largos períodos de oscuridad, mediante un período de iluminación de 30 minutos con luz de diferentes longitudes de onda, y se contaron los esporangios en un volumen fijo de material. Se obtuvieron 4 réplicas para cada una de las cinco longitudes de onda.

```
##        Color N\xfamero.de.esporangios  NA. NA..1 NA..2
## 1       Azul                     7720 7490  7986  7382
## 2      Verde                     7918 7948  7632  8215
## 3   Amarillo                     6495 7101  7412  7006
## 4       Rojo                     4741 4150  5315  4810
## 5 Infrarrojo                     7520 7418  7937  7118
```

Teóricamente solamente la luz roja tiene efecto sobre el número de esporangios.

a) Realizar un análisis para decidir si hay diferencia entre los efectos de las longitudes de onda.
b) Poner a prueba el supuesto teórico.
c) ¿Puede inferirse algún otro resultado?

6.1.1 Contrastes

En R existen varias formas de hacer contrastes. Una de las más prácticas es usar el paquete `emmeans` que además de estimar la medias marginales también permite realizar comparaciones de a pares. La función `eemeans` devuelve por defecto las medias marginales junto con los errores estándar, grados de libertad, e intervalos de confianza.

```
alga.em <- emmeans(alga.aov, ~ Color)
alga.em
```

```
##  Color       emmean       SE df lower.CL upper.CL
##  Amarillo   7003.50 175.4702 15 6629.494 7377.506
##  Azul       7644.50 175.4702 15 7270.494 8018.506
##  Infrarrojo 7498.25 175.4702 15 7124.244 7872.256
##  Rojo       4754.00 175.4702 15 4379.994 5128.006
##  Verde      7928.25 175.4702 15 7554.244 8302.256
##
## Confidence level used: 0.95
```

También puede ser usada para estimar la comparación de a pares. Por defecto, usa el método de Tukey. Se pueden usar otros métodos como Bonferroni (`bonferroni`), Scheffé (`scheffe`), LSD (`none`), y otro más (para más detalles ver `?summary.emmGrid` sección *P-value adjustment*)

```
pairs(alga.em, adjust = "none")
```

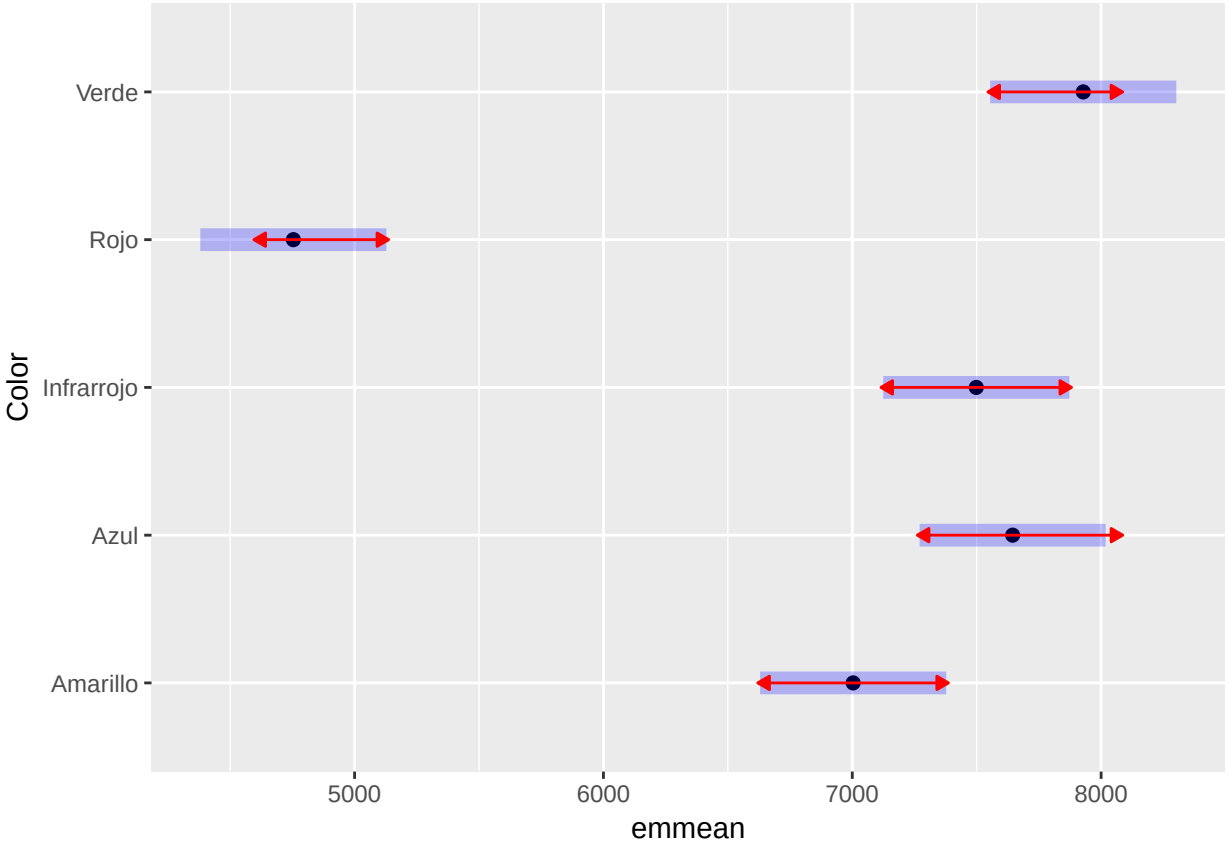
```
##  contrast              estimate       SE df t.ratio p.value
##  Amarillo - Azul        -641.00 248.1523 15  -2.583  0.0208
##  Amarillo - Infrarrojo  -494.75 248.1523 15  -1.994  0.0647
##  Amarillo - Rojo        2249.50 248.1523 15   9.065  <.0001
##  Amarillo - Verde       -924.75 248.1523 15  -3.727  0.0020
##  Azul - Infrarrojo       146.25 248.1523 15   0.589  0.5644
##  Azul - Rojo            2890.50 248.1523 15  11.648  <.0001
##  Azul - Verde           -283.75 248.1523 15  -1.143  0.2708
##  Infrarrojo - Rojo      2744.25 248.1523 15  11.059  <.0001
##  Infrarrojo - Verde     -430.00 248.1523 15  -1.733  0.1036
##  Rojo - Verde          -3174.25 248.1523 15 -12.792  <.0001
```

```
LSD.test(alga.aov, "Color", console = TRUE)
```

```
##
## Study: alga.aov ~ "Color"
##
## LSD t Test for esporangios
##
## Mean Square Error:  123159.2
##
## Color,  means and individual ( 95 %) CI
##
##            esporangios      std r      LCL      UCL  Min  Max
## Amarillo       7003.50 380.7698 4 6629.494 7377.506 6495 7412
## Azul           7644.50 267.7679 4 7270.494 8018.506 7382 7986
## Infrarrojo     7498.25 338.6270 4 7124.244 7872.256 7118 7937
## Rojo           4754.00 477.0891 4 4379.994 5128.006 4150 5315
## Verde          7928.25 238.3868 4 7554.244 8302.256 7632 8215
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of t: 2.13145
##
## least Significant Difference: 528.9242
##
## Treatments with the same letter are not significantly different.
##
##            esporangios groups
## Verde          7928.25      a
## Azul           7644.50      a
## Infrarrojo     7498.25     ab
## Amarillo       7003.50      b
## Rojo           4754.00      c
```

También se pueden graficar los intervalos de confianza de las medias estimadas, y de las comparaciones de entre ellas.

```
plot(alga.em, comparisons = TRUE)
```



Las barras azules son los intervalos de confianza para las medias y las flechas rojas son los intervalos de confianza de las comparaciones entre ellos. Si las flechas no se superponen las diferencias son significativas entre ellos.

También es posible obtener lo que se llama *compact letter display* que es una forma muy práctica de ver comparaciones.

```
cld(alga.em)
```

```
##        Color  emmean       SE df lower.CL upper.CL .group
## 4       Rojo 4754.00 175.4702 15 4379.994 5128.006    1
## 1   Amarillo 7003.50 175.4702 15 6629.494 7377.506     2
## 3 Infrarrojo 7498.25 175.4702 15 7124.244 7872.256     23
## 2       Azul 7644.50 175.4702 15 7270.494 8018.506     23
## 5      Verde 7928.25 175.4702 15 7554.244 8302.256      3
```

Los niveles que no comparten números o letras son significativamente distintos.

También podemos ver cuales son los coeficientes de los contrastes usados.

```
coef(pairs(alga.em))
```

```
##                 Color c.1 c.2 c.3 c.4 c.5 c.6 c.7 c.8 c.9 c.10
## Amarillo     Amarillo   1   1   1   1   0   0   0   0   0    0
## Azul            Azul  -1   0   0   0   1   1   1   0   0    0
## Infrarrojo Infrarrojo   0  -1   0   0  -1   0   0   1   1    0
## Rojo            Rojo    0   0  -1   0   0  -1   0  -1   0    1
## Verde          Verde    0   0   0  -1   0   0  -1   0  -1   -1
```

Ahora, la pregunta que nos hacen es poner a prueba el supuesto teórico de que solo las roja es efectiva. Pode-

mos hacerlo de dos formas distintas. Una por contrastes ortogonales. Es quizás el método más complicado, aunque más poderoso, de hacer. Primero debemos implementar nuestros coeficientes. Los niveles del factor son ordenados por orden alfabético a menos que indiquemos otro orden. Por lo tanto, el orden de los niveles de `Color` es: Amarillo, Azul, Infrarrojo, Rojo, Creamos una matriz de tamaño I x (I-1). Cada columna es un contraste.

```
contraste.algas <- matrix(c(-1, -1, -1 , 4, -1,
                            -1, -1, -1, 0, 3,
                            -1, -1, 2, 0 , 0,
                            1, -1, 0, 0, 0),
                          nrow = 5)
row.names(contraste.algas) <- levels(alga$Color)
colnames(contraste.algas) <- paste("c", 1:4, sep = ".")
contraste.algas
```

```
##            c.1 c.2 c.3 c.4
## Amarillo    -1  -1  -1   1
## Azul        -1  -1  -1  -1
## Infrarrojo  -1  -1   2   0
## Rojo         4   0   0   0
## Verde       -1   3   0   0
```

```
# Comprobar que son ortogonales, fuera de la diagonal debe dar 0
crossprod(contraste.algas)
```

```
##     c.1 c.2 c.3 c.4
## c.1  20   0   0   0
## c.2   0  12   0   0
## c.3   0   0   6   0
## c.4   0   0   0   2
```

Una vez hecho la matriz de coeficientes, la usamos dentro de la función `aov` especificando el argumento `contrasts` que debe ser una lista con nombres igual a los variables explicatorias.

```
alga.aov_or <- aov(esporangios ~ Color, data = alga,
                   contrasts = list(Color = contraste.algas))
```

Luego hay que hacer algo similar para que `summary` muestre esos contrastes. Especificar el argumento `split` que también tiene que ser una lista con los nombres de los variables explicatorias, pero dentro de cada uno hay un vector con nombres donde el número indica que contraste es.

```
summary(alga.aov_or, split = list(Color = c("Rojo vs Todos" = 1,
                                            "Verde vs Amarillo, Azul, Infrarrojo"  = 2,
                                            "Amarillo Azul vs Infrarrojo" = 3,
                                            "Amarillo vs Azul" = 4)))
```

```
##                                             Df   Sum Sq  Mean Sq F value
## Color                                        4 26255709  6563927  53.296
##   Color: Rojo vs Todos                       1 24458084 24458084 198.589
##   Color: Verde vs Amarillo, Azul, Infrarrojo 1   894894   894894   7.266
##   Color: Amarillo Azul vs Infrarrojo         1    80968    80968   0.657
##   Color: Amarillo vs Azul                    1   821762   821762   6.672
## Residuals                                   15  1847388   123159
##                                               Pr(>F)
## Color                                       1.09e-08 ***
##   Color: Rojo vs Todos                      4.67e-10 ***
##   Color: Verde vs Amarillo, Azul, Infrarrojo   0.0166 *
```

```
##   Color: Amarillo Azul vs Infrarrojo           0.4301
##   Color: Amarillo vs Azul                      0.0208 *
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Noten las diferencias de resultados entres las comparaciones múltiples.

La otra forma es usar comparaciones de a pares pero solo usando tratamientos vs control. En este caso nuestro "control" es el color rojo. Se puede hacer usando `emmeans` y resulta mucho más sencillo. La función a usar es `contrast`. Además de indicar el objeto sobre el que hay que realizar los contrastes, también es necesario indicar el método (`method`), y opcionalmente el número de nivel que corresponde al tratamiento control (`ref`)

```
contrast(object = alga.em, method = "trt.vs.ctrl", ref = 4)
```

```
##  contrast          estimate       SE df t.ratio p.value
##  Amarillo - Rojo    2249.50 248.1523 15   9.065  <.0001
##  Azul - Rojo        2890.50 248.1523 15  11.648  <.0001
##  Infrarrojo - Rojo  2744.25 248.1523 15  11.059  <.0001
##  Verde - Rojo       3174.25 248.1523 15  12.792  <.0001
##
## P value adjustment: dunnettx method for 4 tests
```

4. La faciolasis es una enfermedad parasitaria producida por la *Fasciola hepatica* (trematode hepático). Los trematodes adultos viven en el conducto biliar del huésped, donde segregan cantidades significativas de ciertos aminoácidos, en especial prolina; el huésped presenta, como característica, anemia (reducción en los glóbulos rojos de la sangre). Se tomaron 40 ratas Wistar, sanas de aproximadamente igual peso y edad, se dividieron al azar en 4 grupos de 10 ratas cada uno. Se adaptó un aparato para infundir material directamente al conducto biliar de las ratas mediante una cánula. Las ratas del grupo I recibieron 20 minimoles de prolina disuelta en suero fisiológico, las del grupo II recibieron un cóctel consistente siete aminoácidos (excluyendo prolina) segregados por el trematode, también disuelto en suero fisiológico; el grupo III recibió lo mismo que el II más el agregado de 20 milimoles de prolina (simulando a lo segregado por el trematode) y el grupo IV sólo se trató con suero fisiológico. En todos los casos se tomó como variable el número de glóbulos rojos del huésped, expresados en millones por mm3 de sangre. Los resultados se presentan en la siguiente tabla:

   GRUPO.I GRUPO.II GRUPO.III

   1 20 mmol prolina mezcla aa - prolina mezcla aa + prolina 2 6.07 5.69 5.61 3 5.02 5.54 5.40 4 5.69 5.35 5.26 5 5.43 5.11 4.99 6 5.87 5.94 5.44 7 5.55 5.25 5.13 8 5.64 6.02 5.21 9 5.95 5.64 5.52 10 5.20 5.11 4.79 11 5.40 5.04 4.92 GRUPO.IV 1 suero fisiológico 2 7.35 3 7.11 4 6.99 5 6.72 6 7.16 7 6.85 8 6.94 9 7.25 10 6.51 11 6.65

a) Plantear y comprobar todos los supuestos para la validez de las pruebas estadísticas utilizadas.
b) ¿Está asociada la reducción del número de glóbulos rojos de la sangre del huésped con la segregación de aminoácidos por el trematode?
c) ¿Está específicamente asociado a la segregación de prolina?
d) Realice un breve comentario sobre el diseño del experimento.

5. Se realiza una experiencia a fin de comparar tres métodos diferentes para determinar el contenido de oxígeno disuelto en el agua de lagos. Se extrae una muestra aleatoria de 18 muestras de agua de un lago, las cuales se dividen al azar en tres grupos de igual tamaño y cada uno de los grupos es asignado al azar a uno de los métodos que se quiere comparar. Se obtienen los siguientes resultados, expresados en mg/litro:

Método.1 Método.2 Método.3 1 832 1023 8710 2 324 832 1660 3 550 1318 5495 4 617 1995 3981 5 525 832 2138 6 1349 912 3548 7 501 646 5130

a) Comprobar las suposiciones del ANOVA
b) Poner a prueba la hipótesis "No hay efecto del método en la determinación de oxígeno en el agua del lago". Indicar P.
c) Realizar comparaciones entre métodos, utilizando todos los métodos de contraste conocidos e indicar cuáles serían los adecuados a este problema particular.
d) Hallar la potencia de la prueba para alguna hipótesis alternativa.
e) Estimar el tamaño de la muestra (¿de qué?) con la que debería trabajar para tener una potencia del 95%, con una probabilidad de cometer error de Tipo I del 5%.

6. En un estudio sobre viabilidad, se aíslan tres parejas de *Drosophila melanogaster* en 10 frascos y se hace un recuento del número de huevos al cabo de 8 días. Esta experiencia se repite 4 veces con parejas distintas. Los resultados obtenidos son:

```
mosca
```

```
##    Serie.1 Serie.2 Serie.3 Serie.4
## 1       47      28      32      50
## 2       36      31      41      44
## 3       22      32      44      67
## 4       69      45      17      63
## 5       68      72      96      87
## 6       57     101      20      74
## 7       37      55      45      21
## 8      108      27      55      54
## 9       29      49      36      91
## 10      72      36      72      72
```

a. ¿Es posible reunir las cuatro series en una sola para efectuar un análisis conjunto de la viabilidad? Trabajar con $= 0.05$
b. Hallar la potencia de la prueba realizada cuando se dan ciertas alternativas.
c. Estimar el tamaño de la muestra con que debería trabajar en cada tratamiento para tener una potencia mayor del 95%.

Chapter 7

ANOVA DE DOS FACTORES

7.1 Ventajas de los estudios multifactoriales

7.1.1 Eficiencia:

7.1.1.1 Cantidad de Información: Permiten estudiar la *interacción* de los factores.

Validez de las decisiones: los experimentos multifactoriales también pueden robustecer la validez de las decisiones.

Comentarios:

1. Los análisis multifactoriales permiten una evaluación efectiva de los efectos de la interacción y economiza el número de casos requeridos para el análisis.
2. Experimentos involucrando muchos factores, cada uno con numerosos niveles, se vuelven complejos, costosos e insumen tiempo.

7.2 Elementos del Modelo

Ejemplo: Consideremos un estudio de dos factores, en el cual son de interés los efectos del sexo y la edad en el aprendizaje de una tarea. El factor edad lo definimos en términos de sólo tres niveles (adolescente, adulto joven, anciano).

La respuesta para un dado tratamiento, en un estudio de dos factores, es indicada por µij, donde i hace referencia al nivel del factor A ($i = 1, 2, \ldots, I$) y j se refiere al nivel del factor B ($j = 1, 2, \ldots, J$).

Table 7.1: Tiempo de aprendizaje (en minutos) en mujeres de y hombres de tres edades. Caso sin interacción y sin efecto del factor *Sexo*.

SEXO	EDAD			
	Adolescente (j=1)	**Adulto joven (j=2)**	**Anciano (j=3)**	$\mu_{\mathbf{i}\bullet}$
Masculino ($i = 1$)	9 (μ_{11})	11 (μ_{12})	16 (μ_{13})	12 ($\mu_{1\bullet}$)
Femenino ($i = 2$)	9 (μ_{21})	11 (μ_{22})	16 (μ_{23})	12 ($\mu_{2\bullet}$)

SEXO	EDAD			
$\mu_{\bullet j}$	9 ($\mu_{\bullet 1}$)	11 ($\mu_{\bullet 2}$)	16 ($\mu_{\bullet 3}$)	12 ($\mu_{\bullet\bullet}$)

Indicamos

$$\mu_{\bullet j} = \frac{\sum_{i=1}^{I} \mu_{ij}}{I}$$

y

$$\mu_{i\bullet} = \frac{\sum_{j=1}^{J} \mu_{ij}}{J}$$

La media general se indica cómo $\mu_{\bullet\bullet}$, y es definida en las siguientes formas equivalentes:

$$\mu_{\bullet\bullet} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}}{ij}$$
$$\mu_{\bullet\bullet} = \frac{\sum_{i=1}^{I} \mu_{i\bullet}}{I}$$
$$\mu_{\bullet\bullet} = \frac{\sum_{j=1}^{J} \mu_{\bullet j}}{J}$$

7.2.1 Efectos principales

Definimos el efecto principal del factor A al i-ésimo nivel, como:

$$\alpha_i = \mu_{i\bullet} - \mu_{\bullet\bullet}$$

En el ejemplo:

$$\alpha_1 = \mu_{1\bullet} - \mu_{\bullet\bullet} = 912 = -3$$

De forma similar, el efecto principal del j-ésimo nivel del facto B se define:

$$\beta_j = \mu_{\bullet j} - \mu_{\bullet\bullet}$$

En el ejemplo:

$$\beta_1 = \mu_{\bullet 1} - \mu_{\bullet\bullet} \bullet\bullet = 1212 = 0$$

Se sigue que:

$$\sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0$$

Así, la suma de los efectos principales para cada factor es cero.

7.2.1.1 Aditividad de los efectos de los factores

En general, si los efectos son aditivos se tiene:

$$\mu_{ij} = \mu_{\bullet\bullet} + \alpha_i + \beta_j$$

lo que se puede expresar de forma equivalente, usando la definición de α_i y de β_j, como:

$$\mu_{ij} = \mu_{\bullet\bullet} + \mu_{i\bullet} + \mu_{\bullet j}$$

En el ejemplo

$$\mu_{11} = \mu_{\bullet\bullet} + \alpha_I + \beta_j = 12 + 0 + (-3) = 9$$

Cuando todos los tratamientos pueden ser expresados en esta forma, se dice que los factores principales *no interactúan*, o que los efectos de los factores son *aditivos*.

7.3 Representación gráfica

Una de las mejores formas para representar este tipo de datos es un gráfico de líneas y puntos. En el eje de las abscisas va una de las variables y en el eje de las ordenadas la variable de respuesta. Se grafica cada observación o media como punto y se unen los puntos de los niveles que son iguales para la otra variable. Este tipo de gráfico permite ver si las lineas son paralelas. Si lo son indica que los efectos de los factores son aditivos. La falta de paralelismo puede estar indicando que hay interacción.

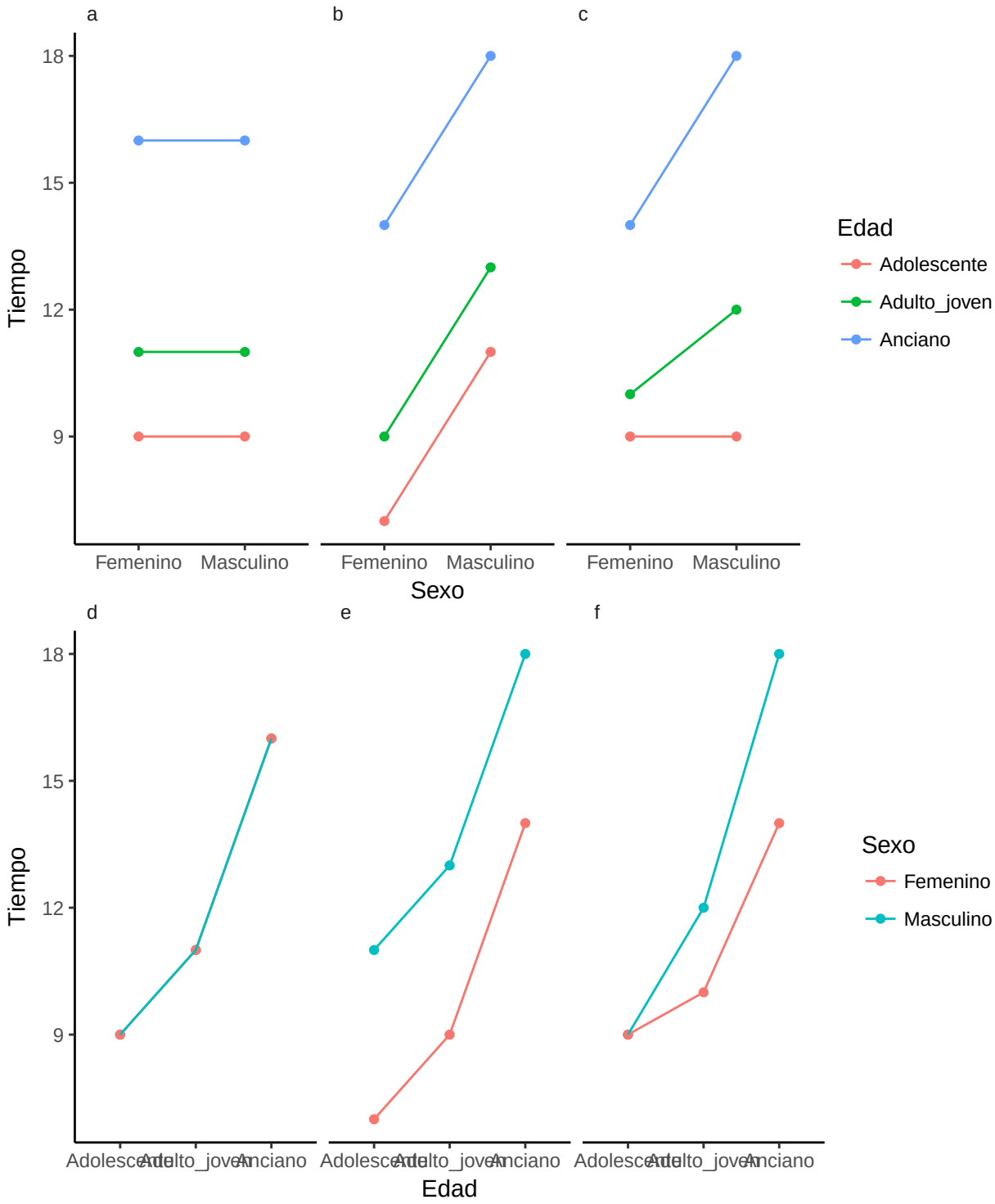


Table 7.2: Tiempo de aprendizaje (en minutos) en mujeres de y hombres de tres edades. Caso sin interacción y efecto del factor *Sexo*.

SEXO	EDAD			
	Adolescente ($j = 1$)	**Adulto joven** ($j = 2$)	**Anciano** ($j = 3$)	$\mu_{i\bullet}$

SEXO	EDAD			
Masculino ($i = 1$)	11 (μ_{11})	13 (μ_{12})	18 (μ_{13})	14 ($\mu_{1\bullet}$)
Femenino ($i = 2$)	7 (μ_{21})	9 (μ_{22})	14 (μ_{23})	10 ($\mu_{2\bullet}$)
$\mu_{\bullet j}$	9 ($\mu_{\bullet 1}$)	11 ($\mu_{\bullet 2}$)	16 ($\mu_{\bullet 3}$)	12 ($\mu_{\bullet\bullet}$)

7.4 Interacción

$$\mu_{ij} = \mu_{\bullet\bullet} + \alpha_i + \beta_j$$

Si esto se cumple, los efectos serán aditivos; de lo contrario los factores interactúan.

Table 7.3: Tiempo de aprendizaje (en minutos) en mujeres de y hombres de tres edades. Caso con interacción.

SEXO	EDAD				
	Adolescente ($j = 1$)	**Adulto joven** ($j = 2$)	**Anciano** ($j = 3$)	$\mu_{i\bullet}$	
Masculino ($i = 1$)	9 (μ_{11})	12 (μ_{12})	18 (μ_{13})	13 ($\mu_{1\bullet}$)	1(α_1)
Femenino ($i = 2$)	9 (μ_{21})	10 (μ_{22})	14 (μ_{23})	11 ($\mu_{2\bullet}$)	-1 (α_2)
$\mu_{\bullet j}$	9 ($\mu_{\bullet 1}$) -3 (β_1)	11 ($\mu_{\bullet 2}$) -1 (β_2)	16 ($\mu_{\bullet 3}$) 4 (β_3)	12 ($\mu_{\bullet\bullet}$)	

Para el ejemplo de la tabla, es claro que los efectos de los factores interactúan, por ejemplo:

$$\mu_{\bullet\bullet} + \alpha_1 + \beta_1 = 12 + 1 + (-3) = 10$$

mientras que $\mu_{11} = 9$.

La diferencia entre la media del tratamiento μ_{ij} y el valor $(\mu_{\bullet\bullet} + \alpha_i + \beta_j)$ es llamada la *interacción* del i-ésimo nivel del factor A con el j-ésimo nivel del factor B, se simboliza $(\alpha\beta)_{ij}$ y la definimos como:

$$(\alpha\beta)_{ij} = \mu_{ij} - (\mu_{\bullet\bullet} + \alpha_i + \beta_j)$$

Reemplazando α_i y β_j por su definición, se obtiene la siguiente expresión alternativa:

$$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{I\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet}$$

Por ejemplo, la interacción para:

$$\begin{aligned}(\alpha\beta)_{13} &= \mu_{13}\,(\mu_{\bullet\bullet} + \alpha_1 + \beta_3) \\ &= 18\,(12 + 1 + 4) \\ &= 1\end{aligned}$$

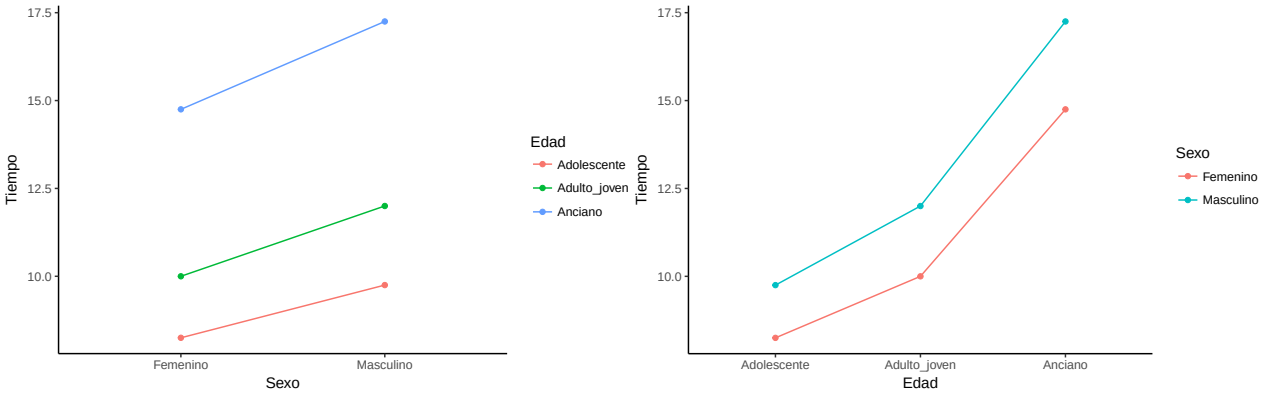


Figure 7.1: Gráfico de interacción para las medias del tiempo de aprendizaje. Los datos graficados corresponden a la Tabla 7.5

Table 7.4: Efectos de $\alpha\beta_{ij}$.

	$j=1$	$j=2$	$j=3$	Promedio
$i=1$	-1	0	1	0
$i=2$	1	0	-1	0
Promedio	0	0	0	0

7.4.1 Interacciones no importantes

Muchas veces se detectan interacciones pero estas son pequeñas y no cambian las conclusiones que se pueden sacar sobre el efecto de los factores principales.

Table 7.5: Tiempo de aprendizaje (en minutos) en mujeres de y hombres de tres edades. Caso con interacción no importante.

SEXO	EDAD			
	Adolescente ($j=1$)	**Adulto joven** ($j=2$)	**Anciano** ($j=3$)	$\mu_{i\bullet}$
Masculino ($i=1$)	9.75 (μ_{11})	12 (μ_{12})	17.25 (μ_{13})	14 ($\mu_{1\bullet}$)
Femenino ($i=2$)	8.25 (μ_{21})	10 (μ_{22})	14.75 (μ_{23})	10 ($\mu_{2\bullet}$)
$\mu_{\bullet j}$	9 ($\mu_{\bullet 1}$)	11 ($\mu_{\bullet 2}$)	16 ($\mu_{\bullet 3}$)	12 ($\mu_{\bullet\bullet}$)

7.4.1.1 Interacciones transformables y no transformables

Efectos de los factores multiplicativos

Consideremos el caso donde los efectos de los factores son multiplicativos, en lugar de aditivos:

$\mu_{ij} = \mu_{\bullet\bullet}\alpha_i\beta_j$

Estas interacciones pueden ser eliminadas aplicando la transformación logarítmica:

$$\log(\mu_{ij}) = \log(\mu_{\bullet\bullet}) + \log(\alpha_i) + \log(\beta_j)$$

o sea

$$u'_{ij} = u'_{\bullet\bullet} + \alpha_i^{'} + \beta_j^{'}$$

$$\mu_{ij}^{'} = log\mu_{ij} \quad \mu_{\bullet\bullet}^{'} = log\mu_{\bullet\bullet} \quad \alpha_i^{'} = \log\alpha_i \quad \beta_j^{'} = \log\beta_j$$

Entonces se usa la variable $Y' = \log Y$

Cuando una simple transformación de Y remueve los efectos de la interacción o los hace poco importantes, decimos que la interacción es *transformable.*

Interacciones multiplicativas

Otro ejemplo de interacciones transformables aparece cuando cada efecto de interacción es igual al producto de funciones de los efectos principales.

$$\mu_{ij} = \alpha_i + \beta_j + 2\sqrt{\alpha_i}\sqrt{\beta_j}$$

o lo que es lo mismo

$$\mu_{ij} = \left(\sqrt{\alpha_i}\sqrt{\beta_j}\right)^2$$

Si se aplica la transformación raíz cuadrada, se obtiene:

$$\mu_{ij}^{'} = \alpha_i^{'} + \beta_j^{'}$$

donde:

$$u^{'}{}_{ij} = \sqrt{\mu_{ij}} \quad \alpha_i^{'} = \sqrt{\alpha_i} \quad \beta_j^{'} = \sqrt{\beta_i}$$

Las transformaciones que convierten las interacciones importantes en no importantes son: cuadrado, raíz cuadrada, logaritmo y recíproca.

Table 7.6: Ejemplo de transformacion de medias de tratamientos: a) Medias de tratamientos escala original b) Medias de tratamientos después de la transformación $\surd$

factor A	factor B	
	$j=1$	$j=2$
$i=1$	16	64
$i=2$	49	121
$i=3$	64	144

factor A	factor B	
	$j=1$	$j=2$
$i=1$	4	8

factor A	factor B	
$i = 2$	7	11
$i = 3$	8	12

7.5 MODELO I PARA ESTUDIOS DE DOS FACTORES

El factor A es estudiado en I niveles, y estos no representan una muestra aleatoria de todos los niveles posibles de A. De manera equivalente el factor B se estudia sobre J niveles. Todas las IJ combinaciones de los niveles de los factores son incluidas en el análisis. El número de casos para cada uno de los IJ tratamientos es el mismo, lo indicamos con n, y es necesario que $n > 1$. Entonces el número total de casos es:

$$N = IJn$$

7.5.1 Modelo de las medias de celdas

Se puede expresar el modelo de niveles del factor fijos en términos de las medias de los tratamientos.

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

$$i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, n$$

7.5.1.1 Características importantes del Modelo

1. $E(Y_{ijk}) = \mu_{ij}$
2. $Var(Y_{ijk}) = Var(\varepsilon_{ijk}) = \sigma^2$
3. Y_ijk son independientes y se distribuyen $N\left(0, \sigma^2\right)$
4. El modelo de ANOVA es un modelo lineal.
5. El modelo de ANOVA de dos factores es similar al de un factor. Normalidad, independencia de los términos del error y constancia de la varianza para los términos del error son propiedades de ambos modelos.

7.5.2 Modelo de los efectos de los factores

Una forma equivalente de enunciar el modelo se obtiene de la definición de interacción:

$$(\alpha\beta)_{ij} = \mu_{ij}(\mu_{\bullet\bullet} + \alpha_i + \beta_j)$$

Reordenando términos se obtiene:

$$\mu_{ij} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

donde:

$$\mu_{\bullet\bullet} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}}{IJ}$$

$$\alpha_i = \mu_{i\bullet} - \mu_{\bullet\bullet}$$

$$\beta_j = \mu_{\bullet j} - \mu_{\bullet\bullet}$$

$$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} - \mu_{\bullet\bullet}$$

La media de la celda µij para cualquier tratamiento puede ser vista como la suma de cuatro componentes de los efectos de los factores. Específicamente:

1. Una constante general $\mu_{\bullet\bullet}$.
2. El efecto principal α_i del factor A en el i-ésimo nivel.
3. El efecto principal β_j del factor B en el j-ésimo nivel.
4. El efecto de la interacción $(\alpha\beta)_{ij}$

Reemplazando μ_{ij} en el modelo de las medias de las celdas, se obtiene:

$$Y_{ijk} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

donde:

$\mu_{\bullet\bullet}$ es una constante
α_i son constantes sujetas a la restricción $\sum \alpha_i = 0$
β_j son constantes sujetas a la restricción $\sum \beta_j = 0$
$(\alpha\beta)_{ij}$ son constantes sujetas a las restricciones

$$\sum_i^I (\alpha\beta)_{ij} = 0 \quad \sum_j^J (\alpha\beta)_{ij} = 0$$

ε_{ijk} son independientes y se distribuyen $N(0,\ \sigma^2)$

7.5.3 ANOVA (MODELO I)

7.5.3.1 Notación

$$\overline{Y}_{ij\bullet} = \frac{\sum_{k=1}^{n} Y_{ijk}}{n}$$

$$\overline{Y}_{i\bullet\bullet} = \frac{\sum_{j}^{J}\sum_{k=1}^{n} Y_{ijk}}{\mathrm{Jn}}$$

$$\overline{Y}_{\bullet j\bullet} = \frac{\sum_{i}^{I}\sum_{k=1}^{n} Y_{ijk}}{\mathrm{In}}$$

$$\overline{Y}_{\bullet\bullet\bullet} = \frac{\sum_{i}^{I}\sum_{j}^{J}\sum_{k=1}^{n} Y_{ijk}}{\mathrm{IJn}}$$

7.5.3.2 Ajuste del modelo

7.5.3.2.1 Modelo de las medias de celda

El cuadrado a minimizar es el siguiente:

$$\sum\sum\sum \left(Y_{ijk} - \mu_{ij}\right)^2$$

Minimizando se obtiene:

$$\hat{\mu}_{ij} = \overline{Y}_{ij\bullet}$$

Los residuos se definen como la diferencia entre los valores observados y los estimados:

$$\varepsilon_{ijk} = Y_{ijk} - \overline{Y}_{ij\bullet}$$

7.5.3.2.2 Modelo de los efectos del factor

El cuadrado a minimizar es el siguiente:

$$\sum\sum\sum \left(Y_{ijk} - \mu_{\bullet\bullet} - \alpha_i - \beta_j - (\alpha\beta)_{ij}\right)^2$$

sujeto a las restricciones:

$$\sum_i^I \alpha_i = 0 \quad \sum_j^J \beta_j = 0 \quad \sum_i^I (\alpha\beta)_{ij} = 0 \quad \sum_j^J (\alpha\beta)_{ij} = 0$$

Cuando se minimiza se obtienen los siguientes estimadores de mínimos cuadrados:

Parámetro	Estimador
$\mu_{\bullet\bullet}$	$\overline{Y}_{\bullet\bullet\bullet}$
$\alpha_i = \mu_{i\bullet} - \mu_{\bullet\bullet}$	$\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}$
$\beta_j = \mu_{\bullet j} - \mu_{\bullet\bullet}$	$\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}$
$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet}$	$\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}$

7.5.3.2.3 Descomposición de la suma de cuadrados total

Para una observación, se puede descomponer la desviación con respecto a la media total, en dos partes:

$$\underbrace{\left(Y_{ijk} - \overline{Y}_{\bullet\bullet\bullet}\right)}_{\text{Desviacion Total}} = \underbrace{\left(\overline{Y}_{ij\bullet} - \overline{Y}_{\bullet\bullet\bullet}\right)}_{\text{Desviacion de los tratamientos}} + \underbrace{\left(Y_{ijk} - \overline{Y}_{ij\bullet}\right)}_{\text{Desviacion de las observaciones}}$$

También se puede descomponer la desviación estimada de la media de los tratamientos en:

$$\underbrace{\overline{Y}_{ij\bullet} - \overline{Y}_{\bullet\bullet\bullet}}_{\substack{\text{Desviacion de} \\ \text{los tratamientos}}} = \underbrace{\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}}_{\substack{\text{Efecto principal} \\ \text{A}}} + \underbrace{\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}}_{\substack{\text{Efecto principal} \\ \text{B}}} + \underbrace{\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}}_{\substack{\text{Efecto de} \\ \text{la interaccion}}}$$

###TABLA DE ANOVA PARA DOS FACTORES. MODELO I

Más de una observación por celda

Fuente de variación	SC	GL	CM	E(CM)
Entre tratamientos	$SC_E = n\sum_{ij}\overline{Y}^2_{ij\bullet} - N\overline{Y}^2_{\bullet\bullet\bullet}$	$IJ-1$	$\frac{SC_E}{IJ-1}$	$\sigma^2 + \frac{n}{IJ-1}\sum\sum\left(\mu_{ij}-\mu_{\bullet\bullet}\right)^2$
A	$SC_A = Jn\sum_i \overline{Y}^2_{i\bullet\bullet} - N\overline{Y}^2_{\bullet\bullet\bullet}$	$I-1$	$\frac{SC_A}{I-1}$	$\sigma^2 + \frac{\mathrm{Jn}}{I-1}\sum\left(\mu_{i\bullet}-\mu_{\bullet\bullet}\right)^2$
B	$SC_B = In\sum_j \overline{Y}^2_{\bullet j\bullet} - N\overline{Y}^2_{\bullet\bullet\bullet}$	$J-1$	$\frac{SC_B}{J-1}$	$\sigma^2 + \frac{\mathrm{In}}{J-1}\sum\left(\mu_{\bullet j}-\mu_{\bullet\bullet}\right)^2$
AB (Interacción)	$SC_E - SC_A - SC_B$	$(I-1)(J-1)$	$\frac{SC_{AB}}{(I1)(J-1)}$	$\sigma^2 + \frac{n}{(I-1)(J-1)}\sum\sum\left(\mu_{ij}-\mu_{i\bullet}-\mu_{\bullet j}+\mu_{\bullet\bullet}\right)^2$
Error	$SC_T - SC_E$	$N-IJ$	$\frac{SC_D}{N-IJ}$	σ^2
Total	$\sum_i\sum_j\sum_k Y^2_{ijk} - N\overline{Y}^2_{\bullet\bullet\bullet}$	$N-1$		

7.6 Prueba de F

7.6.0.1 Prueba para la interacción

$$H_0 : \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} = 0 \ \forall\ i, j$$
$$H_a : \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} \neq 0 \ para\ algun\ i, j$$

o

$$H_0 : \text{todos los } (\alpha\beta)_{ij} = 0$$
$$H_a : \text{no todos los } (\alpha\beta)_{ij} = 0$$

La prueba estadística apropiada es:

$$F^* = \frac{CM_{AB}}{CM_D}$$

Recordamos que bajo H_0 F^* se distribuye según una $F_{1-\alpha;(I-1)(J-1),(N-IJ)}$.

Entonces:

- Sí $F^* \leq F_{1-\alpha;(I-1)(J-1),(N-IJ)}$, no se rechaza H_0
- Sí $F^* > F_{1-\alpha;(I-1)(J-1),(N-IJ)}$, se rechaza H_0

7.6.0.2 Prueba para los efectos principales

Estas pruebas se realizan cuando no existe interacción.

Para el factor A:

$$H_0 : \mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{I\bullet}$$
$$H_a : No\ todos\ los\ \mu_{i\bullet} \text{ son iguales}$$

o

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$$
$$H_a : \text{No todos los } \alpha_i \text{ iguales a cero}$$

Se usa el estadístico

$$F^* = \frac{CM_A}{CM_D}$$

Dado que F^*, bajo H_0, se distribuye según una $F_{1-\alpha;(I-1)(J-1),(N-IJ)}$.

Entonces:

- Sí $F^* \leq F_{1-\alpha;(I-1),(N-IJ)}$, no se rechaza H_0
- Sí $F^* > F_{1-\alpha;(I-1),(N-IJ)}$, se rechaza H_0

Para el factor B:

$$H_0 : \mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet J}$$
$$H_a : \text{No todos los } \mu_{\bullet j} \text{ son iguales}$$

o

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_J = 0$$
$$H_a : \text{No todos los } \beta_j \text{ iguales a cero}$$

El estadístico es

$$F^* = \frac{CM_B}{CM_D}$$

y la regla de decisión es:

- Sí $F^* \leq F_{1-\alpha;(J-1),(N-IJ)}$, no se rechaza H_0
- Sí $F^* > F_{1-\alpha;(J-1),(N-IJ)}$, se rechaza H_0

Ejemplo: El asma bronquial es una enfermedad alérgica cuya virulencia depende de la estación. Se desea comparar tres fármacos antihistamínicos A, B y C, en las cuatro estaciones del año. Se toma una muestra de 48 personas con asma crónico de intensidad análoga, que se divide en 12 grupos, uno para cada fármaco y estación, a razón de 4 enfermos por grupo. Los resultados se evaluaron en una escala objetiva que iba de 100, y fueron los siguientes:

Estación	Fármaco A	Fármaco B	Fármaco C
Primavera	23, 28, 32, 18	56, 58, 53, 55	42, 41, 36, 37
Verano	32, 41, 43, 48	64, 58, 67, 72	51, 53, 55, 60
Otoño	18, 16, 21, 10	48, 50, 47, 47	28, 31, 23, 33
Invierno	30, 40, 33, 47	60, 61, 63, 59	56, 60, 61, 55

a) Determinar si existen diferencias entre los fármacos A, B y C y entre las estaciones.

b) ¿Es significativa la interacción?

Table 7.11: Resumen de los datos de tres fármacos contra el asma en las cuatro estaciones del año.

Estación	Fármaco	n	Suma	Media	Varianza
Invierno	A	4	150	37.500	57.666667
Invierno	B	4	243	60.750	2.916667
Invierno	C	4	232	58.000	8.666667
Otoño	A	4	65	16.250	21.583333
Otoño	B	4	192	48.000	2.000000
Otoño	C	4	115	28.750	18.916667
Primavera	A	4	101	25.250	36.916667
Primavera	B	4	222	55.500	4.333333
Primavera	C	4	156	39.000	8.666667
Verano	A	4	164	41.000	44.666667
Verano	B	4	261	65.250	34.250000
Verano	C	4	219	54.750	14.916667
Total	A	16	480	30.000	135.866667
Total	B	16	918	57.375	52.650000
Total	C	16	722	45.125	160.650000

Tabla de ANOVA.

7.7 Contrastes

7.7.1 Entre Filas

$$\hat{f} = \sum c_i \overline{Y}_{i\bullet\bullet}$$

Bonferroni y Scheffé

$$\varepsilon = \frac{\left|\hat{f}\right|}{\sqrt{CM_D \ \frac{\sum c_i^2}{\text{Jn}}}}$$

7.7.1.1 Planeados:

7.7.1.1.1 Bonferroni

$$VC = t_{1-\frac{\alpha}{2m};GL_D}$$

7.7.1.2 No Planeados:

7.7.1.2.1 Scheffé

$$VC = S = \sqrt{(I-1)\, F_{(I-1);GL_D;1-\alpha}}$$

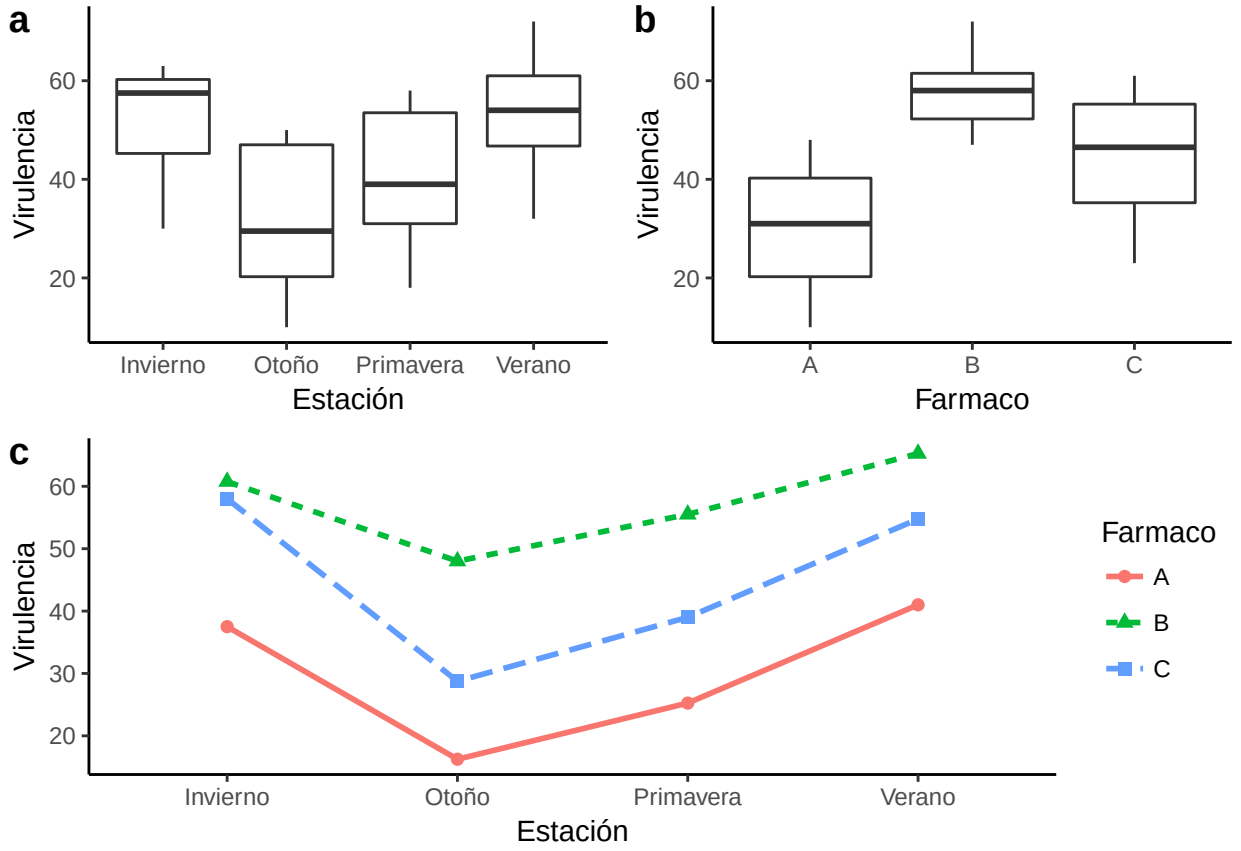


Figure 7.2: Efecto de tres fármacos contra el asma en las cuatro estaciones del año. a – gráfico de cajas y barras para las 4 estaciones, n = 12. b – gráfico de cajas y barras para los 3 fármacos, n = 16. c – gráfico de interacción fármaco x estación, n = 4.

7.7.1.2.2 Tukey

$$\frac{\overline{Y}_{i\bullet\max} - \overline{Y}_{i\bullet\min}}{S_{\overline{Y}}} \sim q_{I;N-IJ}$$

$$S_{\overline{y}} = \sqrt{\frac{CM_D}{\text{Jn}}}$$

7.7.1.2.3 Ortogonales

$$SC = \frac{\hat{f}^2}{\frac{\sum c_i^2}{\text{Jn}}}$$

7.7.2 Entre columnas

$$\hat{f} = \sum c_j \overline{Y}_{\bullet j\bullet}$$

Bonferroni y Scheffé

$$\varepsilon = \frac{\left|\hat{f}\right|}{\sqrt{CM_D \ \frac{\sum c_j^2}{\text{In}}}}$$

7.7.2.1 Planeados:

7.7.2.1.1 Bonferroni

$$VC = t_{1-\frac{\alpha}{2m};GL_D}$$

7.7.2.2 No Planeados:

7.7.2.2.1 Scheffé

$$VC = S = \sqrt{(J-1)\, F_{(J-1);GL_D;1-\alpha}}$$

7.7.2.2.2 Tukey

$$\frac{\overline{Y}_{\bullet j \ \max} - \overline{Y}_{\bullet j \ \min}}{S_{\overline{Y}}} \sim q_{j;N-IJ}$$

$$S_{\overline{Y}=\sqrt{\frac{CM_D}{In}}}$$

7.7.2.2.3 Ortogonales

$$SC = \frac{\hat{f}^2}{\frac{\sum c_j^2}{\text{In}}}$$

7.7.3 Interacción

$$\hat{f} = \sum c_{ij} \overline{Y}_{ij\bullet}$$

Bonferroni y Scheffé

$$\varepsilon = \frac{\left|\hat{f}\right|}{\sqrt{CM_D \ \frac{\sum c_{ij}^2}{n}}}$$

7.7.3.1 Planeados:

7.7.3.1.1 Bonferroni

$$VC = t_{1-\frac{\alpha}{2m};GL_D}$$

7.7.3.2 No Planeados:

7.7.3.2.1 Scheffé

$$VC = S = \sqrt{(IJ-1)\, F_{(IJ-1);N-IJ;1-\alpha}}$$

7.7.3.2.2 Tukey

$$\frac{\overline{Y}_{ij\,\max} - \overline{Y}_{ij\,\min}}{S_{\overline{Y}}} \sim q_{ij;N-IJ}$$

$$S_{\overline{Y}} = \sqrt{\frac{CM_D}{n}}$$

7.7.3.2.3 Ortogonales

$$SC = \frac{\hat{f}^2}{\frac{\sum c_{ij}^2}{n}}$$

$$f = \sum c_{ij}\mu_{ij} \quad \hat{f} = \sum c_{ij}\overline{Y}_{ij}$$

Ejemplo (continuación del anterior)

Expansión de los contrastes ortogonales

Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Estación	3	4132	1377	64.69	1.425e-14
Estación: Invierno vs Verano	1	15.04	15.04	0.7065	0.4062
Estación: Otoño vs Verano	1	3828	3828	179.8	1.442e-15
Estación: Primavera vs Verano	1	289	289	13.57	0.0007492
Farmaco	2	6017	3009	141.3	9.013e-18
Farmaco: A vs B	1	1830	1830	85.95	4.507e-11
Farmaco: B vs C	1	4187	4187	196.7	3.698e-16
Estación:Farmaco	6	338.8	56.47	2.652	0.03106
Estación:Farmaco: Invierno vs Verano.A vs B	1	45.56	45.56	2.14	0.1522
Estación:Farmaco: Otoño vs Verano.A vs B	1	28.52	28.52	1.34	0.2547
Estación:Farmaco: Primavera vs Verano.A vs B	1	5.042	5.042	0.2368	0.6295
Estación:Farmaco: Invierno vs Verano.B vs C	1	25.52	25.52	1.199	0.2809
Estación:Farmaco: Otoño vs Verano.B vs C	1	189.1	189.1	8.88	0.005141
Estación:Farmaco: Primavera vs Verano.B vs C	1	45.12	45.12	2.119	0.1541
Residuals	36	766.5	21.29	NA	NA

Contrastes planeados:

Invierno vs Verano

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Estación	3	4132	1377	64.69	1.425e-14
Estación: Invierno vs Verano	1	15.04	15.04	0.7065	0.4062
Farmaco	2	6017	3009	141.3	9.013e-18
Estación:Farmaco	6	338.8	56.47	2.652	0.03106
Residuals	36	766.5	21.29	NA	NA

B vs A-C

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Estación	3	4132	1377	64.69	1.425e-14
Farmaco	2	6017	3009	141.3	9.013e-18
Farmaco: B vs A-C	1	4187	4187	196.7	3.698e-16
Estación:Farmaco	6	338.8	56.47	2.652	0.03106
Residuals	36	766.5	21.29	NA	NA

C Otoño vs C Otros

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FxE	11	10488	953.5	44.78	1.178e-17
FxE: C Otoño vs C Otros	1	1430	1430	67.17	9.502e-10
Residuals	36	766.5	21.29	NA	NA

Contrastes no planeados: LSD vs Tukey

contrast	estimate	SE	df	t.ratio	p.value.lsd	p.value.tukey
Invierno,A - Otoño,A	21.25	3.263	36	6.513	1.442e-07	8.598e-06

contrast	estimate	SE	df	t.ratio	p.value.lsd	p.value.tukey
Invierno,A - Primavera,A	12.25	3.263	36	3.754	0.0006132	0.02591
Invierno,A - Verano,A	-3.5	3.263	36	-1.073	0.2905	0.9941
Invierno,A - Invierno,B	-23.25	3.263	36	-7.126	2.247e-08	1.368e-06
Invierno,A - Otoño,B	-10.5	3.263	36	-3.218	0.002731	0.09413
Invierno,A - Primavera,B	-18	3.263	36	-5.517	3.076e-06	0.0001732
Invierno,A - Verano,B	-27.75	3.263	36	-8.505	3.896e-10	2.439e-08
Invierno,A - Invierno,C	-20.5	3.263	36	-6.283	2.914e-07	1.72e-05
Invierno,A - Otoño,C	8.75	3.263	36	2.682	0.01099	0.2755
Invierno,A - Primavera,C	-1.5	3.263	36	-0.4597	0.6485	1
Invierno,A - Verano,C	-17.25	3.263	36	-5.287	6.237e-06	0.0003443
Otoño,A - Primavera,A	-9	3.263	36	-2.758	0.009071	0.2402
Otoño,A - Verano,A	-24.75	3.263	36	-7.586	5.687e-09	3.503e-07
Otoño,A - Invierno,B	-44.5	3.263	36	-13.64	8.623e-16	0
Otoño,A - Otoño,B	-31.75	3.263	36	-9.731	1.281e-11	8.121e-10
Otoño,A - Primavera,B	-39.25	3.263	36	-12.03	3.581e-14	2.068e-12
Otoño,A - Verano,B	-49	3.263	36	-15.02	4.457e-17	0
Otoño,A - Invierno,C	-41.75	3.263	36	-12.8	5.846e-15	1.286e-13
Otoño,A - Otoño,C	-12.5	3.263	36	-3.831	0.0004921	0.02126
Otoño,A - Primavera,C	-22.75	3.263	36	-6.973	3.567e-08	2.162e-06
Otoño,A - Verano,C	-38.5	3.263	36	-11.8	6.254e-14	3.79e-12
Primavera,A - Verano,A	-15.75	3.263	36	-4.827	2.545e-05	0.001336
Primavera,A - Invierno,B	-35.5	3.263	36	-10.88	6.221e-13	3.958e-11
Primavera,A - Otoño,B	-22.75	3.263	36	-6.973	3.567e-08	2.162e-06
Primavera,A - Primavera,B	-30.25	3.263	36	-9.271	4.507e-11	2.846e-09
Primavera,A - Verano,B	-40	3.263	36	-12.26	2.063e-14	1.089e-12
Primavera,A - Invierno,C	-32.75	3.263	36	-10.04	5.622e-12	3.572e-10
Primavera,A - Otoño,C	-3.5	3.263	36	-1.073	0.2905	0.9941
Primavera,A - Primavera,C	-13.75	3.263	36	-4.214	0.0001606	0.007619
Primavera,A - Verano,C	-29.5	3.263	36	-9.041	8.542e-11	5.382e-09
Verano,A - Invierno,B	-19.75	3.263	36	-6.053	5.903e-07	3.443e-05
Verano,A - Otoño,B	-7	3.263	36	-2.145	0.03874	0.596
Verano,A - Primavera,B	-14.5	3.263	36	-4.444	8.097e-05	0.004013
Verano,A - Verano,B	-24.25	3.263	36	-7.432	8.97e-09	5.505e-07
Verano,A - Invierno,C	-17	3.263	36	-5.21	7.891e-06	0.0004325
Verano,A - Otoño,C	12.25	3.263	36	3.754	0.0006132	0.02591
Verano,A - Primavera,C	2	3.263	36	0.613	0.5437	1
Verano,A - Verano,C	-13.75	3.263	36	-4.214	0.0001606	0.007619
Invierno,B - Otoño,B	12.75	3.263	36	3.908	0.0003943	0.0174
Invierno,B - Primavera,B	5.25	3.263	36	1.609	0.1163	0.8945
Invierno,B - Verano,B	-4.5	3.263	36	-1.379	0.1763	0.9603
Invierno,B - Invierno,C	2.75	3.263	36	0.8428	0.4049	0.9993
Invierno,B - Otoño,C	32	3.263	36	9.808	1.041e-11	6.606e-10
Invierno,B - Primavera,C	21.75	3.263	36	6.666	9.038e-08	5.421e-06
Invierno,B - Verano,C	6	3.263	36	1.839	0.07419	0.7864
Otoño,B - Primavera,B	-7.5	3.263	36	-2.299	0.02744	0.4955
Otoño,B - Verano,B	-17.25	3.263	36	-5.287	6.237e-06	0.0003443
Otoño,B - Invierno,C	-10	3.263	36	-3.065	0.004112	0.1312
Otoño,B - Otoño,C	19.25	3.263	36	5.9	9.457e-07	5.468e-05
Otoño,B - Primavera,C	9	3.263	36	2.758	0.009071	0.2402
Otoño,B - Verano,C	-6.75	3.263	36	-2.069	0.04581	0.6463
Primavera,B - Verano,B	-9.75	3.263	36	-2.988	0.00503	0.1539
Primavera,B - Invierno,C	-2.5	3.263	36	-0.7662	0.4485	0.9997

contrast	estimate	SE	df	t.ratio	p.value.lsd	p.value.tukey
Primavera,B - Otoño,C	26.75	3.263	36	8.198	9.42e-10	5.869e-08
Primavera,B - Primavera,C	16.5	3.263	36	5.057	1.262e-05	0.0006808
Primavera,B - Verano,C	0.75	3.263	36	0.2299	0.8195	1
Verano,B - Invierno,C	7.25	3.263	36	2.222	0.03266	0.5455
Verano,B - Otoño,C	36.5	3.263	36	11.19	2.858e-13	1.81e-11
Verano,B - Primavera,C	26.25	3.263	36	8.045	1.471e-09	9.141e-08
Verano,B - Verano,C	10.5	3.263	36	3.218	0.002731	0.09413
Invierno,C - Otoño,C	29.25	3.263	36	8.965	1.059e-10	6.664e-09
Invierno,C - Primavera,C	19	3.263	36	5.823	1.197e-06	6.889e-05
Invierno,C - Verano,C	3.25	3.263	36	0.9961	0.3259	0.9968
Otoño,C - Primavera,C	-10.25	3.263	36	-3.141	0.003355	0.1114
Otoño,C - Verano,C	-26	3.263	36	-7.969	1.84e-09	1.142e-07
Primavera,C - Verano,C	-15.75	3.263	36	-4.827	2.545e-05	0.001336

Tabla comparaciones múltiples LSD (triángulo superior), Tukey (triángulo inferior) y medias (diagonal)

	Invierno.A	Otoño.A	Primavera.A	Verano.A	Invierno.B	Otoño.B	Primavera.B	Verano.B	Invierno.C	Otoño.C	Primavera.C	Verano.C
Invierno.A	*37.5*	**1.442e-07**	**0.0006132**	0.2905	**2.247e-08**	**0.002731**	**1.076e-06**	**3.896e-10**	**2.914e-07**	**0.01099**	0.6485	**6.237e-06**
Otoño.A	**8.598e-06**	*16.25*	**0.009071**	**5.687e-09**	**8.623e-16**	**1.281e-11**	**3.581e-14**	**4.457e-17**	**5.846e-15**	**0.0004982**	**2.567e-08**	**6.254e-14**
Primavera.A	0.02591	**8.121e-10**	*25.25*	**2.545e-05**	**6.221e-13**	**3.567e-08**	**4.507e-11**	**2.063e-14**	**5.622e-12**	0.2905	**0.0001606**	**6.542e-11**
Verano.A	1.2402	**2.162e-06**	**2.439e-08**	*41*	**5.903e-07**	**0.03874**	**8.097e-05**	**8.97e-09**	**7.891e-06**	**0.0006132**	0.2437	**0.0001606**
Invierno.B	0.3941	0.596	**0**	**3.572e-10**	*60.75*	**0.0003943**	0.3163	0.1763	0.4049	**1.041e-11**	**9.038e-08**	0.07419
Otoño.B	**3.503e-07**	**0.01744**	**1.089e-12**	**0.0004325**	0.5941	*48*	**0.02744**	**6.237e-06**	**0.004112**	**2.457e-07**	**0.009071**	**0.04581**
Primavera.B	**0.0133**	**0.0001753**	**2.305e-07**	0.9993	**0.02591**	1	*55.5*	**0.00503**	0.4485	**9.42e-10**	**1.262e-05**	0.8195
Verano.B	**3.368e-06**	**2.068e-12**	0.9603	0.1312	**6.606e-10**	**2.162e-06**	**0.0006808**	*35.25*	**0.03266**	**2.858e-13**	**1.471e-09**	**0.002731**
Invierno.C	**0**	**2.846e-09**	**0.0003443**	0.9997	**5.468e-05**	**0.007619**	**9.141e-08**	**3.79e-12**	*58*	**1.059e-10**	**1.197e-06**	0.3259
Otoño.C	**3.958e-11**	**0.004013**	0.1539	0.5455	**5.869e-08**	1	**6.889e-05**	**5.382e-09**	0.6463	*28.75*	**0.003355**	**1.84e-09**
Primavera.C	**3.44e-05**	0.8945	**1.72e-05**	0.2755	**1.81e-11**	**5.421e-06**	0.1114	**0.007619**	1	0.9968	*39*	**2.545e-05**
Verano.C	0.09413	0.4955	**1.286e-13**	**0.02126**	**6.664e-09**	0.2402	**0.0003443**	0.7864	0.09413	**1.142e-07**	**0.001336**	*54.75*

7.8 Potencia de la prueba F

La potencia de la prueba F para la interacción, los efectos del factor principal A y los efectos del factor principal B puede ser evaluada de manera similar al caso del análisis de un solo factor *A* través de los gráficos de Pearson-Hartley. El parámetro de no centralidad Φ y los grados de libertad, para cada uno de estos casos son los siguientes:

7.8.1 Interacción

$$\Phi = \frac{1}{\sigma}\sqrt{\frac{n\sum\sum(\alpha\beta)_{ij}^2}{(I-1)(J-1)+1}} = \frac{1}{\sigma}\sqrt{\frac{n\sum\sum(\mu_{ij}-\mu_{i\bullet}-\mu_{\bullet j}+\mu_{\bullet\bullet})^2}{(I-1)(J-1)+1}}$$

$$\nu_1 = (I-1)(J-1) \quad \nu_2 = N - IJ$$

7.8.2 Prueba para el factor principal A:

$$\Phi = \frac{1}{\sigma}\sqrt{\frac{nJ\sum(\alpha)_i^2}{I}} = \frac{1}{\sigma}\sqrt{\frac{nJ\sum(\mu_{i\bullet}-\mu_{\bullet\bullet})^2}{I}}$$

$$\nu_1 = I - 1 \quad \nu_2 = N - IJ$$

7.8.3 Prueba para el factor principal B:

$$\Phi = \frac{1}{\sigma}\sqrt{\frac{nI\sum(\beta)_j^2}{J}} = \frac{1}{\sigma}\sqrt{\frac{nI\sum(\mu_{\bullet j}-\mu_{\bullet\bullet})^2}{J}}$$

$$\nu_1 = J - 1 \quad \nu_2 = N - IJ$$

Ejemplo: Cálculo de para la interacción:

$$\Phi = 1.50778432$$

$$\nu_1 = 6$$

$$\nu_2 = 36$$

$$P = 0.8$$

7.9 CASO DE UNA OBSERVACIÓN POR TRATAMIENTO

7.9.1 Modelo sin interacción

El modelo de ANOVA con niveles del factor fijos y sin interacción, para el caso en que n=1, es:

$$Y_{ij} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + \varepsilon_{ij}$$

Dado que el valor esperado del CM_{AB} es 2, en la prueba estadística F^* para ver la significación de los efectos principales se utiliza ahora el CM_{AB} en el denominador, en lugar del CM_D:

Efectos del factor principal A: $F^* = \frac{CM_A}{CM_{AB}}$

Efectos del factor principal B: $F^* = \frac{CM_B}{CM_{AB}}$

De manera similar para realizar contrastes, se reemplaza el CM_D por el CM_{AB} y se modifican los grados de libertad.

Tabla de ANOVA para el Modelo con Niveles Fijos del Factor sin Interacción, n = 1

Fte. de Variación	SC	GL	CM	E(CM)
factor A	$J\sum\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2$	$I-1$	$\frac{SC_A}{I-1}$	$\sigma^2+\frac{J}{I-1}\sum\left(\mu_{i\bullet}-\mu_{\bullet\bullet}\right)^2$
factor B	$I\sum\left(\overline{Y}_{\bullet j}-\overline{Y}_{\bullet\bullet}\right)^2$	$J-1$	$\frac{SC_B}{J-1}$	$\sigma^2+\frac{I}{J-1}\sum\left(\mu_{\bullet j}-\mu_{\bullet\bullet}\right)^2$
Error	$\sum\sum\left(Y_{ij}-\overline{Y}_{i\bullet}-\overline{Y}_{\bullet j}+\overline{Y}_{\bullet\bullet}\right)^2$	$(I-1)(J-1)$	$\frac{SC_{AB}}{(I-1)(J-1)}$	σ^2
Total	$\sum\sum\left(Y_{ij}-\overline{Y}_{\bullet\bullet}\right)^2$	$N-1$		

7.9.2 Prueba de Tukey (Aditividad)

Supongamos que se asume que:

$$(\alpha\beta)_{ij}=D\alpha_i\beta_j$$

donde D es alguna constante.

Usando la expresión del modelo de ANOVA con interacciones para el caso de n = 1, se tiene:

$$Y_{ij}=\mu_{\bullet\bullet}+\alpha_i+\beta_j+D\alpha_i\beta_j+\varepsilon_{ij}$$

Asumiendo que los otros parámetros son conocidos, el estimador de mínimos cuadrados de D es:

$$\hat{D}=\frac{\sum_i\sum_j\alpha_i\beta_jY_{ij}}{\sum_i\alpha_i^2\ \sum_j\beta_j^2}$$

El estimador de α_i es $\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}$, y de β_j es $\overline{Y}_{\bullet j}-\overline{Y}_{\bullet\bullet}$. Reemplazando los parámetros en $\hat{D}$ por sus estimadores, se obtiene:

$$\hat{D}=\frac{\sum_i\sum_j\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)\left(\overline{Y}_{\bullet j}-\overline{Y}_{\bullet\bullet}\right)Y_{ij}}{\sum_i\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2\sum_j\left(\overline{Y}_{j\bullet}-\overline{Y}_{\bullet\bullet}\right)^2}$$

Sustituyendo las estimaciones en $\sum\sum D^2\alpha_i^2\beta_j^2$, se obtiene la suma de cuadrados de la interacción:

$$\begin{aligned}SC_{AB}^{*}&=\sum_i\sum_j\hat{D}^2\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2\left(\overline{Y}_{\bullet j}-\overline{Y}_{\bullet\bullet}\right)^2\\&=\frac{\left[\sum_i\sum_j\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)\left(\overline{Y}_{\bullet j}-\overline{Y}_{\bullet\bullet}\right)Y_{ij}\right]^2}{\sum_i\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2\sum_j\left(\overline{Y}_{\bullet j}-\overline{Y}_{\bullet\bullet}\right)^2}\end{aligned}$$

La descomposición de la SCT para este caso especial de interacción es:

$$SC_T = SC_A + SC_B + SC^*_{AB} + SC^*_R$$

donde SC^*_R es la suma de cuadrados residual:

$$SC^*_R = SC_T - SC_A - SC_B - SC^*_{AB}$$

si $D = 0$, la prueba estadística:

$$F^* = \frac{\text{SC}^*_{AB}}{1} \div \frac{\text{SC}^*_R}{IJ - I - J} \sim F_{1;IJ-I-J}$$

Las hipótesis a poner a prueba son, entonces:

$H_0 : D = 0$ (no hay interacción)

$H_a : D \neq 0$ (hay interacción)

La regla de decisión es:

- Si $F^* \leq F_{1-\alpha;1;IJ-I-J}$, no se rechaza H_0.
- Si $F^* > F_{1-\alpha;1;IJ-I-J}$, se rechaza H_0.

Ejemplo: Se estudió la razón de la superficie a peso seco, para tres condiciones de sombra y tres especies de cítricos. Los resultados se presentan en la tabla:

	Naranja	Toronja	Mandarina
Sol	112	90	123
Media sombra	86	73	89
Sombra	80	62	81

Se desea saber si hay diferencias entre las especies y entre las condiciones de iluminación.

Prueba de Tukey

Se calculan las medias de los niveles de cada factor

	Naranja	Toronja	Mandarina	$\overline{Y}_{i\bullet}$
Sol	112	90	123	108.333333
Media sombra	86	73	89	82.6666667
Sombra	80	62	81	74.3333333
$\overline{Y}_{\bullet j}$	92.6666667	75	97.6666667	$\overline{Y}_{\bullet\bullet} = 88.4444444$

Se calcula la SC^*_{AB}

	Naranja	Toronja	Mandarina	$\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$
Sol	112	90	123	19.8888889
Media sombra	86	73	89	-5.77777778
Sombra	80	62	81	-14.1111111
$\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}$	4.22222222		9.22222222	
		13.4444444		

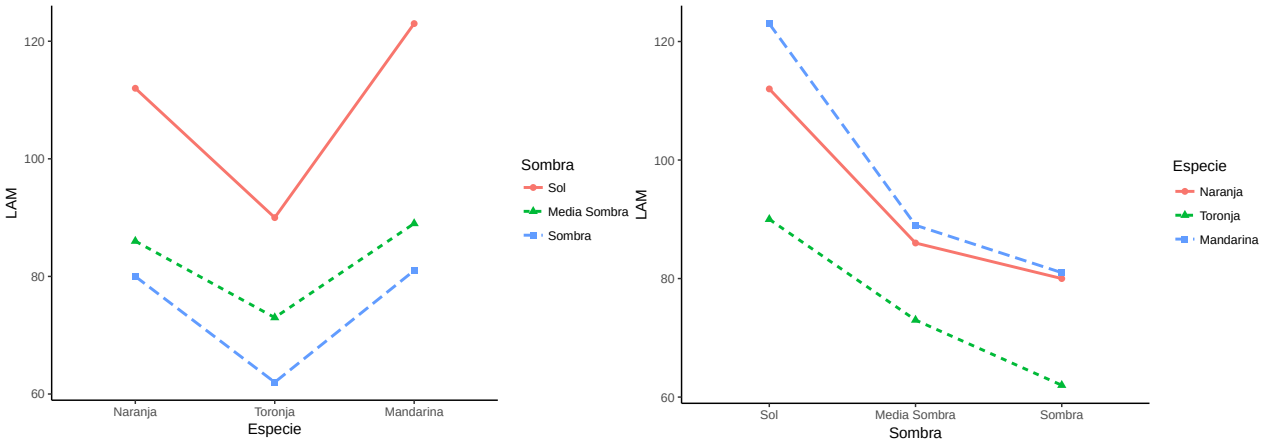


Figure 7.3: Graficos de perfiles de la razón de área foliar a peso de hoja para tres especies de citrícos bajo tres condiciones de luz.

$(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})Y_{ij}$		
9405.23457	-	22560.6296
		24065.5556
-2097.97531		5670.5679
		4742.2716
-4766.41975		11762.3951
		10541

$$\sum_i \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 = 628.074074$$

$$\sum_j \left(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}\right)^2 = 283.62963$$

$$SC^*_{AB} = 56.96674$$

Cálculo de SC^*_R

$$\begin{aligned} SC^*_R &= SC_T - SC_A - SC_B - SC^*_{AB} \\ &= 2822.2222 - 1884.22217 - 850.888916 - 56.96674 \\ &= 30.1443745 \end{aligned}$$

Cálculo de F^*, para poner a prueba la hipótesis de no interacción

$$F^* = \frac{SC^*_{AB}}{1} \div \frac{SC^*_R}{IJ - I - J} = 56.96674 \div \frac{30.1443745}{3 \times 3 - 3 - 3} = 5.66939016$$

El valor obtenido se compara con $F_{0.05;1,3} = 10.1279625$; con lo cual no se rechaza la hipótesis nula. El valor de p es 0.09752552.

En un gráfico de perfiles se observa:

Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Especie	2	850.9	425.4	19.54	0.008625
Sombra	2	1884	942.1	43.26	0.001953
Residuals	4	87.11	21.78	NA	NA

Análisis de varianza de dos factores con una sola muestra por grupo

RESUMEN	*Cuenta*	*Suma*	*Promedio*	*Varianza*
Sol	3	325	108.333333	282.333333
media sombra	3	248	82.6666667	72.3333333
sombra	3	223	74.3333333	114.333333
Naranja	3	278	92.6666667	289.333333
Toronja	3	225	75	199
Mandarina	3	293	97.6666667	497.333333

ANÁLISIS DE VARIANZA

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Filas	1884.222222	2	942.111111	43.26020408	0.00195266	6.944276265
Columnas	850.8888889	2	425.444444	19.53571429	0.00862465	6.944276265
Error	87.11111111	4	21.7777778			
Total	2822.222222	8				

7.10 MODELO II Y MODELO III PARA ESTUDIOS DE DOS FACTORES

7.10.1 Modelo aleatorio (Modelo II)

El modelo aleatorio para estudios de dos factores con igual tamaño muestral, n, es:

$$Y_{ijk} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

donde:

$\mu_{\bullet\bullet}$ es una constante

α_i, β_j , $(\alpha\beta)_{ij}$ son variables aleatorias independientes con distribución normal con media cero y varianzas σ^2_α, σ^2_β, $\sigma^2_{\alpha\beta}$ respectivamente.

ε_{ijk} son independientes $N(0, \sigma^2)$

α_i, β_j , $(\alpha\beta)_{ij}$, ε_{ijk} son independientes de a pares

$$i = 1, \ \ldots, \ I; \ j = 1, \ \ldots, \ J; \ k = 1, \ \ldots, \ n$$

Para este modelo de ANOVA el valor esperado de los Y_{ijk} es:

$$E\left(Y_{ijk}\right) = \mu_{\bullet\bullet}$$

y la varianza de los Y_{ijk}, indicada como σ^2_Y, es:

$$\text{Var}\left(Y_{ijk}\right) = \sigma^2_Y = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta} + \sigma^2$$

7.10.2 Modelo Mixto (Modelo III)

Para este caso, con igual tamaño muestral, el modelo de ANOVA es:

$$Y_{ijk} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

donde:

$\mu_{\bullet\bullet}$ es una constante

α_i son constantes sujetas a la restricción $\sum \alpha_i = 0$

β_j son independientes $N(0, \sigma^2_\beta)$

$(\alpha\beta)_{ij}$ son $N\left(0, \frac{I-1}{I}\sigma^2_{\alpha\beta}\right)$, sujetas a las restricciones:
a) $\sum (\alpha\beta)_{ij} = 0 \; \forall \; j$; b) $\text{Cov}\left[(\alpha\beta)_{ij}\,;(\alpha\beta)_{i'j}\right] = -\frac{1}{I}\sigma^2_{\alpha\beta} \; \forall \;\; i \neq i'$

ε_{ijk} son independientes y se distribuyen $N\left(0, \sigma^2\right)$

β_j, $(\alpha\beta)_{ij}$ y ε_{ijk} son independientes de a pares

$$i = 1, \; \ldots, \; I; \; j = 1, \; \ldots, \; J; \; k = 1, \; \ldots, \; n$$

Para el modelo mixto, el valor esperado de Y_{ijk} es:

$$E(Y_{ijk}) = \mu_{\bullet\bullet} + \alpha_i$$

y la varianza de Y_{ijk} es:

$$\text{Var}\left(Y_{ijk}\right) = \sigma^2_Y = \sigma^2_\beta + \frac{I-1}{I}\sigma^2_{\alpha\beta} + \sigma^2$$

7.10.3 Pruebas estadísticas

Table 7.26: Pruebas estadísticas para los Modelos Aleatorios y Mixto

Prueba para presencia de efectos de …	Modelo Fijo (A y B fijos)	Modelo Aleatorio (A y B aleatorio)	Modelo Mixto (A fijo y B aleatorio)
factor A	CM_A/CM_D	CM_A/CM_{AB}	CM_A/CM_{AB}
factor B	CM_B/CM_D	CM_B/CM_{AB}	CM_B/CM_D
Interacción AB	CM_{AB}/CM_D	CM_{AB}/CM_D	CM_{AB}/CM_D

7.10.4 Estimación de los componentes de la varianza

Con un modelo aleatorio, por ejemplo σ^2_α puede ser estimado por:

$$E\left(CM_A\right) - E\left(CM_{AB}\right) = \sigma^2 + nJ\sigma^2_\alpha + n\sigma^2_{\alpha\beta} - \left(\sigma^2 + n\sigma^2_{\alpha\beta}\right) = nJ\sigma^2_\alpha$$

De esta forma un estimador insesgado de σ^2_α es:

$$S^2_\alpha = \frac{CM_A - CM_{AB}}{\text{nJ}}$$

En un modelo mixto con el factor A fijo y el factor B aleatorio, se obtendrá:

$$S^2_I = \frac{CM_B - CM_D}{\text{nI}}$$

Table 7.27: Esperanza de los cuadrados medios en estudios de dos factores

CM	GL	Niveles del Factor Fijos (A y B fijos)	Niveles del factor Aleatorios (A y B aleatorios)	Niveles del Factor Mixtos (A fijo, B aleatorio)
CM_A	$I-1$	$\sigma^2 + nJ\frac{\sum \alpha_i^2}{I-1}$	$\sigma^2+nJ\sigma^2_\alpha+n\sigma^2_{\alpha\beta}$	$\sigma^2 + nJ\frac{\sum \alpha_i^2}{I-1} + n\sigma^2_{\alpha\beta}$
CM_B	$J-1$	$\sigma^2 + nI\frac{\sum \beta_j^2}{J-1}$	$\sigma^2+nI\sigma^2_\beta+n\sigma^2_{\alpha\beta}$	$\sigma^2 + nI\sigma^2_\beta$
CM_{AB}	$(I-1)(J-1)$	$\sigma^2 + nJ\frac{\sum (\alpha\beta)_{ij}^2}{(I-1)(J-1)}$	$\sigma^2 + n\sigma^2_{\alpha\beta}$	$\sigma^2 + n\sigma^2_{\alpha\beta}$
CM_D	$N-IJ$	σ^2	σ^2	σ^2

Ejemplos: 1) Se quiere estudiar el efecto de diferentes longitudes de onda (tratamiento) sobre la fertilidad en *D. Melanogaster*. Para ello se eligieron al azar 25 cepas y se sometieron a cuatro longitudes de onda diferentes elegidas también al azar. Para combinación longitud de onda-cepa se seleccionaron 12 hembras al azar y se registró la cantidad total de huevos puestos al cabo del cuarto día de postura, con los siguientes resultados:

	GL	CM	E(CM)
Entre cepas A	24	3243.00	
Entre tratamientos B	3	466.59	
Interacción AB	72	459.00	
Dentro (Error)	1100	231.00	

1. Completar el cuadro con la columna de las E(CM)
2. Hacer las pruebas de hipótesis correspondientes.
3. Calcular los estimadores de los componentes de la varianza.

a) y c)

CM	**Varianzas**	E(CM)	% de la Varianza
3243.00	58.0000	3243 = 231 + 12 * 4 * 58 + 12 * 19	18.82962211
466.59	0.0253	466.59 = 231 + 12 * 25 * 0.0253 * 12 * 19	0.00821361
459.00	19.0000	459 = 231 + 12 * 19	6.16832448
231.00	231.0000	231	74. 99383979

b)

Fte. de Variación	**GL**	**CM**	F*	p
Entre cepas A	24	3243.00	7.06535948	4.8134E -11
Entre tratamientos B	3	466.59	1.01653595	0.39049455
Interacción AB	72	459.00	1.98701299	3.9506E-06
Dentro (Error)	1100	231.00		

2) Para comparar la calidad periodística de tres periódicos A, B y C de ámbito nacional, se eligieron al azar 15 ciudades grandes del país. De cada ciudad se tomó una muestra de 10 lectores de A, 10 lectores de B y 10 lectores de C. A cada lector se le formularon una serie de preguntas cuyo resultado era una puntuación que indicaba la calidad del periódico. Tratados los datos mediante un análisis de la varianza, se obtuvo la siguiente tabla:

Fuente de Variación	GL	SC
Entre periódicos	2	14682
Entre ciudades	14	32712
Interacción	28	52570
Error (Dentro)	405	480308

1. Indicar que tipo de diseño se utilizó para realizar el análisis.
2. Estudiar si existe diferencia entre periódicos, entre ciudades y la significación de la interacción.

Es un diseño mixto de dos factores con interacción. El factor periódico es fijo, mientras que el factor ciudad es de efectos aleatorios (las ciudades han sido elegidas al azar).

Fuente de Variación	**GL**	**SC**	**CM**	**F***	**p**
Entre periódicos	2	14682	7341.000000	3.909987	0.03180
Entre ciudades	14	32712	2336.571429	0.068106	0.99999
Interacción	28	52570	1877.500000	1.583125	0.03119
Error (Dentro)	405	480308	1185.945679		

Un estimador de la varianza del factor ciudades es:

$$S^2_\beta = \frac{CM_B - CM_D}{nI} = \frac{2336.571429 - 1185.945679}{10 \times 3} = 38.35419165$$

Chapter 8

Prueba de Wilcoxon-Mann-Whitney para dos pruebas independientes

Esta prueba es el análogo más directo de la prueba de t para dos muestras independientes. Existen dos variantes de esta prueba que utilizan distintos estadísticos (W y U), cada uno de los cuales posee su propia tabla de valores críticos. Sin embargo, ambas variantes dan idénticos resultados debido a que son formas diferentes de utilizar y evaluar de igual manera la misma información. Cuando se emplea el estadístico U esta prueba suele ser denominada también como Prueba U de Mann-Whitney.

8.1 Datos

Para utilizar esta prueba es necesario contar con dos muestras aleatorias de las dos poblaciones a comparar, tal que $(X_1, X_2, , X_{N1})$ es la muestra aleatoria de tamaño N_1 de la población 1 y $(Y_1, Y_2, , Y_{N2})$ es la muestra aleatoria de tamaño N_2 de la población 2. La función de distribución en probabilidades de la población 1 es $F(x)$ y la función de distribución en probabilidades de la población 2 es $G(x)$.

8.2 Supuestos

Los supuestos de esta prueba son: 1. Ambas muestras son muestras aleatorias de las respectivas poblaciones. 2. Los datos son independientes tanto dentro de cada muestra como entre muestras. 3. Las mediciones han sido tomadas utilizando al menos una escala ordinal. 4. Si existen diferencias en las funciones de distribución de ambas poblaciones, estas diferencias están asociadas a la localización de las distribuciones. Esto significa que $F(x) = G(x + c)$, donde c es una constante. Este supuesto sólo es necesario cuando la hipótesis a poner a prueba está asociada a las $E(X)$ y $E(Y)$.

8.3 Procedimiento básico

Los datos combinados de ambas muestras deben ser ordenados de menor a mayor. A estos datos ordenados se les deben asignar rangos desde 1 a $N1 + N2$. En caso de existir empates, debe asignarse el rango promedio de los rangos correspondientes. Luego de la asignación de los rangos, deben obtenerse las sumas de los rangos de cada muestra como:

Table 8.1: Datos de dos poblaciones de ejemplo

Pob1	Pob2
0	5
1	6
1	3
2	2
3	2
2	0
2	2
4	3
3	6
0	3

$$R_1 = \sum_{i=1}^{N_1} R(X_i)$$
$$R_2 = \sum_{i=1}^{N_2} R(Y_i)$$

donde R_1 es la suma de los rangos asignados a la muestra de la población 1, R_2 es la suma de los rangos asignados a la muestra de la población 2, $R(X_i)$ es el rango asignado al i-ésimo dato de la muestra de la población 1 y $R(Y_i)$ es el rango asignado al i-ésimo dato de la muestra de la población 2. Igualmente, los valores R_1 y R_2 se encuentran relacionados, de tal manera que: $R_1 + R_2 = (N_1 + N_2)(N_1 + N_2 + 1)/2$ Esto implica que conocidos N_1, N_2 y uno de los R_i es posible hallar el otro.

8.3.1 Ejemplo

Supongamos que obtenemos los siguientes valores para la población 1 y 2:

A continuación debemos reunir esos datos en una solo conjunto, agregando una identificación [1]

A continuación se ordenan y se la asigna un número de 1 hasta $N_1 + N_2$ según el orden (columna `Rango`)[2]. Los empates deben tratarse de forma especial ya que no hay forma de decidir que número va primero. Por lo tanto se promedian los valores de sus rangos (columna `Rango_Empates`)

```
datos_long <- datos_long %>%
  mutate(Rango = rank(Valor, ties.method = "r")) %>%
  arrange(Rango) %>%
  mutate(Rango_Empates = rank(Valor))
kable(datos_long)
```

[1] Esto se puede hacer sin muchas vueltas con la función `gather()` del paquete `tidyr`.
[2] Ver la función `rank()`

Table 8.2: Datos de ambas poblaciones juntas

Pob	Valor
Pob1	0
Pob1	1
Pob1	1
Pob1	2
Pob1	3
Pob1	2
Pob1	2
Pob1	4
Pob1	3
Pob1	0
Pob2	5
Pob2	6
Pob2	3
Pob2	2
Pob2	2
Pob2	0
Pob2	2
Pob2	3
Pob2	6
Pob2	3

Pob	Valor	Rango	Rango_Empates
Pob1	0	1	2.0
Pob2	0	2	2.0
Pob1	0	3	2.0
Pob1	1	4	4.5
Pob1	1	5	4.5
Pob2	2	6	8.5
Pob1	2	7	8.5
Pob2	2	8	8.5
Pob1	2	9	8.5
Pob1	2	10	8.5
Pob2	2	11	8.5
Pob2	3	12	14.0
Pob1	3	13	14.0
Pob2	3	14	14.0
Pob1	3	15	14.0
Pob2	3	16	14.0
Pob1	4	17	17.0
Pob2	5	18	18.0
Pob2	6	19	19.5
Pob2	6	20	19.5

8.4 Estadísticos

8.4.1 Variante Wilcoxon (W)

8.4.1.1 Sin empates

Si no existen empates o si estos son pocos, el estadístico a calcular es:

$$W = R_1$$

Los valores críticos para este estadístico (w_p) pueden obtenerse de la Tabla A7. En esta tabla se presentan los valores críticos (cuantiles) para $1 - \alpha$ en el rango 0.001-0.5. Los valores para el rango 0.5-0.999 pueden obtenerse como:

$$w_{1-p} = N_1(N_1 + N_2 + 1) - w_p$$

La Tabla A7 permite obtener los valores críticos para muestras donde $N_1 \leq 20$ y $N_2 \leq 20$. En el caso de que N_1 y/o N_2 fueran mayores que 20, los valores críticos pueden obtenerse mediante una aproximación normal de la forma:

$$w_p \cong \frac{N_1(N_1 + N_2 + 1)}{2} + z_p\sqrt{\frac{N_1 N_2(N_1 + N_2 + 1)}{12}}$$

8.4.1.2 Con empates

Si existen numerosos empates, el estadístico a calcular es:

$$W_1 = \frac{W - \frac{N_1(N_1+N_2+1)}{2}}{\sqrt{\frac{N_1 N_2}{(N_1+N_2)(N_1+N_2-1)}\sum_{i=1}^{N_1+N_2} R_i^2 - \frac{N_1 N_2 (N_1+N_2+1)^2}{4(N_1+N_2-1)}}}$$

donde R_i^2 son los $R(X_i)$ y $R(Y_i)$ elevados al cuadrado. Este estadístico se distribuye según una distribución normal estándar.

8.4.2 Variante Mann-Whitney (U)

Se calculan dos estadísticos, el U y el U. Ambos estadísticos se obtienen de igual forma como:

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1$$
$$U' = N_2 N_1 + \frac{N_2(N_2+1)}{2} - R_2$$

Además, cualquiera de ellos puede obtenerse a partir del otro teniendo en cuenta que la relación entre ellos es:

$$U = N_1 N_2 - U'$$

Cómo puede verse a partir de las ecuaciones que definen a U y U, puede considerarse que U es el estadístico asociado a la población 1, mientras que U es el estadístico asociado a la población 2. Igualmente, dado que la asignación de población 1 y población 2 es puramente arbitraria, los estadísticos U y U pueden corresponder indistintamente a cualquiera de las dos muestras a comparar. Por una razón exclusivamente operativa y relacionada con la tabla de valores críticos que se utilizará, se considera como población 1 a la que posea el mayor tamaño muestral y como población 2 a la que posea el menor tamaño muestral ($N_1 < N_2$). Los valores

críticos para los estadísticos U o U' pueden obtenerse utilizando la Tabla U para muestras donde $N_1 \leq 20$ y $N_2 \leq 20$. Para tamaños muestrales mayores, pueden utilizarse aproximaciones normales. Dependiendo de la existencia o no de empates los estadísticos a calcular son:

8.4.2.1 Sin empates

$$Z = \frac{U - \frac{(N_1 N_2)}{2}}{\sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}}$$

8.4.2.2 Con empates

Previamente al cálculo del estadístico y considerando la existencia de E grupos de datos empatados debe obtenerse el valor $\sum e$ tal que:

$$\sum e = \sum_{i=1}^{E} (\varepsilon_i^3 - \varepsilon_i)$$

donde ε_i es el número de datos empatados en el grupo i-ésimo de datos empatados.

Una vez obtenido el $\sum e$, el estadístico corregido por empates (z_c) puede obtenerse como:

$$Z_c = \frac{U - \frac{(N_1 N_2)}{2}}{\sqrt{\left[\frac{N_1 N_2}{(N_1+N_2)^2 - (N_1+N_2)}\right]\left[\frac{(N_1+N_2)^3 - (N_1+N_2) - \sum e)}{12}\right]}}$$

Ambos estadísticos (z y z_c) se distribuyen de acuerdo a una distribución normal estándar.

8.5 Hipótesis

En términos generales, la prueba de Wilcoxon-Mann-Whitney se utiliza para poner a prueba hipótesis sobre las distribuciones $F(x)$ y $G(x)$. Como la prueba es sensible a diferencias en las tendencias centrales puede emplearse para poner a prueba hipótesis sobre $E(X)$ y $E(Y)$.

8.5.1 Prueba a dos colas

$H_0 : F(x) = G(x)$ para todo x ó $E(X) = E(Y)$
$H_a : F(x)G(x)$ para algún x ó $E(X)E(Y)$

8.5.1.1 Variante Wilcoxon*

Utilizando el estadístico W los criterios de decisión son:
Si $W \leq w_{\alpha/2}$ ó $W \geq w_{1-\alpha/2}$ Entonces **Rechazo** H_0
Si $w_{\alpha/2} < W < w_{1-\alpha/2}$ Entonces **No rechazo** H_0
Para evitar calcular el $w_{1-\alpha/2}$, puede calcularse un estadístico W como:

$$W' = N_1(N_1 + N_2 + 1) - W$$

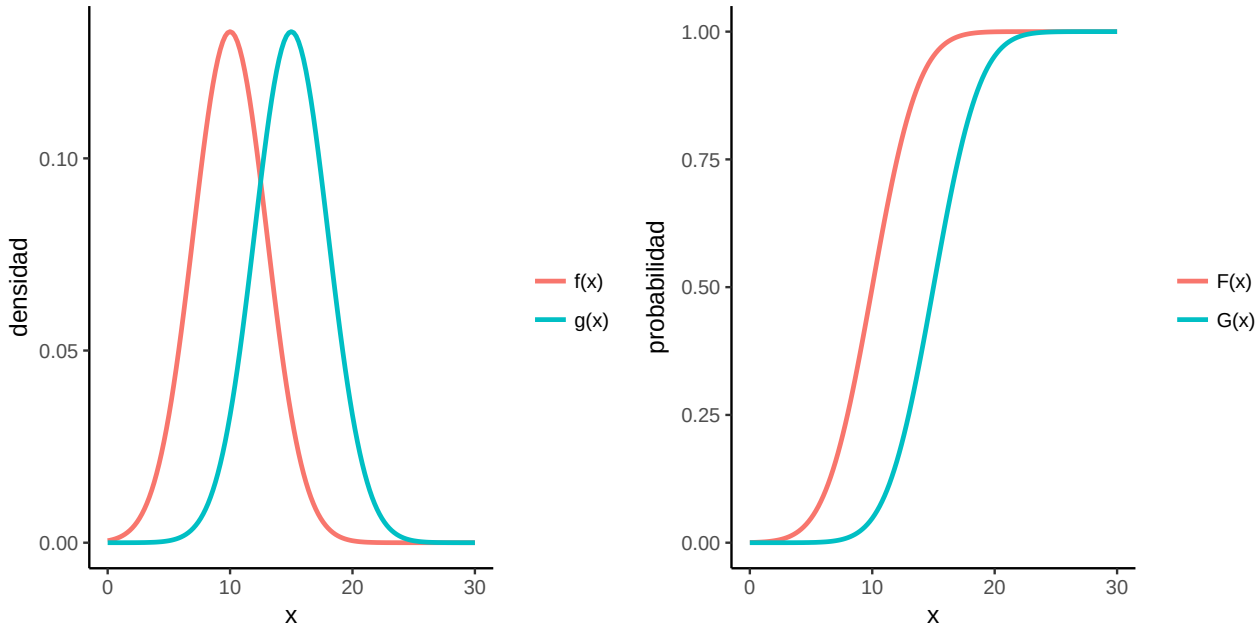


Figure 8.1: Funciones de densidad $[f(x)$ y $g(x)]$ y funciones de distribución $[F(x)$ y $G(x)]$ para dos variables aleatorias que se distribuyen normalmente con $\sigma^2 = 3$ y que sólo difieren en el valor de $\mu = 10$ y $\mu = 15$.

quedando definidos los criterios de decisión como:

Si $W \leq w_{\alpha/2}$ ó $W \leq w_{\alpha/2}$ Entonces **Rechazo** H_0
Si $W > w_{\alpha/2}$ y $W > w_{\alpha/2}$ Entonces **No rechazo** H_0

Utilizando el estadístico $W1$ los criterios de decisión son:

Si $2 * [1 - P(ZW_1)] \leq \alpha$ Entonces **Rechazo** H_0 Si $2 * [1 - P(ZW_1)] > \alpha$ Entonces **No rechazo** H_0

8.5.1.2 Variante Mann-Whitney

Definiendo como estadístico U al mayor entre U y U, los criterios de decisión son:
Si $UU_{N1,N2,\alpha/2}$ Entonces **Rechazo** H_0
Si $U < U_{N1,N2,\alpha/2}$ Entonces **No rechazo** H_0
Donde $U_{N1,N2,\alpha/2}$ es el valor crítico obtenido de la Tabla U.

Utilizando la aproximación normal, ya sea sin empates (z) o con corrección por empates (z_c), los criterios de decisión son:
Si $2 * [1 - P(Zz)]\alpha$ Entonces **Rechazo** H_0
Si $2 * [1 - P(Zz)] > \alpha$ Entonces **No rechazo** H_0

8.5.2 Prueba de una cola a la izquierda

H_0: $F(x) = G(x)$ para todo x ó $E(X) = E(Y)$ H_a: $F(x) > G(x)$ para algún x ó $E(X) < E(Y)$

Para visualizar la relación entre las hipótesis planteadas en términos de las funciones de distribución y las hipótesis planteadas en términos de las esperanzas puede utilizarse la 8.1. En esta figura muestra las funciones de densidad y de distribución para dos variables normales que difieren solamente en el valor de su media.

8.5.2.1 Variante Wilcoxon

Utilizando el estadístico W los criterios de decisión son:
Si Ww_α Entonces **Rechazo** H_0
Si $W > w_\alpha$ Entonces **No rechazo** H_0

Utilizando el estadístico W_1 los criterios de decisión son:
Si $P(ZW_1)\alpha$ Entonces **Rechazo** H_0
Si $P(Z > W_1) > \alpha$ Entonces **No rechazo** H_0

8.5.2.2 Variante Mann-Whitney

Definiendo como estadístico U al U, los criterios de decisión son:
Si $UU_{N1,N2,\alpha}$ Entonces **Rechazo** H_0
Si $U < U_{N1,N2,\alpha}$ Entonces **No rechazo** H_0
Donde $U_{N1,N2,\alpha}$ es el valor crítico obtenido de la Tabla U.

Utilizando la aproximación normal, ya sea sin empates (z) o con corrección por empates (z_c), los criterios de decisión son:
Si $1 - P(Z \leq z) \leq \alpha$ ó $1 - P(Zz_c) \leq \alpha$ Entonces **Rechazo** H_0
Si $1 - P(Zz) > \alpha$ ó $1 - P(Zz_c) > \alpha$ Entonces **No rechazo** H_0

8.5.3 Prueba de una cola a la derecha

H_0: $F(x) = G(x)$ para todo x ó $E(X) = E(Y)$
H_a: $F(x) < G(x)$ para algún x ó $E(X) > E(Y)$

8.5.3.1 Variante Wilcoxon

Utilizando el estadístico W los criterios de decisión son:
Si $Ww_{1-\alpha}$ Entonces **Rechazo** H_0
Si $W < w_{1-\alpha}$ Entonces **No rechazo** H_0

También puede utilizarse el estadístico W, quedando entonces los criterios de decisión como:
Si $W'w_\alpha$ Entonces **Rechazo** H_0
Si $W' > w_\alpha$ Entonces **No rechazo** H_0

Utilizando el estadístico W_1 los criterios de decisión son:
Si $P(ZW_1)\alpha$ Entonces **Rechazo** H_0
Si $P(ZW_1) > \alpha$ Entonces **No rechazo** H_0

8.5.3.2 Variante Mann-Whitney

Definiendo como estadístico U al U, los criterios de decisión son:
Si $UU_{N1,N2,\alpha}$ Entonces **Rechazo** H_0
Si $U < U_{N1,N2,\alpha}$ Entonces **No rechazo** H_0
Donde $U_{N1,N2,\alpha}$ es el valor crítico obtenido de la Tabla U.

Utilizando la aproximación normal, ya sea sin empates (z) o con corrección por empates (z_c), los criterios de decisión son:
Si $1 - P(Zz)\alpha$ ó $1 - P(Zz_c)\alpha$ Entonces **Rechazo** H_0
Si $1 - P(Zz) > \alpha$ ó $1 - P(Zz_c) > \alpha$ Entonces **No rechazo** H_0

Table 8.3: Largo total (cm) de las raneyas consumidas por machos y hembras del lobo marino

Machos	Hembras
15	9
10	8
14	5
17	5
16	5
11	9
15	9
20	10
13	13
	13
	15
	5
	12
	6

8.6 Ejemplo 2

Un ecólogo desea comparar las tallas de raneya (pez demersal bentónico) consumidas por machos y hembras del lobo marino. Los largos totales de las raneyas fueron estimados a partir de los otolitos hallados en los estómagos. Debido a que las hembras son predadores más costeros que los machos y las áreas costeras son áreas de cría para diversas especies de peces, el biólogo especula que las raneyas consumidas por las hembras deberían ser más pequeñas que las consumidas por los machos. Los datos obtenidos fueron:

Teniendo en cuenta que el tamaño muestral de las hembras es mayor que el de los machos, se considera como población 1 (X) a las hembras y como población 2 (Y) a los machos. En función de esto, las hipótesis para este enunciado son:

H_0: $F(x) = G(x)$
H_a: $F(x) > G(x)$ ó $E(X) < E(Y)$

Los cálculos realizados son:

Sexo	LT	R_i	e
Hembras	5	2.5	60
Hembras	5	2.5	0
Hembras	5	2.5	0
Hembras	5	2.5	0
Hembras	6	5.0	0
Hembras	8	6.0	0
Hembras	9	8.0	24
Hembras	9	8.0	0
Hembras	9	8.0	0
Machos	10	10.5	6
Hembras	10	10.5	0
Machos	11	12.0	0
Hembras	12	13.0	0
Machos	13	15.0	24
Hembras	13	15.0	0
Hembras	13	15.0	0
Machos	14	17.0	0
Machos	15	19.0	24
Machos	15	19.0	0
Hembras	15	19.0	0
Machos	16	21.0	0
Machos	17	22.0	0
Machos	20	23.0	0

Variante Wilcoxon		
W	117,5	
W'	218,5	
1 cola	w14;9;0,05	142
2 colas	w14;9;0,025	137
	w14;9;0,975	199
W1	-3,19943151	
1 cola	Valor p	0,00068856
2 colas	Valor p	0,00137712

Variante Mann-Whitney		
2 colas	U	113,5
	$U_{14;9;0,025}$	95
1 cola	U	113,5
	$U_{14;9;0,05}$	90
	z	3,18120098
1 cola	Valor p	0,00073339
2 colas	Valor p	0,00146679
	z_c	3,19943151
1 cola	Valor p	0,00068856
2 colas	Valor p	0,00137712

De acuerdo a los resultados obtenidos, es posible rechazar la hipótesis nula.
Considerando la hipótesis a dos colas, es posible rechazar la hipótesis nula.

Esta prueba puede realizarse con R. Se encuentra en la función wilcox.test() en el paquete stats.

Hay que tener en cuenta que la definición del estadístico W en R es diferente. En este caso es la suma de rangos de la primer población *menos* el mínimo. Por lo tanto, el estadístico es el mismo que el U de Mann-Whitney.

Los resultados para el ejemplo son:

```
wilcox.test(datos_ejemplo2$Machos, datos_ejemplo2$Hembras, alternative = "greater")
```

```
## Warning in wilcox.test.default(datos_ejemplo2$Machos,
## datos_ejemplo2$Hembras, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  datos_ejemplo2$Machos and datos_ejemplo2$Hembras
## W = 113.5, p-value = 0.0007681
## alternative hypothesis: true location shift is greater than 0
```

Nótese que la prueba da el valor de la variante Mann-Whitney. Las mismas consideraciones que se hicieron para la prueba de rangos de Wilcoxon en la sección anterior deben hacerse aquí. Es decir, especificar la cola a usar y rechazar cuando p

8.7 Prueba de Wilcoxon de rangos con signo para muestras apareadas

Esta prueba implica comparar una serie de pares ordenados $(x_i,\ y_i)$ obtenidos de una serie de variables aleatorias bivariadas $(X_i,\ Y_i)$. Básicamente la prueba se basa en la asignación de rangos a las diferencias obtenidas D_i para cada uno de los pares ordenados de la muestra. Por lo tanto, dada una muestra aleatoria de N pares ordenados, se calculan las diferencias como:

$$D_i = x_i - y_i \text{ para } i = 1, 2, 3, \ldots, N$$

Una vez obtenidas las D_i, se descartan todas las diferencias donde $D_i = 0$. De esta forma el tamaño muestral real para la prueba (n) se reduce de tal forma que $n \leq N$. Posteriormente se ordenan las D_i remanentes de menor a mayor teniendo en cuenta el valor absoluto de las diferencias $(|D_i|)$. Luego, se asignan los rangos a las $|D_i|$ desde 1 hasta n. En aquellos casos donde las $|D_i|$ son iguales (empates), se les asigna a todas las diferencias empatadas el promedio de los rangos correspondientes. Finalmente, se les asigna a los rangos el signo correspondiente a las D_i obteniéndose de esta forma los rangos con signo R_i.

8.7.1 Supuestos

La prueba de Wilcoxon tiene los siguientes supuestos:

La distribución de los D_i es simétrica.

Las D_i son mutuamente independientes.

Todas las D_i tiene igual media.

Las D_i están medidas utilizando como mínimo una escala de intervalo.

8.7.2 Estadísticos

Para esta prueba se calculan dos estadísticos, el T^+ y el T^-. Ambos estadísticos se calculan como:

$$\begin{array}{ll} T^+ = |\sum R_i| & \text{si } R_i > 0 \\ T^- = |\sum R_i| & \text{si } R_i < 0 \end{array}$$

Los valores críticos para estos estadísticos wp se obtienen de una tabla construida *ad hoc* para esta prueba (Tabla A12). En esta tabla se presentan los valores críticos (cuantiles) para $1-\alpha$ en el rango 0.005-0.5. Los valores para el rango 0.5-0.995 pueden obtenerse como:

$$w_p = \frac{n\,(n+1)}{2} - w_{1-p}$$

presentándose también en 12 el primer término de esta diferencia para facilitar el cálculo de los valores críticos. El n es el tamaño muestral efectivo de la prueba.

Para $n > 50$ o para aquellos casos con numerosos empates, puede obtenerse un estadístico que se distribuye en forma aproximadamente normal y que se calcula como:

$$Z = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$$

8.7.3 Hipótesis

8.7.3.1 Prueba a dos colas

La prueba a dos colas implica poner a prueba las siguientes hipótesis:

$H_0:\ E(D) = 0$ ó $E(X) = E(Y)$

$H_a:\ E\,(D) \neq 0$ ó $E\,(X) \neq E(Y)$

Para poner a prueba esta hipótesis se utiliza el menor de los estadísticos T+ y T-, comparando el estadístico con los valores críticos $w_{\alpha/2}$ y $w_{1-\alpha/2}$.

Si $T^{\pm} > \ w_{1-\frac{\alpha}{2}}$ ó $T^{\pm} < w_{\frac{\alpha}{2}}$ Entonces Rechazo H_0

Si $w_{\frac{\alpha}{2}} \ < T^{\pm} < \ w_{1-\frac{\alpha}{2}}$ Entonces No rechazo H_0

Empleando la aproximación normal, si

Si $2 * [1 - P\,(z \leq |Z|)] \leq \alpha$ Entonces Rechazo H_0

Si $2 * [1 - P\,(z \leq |Z|)] > \alpha$ Entonces No rechazo H_0

8.7.3.2 Prueba de una cola a la izquierda

La prueba de una cola a la izquierda implica poner a prueba las siguientes hipótesis:

$H_0:\ E\,(D) \geq 0$ ó $E\,(X) > E(Y)$

$H_a:\ E(D) < 0$ ó $E(X) < E(Y)$

Para poner a prueba esta hipótesis se utiliza el estadístico T+, comparando el estadístico con el valor crítico w_α

Si $T^+ < w_\alpha$ Entonces Rechazo H_0

Si $T^+ \geq$ wα Entonces No rechazo H_0

Empleando la aproximación normal, si

Si $1 - P\left(z \leq |Z|\right) \leq \alpha$ Entonces Rechazo H_0

Si $1 - P\left(z \leq |Z|\right) > \alpha$ Entonces No rechazo H_0

8.7.3.3 Prueba de una cola a la derecha

La prueba de una cola a la derecha implica poner a prueba las siguientes hipótesis:

$H_0: \ E\left(D\right) \leq 0$ ó $E\left(X\right) \leq E(Y)$

$H_a: \ E(D) > 0$ ó $E(X) > E(Y)$

Para poner a prueba esta hipótesis se utiliza el estadístico T+, comparando el estadístico con el valor crítico w1-

Si $T^+ > \ w_{1-\alpha}$ Entonces Rechazo H_0

Si $T^+ \leq w_{1-\alpha}$Entonces No rechazo H_0

Empleando la aproximación normal, si

Si $1 - P\left(z \leq |Z|\right) \leq \alpha$ Entonces Rechazo H_0

Si $1 - P\left(z \leq |Z|\right) > \alpha$ Entonces No rechazo H_0

Ejemplo 1: Una especie de ave pone dos huevos por temporada reproductiva. Un etólogo desea evaluar si el primer pichón que eclosiona presenta un comportamiento menos agresivo que el segundo. Para ello, tomó 12 nidos al azar y marcó a los pichones para poder identificar cuál de ellos fue el primero en eclosionar. Asimismo, registró el número de "peleas entre hermanos" iniciadas por cada pichón. Este registro fue iniciado a partir del momento en que ambos pichones estaban en condiciones de iniciar una pelea y finalizó cuando alguno de los pichones abandonó el nido. Los datos obtenidos fueron:

Table 8.6: Número de peleas iniciadas

Pichón 1 X	Pichón 2 Y
96	91
65	77
80	71
72	87
76	77
72	72
81	88
88	86
90	91
64	68
77	71
65	70

$H_0: \ E\left(X\right) \geq E(Y)$ ó $E\left(D\right) > 0$

$H_a: \ E\left(X\right) < E(Y)$ ó $E\left(D\right) < 0$

Cálculo de los estadísticos:

Orden	X	Y	D_i	$\|D_i\|$	Rango	R_i
1	72	72	0	0	-	-
2	76	77	-1	1	1.5	-1.5
3	90	91	-1	1	1.5	-1.5
4	88	86	2	2	3	3
5	64	68	-4	4	4	-4
6	96	91	5	5	5.5	5.5
7	65	70	-5	5	5.5	-5.5
8	77	71	6	6	7	7
9	81	88	-7	7	8	-8
10	80	71	9	9	9	9
11	65	77	-12	12	10	-10
12	72	87	-15	15	11	-11
T+	24.5		Valores	1 cola	w_{0.05}	14
T-	41.5		críticos	2 colas	w_{0.025}	11
n	11				w_{0.975}	55
Z	-0.756			$P(z <= \|Z\|)$	0.775	
				$1 - P(z \leq \|Z\|) = p$ (1 cola)	0.225	
				$2 * [1 - P(z \leq \|Z\|)] = p$ (2 colas)	0.449	

La conclusión de esta prueba es no rechazar la H_0, ya que el $T^+ > w_{0.05}$. Utilizando la aproximación normal, la conclusión es similar (p=0.225).

La prueba de Wilcoxon a dos colas puede realizarse con R. Esta prueba se encuentra en la función `wilcox.test()` en el paquete stats. Los resultados para el ejemplo son:

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  X and Y
## V = 24.5, p-value = 0.2382
## alternative hypothesis: true location shift is less than 0
```

Nótese que en esta tabla de resultados se presenta como estadístico T al menor entre T^+ y T^-, y se nombra como V. Por otro lado, debemos indicar que cola queremos usar en el parámetro alternative para que nos calcule la probabilidad. Si queremos usar dos colas, la opción por defecto "two.sided" es la que queremos. Entonces la probabilidad es calculada como $2 * [P(w \leq W)]$, donde w es el valor del estadístico y W es la distribución de Wilcoxon. En cambio, si queremos la cola derecha, la opción a usar es "less"; si queremos la cola izquierda la opción es "greater". En el primer caso se calcula la $P(w \leq W)$; mientras que en el segundo caso es calculada su complemento: 1-P $(w \leq W)$. En todos los casos debemos comparar p con nuestro α y la regla de decisión será: Si $p \leq \alpha$ Rechazo H_0.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --

## ✓ ggplot2 2.2.1.9000     ✓ purrr   0.2.4
## ✓ tibble  1.4.2          ✓ dplyr   0.7.4
## ✓ tidyr   0.8.0          ✓ stringr 1.3.0
## ✓ readr   1.1.1          ✓ forcats 0.3.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
solucion <- FALSE
```

Chapter 9

Ejercicios de dos muestras no paramétrico

9.1 Reproduciendo el algoritmo manualmente

Si bien el algoritmo está implementado en R, resulta importante entender como funciona. Una buena forma es realizar los cálculos manualmente. No vamos a hacer todo manualmente, vamos a aprovechar algunas funciones para hacer los pasos más tediosos.

Usando los datos del ejemplo de los lobos marinos:

```
library(tidyverse)
lobos_marinos <- tibble(Machos = c(15, 10, 14, 17, 16, 11, 15, 20, 13, rep(NA, 5)),
           Hembras = c(9, 8, 5, 5, 5, 9, 9, 10, 13, 13, 15, 5, 12,6))
lobos_marinos
```

```
## # A tibble: 14 x 2
##    Machos Hembras
##     <dbl>   <dbl>
##  1    15.      9.
##  2    10.      8.
##  3    14.      5.
##  4    17.      5.
##  5    16.      5.
##  6    11.      9.
##  7    15.      9.
##  8    20.     10.
##  9    13.     13.
## 10    NA      13.
## 11    NA      15.
## 12    NA       5.
## 13    NA      12.
## 14    NA       6.
```

Recordemos que primero cargué el paquete `tidyverse` que nos va facilitar muchas funciones para trabajar con los datos. Luego creé el objeto `lobos_marinos` que contiene los datos de la longitudes de raneyas consumidas por machos y hembras. Notarás que usé una función nueva, `rep()`. Lo que hace es repetir la secuencia que quieras el número de veces indicados. En este caso `NA` (*not available*, no disponible) cinco veces. Esto lo hice porque los *data frame* son estructuras rectangulares de datos y *todas las columnas deben*

tener el mismo número de observaciones. Luego, se puede ver el resultado escribiendo el nombre del objeto.

Ahora ya tengo los datos, pero para ordenarlos debo tener los largos en una columna, sin perder de vista que datos corresponden a machos y cuales a hembras.

```
lobos_marinos %>%
  gather(key = Sexo, value = LT)
```

```
## # A tibble: 28 x 2
##    Sexo      LT
##    <chr>  <dbl>
##  1 Machos   15.
##  2 Machos   10.
##  3 Machos   14.
##  4 Machos   17.
##  5 Machos   16.
##  6 Machos   11.
##  7 Machos   15.
##  8 Machos   20.
##  9 Machos   13.
## 10 Machos   NA
## # ... with 18 more rows
```

La función `gather()` se encarga de juntar los datos. Por defecto junta todas las columnas y devuelve dos. Una es la `key` (clave) que es el nombre de la columna a la cual pertence a el valor de la columna `value`. Aquí le he dado como nombres `Sexo` a la columna `key` y `LT` (Largo Total) a la de `value`. Todo muy bonito, pero hay un problema. Tenemos los valores `NA` que no necesitamos. Los podemos quitar con `na.omit()` que va eliminar todas las filas que contengan `NA`.

```
lobos_marinos %>%
  gather(key = Sexo, value = LT) %>%
  na.omit()
```

```
## # A tibble: 23 x 2
##    Sexo       LT
##    <chr>   <dbl>
##  1 Machos    15.
##  2 Machos    10.
##  3 Machos    14.
##  4 Machos    17.
##  5 Machos    16.
##  6 Machos    11.
##  7 Machos    15.
##  8 Machos    20.
##  9 Machos    13.
## 10 Hembras    9.
## # ... with 13 more rows
```

Ya no están las filas con `NA`. Hay que tener mucho cuidado cuando usamos `na.omit` porque esto elimina la fila con tan solo un valor de `NA` y puede que falte un dato pero ¡los otros sirvan!

Ya están los datos en el formato adecuado. Ese formato se conoce como *largo*; una fila corresponde a una observación. Ahora podemos comenzar con el algoritmo que está la sección de teoría 8.3.

El primer paso es rankear los datos. Recordemos que si hay empates se debe poner el valor promedio de los rangos de cada uno. Es decir que si tenemos cada dos datos son 10 y suponiendo que a cada uno le asignamos el rango 11 y 12. Entonces el valor que le corresponde es el promedio de 11 y 12, 11.5. Afortunadamente, la función `rank()` tiene un argumento para indicar como queremos definir los empates. Lo que pide el algoritmo

es la media que equivale colocar el argumento `ties = "average"`.

```
lobos_marinos <- lobos_marinos %>%
  gather(key = Sexo, value = LT) %>%
  na.omit() %>%
  mutate(rango = rank(x = LT, ties = "average"))
lobos_marinos
```

```
## # A tibble: 23 x 3
##    Sexo       LT rango
##    <chr>   <dbl> <dbl>
##  1 Machos    15. 19.0
##  2 Machos    10. 10.5
##  3 Machos    14. 17.0
##  4 Machos    17. 22.0
##  5 Machos    16. 21.0
##  6 Machos    11. 12.0
##  7 Machos    15. 19.0
##  8 Machos    20. 23.0
##  9 Machos    13. 15.0
## 10 Hembras    9.  8.00
## # ... with 13 more rows
```

Usé la función `mutate()` que agrega una columna a un *data frame*. Bien, ya tengo los rangos y he resuelto el problema de los empates en un solo paso gracias a la función `rank()`. Guardé los resultados con el mismo nombre porque no modifiqué los datos originales, sino que solo agregué variables nuevas.

Ahora podemos calcular el estadístico W que igual a la suma de los rangos de la población 1. Recordemos que la definición de población 1 y 2 es totalmente arbitraria.

```
lobos_marinos %>%
  group_by(Sexo) %>%
  summarise(W = sum(rango),
            N = n())
```

```
## # A tibble: 2 x 3
##   Sexo        W     N
##   <chr>   <dbl> <int>
## 1 Hembras  118.    14
## 2 Machos   158.     9
```

Aquí hay dos nuevas funciones. La primera, `group_by()`, agrupa los valores de acuerdo a la/s columna/s especificadas. En este caso `Sexo`. En segundo lugar, la función `summarize` resume los datos según las funciones que especifiquemos. Aquí, sumamos los valores de la columna rango según la columna sexo.

Estos valores que calculamos sirven si los empates son pocos. ¿Pero realmente son pocos? Podemos comprobarlo con la ayuda de otra función:

```
frecuencias <- table(lobos_marinos$LT)
sum(frecuencias[frecuencias > 1])/sum(frecuencias)
```

```
## [1] 0.6521739
```

La función `table()` calcula las frecuencias de los valores y usé esas frecuencias para ver cuantos de esos se repiten, es decir que la frecuencia es mayor a 1. Y dividí la suma de los frecuencias de valores repetidos por el N total. No tienen que entender todo el código de arriba, pero eso es lo que hice. Entonces, el 65% de los valores están empatados. Es un porcentaje alto de los datos. Lo que justifica calcular el [estadístico mucho mas complejo][w-con-empates].

Es una formula larga que puede dar lugar a errores. Hay varios valores que se usan varias veces y conviene calcularlos antes.

```
lobos_marinos_W_N <- lobos_marinos %>%
  group_by(Sexo) %>%
  summarise(W = sum(rango),
            N = n())

N1 <- lobos_marinos_W_N$N[1]
N2 <- lobos_marinos_W_N$N[2]

N1N2 <- N1 * N2
N1_N2 <- N1 + N2

W <- lobos_marinos_W_N$W[1]
Ri2 <- sum(lobos_marinos$rango^2)

(W - N1*(N1_N2+1)/2) /
  sqrt((N1N2/(N1_N2*(N1_N2-1))*Ri2 - (N1N2*(N1_N2+1)^2)/(4*(N1_N2-1))))
```

```
## [1] -3.199432
```

```
wilcox_empates <- function(r1, r2){
  W <- sum(r1)
  r <- c(r1, r2)
  N1 <- length(r1)
  N2 <- length(r2)

  N1N2 <- N1 * N2
  N1_N2 <- N1 + N2
  Ri2 <- sum(r^2)

  (W - N1*(N1_N2+1)/2) /
  sqrt((N1N2/(N1_N2*(N1_N2-1))*Ri2 - (N1N2*(N1_N2+1)^2)/(4*(N1_N2-1))))
}

rango_machos <- lobos_marinos %>% filter(Sexo == "Machos") %>% pull(rango)
rango_hembras <- lobos_marinos %>% filter(Sexo == "Hembras") %>% pull(rango)

wilcox_empates(rango_machos, rango_hembras)
```

```
## [1] 3.199432
```

Lo primero es probar que lo que pretendemos convertir en función funcione como esperamos fuera. Luego, es más fácil convertir el código en función. No vamos a entrar en detalles ahora sobre las funciones. Solo tienen que saber que se crean con `function`, se definen los argumentos, el código a ejecutar y se guarda como un objeto.

Ejercicio 9.1. 1. Notarán que los resultados son iguales pero de signo opuesto. ¿Por qué?

2. Implementen el algoritmo para calcular la U de Mann-Whitney.

9.2 Funciones no paramétricas en *R*

Los ejercicios a continuación deben realizarse utilizando pruebas no paramétricas. La prueba de Wilcoxon o Mann-Whitney para dos muestras o la prueba de rangos con signos de Wilcoxon para muestras apareadas.

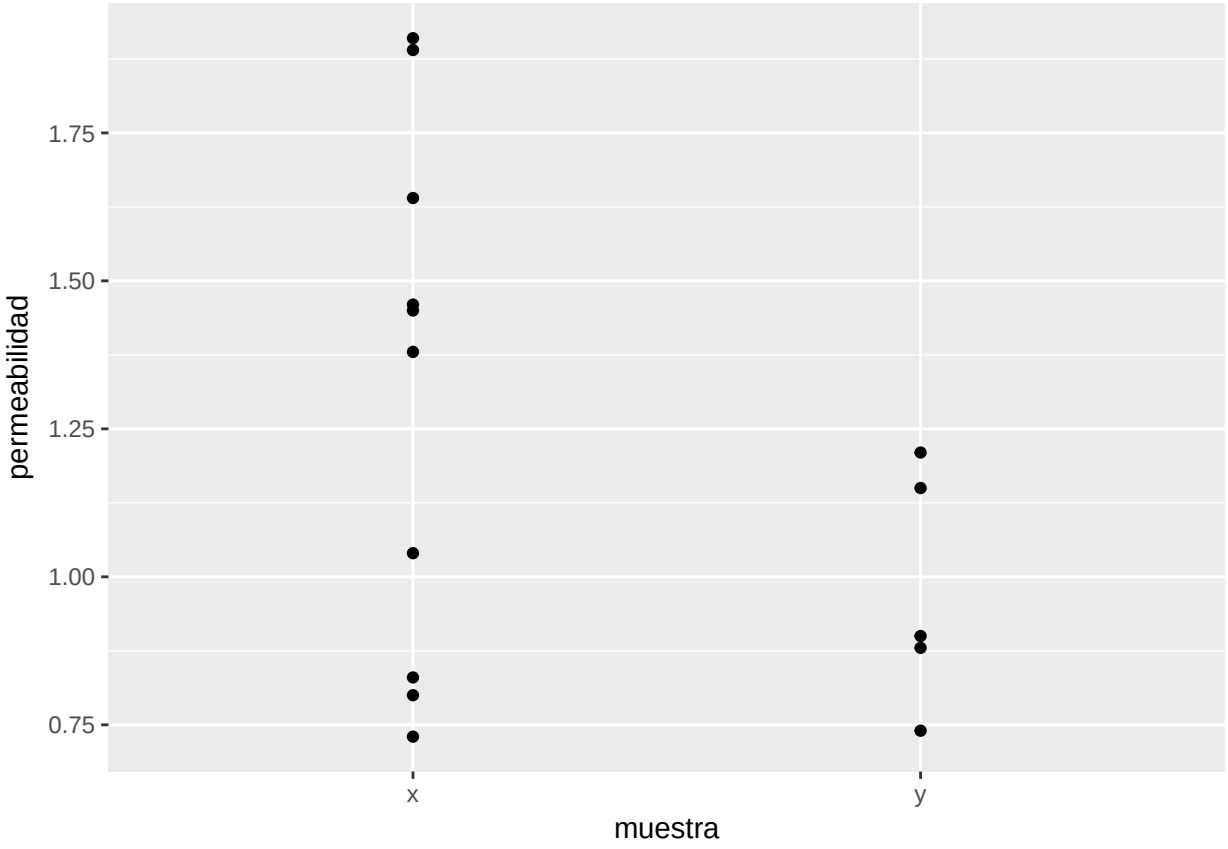
En *R*, la función `wilcox.test()` realiza esta pruebas para dos muestras o muestras apareadas. Por ejemplo, dados estos datos de constante de permeabilidad de la membrana corioaminiótica humana a término (`x`) y entre 12 y 26 semanas de gestación (`y`). La hipótesis alternativa de interés es que hay mayor permabilidad de esta membrana al término del embarazo.

```
x <- c(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, 0.73, 1.46)
y <- c(1.15, 0.88, 0.90, 0.74, 1.21)

perm <- data.frame(muestra = rep(c("x", "y"), times = c(length(x), length(y))),
           permeabilidad = c(x, y))
```

Los datos son de Hollander & Wolfe (1973), 69f. Aquí, los guardamos como `x` e `y`. Luego los reunimos en un *data frame* para poder gráficar con `ggplot`. Para esto queremos que quede en una columna los valores, que se puede realizar con función `c()` que concatena los valores que le indiquemos. Y además, tenemos que agregar una columna que indique si ese valor es de la muestra `x` o de `y`. Se podría usar `c()` y repetir tantas veces como sea necesario x e y manualmente. Pero es aburrido y seguramente nos equivocaremos. Y si para algo se hicieron las máquinas es para que nosotros no tengamos que hacer las tareas aburridas (o que al menos nos lleven menos tiempo). Para que lleve menos tiempo usamos la función `rep()`que repite el una secuencia la cantidad de veces que indiquemos (el argumento `times`). Para obtener la cantidad de veces que cada una debe repetirse usamos la función `length()` que nos devuelve la longitud cada uno de esos objetos.

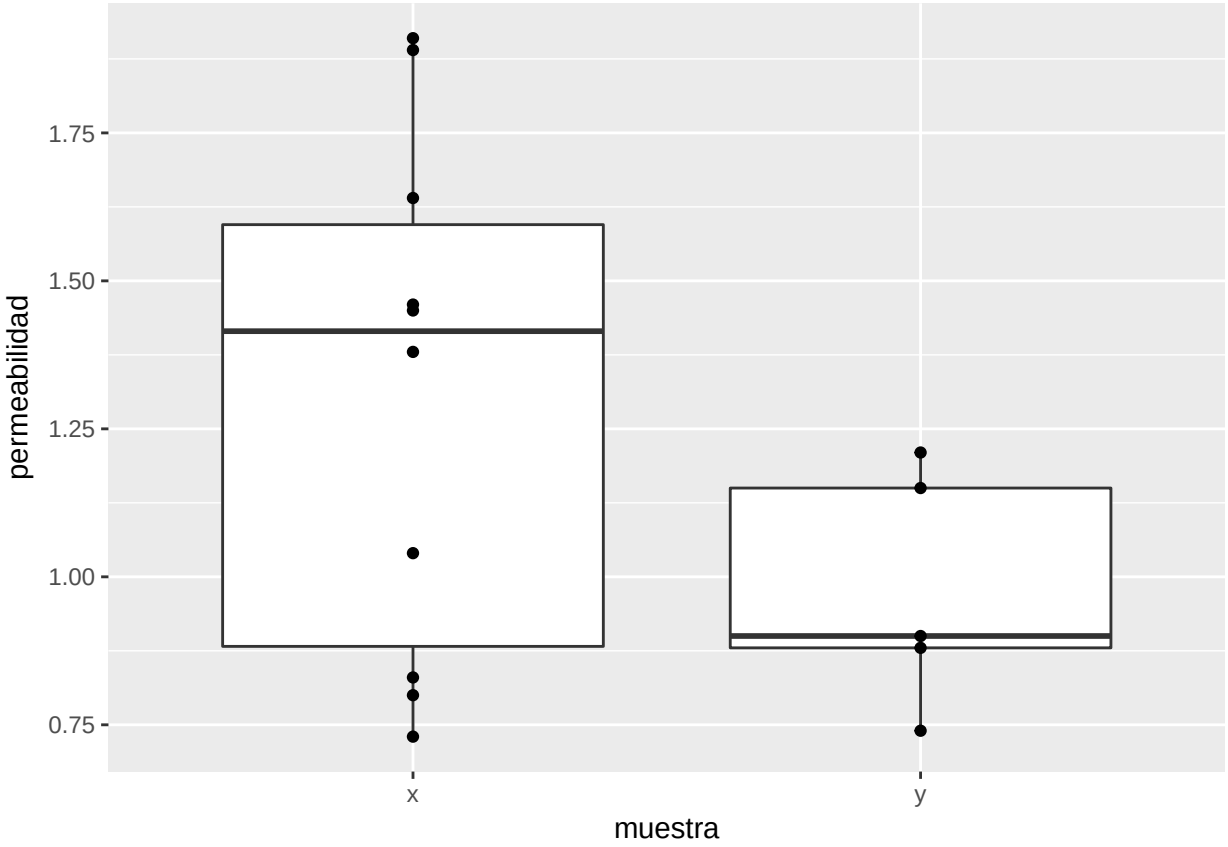
Siempre es recomendable graficar los datos antes de analizarlos. Nos va a revelar mucha información y posibles problemas solo con mirarlos. Hay que graficar los datos tal cual están, ya que si solo graficamos las medidas de resúmen típcas hay información que puede quedar escondida.

```
ggplot(perm, aes(muestra, permeabilidad)) +
  geom_point()
```



Aunque también podemos combinar este gráfico con un boxplot y obtener un poco más de información.

```
ggplot(perm, aes(muestra, permeabilidad)) +
  geom_boxplot() +
  geom_point()
```



Graficamente, parece ser que la permeabilidad es mayor a término que en menores semanas de gestación. Pero, ¿es significativa estadísticamente?. Para responder esta pregunta vamos a usar la prueba de Wilcoxon para dos muestras.

La función `wilcox.test` tiene varios argumentos. Por un lado, debemos especificar los objetos que contienen los datos, `x` e `y` en nuestro caso. Y en luego, la hipótesis alternativa, la distribución de `x` es mayor (*greater*), menor (*less*) o distinta (mayor o menor, *two.sided*) que `y`. Es importante tener esto en claro. Siempre se trata del primero versus el segundo.

```
wilcox.test(x, y, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test
##
## data:  x and y
## W = 35, p-value = 0.1272
## alternative hypothesis: true location shift is greater than 0
```

Dado que $P[X > x] = 0.1272$ no rechazamos la hipótesis nula.

9.3 Fórmulas

Otra forma muy común de analizar datos es utilizando fórmulas. Estas no son fórmulas algebraicas como ya vimos en la parte de gráficos con `ggplot2`. Las formulas en R describen como se relacionan las variables. Por un lado, tenemos las variables dependientes y por otro las independientes y estás separadas por la virguilla (~). Por ejemplo: `variable_dependiente ~ variable_independiente`. Por ahora solo vamos a trabajar

con una sola variable dependiente e independiente, pero es posible describir todo tipo de modelos mediante esta interfaz. El ejemplo anterior puede ser analizado usando la interfaz de fórmula:

```
# Debemos usar el argumento data para que la función sepa donde estan los datos.
# Prueben que sucede si no está.

wilcox.test(formula = permeabilidad ~ muestra, alternative = "greater", data = perm)
```

```
##
##  Wilcoxon rank sum test
##
## data:  permeabilidad by muestra
## W = 35, p-value = 0.1272
## alternative hypothesis: true location shift is greater than 0
```

Presten atención a que hay que agregar el argumento `data`. Sin este argumento la función `wilcox.test()` no sabe donde buscar los datos.

Como ver, los resultados son iguales. Pero hay que tener cuidados. El orden de los datos puede no ser el mismo que usaron en la versión sin fórmula. Esto se debe a que los niveles de los factores se asignan por orden alfabético si no los especificamos. Entonces, el orden puede ser distinto, y recordemos que la prueba especifica el primero vs el segundo. Por ejemplo, si vemos los niveles de muestra, no dice que tiene dos niveles, `x` y luego `y`.

```
perm$muestra
```

```
##  [1] x x x x x x x x x x y y y y y
## Levels: x y
```

Veamos que pasa si cambiamos el orden de los niveles:

```
perm <- perm %>%
  mutate(muestra = factor(x = muestra, levels = c("y", "x") ))
perm$muestra
```

```
##  [1] x x x x x x x x x x y y y y y
## Levels: y x
```

Ahora volvamos a hacer el análisis igual que antes:

```
wilcox.test(formula = permeabilidad ~ muestra, alternative = "greater", data = perm)
```

```
##
##  Wilcoxon rank sum test
##
## data:  permeabilidad by muestra
## W = 15, p-value = 0.8968
## alternative hypothesis: true location shift is greater than 0
```

¡Noten que los resultados son muy distintos!

9.4 Problemas

En todos los casos indicar la hipótesis nula y la alternativa. Graficar y realizar la prueba.

Los datos de los problemas ya se encuentran guardados. Hay que cargarlos con:

```
load("data/dos_muestras_np.RData")
```

1.- Un investigador, trabajando con una especie de ratones de campo, desea saber si los ejemplares provenientes del valle son de similar tamaño a los provenientes de la meseta. Para ello realizó capturas de ratones en ambos ambientes, midiendo el peso en gr de los ejemplares capturados. Los datos obtenidos fueron:

```
ratones
```

```
## # A tibble: 35 x 3
##     Peso Ambiente   row
##    <dbl> <fct>    <int>
##  1   38. Valle        1
##  2   47. Valle        2
##  3   50. Valle        3
##  4   51. Valle        4
##  5   39. Valle        5
##  6   39. Valle        6
##  7   44. Valle        7
##  8   40. Valle        8
##  9   46. Valle        9
## 10   50. Valle       10
## # ... with 25 more rows
```

¿Qué conclusión se puede sacar con estos datos?

2.- Una empresa pesquera desea evaluar si existen diferencias entre dos jefes de planta que trabajan en uno de sus buques factoría. De acuerdo a lo expresado por el capitán del buque, el Jefe 1 aprovecha mejor la captura que el Jefe 2. Para estudiar esta cuestión embarcaron a ambos jefes de planta en un mismo viaje de pesca y les asignaron aleatoriamente a cada uno de ellos los lances que debían procesar. En cada lance, un empleado imparcial de control de calidad registraba el porcentaje de descarte producido a partir de la captura. Los resultados obtenidos fueron:

```
jefes
```

```
## # A tibble: 35 x 3
##    Descarte  Jefe   row
##       <dbl> <dbl> <int>
##  1      17.    1.     1
##  2      10.    1.     2
##  3      15.    1.     3
##  4      22.    1.     4
##  5      15.    1.     5
##  6       9.    1.     6
##  7      20.    1.     7
##  8      14.    1.     8
##  9      21.    1.     9
## 10      19.    1.    10
## # ... with 25 more rows
```

¿Tiene razón el capitán?

Definición 9.1. Más adelante no podrán usar la formulas para hacer las pruebas apareadas. Para hacer comparaciones de datos apareados cada unidad muestral debe estar en una fila, con dos columnas: una para antes y otra para el después. Sin embargo, esta forma no permite utilizar el argumento `data` para indicar donde se encuentran los datos. Por eso, hay que hacer que estén disponibles para la función. Una forma es usar el operador de exposición `%$%`; expone los nombres del objeto que está a la izquierda del mismo:

```
datos %$% wilcox.test(x = x, y = y, paired = TRUE)
```

Además, vemos que debemos agregar el argumento `paired = TRUE` para indicar que los datos son apareados.

3.- En un estudio clínico, se desea evaluar si una cierta droga disminuye o no la concentración de un virus en sangre. Para ello se utilizaron 17 cobayos infectados, registrándose previamente al inicio de la experiencia la concentración del virus en sangre. Luego de finalizado el tratamiento con la droga, se volvió a estudiar la concentración del virus en los cobayos. Los resultados obtenidos fueron:

```
cobayos
```

```
## # A tibble: 17 x 3
##    Ejemplar Antes Despues
##       <dbl> <dbl>   <dbl>
##  1       1.    4.      6.
##  2       2.   16.     12.
##  3       3.   18.     14.
##  4       4.    9.      7.
##  5       5.   18.     18.
##  6       6.   11.     11.
##  7       7.    3.      2.
##  8       8.   16.     14.
##  9       9.   13.     10.
## 10      10.    9.      6.
## 11      11.    6.      1.
## 12      12.   13.     12.
## 13      13.   19.     19.
## 14      14.    4.      4.
## 15      15.    5.      7.
## 16      16.   17.     15.
## 17      17.   15.     14.
```

¿Qué conclusiones puede Ud. sacar acerca de la efectividad del tratamiento?

4 - Un ecólogo desea evaluar en una especie de foca si el éxito reproductivo de las hembras está asociado al sexo de sus crías. Para ello utilizó información de una población que ha sido seguida durante varias generaciones, registrando para 15 hembras el número de nietos producidos por sus hijos e hijas. Los datos fueron:

```
focas
```

```
## # A tibble: 15 x 3
##    Ejemplar Hembra Macho
##       <dbl>  <dbl> <dbl>
##  1       1.   128.  120.
##  2       2.    95.  111.
##  3       3.   104.  119.
##  4       4.    99.  111.
##  5       5.   111.  120.
##  6       6.    93.  109.
##  7       7.   132.  108.
##  8       8.   129.  130.
##  9       9.   127.  130.
## 10      10.   100.  119.
## 11      11.   122.  105.
## 12      12.   124.  132.
## 13      13.    94.  127.
## 14      14.    96.  126.
## 15      15.   127.  121.
```

¿Depende el éxito reproductivo de las hembras del sexo de sus hijos?

5.- Una especie de ave pone dos huevos por temporada reproductiva. Se ha visto que de los dos pichones el primero en eclosionar tiene mayores probabilidades de sobrevivir. Un biólogo desea establecer si esta situación está relacionada con el peso del pichón al momento de la eclosión. Para ello registró el peso de los pichones al momento de la eclosión del huevo, obteniendo los siguientes datos:

```
pichones
```

```
## # A tibble: 17 x 3
##     Nido Primer Segundo
##    <dbl>  <dbl>   <dbl>
##  1    1.   102.    110.
##  2    2.    86.     95.
##  3    3.   112.    117.
##  4    4.    85.    119.
##  5    5.    91.    117.
##  6    6.   101.     94.
##  7    7.   102.    102.
##  8    8.   111.     96.
##  9    9.   116.    103.
## 10   10.   114.    120.
## 11   11.    83.    102.
## 12   12.    85.     98.
## 13   13.   105.    118.
## 14   14.    95.    106.
## 15   15.   107.     94.
## 16   16.    94.    108.
## 17   17.    99.    102.
```

¿Cuál es la conclusión que debería sacar el biólogo?

6.- Un biólogo está estudiando el efecto del aprendizaje en la habilidad de los osos para capturar peces. Para ello registra el porcentaje de éxitos de captura durante una semana de 17 ositos cuando comienzan a pescar y repite el análisis 6 meses después. Los resultados obtenidos fueron:

```
osos
```

```
## # A tibble: 17 x 3
##      Oso Tiempo0 Tiempo6
##    <dbl>   <dbl>   <dbl>
##  1    1.     42.     75.
##  2    2.     47.     61.
##  3    3.     58.     63.
##  4    4.     40.     74.
##  5    5.     56.     65.
##  6    6.     60.     54.
##  7    7.     45.     58.
##  8    8.     72.     53.
##  9    9.     63.     66.
## 10   10.     79.     77.
## 11   11.     43.     65.
## 12   12.     45.     52.
## 13   13.     79.     64.
## 14   14.     45.     65.
## 15   15.     42.     62.
## 16   16.     49.     54.
## 17   17.     53.     68.
```

¿Mejora la capacidad de captura de los ositos con la experiencia?

7.- Los pingüinos de Magallanes hacen sus nidos en cuevas en las laderas o bajo de los arbustos. Un biólogo sostiene que el éxito reproductivo de las hembras que nidifican en las laderas es mayor que el de aquellas que lo hacen bajo los arbustos. Para poner a prueba esta hipótesis utilizó datos del número de pichones vivos que tuvieron durante su vida hembras que nidificaron en laderas y en arbustos. Las hembras pudieron identificarse debido a que fueron anilladas de pichones y no se registraron cambios en el tipo de nido que utilizaron a lo largo de la vida. Los datos obtenidos fueron:

Número de pichones producidos por hembras de pingüino de Magallanes a lo largo de su vida, discriminado por el tipo de nido que utilizaron.

```
pinguinos
```

```
## # A tibble: 32 x 3
##    Pichones Nido      row
##       <dbl> <fct>   <int>
##  1      10. Arbusto     1
##  2      13. Arbusto     2
##  3       7. Arbusto     3
##  4       7. Arbusto     4
##  5      11. Arbusto     5
##  6      11. Arbusto     6
##  7       8. Arbusto     7
##  8      11. Arbusto     8
##  9       8. Arbusto     9
## 10      12. Arbusto    10
## # ... with 22 more rows
```

¿Está Ud. de acuerdo con el biólogo?

8.- Estudiando la dieta de un delfín y del lobo marino, un biólogo desea establecer si las tallas consumidas de calamares por estos predadores son similares. Utilizando regresiones alométricas estimó los largos dorsales del manto (LDM) a partir de los picos hallados en los contenidos estomacales. Los datos obtenidos fueron:

Tallas de calamares (LDM, cm) consumidos por delfines y lobos marinos.

```
LDM
```

```
## # A tibble: 37 x 3
##      LDM Especie   row
##    <dbl> <fct>   <int>
##  1  26.0 Delfín      1
##  2  21.0 Delfín      2
##  3  24.6 Delfín      3
##  4  20.9 Delfín      4
##  5  26.4 Delfín      5
##  6  23.9 Delfín      6
##  7  25.6 Delfín      7
##  8  24.2 Delfín      8
##  9  20.4 Delfín      9
## 10  23.3 Delfín     10
## # ... with 27 more rows
```

¿Qué puede concluir sobre las tallas de los calamares consumidos por los delfines y lobos marinos?

9.- Para determinar si una droga es eficaz para disminuir la concentración de un virus en sangre, se seleccionaron al azar ratones infectados y se les inyectó la droga a evaluar. Otro grupo de ratones infectados fue utilizado como control empleándose un placebo en lugar de droga. Luego del experimento se midió la

concentración del virus en sangre utilizando una escala apropiada. Los resultados fueron: Concentración del virus en sangre de los ratones tratados y del grupo control

```
droga
```

```
## # A tibble: 24 x 3
##    Virus Tratamiento   row
##    <dbl> <fct>       <int>
##  1  34.4 Control         1
##  2  39.8 Control         2
##  3  26.6 Control         3
##  4  33.3 Control         4
##  5  37.3 Control         5
##  6  30.3 Control         6
##  7  16.2 Control         7
##  8  21.1 Control         8
##  9  43.4 Control         9
## 10  26.0 Control        10
## # ... with 14 more rows
```

¿Es eficaz la droga para disminuir la concentración del virus en sangre?

10.- Se desea establecer si las poblaciones bonaerense y patagónica de anchoita presentan similares niveles de parasitosis por nematodes en músculo. Para ello se tomaron muestras aleatorias de anchoitas de ambas poblaciones y se determinó para cada ejemplar el número de larvas de nematodes alojadas en el músculo. Los resultados fueron: Número de larvas de nematodes en el músculo de anchoitas discriminadas por poblaciones.

```
nematodes
```

```
## # A tibble: 41 x 3
##    Larvas Anchoita     row
##     <dbl> <fct>      <int>
##  1    14. Patagónica     1
##  2    36. Patagónica     2
##  3    26. Patagónica     3
##  4    23. Patagónica     4
##  5    14. Patagónica     5
##  6    36. Patagónica     6
##  7    26. Patagónica     7
##  8    23. Patagónica     8
##  9    14. Patagónica     9
## 10    26. Patagónica    10
## # ... with 31 more rows
```

¿Existen diferencias entre poblaciones de anchoita con respecto a la parasitosis por nematodes en músculo?

11.- Un ecólogo desea determinar si la eficiencia de captura de dos especies de araña es similar. Para ello realizó un experimento donde seleccionó al azar ejemplares de cada especie, les permitió confeccionar sus telas y luego introdujo una mosca en cada caja. El ecólogo determinó para cada araña el tiempo en segundos que tardó en capturar la mosca. Los resultados fueron: Tiempo de captura de la mosca en segundos.

```
arana
```

```
## # A tibble: 36 x 3
##    Tiempo Especie   row
##     <dbl>   <dbl> <int>
##  1   159.      1.     1
##  2   143.      1.     2
##  3    90.      1.     3
```

```
##  4   130.      1.     4
##  5   148.      1.     5
##  6   150.      1.     6
##  7   161.      1.     7
##  8   166.      1.     8
##  9   164.      1.     9
## 10    87.      1.    10
## # ... with 26 more rows
```

¿Existen diferencias en las eficiencias de captura entre las especies de araña?

12. Un productor de fruta fina está convencido que la producción en el Bolsón es más alta que en Esquel. Para ello tomó datos producción de distintas parcelas en ambas localidades. Sabiendo que el azar tenía algo que ver con la estadística, se preocupó de seleccionar al azar las parcelas. Los datos obtenidos fueron: Producción de fruta fina (kg) en cada parcela, discriminada por localidad

```
fruta
```

```
## # A tibble: 35 x 3
##    Producción Localidad   row
##         <dbl> <fct>     <int>
##  1       56.6 Esquel        1
##  2       90.7 Esquel        2
##  3       29.7 Esquel        3
##  4       30.0 Esquel        4
##  5       61.4 Esquel        5
##  6       46.0 Esquel        6
##  7       61.3 Esquel        7
##  8       59.0 Esquel        8
##  9       61.4 Esquel        9
## 10       52.1 Esquel       10
## # ... with 25 more rows
```

Una vez tomados los datos, el productor acude a Ud. en busca de ayuda para responder su pregunta. ¿Qué le dirá al productor?

Chapter 10

ANOVA No Paramétrico

10.1 Pruebas para varias muestras independientes

Las pruebas no paramétricas para varias muestras independientes son, conceptualmente, una extensión de las pruebas para dos muestras. Este tipo de pruebas tienen como análogo paramétrico al ANOVA de una vía. Esencialmente, este tipo de pruebas comparan k muestras y pretenden determinar si las k muestras son similares entre sí.

10.1.1 Prueba de la mediana

10.1.1.1 Datos

Para utilizar esta prueba es necesario contar con k muestras aleatorias de las k poblaciones a comparar. Cada muestra tiene un tamaño n_i de tal forma que el tamaño muestral total (N) puede obtenerse como:

$$N = n_1 + n_2 + \ldots + n_k$$

10.1.1.2 Supuestos

- Cada una de las k muestras es una muestra aleatoria de la población respectiva.
- Las k muestras son independientes entre sí.
- Las mediciones están realizadas en, al menos, una escala ordinal.
- Si todas las poblaciones a comparar tienen idéntica mediana, la probabilidad p de que una observación cualquiera exceda el valor de la mediana es la misma para todas las poblaciones. Nótese que este supuesto no implica que las funciones de distribución de cada una de las poblaciones deban ser las mismas, ni que estas funciones sean simétricas con respecto a la mediana.

10.1.1.3 Procedimiento

Se toman los N datos y se calcula la mediana general (Gran Mediana, M).

Se clasifican los datos de cada muestra teniendo en cuenta si son mayores que la gran mediana o menores o iguales que este parámetro ($X_{\text{ij}} > M$ ó $X_{\text{ij}} \leq M$).

Utilizando esta clasificación, se construye una tabla de contingencia de 2xk de la forma:

	Población				Marginal
	1	2	...	K	
$X\text{ij} > M$	O_{11}	O_{12}	...	O_{1k}	a
$X\text{ij} \leq M$	O_{21}	O_{22}	...	O_{2k}	b
Marginal (n_i)	n_1	n_2	...	n_k	N

Se evalúa la hipótesis utilizando los estadísticos adecuados para las tablas de contingencia (χ^2 o G) y teniendo en cuenta las restricciones aplicables a este tipo de pruebas. Usualmente, cuando se realiza la prueba de la mediana se tiende a emplear el estadístico χ^2.

10.1.1.4 Hipótesis

H_0 : las k poblaciones tienen idéntica mediana.

H_a : alguna de las k poblaciones presenta una mediana diferente.

Los criterios de decisión para esta prueba son los correspondientes a las tablas de contingencia. En este caso, dado que el número de filas es siempre igual a 2, los grados de libertad para el χ^2 son $k-1$. El estadístico obtenido debe ser comparado con un $\chi^2_{\alpha;k-1}$. Los criterios de decisión son:

Si $\chi^2 > \chi^2_{\alpha;k-1}$ entonces Rechazo H_0

Si $\chi^2 \leq \chi^2_{\alpha;k-1}$ entonces No rechazo H_0

En forma equivalente, se puede emplear el Valor p para tomar la decisión. El Valor p se calcula como:

$$P\left(\chi^2_{k-1} \geq \chi^2\right) = p$$

Esta probabilidad puede obtenerse en R utilizando la función `pchisq()`.

De esta forma, el criterio de decisión empleando el Valor p queda definido como:

Si $p \leq \alpha$ entonces Rechazo H_0

Si $p > \alpha$ entonces No rechazo H_0

10.1.1.5 Contrastes

Si la prueba resultó significativa, entonces es de utilidad determinar cuáles son las poblaciones que tienen medianas diferentes. Se utiliza la misma prueba de la mediana, pero considerando $k = 2$. Así, se puede construir una tabla de 2x2 y comparar a las poblaciones de a pares. Cabe aclarar que para cada uno de estos contrastes de a pares debe calcularse la gran mediana.

Aunque este método permite realizar contrastes, algunas consideraciones deben ser realizadas. En primer lugar, al definir una tabla de 2x2, deberán realizarse las correcciones de continuidad pertinentes ya que la tabla tiene un grado de libertad.

10.1.1.6 Comentarios

La prueba de la mediana para dos poblaciones independientes, cuando se la compara con la prueba de t y se cumplen los supuestos de esta última prueba, tiene una eficiencia relativa del 95% para valores pequeños de N (6-10) y disminuye a medida que N se incrementa hasta alcanzar una eficiencia relativa asintótica del 63%.

En el caso de la extensión de la prueba para k poblaciones, cuando se cumplen los supuestos del ANOVA de un factor, la eficiencia relativa asintótica es del 64%. Sin embargo, cuando ambas pruebas se comparan utilizando poblaciones no-normales, por ejemplo, la distribución es exponencial, la eficiencia relativa asintótica de la prueba de la mediana con respecto al ANOVA es del 200%.

Ejemplo 1: Prueba de la mediana.

Se desea evaluar la eficiencia de 4 tratamientos de combinaciones de fertilizantes en la producción de maíz. Para ello se seleccionaron parcelas de similares características y se le asignó al azar a cada una de ellas un tipo de fertilizante. Luego de la cosecha se determinó la producción de cada parcela en kg de mazorcas. Los resultados fueron:

Table 10.2: Producción de maíz (kg) en las parcelas tratadas con 4 tratamientos de fertilizantes diferentes.

Tratamiento 1	Tratamiento 2	Tratamiento 3	Tratamiento 4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

H_0 : La mediana de la producción es la misma con los 4 tratamientos.

H_a :La mediana de la producción con alguno de los tratamientos es diferente

Los tamaños muestrales (n_i) de cada tratamiento fueron:

$$\text{Tratamiento 1}: \ 9 \text{Tratamiento 2}: \ 10 \text{Tratamiento 3}: \ 7 \text{Tratamiento 4}: \ 8$$

La gran mediana (M) para los 34 datos obtenidos fue:

$$M = 89$$

La tabla de datos observados y esperados construida a partir de los datos y considerando el valor de M fue:

Observado	Trat 1	Trat 2	Trat 3	Trat 4	Marginal
> Mediana	6	3	7	0	16
<=Mediana	3	7	0	8	18
Marginal	9	10	7	8	34
Esperado	Trat 1	Trat 2	Trat 3	Trat 4	
> Mediana	4,235	4,706	3,294	3,765	
<=Mediana	4,765	5,294	3,706	4,235	

Utilizando un estadístico χ^2 de la forma:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Se obtiene:

χ^2	17,543
gl	3
Valor p	0,00054637

Por lo tanto, rechazo la H_0. Existe algún tratamiento cuya mediana es diferente de las demás.

Se puede utilizar la función `Median.test()` en R. La tabla de resultados obtenida es:

```
trt <- c(rep(1, 9), rep(2, 10), rep(3, 7), rep(4, 8))
y <- c(83, 91, 94, 89, 89, 96, 91, 92, 90, 91, 90, 81, 83, 84, 83, 88, 91, 89,
       84, 101, 100, 91, 93, 96, 95, 94, 78, 82, 81, 77, 79, 81, 80, 81)
Median.test(y, trt, group = FALSE)
```

```
##
## The Median Test for y ~ trt
##
## Chi Square = 17.54306   DF = 3   P.Value 0.00054637
## Median = 89
##
##   Median  r Min Max   Q25   Q75
## 1   91.0  9  83  96 89.00 92.00
## 2   86.0 10  81  91 83.25 89.75
## 3   95.0  7  91 101 93.50 98.00
## 4   80.5  8  77  82 78.75 81.00
##
## Post Hoc Analysis
##
##         median     chisq pvalue signif.
## 1 and 2   89.0  2.554444 0.1100
## 1 and 3   92.5  6.349206 0.0117       *
## 1 and 4   83.0 13.432099 0.0002     ***
## 2 and 3   91.0 13.246753 0.0003     ***
## 2 and 4   82.5 14.400000 0.0001     ***
## 3 and 4   82.0 15.000000 0.0001     ***
```

Debido a que se rechazó H_0, es necesario realizar contrastes para detectar la ubicación de las diferencias.

Los contrastes se realizan automáticamente en R. Los resultados obtenidos se muestras arriba bajo el título "Post Hoc Analysis"

En función de estos resultados, es posible concluir que todos los tratamientos difieren entre sí, exceptuando a los tratamientos 1 y 2, los cuáles no presentan diferencias significativas. Estas diferencias pueden apreciarse en la siguiente figura:

La conclusión final de este análisis sería recomendar al tratamiento 3 como el más adecuado para incrementar la producción de maíz.

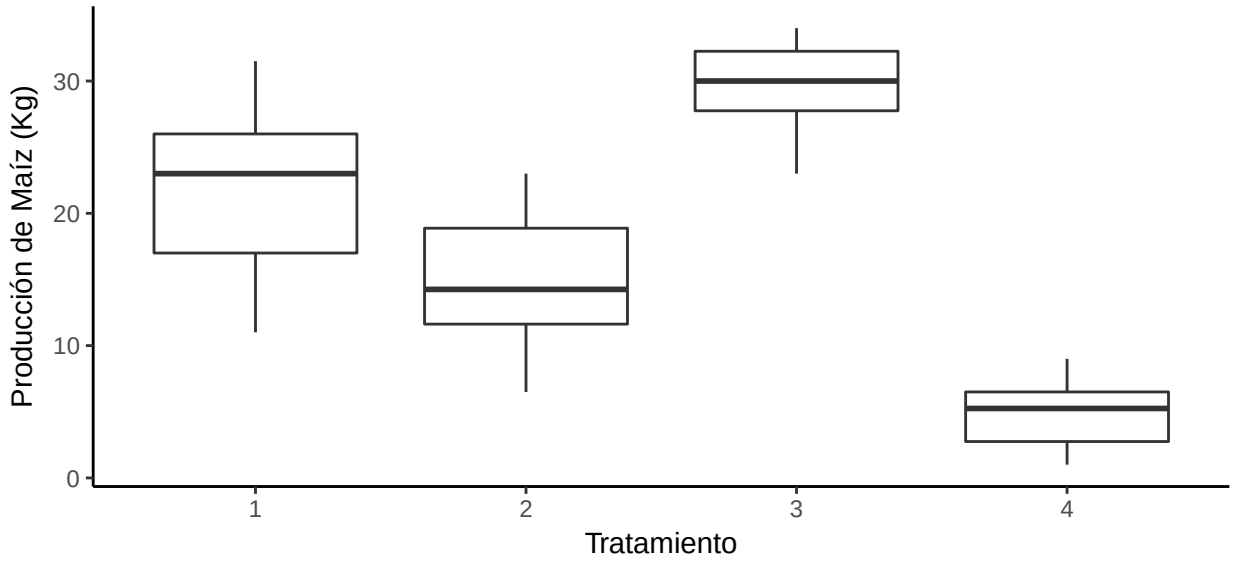


Figure 10.1: Gráfico de cajas y bigotes para la producción de maíz en cuatro tratamientos.

10.2 Prueba de Kruskal-Wallis

La prueba de Kruskal-Wallis es la extensión para k muestras de la prueba de Mann-Whitney. A diferencia de la prueba de la mediana, esta prueba evalúa las funciones de distribución de las poblaciones, aunque es sensible fundamentalmente a diferencias en la tendencia central.

10.2.1 Datos

Los datos necesarios para utilizar esta prueba son k muestras aleatorias de las poblaciones a comparar. Cada muestra tiene un tamaño n_i, obteniéndose el tamaño muestral total (N) como $N = \Sigma n_i$.

10.2.2 Supuestos

- Todas las muestras son muestras aleatorias de las respectivas poblaciones.
- Los datos dentro de cada muestra son independientes y las muestras son independientes entre sí.
- La variable a estudiar está medida, al menos, en una escala ordinal.
- Las k poblaciones tiene idéntica función de distribución o, de lo contrario, una de ellas tiende a presentar valores mayores que las demás.

10.2.3 Procedimiento

Se ordenan y ranquean los N datos de menor a mayor, calculándose cuando es necesario los rangos por empate. De esta forma, cada observación tiene un rango asignado $[R(X_{\text{ij}})]$, pudiéndose calcular las sumas de los rangos de cada población como:

$$R_i = \sum_{j=i}^{n_i} R(X_{ij})$$

Se calcula un término de varianza (S^2) como:

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R\left(X_{\text{ij}}\right)^2 - \frac{N\left(N+1\right)^2}{4}\right)$$

Se calcula el estadístico de prueba (H) como:

$$h = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - N\frac{\left(N+1\right)^2}{4}\right)$$

Este estadístico ya incorpora la corrección por empates y se distribuye aproximadamente según una χ^2 con $k-1$ grados de libertad.

10.2.4 Hipótesis

H_0: Las funciones de distribución de las k poblaciones son idénticas.

H_a: Al menos una de las poblaciones tiende a presentar valores mayores que (al menos) otra de las poblaciones.

Esta hipótesis alternativa suele plantearse también como: Las k poblaciones no tienen idéntica media.

Esto se debe a que la prueba de Kruskal-Wallis ha sido desarrollada para ser más sensible a diferencias en los parámetros de tendencia central que a otros parámetros de la distribución.

Los criterios de decisión para esta prueba son:

Si $H > \chi^2_{\alpha,\ k-1}$ entonces Rechazo H_0

Si $H \leq \chi^2_{\alpha,\ k-1}$ entonces No rechazo H_0

En forma equivalente, se puede emplear el Valor p para tomar la decisión. El Valor p se calcula como:

$$P\left(\chi^2_{k-1} \geq H\right) = p$$

De esta forma, el criterio de decisión empleando el Valor p queda definido como:

Si $p \leq \alpha$ entonces Rechazo H_0

Si $p > \alpha$ entonces No rechazo H_0

10.2.5 Contrastes

Si la prueba resultó significativa, pueden utilizarse varios métodos de contrastes de a pares para poder detectar la ubicación de las diferencias. Nosotros utilizaremos dos. Los denominaremos como *Contraste t* y *Contraste z* en función de los valores críticos que utilizan.

Ambos contrastes utilizan el mismo estadístico, difiriendo entre sí en el valor crítico que emplean. El estadístico se obtiene en ambos casos como:

$$\varepsilon_{\text{ij}} = \left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right|$$

Los valores críticos pueden obtenerse como:

Contraste t

$$VC_{\text{ij}} = t_{N-k;\alpha/2}\sqrt{S^2\frac{(N-1-H)}{N-k}}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

Contraste z

$$VC_{\text{ij}} = Z_{\frac{\alpha}{[k(k-1)]}}\sqrt{\frac{N\,(N+1)}{12}}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

Para ambos contrastes, la regla de decisión es:

Si $\varepsilon_{\text{ij}} > VC_{\text{ij}}$ entonces Rechazo H_0

Si $\varepsilon_{\text{ij}} \leq VC_{\text{ij}}$ entonces No rechazo H_0

Entre ambos métodos, el *Contraste t* es más eficiente, siendo el *Contraste z* mucho más conservativo. La ventaja de este último contraste es su mayor simplicidad de cálculo, ya que no requiere de S2 para la obtención del valor crítico. Asimismo, el *Contraste z* puede ser útil cuando sólo se desean detectar diferencias altamente significativas.

10.2.6 Comentarios

En comparación con el ANOVA de un factor, la eficiencia relativa asintótica (ARE) de la prueba de Kruskal-Wallis nunca es menor al 86.4%. Cuando las poblaciones a comparar son normales, de la prueba de Kruskal-Wallis con respecto al ANOVA es del 95%. Si las distribuciones son uniformes, asciende al 100%, mientras que, si la distribución es exponencial, es del 150%.

Comparando la prueba de Kruskal-Wallis con la prueba de la mediana, es del 150%, 300% y 75% si las distribuciones son normales, uniformes y exponenciales respectivamente.

Ejemplo 1: Prueba de Kruskal-Wallis.

Con el objetivo de comparar ambas pruebas, utilizaremos el ejemplo 1 sobre los tratamientos con fertilizantes en cultivos de maíz, para ejemplificar el uso de la prueba de Kruskal-Wallis.

Luego de ordenar y ranquear, los cálculos parciales y finales obtenidos fueron:

Grupo	n_i	R_i	R_i/n_i	R_i^2/n_i
Trat 1	9	196,50	21,833	4290,250
Trat 2	10	153,00	15,300	2340,900
Trat 3	7	207,00	29,571	6121,286
Trat 4	8	38,50	4,813	185,281
N	34		H	25,6288
k	4		gl	3
$\sum\left[R\left(X_{\text{ij}}\right)^2\right]$	13664		Valor p	1,14057E-05
S^2	98,5303			

Esta prueba puede realizarse utilizando la función `krukal.test()` del paquete stats de R o `kruskal()` del paquete `agricolae`. La tabla de resultados obtenida es:

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  observation by method
## Kruskal-Wallis chi-squared = 25.629, df = 3, p-value = 1.141e-05
```

En función de los resultados obtenidos se puede rechazar la H_0 y concluir que existen diferencias en la producción de maíz entre los tratamientos aplicados.

Para detectar la ubicación de estas diferencias se emplearon ambos métodos de contrastes. Los resultados obtenidos fueron:

Contraste t

Se presentan los R_i/n_i; los n_i y la identificación del tratamiento, tanto en los encabezamientos de filas como de columnas. En el vértice superior izquierdo de la tabla se indican el valor de t y sus grados de libertad, así como el resultado obtenido de la primera raíz en la ecuación del VC_{ij} del contraste (Factor). El $t_{\alpha/2}$ y la primera raíz son los únicos factores en la ecuación del VC_{ij}comunes a todas las comparaciones. En la tabla de resultados se presentan por debajo de la diagonal principal los valores de ε_{ij} y por encima de ésta los VC_{ij}, indicándose con *itálicas* los ε_{ij} que resultaron significativos.

$t_{\alpha/2}$	2,042		21,833	15,300	29,571	4,813
gl	30		9	10	7	8
Factor	4,920		Trat 1	Trat 2	Trat 3	Trat 4
21,833	9	Trat 1		4,617	5,064	4,883
15,300	10	Trat 2	*6,533*		4,952	4,766
29,571	7	Trat 3	*7,738*	*14,271*		5,201
4,813	8	Trat 4	*17,021*	*10,488*	*24,759*	

Contraste z

Se presentan los R_i/n_i; los n_i y la identificación del tratamiento, tanto en los encabezamientos de filas como de columnas. En el vértice superior izquierdo de la tabla se indican el valor del Z y de $1 - \alpha$ / $[k(k-1)]$, así como el resultado obtenido de la primera raíz en la ecuación del VC_{ij} del contraste (Factor). El $Z_{\frac{\alpha}{[k(k-1)]}}$ y la primera raíz son los únicos factores en la ecuación del VC_{ij} comunes a todas las comparaciones. En la tabla de resultados se presentan por debajo de la diagonal principal los valores de ε_{ij} y por encima de ésta los VC_{ij}, indicándose con *itálicas* los ε_{ij} que resultaron significativos.

$Z_{\frac{\alpha}{[k(k-1)]}}$	2,6383		21,833	15,300	29,571	4,813
$1-\alpha/[k(k-1)]$	0,9958		9	10	7	8
Factor	9,9582		Trat 1	Trat 2	Trat 3	Trat 4
21,833	9	Trat 1		11,749	12,947	12,462
15,300	10	Trat 2	6,533		14,043	13,597
29,571	7	Trat 3	7,738	*14,271*		13,136
4,813	8	Trat 4	*17,021*	*10,488*	*24,759*	

Como puede observarse, el *Contraste t* indica que todos los tratamientos difieren entre sí. Por su parte, el *Contraste z* no detecta diferencias significativas en las comparaciones 1 vs 2, 1 vs 3 y 2 vs 4. El *Contraste z* sólo detecta las diferencias más evidentes.

Si se comparan los resultados con la prueba de la mediana, vemos que ambas pruebas detectan claramente las diferencias. Sin embargo, en términos de comparaciones, el *Contraste t* detectó todas las diferencias como significativas, en un punto intermedio quedó la prueba de la mediana donde la comparación 1 vs 2 no se detectó como significativa y en el otro extremo aparece el *Contraste z* que detecta como significativas sólo tres de las seis comparaciones realizadas.

Con R:

Contraste t (`posthoc.kruskal.conover.test()`)

```
## Warning in posthoc.kruskal.conover.test.default(x = observation, g =
## method, : Ties are present. Quantiles were corrected for ties.
```

```
##
##  Pairwise comparisons using Conover's-test for multiple
##                           comparisons of independent samples
##
## data:  observation and method
##
##   1       2       3
## 2 0.0071  -       -
## 3 0.0040  1.9e-06 -
## 4 6.4e-08 9.7e-05 8.8e-11
##
## P value adjustment method: none
```

Contraste z (`posthoc.kruskal.dunn.test()`)

```
## Warning in posthoc.kruskal.dunn.test.default(x = observation, g = method, :
## Ties are present. z-quantiles were corrected for ties.
```

```
##
##  Pairwise comparisons using Dunn's-test for multiple
##                           comparisons of independent samples
##
## data:  observation and method
##
##   1       2       3
## 2 0.15200 -       -
## 3 0.12189 0.00353 -
## 4 0.00042 0.02592 1.4e-06
##
## P value adjustment method: none
```

Chapter 11

DISEÑOS EXPERIMENTALES

11.1 Bloques al azar

11.1.1 Diseño de experimentos

El diseño de un experimento se refiere a la estructura del experimento, en particular a:

1. El grupo de tratamientos incluidos en el estudio.
2. El grupo de unidades experimentales incluidas en el estudio.
3. Las reglas y procedimientos por los cuales los tratamientos son asignados a las unidades experimentales (o viceversa).
4. Las medidas son hechas sobre las unidades experimentales después que los tratamientos han sido aplicados.

Los diseños estadísticos hacen referencia a las reglas y procedimientos a través de los cuales los tratamientos son asignados a las unidades experimentales. El uso de reglas y procedimientos impropios en la asignación de los tratamientos a las unidades experimentales puede traer serias dificultades.

El proceso de medición es otro problema importante en los diseños experimentales. Idealmente, el proceso de medición debería producir medidas insesgadas y precisas. Medidas sesgadas pueden causar serias dificultades en el análisis del estudio. Una fuente importante de sesgo se debe a la falta de reconocimiento de diferencias en el proceso de evaluación.

11.1.2 Elementos de los Diseños de Bloques al Azar

Un diseño de bloques al azar es un diseño aleatorio restringido en el cual las unidades experimentales son distribuidas en grupos homogéneos, llamados *bloques*, y los tratamientos son asignados al azar dentro de los bloques. Es un diseño especialmente utilizado en experimentación agrícola, en el que se desean comparar I tratamientos (por ejemplo, fertilizantes), asignando los tratamientos en J bloques (por ejemplo J fincas), de modo que se reparten los I tratamientos aleatoriamente en cada bloque (los fertilizantes se aplican aleatoriamente en I parcelas de una misma finca). Para una correcta aplicación de este diseño debe haber una máxima homogeneidad dentro de cada bloque, para que el efecto bloque sea el mismo para todos los tratamientos dentro de cada bloque.

En un experimento agrícola, cada bloque está constituido por un número de parcelas que forman una superficie de relativa homogeneidad respecto al resto del campo. Cuando se conoce de gradiente de variabilidad

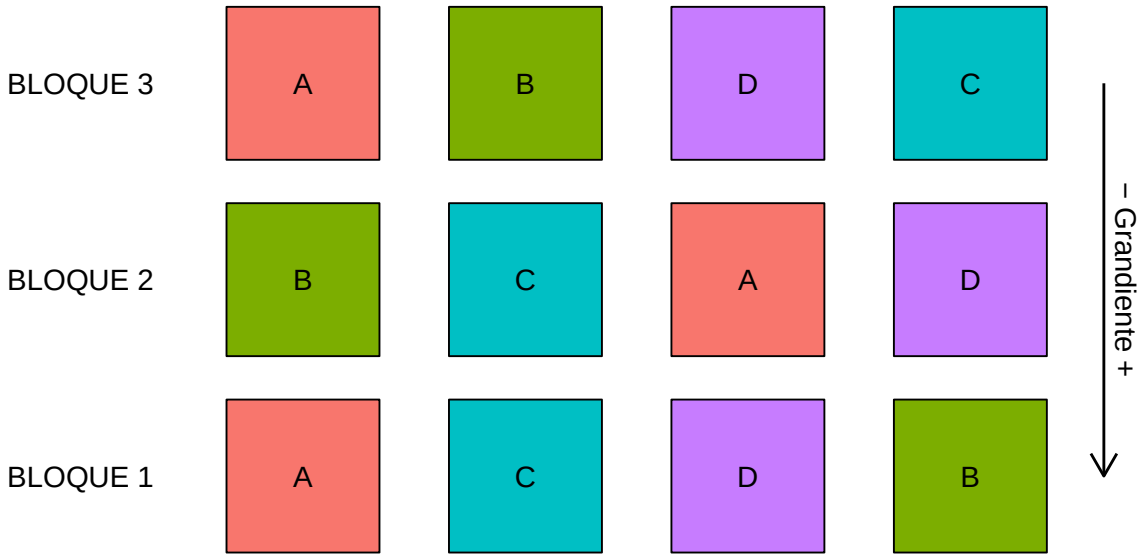


Figure 11.1: Esquema de diseño en bloques al azar. Cada bloque se dividió en cuatro y cada uno de los tratamientos dentro de los bloques se asignó al azar.

del suelo, los bloques deben orientarse perpendicularmente al gradiente y las unidades experimentales deben tener su mayor dimensión en la misma dirección y sentido que dicho gradiente.

En experimentos con animales, cada bloque estará constituido por un número de animales de aproximadamente igual peso, edad, raza, etc. Debe haber diferencias entre bloques.

11.1.3 Criterios para definir los bloques

Es necesaria una definición precisa de la unidad experimental, que dependerá del problema en particular. Una vez definida existen dos tipos de criterios para definir los bloques:

1. Características asociadas con la unidad experimental: edad, inteligencia, educación, áreas geográficas, tamaño de la población, etc.
2. Características asociadas con las características experimentales: observadores, instrumento de medición, tiempo de procesado, etc.

11.1.4 Ventajas y desventajas

1. Ventajas
2. Puede proveer resultados más precisos que un diseño completamente al azar de tamaño comparable.
3. Se pueden estimar los datos de algunas unidades experimentales si se pierden.
4. El análisis estadístico es relativamente simple.
5. Desventajas
6. Las observaciones perdidas dentro de cada bloque requieren cálculos complejos.
7. Los grados de libertad del error experimental no son tan grandes como en un diseño completamente aleatorizado.
8. Se hacen más suposiciones que en un modelo completamente al azar.

11.1.5 Modelo

Se puede pensar a un diseño en bloques al azar como un estudio de dos factores (los bloques y los tratamientos son los factores), con una observación por celda. Si se puede asumir que no hay interacción entre los dos factores, se puede realizar un análisis de los efectos de los factores cuando hay una sola observación por celda y los factores tienen efectos fijos.

Así, el modelo para un diseño en bloques al azar, cuando tanto los bloques como los tratamientos tienen efectos fijos y hay n bloques y r tratamientos, es:

$$Y_{ij} = \mu_{\bullet\bullet} + \rho_i + \tau_j + \varepsilon_{ij}$$

donde:

$\mu_{\bullet\bullet}$. es una constante

$\rho_i$son constantes para el efecto del bloque, sujetas a la restricción $\sum \rho_i = 0$.

τ_j son constantes para los efectos de los tratamientos, sujetas a la restricción $\sum \tau_i = 0$

ε_{ij} son v. a. independientes $N(0, \sigma^2)$

$$i = 1, 2, ..., n; j = 1, 2, ..., r$$

Las Y_{ij} observaciones son independientes y se distribuyen normalmente, con media:

$$E(Y_{ij}) = \mu_{\bullet\bullet} + \rho_i + \tau_j$$

y varianza constante:

$$Var(Y_{ij}) = \sigma^2$$

11.2 Análisis de la varianza y pruebas

Los estimadores de mínimos cuadrados son:

Parámetro	Estimador
$\mu_{\bullet\bullet}$	$\hat{\mu}_{\bullet\bullet} = \overline{Y_{\bullet\bullet}}$
ρ_i	$\hat{\rho}_i = \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$
τ_j	$\hat{\tau}_j = \overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}$

Los valores ajustados serán entonces:

$$\hat{Y}_{ij} = \overline{Y}_{\bullet\bullet} + \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right) + \left(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}\right) = \overline{Y}_{i\bullet} + \overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}$$

y los residuos son:

$$\varepsilon_{ij} = \overline{Y}_{ij} - \hat{Y}_{ij} = \overline{Y}_{ij} - \overline{Y}_{i\bullet} + \overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}$$

11.2.1 Análisis de la varianza

11.2.1.1 Hipótesis

Efectos de los tratamientos fijos	Efectos de los tratamientos aleatorios
H_0: todos $\tau_j = 0$	$H_0 : \sigma_\tau^2 = 0$
H_a: no todos los $\tau_j = 0$	$H_a : \sigma_\tau^2 > 0$

Se usa la misma prueba estadística, tanto si los efectos de los tratamientos son al azar como si son fijos.

$$F^* = \frac{CM_{TR}}{CM_{\text{BL.TR}}}$$

la regla de decisión será entonces:

Sí $F^* \leq F_{\alpha;(r-1),(n-1)(r-1)}$ no se rechaza H_0

Sí $F^* > F_{\alpha;(r-1),(n-1)(r-1)}$ se rechaza H_0

11.2.2 Prueba de Tukey de Aditividad

La suma de cuadrados especial de interacción está dada, en este caso, por:

$$SC^*_{BL\bullet TR} = \frac{\left[\sum_i \sum_j \left(\overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}}\right)\left(\overline{Y_{\bullet j}} - \overline{Y_{\bullet\bullet}}\right) Y_{ij}\right]^2}{\sum_i \left(\overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}}\right)^2 \sum_j \left(\overline{Y_{\bullet j}} - \overline{Y_{\bullet\bullet}}\right)^2}$$

La suma de cuadrados remanente se obtiene como:

$$SC^*_{Rem} = SC_T - SC_{BL} - SC_{TR} - SC^*_{BL\bullet TR}$$

La prueba estadística es.

$$F^* = \frac{SC^*_{BL\bullet TR}}{1} \div \frac{SC^*_{Rem}}{rn - r - n}$$

Table 11.3: Tabla de anova para un diseño de bloques al azar

Fte. de variación	SC	GL	CM	E(CM) Tratamientos fijos	E(CM) Tratamientos aleatorios
Bloques	$r\sum_i \left(\overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}}\right)^2$	$n-1$	$\frac{SC_{BL}}{n-1}$	$\sigma^2 + \frac{r\sum_i \rho_i}{n-1}$	$\sigma^2 + \frac{r\sum_i \rho_i}{n-1}$
Tratamientos	$n\sum_j \left(\overline{Y_{\bullet j}} - \overline{Y_{\bullet\bullet}}\right)^2$	$r-1$	$\frac{SC_{TR}}{r-1}$	$\sigma^2 + \frac{n\sum_j \tau_j}{r-1}$	$\sigma^2 + \backslash n\sigma_\tau^2$
Error	$SC_T - SC_{BL} - SC_{TR}$	$(n1)(r1)$	$\frac{SC^*_{BL\bullet TR}}{(n-1)(r-1)}$	σ^2	σ^2
Total		$nr-1$			

Ejemplo: Para estudiar las diferencias entre tres fertilizantes sobre la producción de papas, se dispuso de 5

fincas, cada una de las cuales se dividió en tres parcelas del mismo tamaño y tipo. Los fertilizantes fueron asignados al azar en las parcelas de cada finca. El rendimiento en toneladas fue:

Finca	Fertilizante 1	Fertilizante 2	Fertilizante 3
1	1	5	8
2	2	8	14
3	7	9	16
4	6	13	18
5	12	14	17

Se desea saber si hay diferencias entre los fertilizantes.

Se trata de un diseño de bloques al azar, cada finca es un bloque.

1. Prueba de aditividad

$$SC^*_{BL\bullet TR} = 0.262665101$$

$$SC^*_{Rem} = 23.60400157$$

$$CM^*_{Rem} = 3.372000224$$

$$F = 0.077895932$$

$$p = 0.78823515$$

Por lo tanto, no se rechaza la hipótesis nula, no hay interacción.

1. ANOVA

RESUMEN	Cuenta	Suma	Promedio	Varianza
1	3	14	4.66666667	12.3333333
2	3	24	8	36
3	3	32	10.6666667	22.3333333
4	3	37	12.3333333	36.3333333
5	3	43	14.3333333	6.33333333
Fertilizante 1	5	28	5.6	19.3
Fertilizante 2	5	49	9.8	13.7
Fertilizante 3	5	73	14.6	15.8

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Filas	171.333333	4	42.8333333	14.3575419	0.00100812	3.83785448
Columnas	202.8	2	101.4	33.9888268	0.00012292	4.45896831
Error	23.8666667	8	2.98333333			

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Total	398	14				

En R

```
##
## Tukey test on 5% alpha-level:
##
## Test statistic: 0.0779
## Critival value: 5.591
## The additivity hypothesis cannot be rejected.
```

Table 11.7: Producción de papas por fertilizante

fertilizante	mean	var	sd
Fertilizante 1	5.6	19.3	4.393
Fertilizante 2	9.8	13.7	3.701
Fertilizante 3	14.6	15.8	3.975

Table 11.8: Producción de papas por finca

Finca	mean	var	sd
1	4.667	12.33	3.512
2	8	36	6
3	10.67	22.33	4.726
4	12.33	36.33	6.028
5	14.33	6.333	2.517

Table 11.9: Prueba de homogeneidad de varianzas de Bartlett.

Test statistic	df	P value
0.1074	2	0.9477

Table 11.10: Prueba de homogeneidad de varianzas de Levene.

	Df	F value	Pr(>F)
group	2	0.06604	0.9364
	12	NA	NA

Table 11.11: ANOVA para la producción de papa con tres fertilizantes con bloques al alzar.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fertilizante	2	202.8	101.4	33.99	0.0001229
Finca	4	171.3	42.83	14.36	0.001008

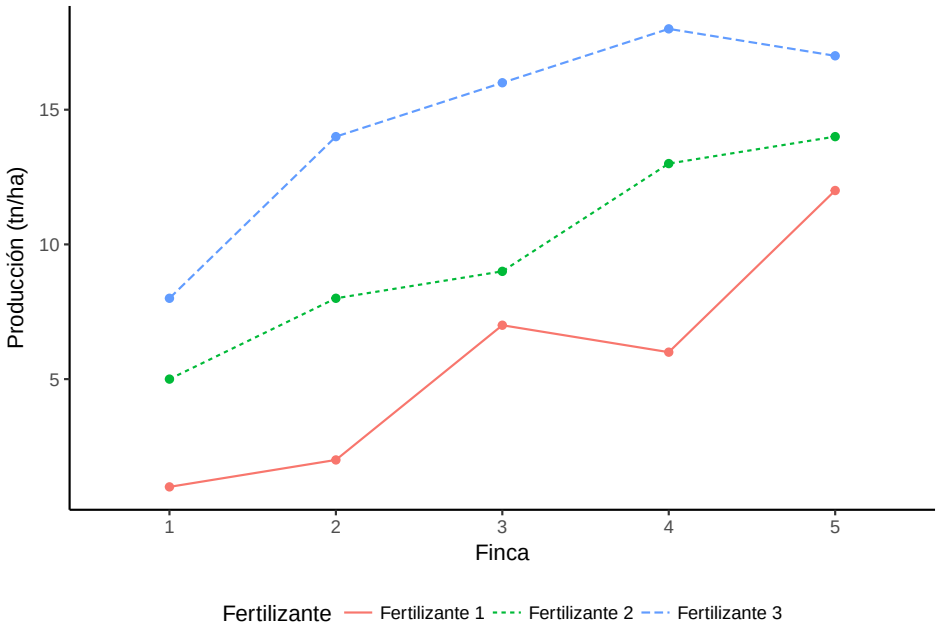


Figure 11.2: Producción de papas (tn/ha) bajo tres fertilizantes en 5 fincas.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	8	23.87	2.983	NA	NA

Table 11.12: Medias marginales, error estándar, grados de libertad, intervalo de confianza inferior y superior del 95%, y agrupamiento según la prueba de Tukey.

fertilizante	lsmean	SE	df	lower.CL	upper.CL	.group
Fertilizante 1	5.6	0.7724	8	3.819	7.381	1
Fertilizante 2	9.8	0.7724	8	8.019	11.58	2
Fertilizante 3	14.6	0.7724	8	12.82	16.38	3

Chapter 12

Regresión

En las ciencias naturales es usual querer explicar una variable con otras. Las variables que se quieren explicar son las variables dependientes y las que se usan para explicar son las llamadas variables explicatorias o también independientes. A estos modelos se los conoce como modelos de regresión. Aunque las variables estén relacionadas esto no implica que haya una relación causal entre ellas. Sin un modelo causal que explique la manera que las variables se relación se está incurriendo una falacia del tipo *cum hoc ergo propter hoc*. Por ejemplo, en la Figura 12.1 se muestra que la relación entre los limones frescos importados desde México (ton) y tasa de mortalidad total en autopistas de EE.UU. Según esta regresión al ¡aumentar la importación de disminuye la tasa de mortalidad! Este resultado carece de lógica ya que no hay una forma en que la importación de limones afecte la mortalidad. Por este motivo hay que ser cuidadoso en cuanto a las conclusiones que se realizan con los resultados.

12.1 Regresión Lineal Simple

La regresión lineal simple se da cuando hay una variable aleatoria con distribución normal y solo una variable predictora. La varible predictora no es una variable aleatoria, sino que puede ser modificada por el investigador. El objetivo de esta técnica es obtener una ecuación lineal que explique el cambio de la variable aleatoria según el cambio de la variable predictora:

$$Y_i = \beta_0 + \beta_1 X_i \tag{12.1}$$

Por ejemplo, el se tiene la edad y la longitud de alas de gorriones de varias edades en la Tabla 12.1. La edad es la variable independiente y la longitud de ala es la variable dependiente. Se busca ajustar un modelo como en el da la Ecuación (12.1). En la Figura 12.2 la recta que mejor ajusta a esto datos en azul. Sin embargo, el ajuste no es perfecto. La diferencia entre el valor ajustado o predicho y el valor observado es el error que se comete. Ese error se simboliza con la letra griega epsilon y debe ser incluido en el modelo:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{12.2}$$

Se asume que $\epsilon_i \sim N\left(0, \sigma^2\right)$ y por lo tanto $\sum \epsilon_i = 0$. El método que se usa para encontrar los estimadores de los β_i consiste en minimizar el cuadrado de los errores. Por eso, el método se lo conoce como mínimos cuadrados o mínimos cuadrados ordinarios. Afortunadamente se pueden resolver de manera analítica por las siguientes ecuaciones

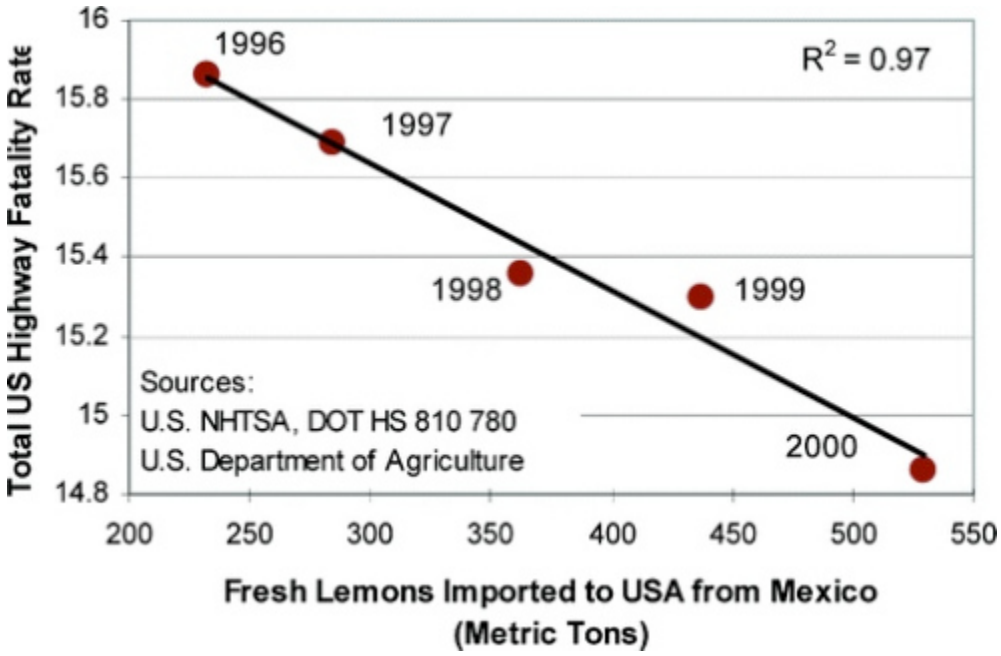


Figure 12.1: (ref:regresion-espuria

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} \tag{12.3}$$
$$b_0 = \bar{Y} - b_1 \bar{X}$$

Table 12.1: Edad en días y longitud de ala en centímetros de 13 gorriones.

Edad	Longitud
3	1.4
4	1.5
5	2.2
6	2.4
8	3.1
9	3.2
10	3.2
11	3.9
12	4.1
14	4.7
15	4.5
16	5.2
17	5

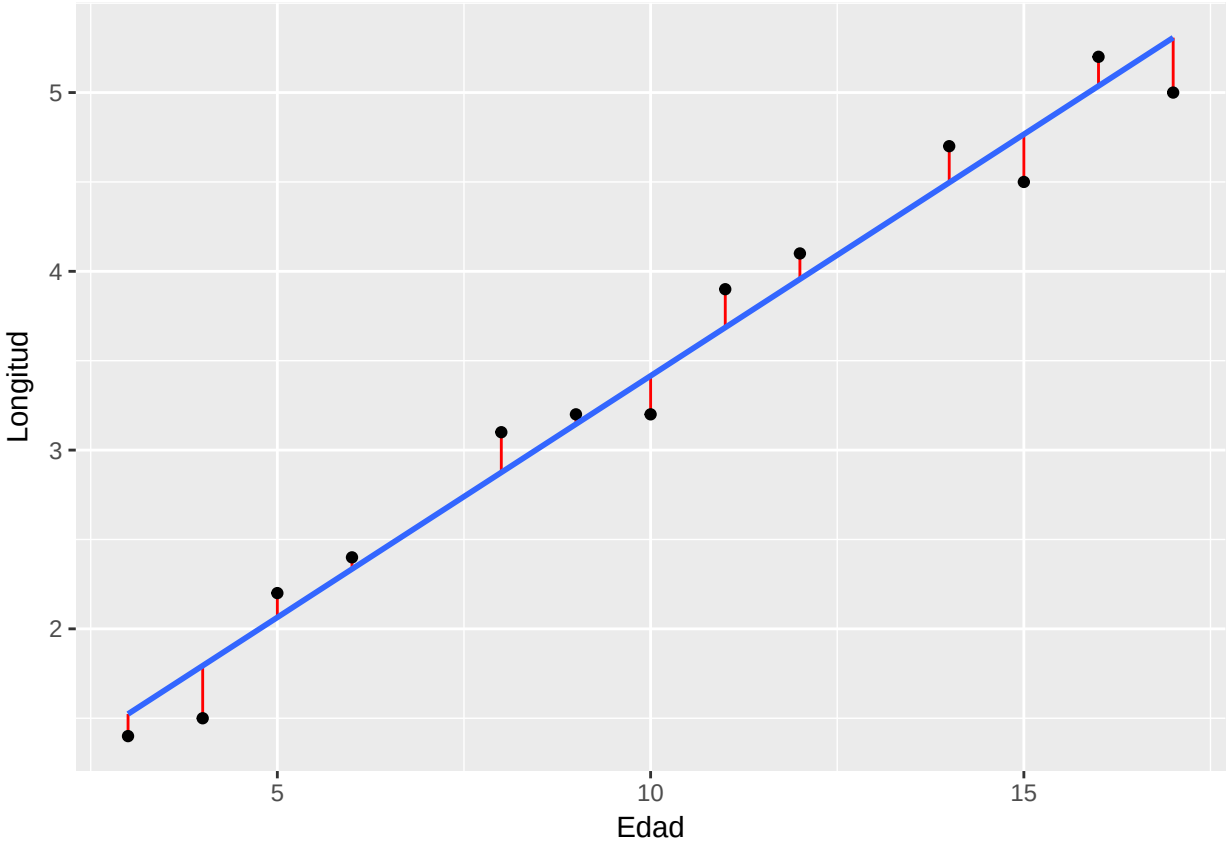


Figure 12.2: Gráfico de dispersión de edad en días y longitud de ala en centímetros de 13 gorriones. En azul línea de regresión estimada con mínimos cuadrados ordinarios. En rojo el error cometido por la regresión entre los valores estimados y los valores observados.

Chapter 13

Ordenación en Espacios Reducidos

En la Ciencias Naturales, generalmente, se poseen varias variables por cada objeto o unidad. Pero en un diagrama de dispersión es solo posible ver dos dimensiones, como máximo se podrán ver tres. Cuando queremos ver cuáles son las tendencias de variación en los objetos con respecto a todas las variables, estos gráficos se quedan cortos. Podríamos ver cada para de combinaciones, pero resulta tedioso y en general no es muy iluminador. Además, que podemos perdernos algunas relaciones interesantes entre varias variables que se manifiestan en más de dos dimensiones.

Lo métodos para ordenación en espacio reducido permiten extraer información sobre la calidad de la proyección y el estudio de las relaciones tanto entre las variables como entre objetos.

Existen varios métodos de ordenación resumidos en la tabla de abajo, más otros no incluidos que no vamos a ver en este curso.

Método	Distancia Preservada	Variables
Análisis de Componentes Principales	Distancia Euclídea	Datos Cuantitativos, Relaciones lineales (cuidado con los doble-ceros)
Análisis de Coordenadas Principales, Escalamiento (multidimensional) métrico, escalamiento clásico	Cualquier medida de distancia	Cuantitativos, semicuantitativos, cualitativos, o mezclados
Análisis de Correspondencias	Distancia χ^2	No-negativos, datos cuantitativos dimensionalmente homogéneos o binarios; abundancia de especies, o datos de presencia/ausencia

Los métodos de ordenación pueden usarse para delinear grupos de objetos cuando la estructura de los datos no es continua (las variables si deben ser continuas). En particular, la ordenación puede ser usada siempre para complementar los análisis de agrupamientos. Esto es así porque mientras en análisis de agrupamiento investiga las relacionas finas entre objetos; la ordenación investiga la variabilidad entera de los datos y extrae los gradientes generales.

En general, se usa la ordenación para estudiar las posiciones relativas de los objetos en un espacio reducido, es decir pasar de un espacio multidimensional a dos o tres dimensiones. Cuando la proyección de los datos en un espacio reducido representa una gran proporción de la variabilidad las distancias entre los objetos sean similares a la que existen en un espacio multidimensional. Cuando las proyecciones no son tan eficientes, la distancia entre los objetos es menor que en el espacio multidimensional. Se pueden dar dos casos: (1) que los objetos estén a distancias *proporcionalmente* similares en los dos espacios, entonces la proyección seguirá siendo útil (2) que las posiciones relativas de los objetos cambien entre los dos espacios, entonces la

proyección es inútil. Por lo tanto, a veces es útil considera la ordenación aun cuando esta represente una pequeña parte de la variación total.

13.1 Análisis de componentes principales

Supongamos que tenemos una distribución multivariada normal, el primer eje principal es la línea que atraviesa la mayor dimensión del elipsoide de densidad que describe la densidad. De la misma manera, los siguientes ejes principales (ortogonales entre sí, es decir en ángulo recto e incorrelados, y sucesivamente más cortos) atraviesan las siguientes dimensiones del elipsoide p-dimensional. Por lo tanto, pueden encontrarse un número p de ejes principales de una matriz de datos de p variables.

Las relaciones entre las variables pueden representarse con una matriz cuadrada $\mathbf{S}_{p\times p}$:

$$\mathbf{S} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Donde la diagonal es la varianza de una variable y fuera de ellas se encuentran las covarianzas

Los ejes principales de una matriz de dispersión $\mathbf{S}$ pueden encontrarse resolviendo la ecuación:

$$(\mathbf{S} - \lambda_k \mathbf{I})\,\mathbf{u}_k = 0$$

Cuya ecuación característica es:

$$|\mathbf{S} - \lambda_k \mathbf{I}| = 0$$

Se usa para computar los autovalores. Los autovectores $\mathbf{u}_k$ asociados a los λ_k se encuentran poniendo los distintos valores de λ_k en la primera ecuación. Estos autovectores son los ejes principales de la matriz de dispersión $\mathbf{S}$. Los componentes principales tienen las siguientes propiedades:

1. Dado que cualquier matriz de dispersión $\mathbf{S}$ es simétrica, sus ejes principales u_k son ortogonales entre sí. Es decir, que representan direcciones linealmente independientes en el elipsoide de densidad de la distribución de objetos.
2. Los autovalores λ_k de una matriz de dispersión $\mathbf{S}$ dan la cantidad de varianza que corresponde a cada uno de los sucesivos ejes principales.
3. Dadas las dos primeras propiedades, el análisis de componentes principales puede resumir, en unas pocas dimensiones, la mayor parte de la variabilidad de una matriz de dispersión con un gran número de variables. También provee de una medida de la variabilidad explicada por cada uno de esos pocos ejes principales independientes.

Un ejemplo sencillo usando solo dos variables, algo que en la práctica nunca sucede, pero resulta útil como ejemplo

$$Y = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix}$$ luego de centrar con las medias de las columnas $[y - \overline{y}] = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$

Calculando la matriz de dispersión:

$$S = \frac{1}{n-1}\left[y - \overline{y}\right]^{'}\left[y - \overline{y}\right] = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

La ecuación característica correspondiente es:

$$|S - \lambda_k I| = \left| \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} - \begin{bmatrix} \lambda_k & 0 \\ 0 & \lambda_k \end{bmatrix} \right| = 0$$

Tiene dos autovalores, $\lambda_1 = 9$ y $\lambda_2 = 2$. La varianza total es la misma, pero particionada de otra manera. La suma de la varianza en la diagonal de $\mathbf{S}$ es $8.2 + 5.8 = 14$, y la suma de los dos autovalores es $(9 + 5) = 14$. El primer componente principal tiene el 64.3% de la varianza ($\lambda_1 = 9$) y el segundo el resto, 35.7%. Hay tantos autovalores como variables, pero cada autovalor tiene cada vez menos varianza. Con los valores de λ_k podemos calcular los autovectores con la ecuación:

$$\left(\mathbf{S} - \lambda_k \mathbf{I}\right)\mathbf{u}_k = 0$$

Una vez que los vectores son normalizados (i.e. la longitud es uno, $\mathbf{u}^{'}\mathbf{u} = 1$) se convierten en columnas de la matriz $\mathbf{U}$:

$$U = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix}$$

Los signos de la matriz son totalmente arbitrarios, si se multiplica por -1 se consigue una imagen especular que es igual de buena representando los datos.

Los autovectores son ortogonales entre sí (incorrelados). Podemos comprobarlo con su producto cruzado $\mathbf{u}_1^{'}\mathbf{u}_2 = (0.8944 \times (-0.4472)) + (0.4472 \times 0.8944) = 0$. Además, los elementos de $\mathbf{U}$ son los cosenos del ángulo entre las variables originales. Usando esta propiedad, se puede ver que los ejes principales especifican una rotación de los ejes de $(0.8944) = 2634^{'}$.

Los elementos de los autovectores también son pesos (*loadings*) de las variables originales. Por lo tanto, la posición del objeto x_i en el primer eje principal está dada por la siguiente función o combinación linear:

$$f_{i1} = \left(y_{i1} - \overline{y}_1\right)u_{11} + \ldots + \left(y_{ip} - \overline{y}_p\right)u_{p1} = \left[y - \overline{y}\right]_i \mathbf{u}_1$$

Los valores de $\left(y_{ij} - \overline{y}_j\right)$ son los valores del objeto i en las variables j centrados y los valores de u_{i1} son lo pesos de las variables en el primer autovector. Las posiciones de los objetos con respecto a los ejes principales están dadas en la matriz F de variables transformadas. Es llamada *matriz de valores de componentes*:

$$\mathbf{F} = \left[y - \overline{y}\right]\mathbf{U}$$

Para el ejemplo esto sería:

$$\mathbf{F} = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix} \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} -3.578 & 0 \\ -1.342 & 2.236 \\ -1.342 & -2.236 \\ 3.130 & 2.236 \\ 3.130 & -2.236 \end{bmatrix}$$

En este caso simple, con solo dos variables, los componentes principales son una representación perfecta de la variabilidad y es solo una rotación de estas dos variables. Cuando hay más de dos variables, como es usual, el análisis de componentes principales también realiza una rotación del sistema de variables-ejes, pero en

un espacio multidimensional. En este caso, los componentes principales I y II definen un plano que permite la representación de la mayor variabilidad. Los objetos son proyectados en este plano de tal forma que conserven, lo más posible, las distancias euclídeas que tienen en el espacio multidimensional de las variables originales.

La posición relativa de los objetos en el espacio p-dimensional rotado de los componentes principales son las mismas que en el espacio p-dimensional de las variables originales. Esto significa que **las distancias euclídeas entre los objetos se conservan a través de la rotación de los ejes**. Esta es una de las propiedades más importantes de análisis de componentes principales.

La calidad de la representación en un espacio euclídeo reducido con m dimensiones puede ser estudiada con la relación

$$\left(\sum_{k=1}^{m} \lambda_k\right) / \left(\sum_{k=1}^{p} \lambda_k\right)$$

Esta relación es equivalente al coeficiente de determinación (R^2) en el análisis de regresión.

El análisis componentes principales se pueden usar para estudiar el rol de las variables en la conformación de los componentes principales. Esto se puede ver en varias maneras: matriz de autovectores, proyección en un espacio reducido (matriz $\mathbf{U\Lambda}^{1/2}$), y proyección en un espacio reducido (matriz $\mathbf{U}$)

1. La matriz de autovectores – Como la matriz U contiene los autovectores normalizados, la diagonal $\mathbf{U'U}$ es igual 1 y los elementos fuera de la marginal es igual 0 porque los autovectores son ortogonales entre sí.

$$\mathbf{U}^{'}\mathbf{U} = \mathbf{I}$$

Por lo tanto, las variables tienen longitud de unidad en el espacio multidimensional y están a 90° entre sí (ortogonales) dado que los autovectores son ortogonales entre sí. Esto es así porque el análisis de componentes es una rotación en el espacio multidimensional. Además, al normalizar los autovectores normaliza los ejes de las variables:

$$\mathbf{U} = \begin{bmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & & \vdots \\ u_{p1} & \cdots & u_{pp} \\ \sqrt{\sum u_{j1}^2} = 1 & \cdots & \sqrt{\sum u_{jp}^2} = 1 \end{bmatrix} \begin{matrix} \sqrt{\sum u_{1k}^2} = 1 \\ \vdots \\ \sqrt{\sum u_{pk}^2} = 1 \\ \end{matrix}$$

Otra forma de estudiar la relación entre los predictores consiste en escalar los autovectores de tal forma que los cosenos de los ángulos entre los ejes de las variables sean **proporcionales** a su *covarianza*. Se logra escalando cada autovector k a una longitud igual a su desvío estándar $\sqrt{\lambda_k}$. La *distancia euclídea* entre objetos **no se conserva** de esta manera.

Usando la matriz diagonal de autovalores $\mathbf{\Lambda}$ se puede computar la nueva de autovectores $\mathbf{U\Lambda}^{\frac{1}{2}}$:

$$\mathbf{U\Lambda}^{\frac{1}{2}} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} \begin{bmatrix} \sqrt{9} & 0 \\ 0 & \sqrt{5} \end{bmatrix} = \begin{bmatrix} 2.6633 & -1.000 \\ 1.3416 & 2.000 \end{bmatrix}$$

En este escalamiento, la relación entre las variables es la misma que en la matriz de dispersión $\mathbf{S}$

$$\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}} = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & \cdots & u_{1p}\sqrt{\lambda_p} \\ \vdots & & \vdots \\ u_{p1}\sqrt{\lambda_1} & \cdots & u_{pp}\sqrt{\lambda_p} \end{bmatrix} \begin{matrix} \sqrt{\sum \left(u_{1k}\sqrt{\lambda_k}\right)^2} = s_1 \\ \vdots \\ \sqrt{\sum \left(u_{pk}\sqrt{\lambda_k}\right)^2} = s_p \end{matrix}$$
$$\sqrt{\sum \left(u_{j1}\sqrt{\lambda_1}\right)^2} = \sqrt{\lambda_1} \quad \cdots \quad \sqrt{\sum \left(u_{jp}\sqrt{\lambda_p}\right)^2} = \sqrt{\lambda_p}$$

1. Proyección de las variables en un espacio reducido (matriz $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$): Lo interesante no es todo el espacio multidimensional, sino la proyección simplificada del mismo en un espacio reducido, en general dos dimensiones. Los elementos $u_{jk}\sqrt{\lambda_k}$ de los autovectores, escalados en $\sqrt{\lambda_k}$, son coordenadas de la proyección de las j variables en los diferentes ejes k. Se grafican como **flechas** ya que son **ejes**. Sus proyecciones son más cortas o iguales que sus longitudes en el espacio multidimensional.

Los ángulos entre las variables son proyecciones de sus verdaderos ángulos de covarianza. Por lo tanto, es importante solo considerar variables que estén bien representadas en el plano de proyección. Una forma de evaluarlo es verificando la longitud de la proyección. En un espacio reducido la longitud de la proyección del eje de la variable es $s_j\sqrt{d/p}$. Esta expresión define la *contribución en equilibrio de la variable* en los varios ejes del espacio multidimensional.

Esta medida nos sirve para comparar con la longitud de la variable y ver si su contribución en mayor o menor de lo esperado bajo la hipótesis de contribuciones iguales en todos los ejes principales. Para nuestro ejemplo, la longitud de las variables:

$$\text{longitud de la primera variable} = \sqrt{2.6833^2 + (-1.000)^2} = 2.8636$$

$$\text{longitud de la segunda variable} = \sqrt{1.3516^2 + 2.000^2} = 2.4083$$

Como el ejemplo tiene solo dos variables, las longitudes son iguales a las contribuciones en equilibrio:

$$\text{proyeccion en equilibrio de la variable } 1 = s_1\sqrt{2/2} = \sqrt{8.2}\sqrt{2/2} = 2.8636$$

$$\text{proyeccion en equilibrio de la variable } 2 = s_2\sqrt{2/2} = \sqrt{5.8}\sqrt{2/2} = 2.4083$$

La matriz de correlación $\mathbf{R}$ está conectada a la matriz de dispersión $\mathbf{S}$ por la diagonal de matriz de desvíos estándar $\mathbf{D}(s)$. El coseno del ángulo α_{jl} entre dos variables y_j e y_l en el espacio multidimensional, está por lo tanto relaciona a la correlación (r_{jl}); puede demostrarse que $\cos(\alpha_{jl}) = r_{jl}$. Este ángulo es igual a la covarianza, porque la estandarización cambió las longitudes de las variables a la unidad. En el ejemplo, la correlación entre las dos variables es igual a $\frac{1.6}{\sqrt{8.2\times5.8}} = 0.232$. El ángulo correspondiente es $0.232 = 7635'$.

Igualmente, el ángulo entre la variable j y el eje principal k, en el espacio multidimensional, es el arco coseno de la correlación entre la variable j y el eje principal k. Esta correlación es el elemento jk de la nueva matriz de autovectores:

$$u_{jk}\sqrt{\lambda_k}/s_j$$

Es decir que la correlación es calculada pesando cada elemento de los autovectores por la relación del desvío estándar del eje principal al de la variable. Para el ejemplo, los desvíos estándar $s_1 = 2.8636$ $s_2 = 2.4083$:

$$\begin{bmatrix} \frac{u_{jk}\sqrt{\lambda_k}}{s_j} \end{bmatrix} = \begin{bmatrix} 0.9370 & -0.3492 \\ 0.5571 & 0.8305 \end{bmatrix} \begin{bmatrix} 2026' & 11026' \\ 5609' & 3351' \end{bmatrix}$$

Una consecuencia importante de esto es que las variables con **correlación más alta**, en valor absoluto, son las que **más contribuyen** a cada autovector. Sin embargo, no se puede hacer la prueba estadística de Pearson para los coeficientes de correlación porque los componentes principales son combinaciones lineales de las variables.

Cuando los ejes de las variables de $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$ están estandarizados a longitud unidad sus proyecciones en el espacio principal **no es recomendada** porque los autovectores re-escalados no son necesariamente ortogonales y pueden tener cualquier longitud.

La matriz de las proyecciones de los ejes de las variables de la matriz puede ser examinada con respecto a los siguientes puntos:

- Las proyecciones de las coordenadas de los ejes de las variables especifican la posición de los ápices de ese eje de variable en el espacio reducido. Se recomienda usar flechas para representarlos.
- La proyección de los $u_{jk}\sqrt{\lambda_k}$ del eje de la variable j en el eje principal k muestra su covarianza con respecto al eje principal, y su signo.
- Verificar variables cuyas **longitudes proyectadas alcancen o excedan** los valores de sus respectivas contribuciones en equilibrio. Los ejes de variables que sean claramente más cortos que esto contribuyen poco a la formación del espacio reducido bajo estudio y, por lo tanto, contribuyen poco a la estructura que se puede ser encontrada para los objetos en ese espacio reducido.
- La **correlación** entre las variables está dada por el **ángulo** entre los ejes de las variables y **no** por la proximidad entre los ápices de los ejes. Hay que tener en cuenta que, la proyección de los ejes en el espacio reducido puede no ser la representación completa de la correlación espacio multidimensional. Puede ser mejor agrupar variables, en el espacio reducido del gráfico, con respecto al espacio multidimensional, realizando un método *de análisis de agrupamiento.*
- Los objetos pueden ser proyectados en ángulo recto sobre los ejes de las variables de acuerdo a sus valores en esas variables. Sin embargo, las distancias entre los objetos **no son aproximaciones de sus distancias euclídeas**.

1. Proyección de las variables en un espacio reducido (matriz **U**): Difiere de lo anterior en los autovectores no han sido escalados a las longitudes de sus desvíos estándar. Los ángulos entre los ejes de las variables y los ejes principales son proyecciones de sus ángulos de rotación. Por ejemplo:

$$U = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} \begin{bmatrix} 2634' & 11634' \\ 6326' & 2634' \end{bmatrix}$$

En esta proyección no es posible interpretar las correlaciones entre las variables ya que siempre son ortogonales en esta representación, donde los autovectores están escalados a 1.

La proyección de u_{jk} de una variable j en el eje principal k es **proporcional** a la **covarianza** de ese descriptor con el eje principal. Se puede comparar la proyección de diferentes ejes de variables en el mismo eje principal. Se puede probar que una proyección isogonal (con ángulos iguales) de p ejes ortogonales de longitud uno da una longitud de $\sqrt{\frac{d}{p}}$ en cada eje de un espacio d-dimensional.

Se puede dibujar un *círculo de equilibrio de las variables* como referencia para evaluar la contribución de cada variable a la conformación del espacio reducido.

Variable centrada j	Escalado de los autovectores	
	$\sqrt{\lambda_k}$	1
Longitud total	s_{j}	1

Variable centrada j	Escalado de los autovectores	
Ángulos en el espacio reducido	Proyección de las covarianzas (correlaciones)	90°, rotaciones rígidas del sistema de ejes
Longitud de la contribución en equilibrio	$s_j\sqrt{d/p}$	Círculo con radio $\sqrt{\frac{d}{p}}$
Proyección en el eje principal k	$u_{jk}\sqrt{\lambda_k}$ (la covarianza con el componente k)	u_{jk} (proporcional a la covarianza con k)
Correlación con el eje principal k	$u_{jk}\sqrt{\lambda_k}/s_j$	$u_{jk}\sqrt{\lambda_k}/s_j$

13.1.1 Biplots

Se le llama biplot al gráfico de componentes principales en donde se grafican al mismo tiempo los ejes de las variables y los objetos en el espacio reducido. Existen dos tipos de gráficos biplots según el escalamiento que se use:

- los de *distancia* se hacen con la yuxtaposición de la matriz $\mathbf{U}$ (los autovectores escalados a longitud unidad) y $\mathbf{F}$ (donde cada componente principal k está escalado a la varianza $= \lambda_k$),
- los de *correlación* usan la matriz $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$ para las variables (cada autovector escalado a longitud $\sqrt{\lambda_k}$) y la matriz $\mathbf{G} = \mathbf{F}\mathbf{\Lambda}^{\frac{1}{2}}$ para los objetos, cuyas columnas tienen varianzas de unidad.

Las principales propiedades de los biplot de distancias son:

1. La distancia entre los objetos del biplot **son aproximaciones** de sus distancias euclídeas en el espacio multidimensional.
2. La proyección del objeto en ángulo recto sobre la variable da la posición aproximada del objeto en esa variable.
3. Dado que las variables tienen longitud 1 en el espacio multidimensional, la longitud de su proyección sobre el espacio reducido indica **cuanto contribuye** a la formación de ese espacio.
4. Los ángulos entre los vectores de las variables no tienen interpretación

En los biplots de correlación:

1. Las distancias entre los objetos del biplot **no son aproximaciones** de sus distancias euclídeas en el espacio multidimensional.
2. La proyección del objeto en ángulo recto sobre la variable da la posición aproximada del objeto en esa variable.
3. Dado que la longitud de las variables es s_j en un espacio multidimensional completo, la longitud de la proyección de la variable en el espacio reducido es una aproximación de su **desvío estándar**.
4. Lo **ángulos** entre las variables reflejan su **correlación**.

En cualquiera de los dos casos, los objetos o variables pueden ser multiplicados por una constante para producir un gráfico claro.

13.2 Componentes principales de una matriz de correlación

También puede realizar este análisis sobre una matriz $\mathbf{R}$ de correlación, ya que las correlaciones son las covarianzas estandarizadas de las variables. La suma de autovalores de S es igual a la suma de varianzas,

mientras que la suma de autovalores de $\mathbf{R}$ es igual p, por lo que los autovalores y por lo tanto los autovectores son diferentes. Esto se debe a que las distancias entre los objetos no son las mismas en los dos casos.

En el caso de las correlaciones, las variables están estandarizadas. Por lo tanto, las distancias entre los objetos son independientes de las unidades de medición, por otro lado, las que están en el espacio original de medida cambian de acuerdo a su cambio en unidad de medida. Cuando todas las variables son del mismo orden de magnitud y tienen las mismas unidades conviene usar la matriz $\mathbf{S}$. En ese caso, los autovectores y los coeficientes de correlación entre las variables y los componen proporcionan información complementaria. El primero da la ponderación de las variables y el segundo cuantifica su importancia relativa. Cuando las variables son de naturaleza diferente, puede ser necesario usar la matriz $\mathbf{R}$ en vez de $\mathbf{S}$.

¿Cuándo usar una $\mathbf{S}$ o $\mathbf{R}$?

- Si uno quiere agrupar los objetos en el espacio reducido ¿El agrupamiento debe hacerse con respecto a las variables originales, por lo tanto, preservando sus diferencias en magnitud? ¿O las variables deberían contribuir de igual forma al agrupamiento de los objetos, independientemente de su varianza? En el segundo caso uno debería proceder con la matriz de correlación

Otra forma de ver esto es:

- Considere que la distancia euclídea es la que se conserva entre los objetos con el análisis de componente principales. ¿Qué es más interesante de interpretar en términos de la configuración espacial de las distancias euclídeas? La covarianza (los datos crudos) o las correlaciones (los datos estandarizados)

Igual que con el caso anterior, el análisis de componentes principales es una rotación del sistema de ejes. Pero, como ahora las variables están *estandarizadas*, los objetos no están posicionados de la misma forma que si las variables fuesen solo *centradas*.

Las conclusiones de lo visto anteriormente no cambian, solo hay que reemplazar matriz de dispersión $\mathbf{S}$ por matriz de correlación $\mathbf{R}$, covarianza por correlación y s_{jl} por r_{jl}. Por lo tanto, las varianzas y desvíos estándar son iguales a 1. Lo que da lugar a ciertas propiedades especiales para la matriz $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$. Primero, $\mathbf{D}(s) = \mathbf{I}$, por lo que $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}} = \mathbf{D}(s)^{-1}\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$, esto significa que los coeficientes $u_{jk}\sqrt{\lambda_k}$ son los coeficientes de correlación entre las variables j y los componentes k. La contribución en equilibrio, en el espacio reducido de $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$, es $s_j\sqrt{\frac{d}{p}} = \sqrt{\frac{d}{p}}$ ($s_j = 1$). Por lo tanto, es posible juzgar si la contribución de una variable es mayor o menor a lo esperado comparando la longitud de las proyecciones a un círculo de equilibrio con radio $\sqrt{\frac{d}{p}}$.

Las propiedades principales de un variable estandarizada se dan en la tabla de abajo.

Variable estandarizada j	Escalado de los autovectores	
	$\sqrt{\lambda_k}$	1
Longitud total	1	1
Ángulos en el espacio reducido	Proyección de las correlaciones	90°, rotaciones rígidas del sistema de ejes
Longitud de la contribución en equilibrio	$\sqrt{d/p}$	Círculo con radio $\sqrt{\frac{d}{p}}$
Proyección en el eje principal k	$u_{jk}\sqrt{\lambda_k}$ (la covarianza con el componente k)	u_{jk} (proporcional a la covarianza con k)
Correlación con el eje principal k	$u_{jk}\sqrt{\lambda_k}$	$u_{jk}\sqrt{\lambda_k}$

13.3 ¿Cuantos componentes son significativos?

Una propiedad de los componentes es que cada uno representa una cantidad cada vez menor de la varianza total. Por lo tanto, un problema es cuantos componentes son significativos en términos biológicos. La misma pregunta vista de otra manera es, cuantas dimensiones tendría que tener el espacio reducido. La mejor manera de ver esto es con diagrama de Shepard. Sin embargo, dado que el análisis de componentes principales es una forma de partición de la varianza uno podría realizar una prueba formal para la varianza asociada con los sucesivos ejes principales.

Existen varias pruebas clásicas para contestar esta pregunta, pero el problema que tienen es que requieren normalidad en las variables, una condición que rara vez se cumple por datos ecológicos.

Hay una regla empírica que sugiere que solo se deben interpretar los componentes principales si su autovalor λ es mayor que la media de los λ. Este es llamado el criterio Kaiser-Guttman.

Otra forma, también empírica, es comparar los valores decrecientes de los autovalores con los valores de modelo de bastón roto. Considere que la varianza es un recurso embebida en un bastón de longitud 1. Si los componentes principales dividieran la varianza al azar entre los ejes principales, las fracciones de la varianza explicada por los ejes principales tendría la misma longitud relativa que las piezas obtenidas al romper el bastón en puntos al azar en tantas piezas como ejes. Si un bastón de unidad es roto en al azar en $p = 2,\ 3,\ \ldots$ piezas, los valores esperados (E) de las longitudes relativas de las piezas sucesivamente menores (j) están dados por la ecuación:

$$E\left(\text{pieza}_j\right) = \frac{1}{p}\sum_{x=j}^{p}\frac{1}{x}$$

Los valores esperados son iguales a la media de las longitudes que fuesen obtenidas al romper el bastón al azar muchas veces y calcular la media de la pieza más grande, la segunda más grande, etc. No tendría sentido interpretar los ejes principales que explican una fracción de la varianza menor o igual que la predicha por el modelo del bastón roto. Puede comprobarse que ejes deben interpretarse consultando una tabla y seleccionando los autovalores que son mayores que las predicciones del modelo. O comparar la suma de los autovalores de 1 hasta k con la suma de los valores de 1 hasta k en el modelo. Esta prueba generalmente selecciona los primeros dos o tres componentes principales.

13.4 Mal uso de los componentes principales

Los errores más comunes son: uso de las variables para las cuales la covarianza no tiene sentido y la interpretación de la relación entre variables, en el espacio reducido, basa en las posiciones relativas de los ápices de los ejes en vez de los ángulos entre ellas.

El análisis de componentes principales fue definido originalmente para el estudio de datos con distribución multinormal, por lo que para usarlo óptimamente es necesario normalizar los datos. Las desviaciones de la normalidad no afectan necesariamente el análisis. Hay que tener cuidado con las distribuciones sesgadas, los primeros ejes principales solo van a separar los pocos objetos con valores extremos del resto, en vez de mostrar los ejes principales de variación de todos los objetos en estudio.

El método debe ser usado con una matriz de varianzas o correlaciones con las siguientes propiedades: a) la matriz S o R ha sido calculada entre variables b) que son cuantitativos y c) para las cuales estimadores validos de la covarianza pueden ser obtenidos. Esto se viola bajo las siguientes condiciones:

1. Una matriz de dispersión no puede ser estimada cuando el número de observaciones n es menor o igual al número de variables p. El número de objetos de ser mayor a p para obtener estimadores validos de la matriz de dispersión. Sin embargo, los primeros ejes principales son poco afectados por

cuando la matriz no es rango completo. Por lo que no debería haber interpretaciones incorrectas de las ordenaciones en espacio reducido.

2. Algunos autores han transpuesto las matriz original y computado correlaciones entre objetos en vez de entre variables. Esto no tiene sentido porque el análisis produce información tanto de los objetos como de las variables. Además, la covarianza entre objetos no tiene sentido. Y la correlación implica estandarización de los vectores, y solo tiene sentido para datos dimensionalmente homogéneos.

3. Las covarianzas y correlaciones solo están definidas para variables cuantitativas. Sin embargo, el análisis de componentes principales es muy robusto a variaciones de precisión de los datos. Las variables pueden ser recodificadas en pocas clases sin cambiar notablemente los resultados. Los coeficientes de correlación usando datos semicuantitativos son equivalentes al coeficiente de correlación de rangos de Spearman.

4. Cuando se calcula en conjuntos de datos con muchos doble ceros, los coeficientes de covarianza o correlación dan ordenaciones que producen estimadores inadecuados de las distancias entre objetos. Con este tipo de datos solo se debe usar componentes principales en gradientes pequeños.