ANOVA

Nicolas Jadan

Visualice sus datos y calcule ANOVA unidireccional en R

Importar los datos a R

```
# Or, if .csv file, use this
my_data <- read.csv("cancer.csv")
```

Comprueba tus datos

Para tener una idea de cómo se ven los datos, usamos la función **sample_n**()[en el paquete **dplyr**]. La función **sample_n**() selecciona aleatoriamente algunas de las observaciones en el marco de datos para imprimir:

```
# Show a random sample
set.seed(1234)
dplyr::sample_n(my_data, 10)
```

```
   group weight
1      M 16.240
2      M 13.610
3      B 11.800
4      B  9.787
5      B 12.180
6      B 12.670
7      M 20.180
8      B 10.710
9      B 11.040
10     B 11.410
```

- Calcular estadísticas de resumen por grupos: recuento, media, sd:

```
# Show the levels
levels(my_data$group)
```

```
NULL
```

```
my_data$group <- ordered(my_data$group,
                         levels = c("B", "M"))
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
group_by(my_data, group) %>%
  summarise(
    count = n(),
    mean = mean(weight, na.rm = TRUE),
    sd = sd(weight, na.rm = TRUE)
  )
```

```
# A tibble: 2 x 4
  group count  mean    sd
  <ord> <int> <dbl> <dbl>
1 B       357  12.1  1.78
2 M       212  17.5  3.20
```

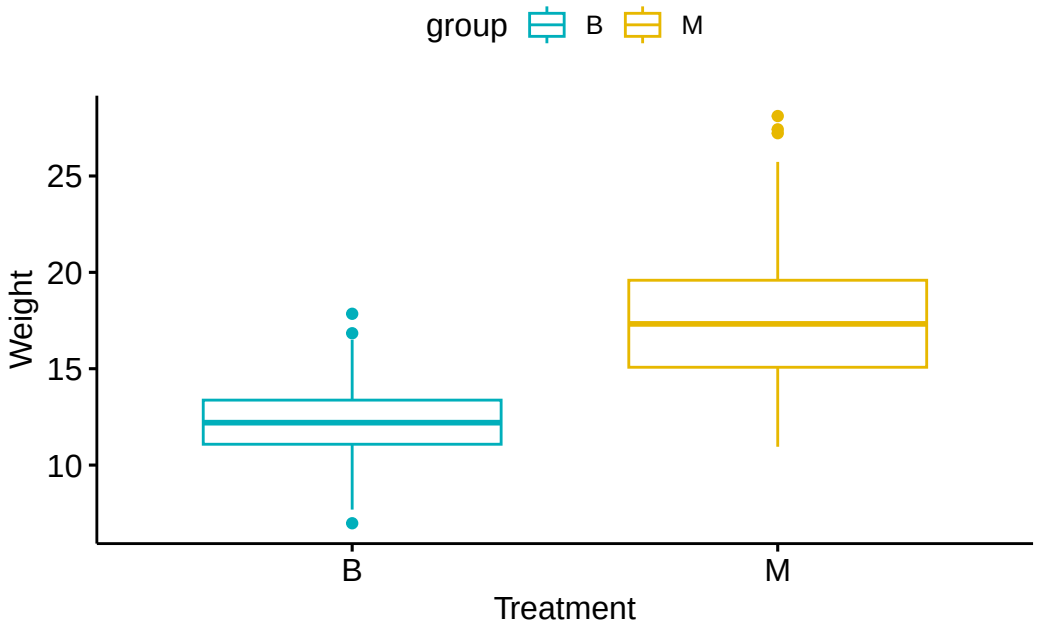
Visualiza tus datos

- Para usar gráficos base de R, lea esto: Gráficos de base de R. Aquí, usaremos el paquete **ggpubr** R para una fácil visualización de datos basada en ggplot2.
- Visualiza tus datos con ggpubr:

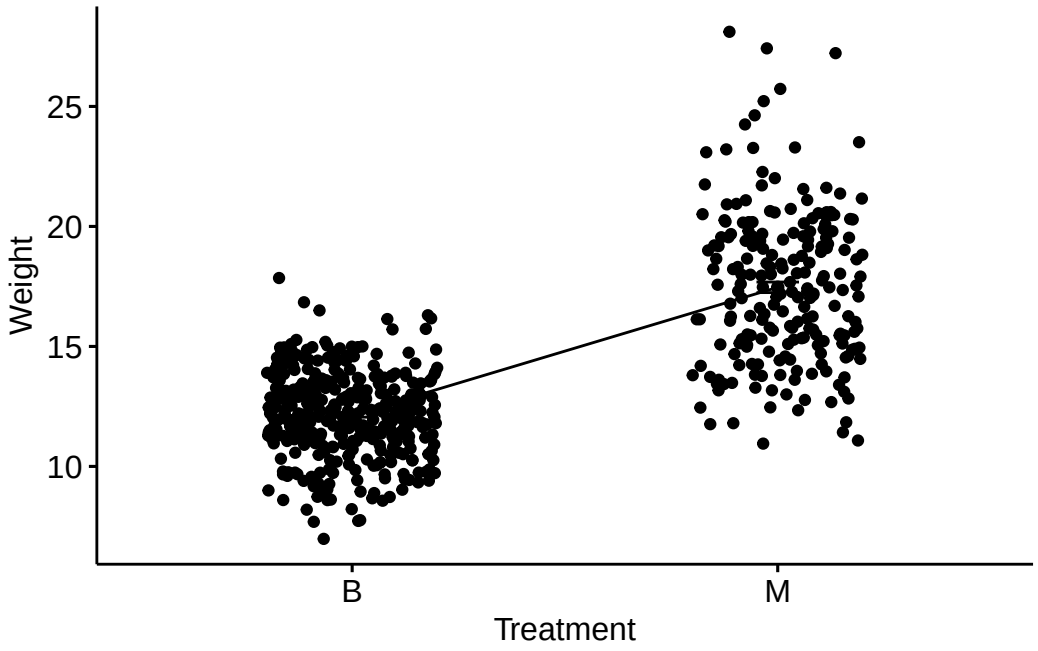
```
# Box plots
# ++++++++++++++++++++
# Plot weight by group and color by group
library("ggpubr")
```

```
Loading required package: ggplot2
```

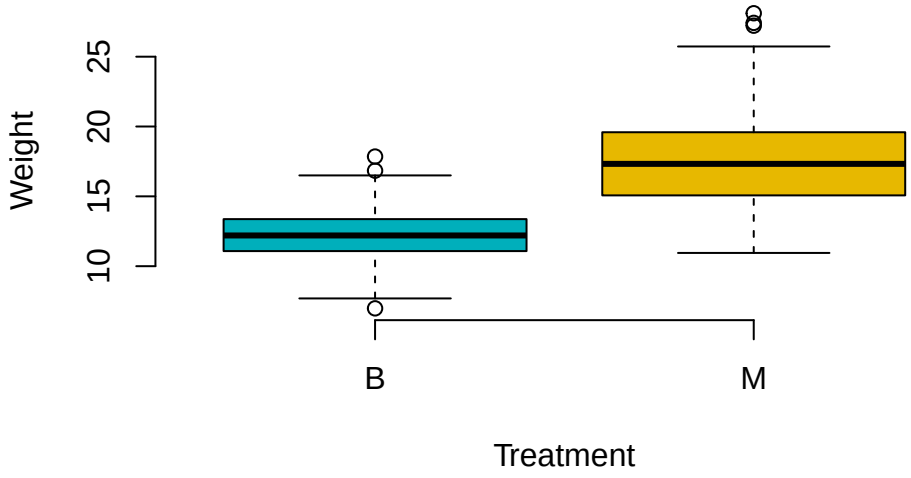
```
ggboxplot(my_data, x = "group", y = "weight",
          color = "group", palette = c("#00AFBB", "#E7B800", "#FC4E07"),
          order = c("B", "M"),
          ylab = "Weight", xlab = "Treatment")
```



```r
# Mean plots
# ++++++++++++++++++++
# Plot weight by group
# Add error bars: mean_se
# (other values include: mean_sd, mean_ci, median_iqr, ....)
library("ggpubr")
ggline(my_data, x = "group", y = "weight",
       add = c("mean_se", "jitter"),
       order = c("B", "M"),
       ylab = "Weight", xlab = "Treatment")
```




```r
# Box plot
boxplot(weight ~ group, data = my_data,
        xlab = "Treatment", ylab = "Weight",
        frame = FALSE, col = c("#00AFBB", "#E7B800", "#FC4E07"))
```




```
# plotmeans
library("gplots")
```

```
Attaching package: 'gplots'
```

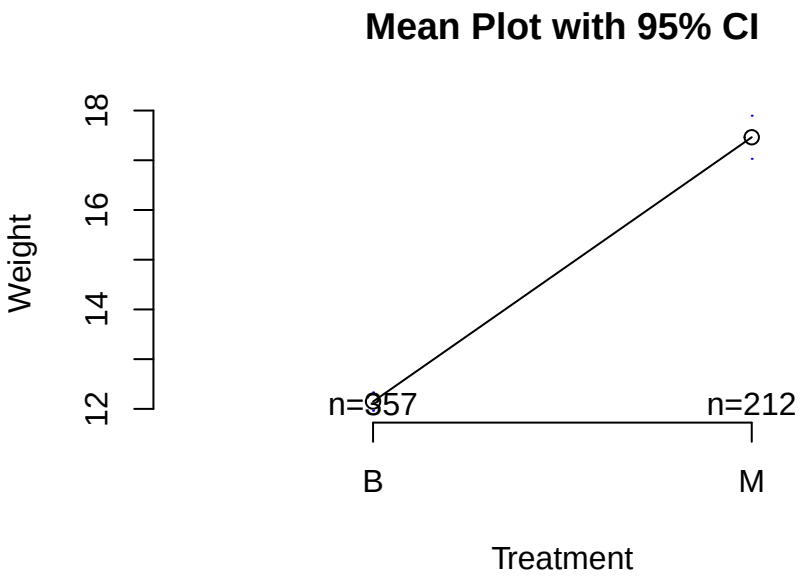
```
The following object is masked from 'package:stats':

    lowess
```

```
plotmeans(weight ~ group, data = my_data, frame = FALSE,
          xlab = "Treatment", ylab = "Weight",
          main="Mean Plot with 95% CI")
```

```
Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, :
zero-length arrow is of indeterminate angle and so skipped
```

```
Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, :
zero-length arrow is of indeterminate angle and so skipped
```

```
Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, :
zero-length arrow is of indeterminate angle and so skipped

Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, :
zero-length arrow is of indeterminate angle and so skipped

Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
graphical parameter

Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" is not
a graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
graphical parameter
```



Calcular la prueba ANOVA unidireccional

La función R **aov**() se puede utilizar para responder a esta pregunta. La función **summary.aov**() se utiliza para resumir el modelo de análisis de varianza.

```r
# Compute the analysis of variance
res.aov <- aov(weight ~ group, data = my_data)
# Summary of the analysis
summary(res.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
group         1   3759    3759     647 <2e-16 ***
Residuals   567   3295       6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretar el resultado de las pruebas ANOVA unidireccionales

Como el valor p es menor que el nivel de significancia 0,05, podemos concluir que existen diferencias significativas entre los grupos resaltados con "*" en el resumen del modelo.

Comparación múltiple por pares entre las medias de los grupos

En la prueba ANOVA unidireccional, un valor p significativo indica que algunas de las medias grupales son diferentes, pero no sabemos qué pares de grupos son diferentes.

Es posible realizar múltiples comparaciones por pares, para determinar si la diferencia media entre pares específicos de grupo es estadísticamente significativa.

Comparaciones múltiples por pares de Tukey

Como la prueba ANOVA es significativa, podemos calcular Tukey **HSD (Tukey** Honest Significant Differences, función R: **TukeyHSD**()) para realizar múltiples comparaciones por pares entre las medias de los grupos.

La función **TukeyHD()** toma el ANOVA instalado como argumento.

```r
TukeyHSD(res.aov)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight ~ group, data = my_data)
```

```
$group
        diff      lwr      upr p adj
M-B 5.316306 4.905781 5.726832     0
```

- **diff**: diferencia entre las medias de los dos grupos
- **LWR**, **UPR**: el punto final inferior y superior del intervalo de confianza al 95% (predeterminado)
- **p adj**: valor p después del ajuste para las comparaciones múltiples.

Comparaciones múltiples usando el paquete multcomp

Es posible usar la función **glht**() [en el paquete **multcomp**] para realizar múltiples procedimientos de comparación para un ANOVA. **GLHT** significa pruebas generales de hipótesis lineales. El formato simplificado es el siguiente:

```
#summary(glht(res.aov, linfct = mcp(group = "Tukey")))
```

Prueba t de Pairewise

La función **pairewise.t.test**() también se puede utilizar para calcular comparaciones por pares entre niveles de grupo con correcciones para pruebas múltiples.

```
pairwise.t.test(my_data$weight, my_data$group,
                 p.adjust.method = "BH")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  my_data$weight and my_data$group

  B
M <2e-16

P value adjustment method: BH
```

El resultado es una tabla de valores p para las comparaciones por pares. Aquí, los valores p han sido ajustados por el método de Benjamini-Hochberg.

Verifique los supuestos de ANOVA: ¿validez de la prueba?

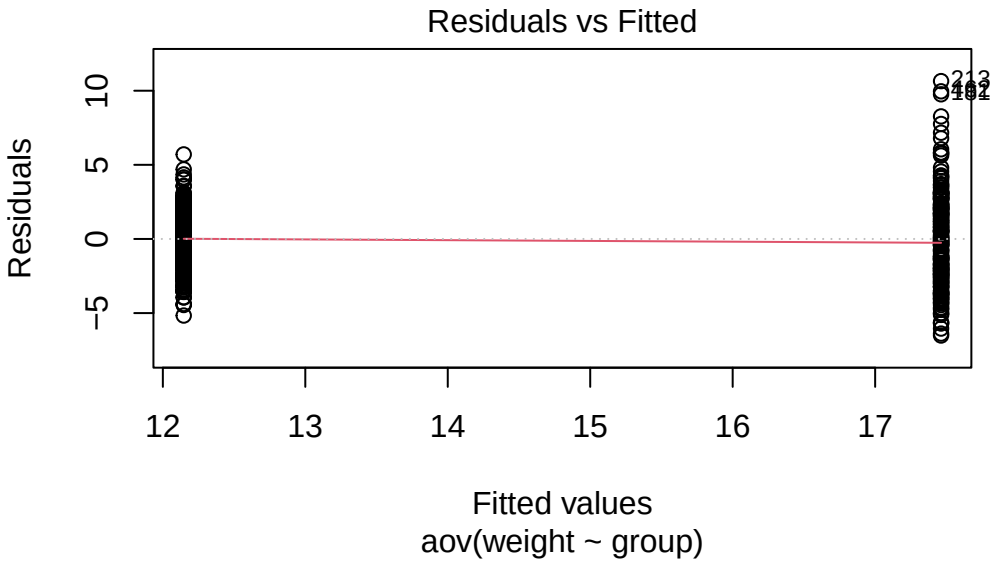
La prueba ANOVA asume que los datos se distribuyen normalmente y la varianza entre los grupos es homogénea. Podemos comprobarlo con algunas gráficas diagnósticas.

Comprobar la homogeneidad de la hipótesis de varianza

La **gráfica de residuos versus ajustes** se puede utilizar para verificar la homogeneidad de las varianzas.

En la siguiente gráfica, no hay relaciones evidentes entre los residuos y los valores ajustados (la media de cada grupo), lo cual es bueno. Por lo tanto, podemos asumir la homogeneidad de las varianzas.

```
# 1. Homogeneity of variances
plot(res.aov, 1)
```



Recomendamos la **prueba de Levene**, que es menos sensible a las desviaciones de la distribución normal. Se utilizará la función **leveneTest**() [en **el paquete car**]:

```
library(car)
```

```
Loading required package: carData


Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode
```

```r
leveneTest(weight ~ group, data = my_data)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   1  90.477 < 2.2e-16 ***
      567
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De la salida anterior podemos ver que el valor p no es menor que el nivel de significación de 0.05. Esto significa que no hay evidencia que sugiera que la varianza entre los grupos sea estadísticamente significativamente diferente. Por lo tanto, podemos asumir la homogeneidad de las varianzas en los diferentes grupos de tratamiento

Relajar la homogeneidad de la hipótesis de varianza

La prueba clásica de ANOVA unidireccional requiere una suposición de varianzas iguales para todos los grupos. En nuestro ejemplo, la homogeneidad de la suposición de varianza resultó estar bien: la prueba de Levene no es significativa.

Un procedimiento alternativo (es decir: **Welch one-way** test), que no requiere que la suposición se haya implementado en la función **oneway.test**().

- **Prueba de ANOVA sin suposición de varianzas iguales**

```r
oneway.test(weight ~ group, data = my_data)
```

```
    One-way analysis of means (not assuming equal variances)

data:  weight and group
F = 493.23, num df = 1.00, denom df = 289.71, p-value < 2.2e-16
```

- **Pruebas t por pares sin suposición de varianzas iguales**

```r
pairwise.t.test(my_data$weight, my_data$group,
                 p.adjust.method = "BH", pool.sd = FALSE)
```

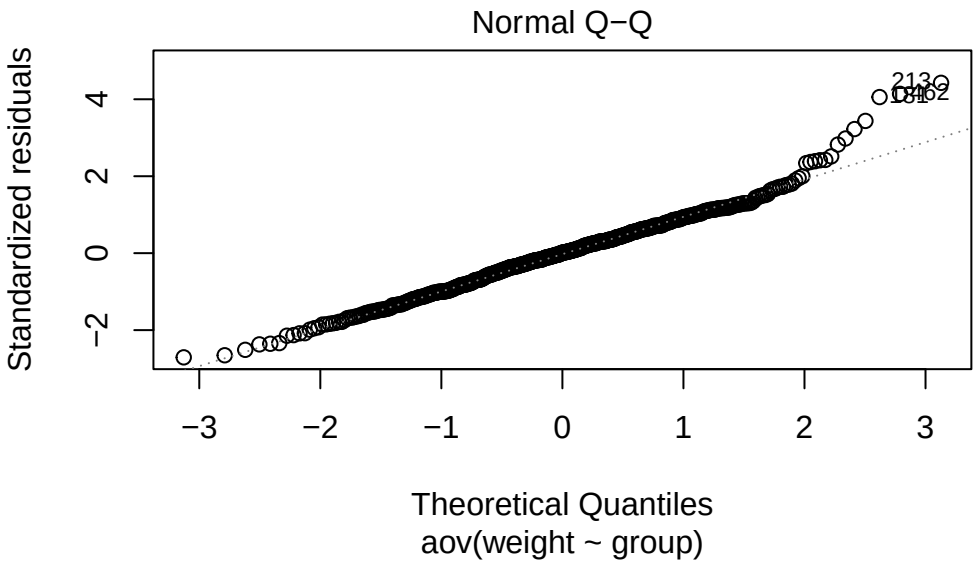
```
    Pairwise comparisons using t tests with non-pooled SD

data:  my_data$weight and my_data$group

  B
M <2e-16

P value adjustment method: BH
```

Comprobar el supuesto de normalidad

Diagrama de normalidad de residuos. En la siguiente gráfica, los cuantiles de los residuos se representan contra los cuantiles de la distribución normal. También se traza una línea de referencia de 45 grados.

La gráfica de probabilidad normal de los residuos se utiliza para comprobar la suposición de que los residuos están distribuidos normalmente. Debe seguir aproximadamente una línea recta.

```r
# 2. Normality
plot(res.aov, 2)
```



Como todos los puntos caen aproximadamente a lo largo de esta línea de referencia, podemos asumir la normalidad.

La conclusión anterior está respaldada por la **prueba de Shapiro-Wilk** en los residuos ANOVA (W = 0.98151, p = 1.292e-06

) que no encuentra indicios de que se viole la normalidad.

```
# Extract the residuals
aov_residuals <- residuals(object = res.aov )
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

```
    Shapiro-Wilk normality test

data:  aov_residuals
W = 0.98151, p-value = 1.292e-06
```

Tenga en cuenta que, una alternativa no paramétrica al ANOVA unidireccional es la **prueba de suma de rangos** de **Kruskal-Wallis**, que se puede usar cuando no se cumplen los supuestos de ANNOVA.

```r
kruskal.test(weight ~ group, data = my_data)
```

```
    Kruskal-Wallis rank sum test

data:  weight by group
Kruskal-Wallis chi-squared = 305, df = 1, p-value < 2.2e-16
```