Análisis de Medias en R

Contents

Introducción:

Un analisis estadistico de medias consiste en el empleo de métodos de estadistica inferencial para analizar a nivel poblacional la media de una o mas variables.

En este articulo se hará una revisión de distintos metodos estadisticos para llevar a cabo un analisis de medias.

Se recomienda haber leido previamente el siguiente articulo: https://rpubs.com/FabioScielzoOrtiz/Metodologia_Contrastes_de_Hipotesis

Contrastes de Hipotesis para la media de una variable en un grupo

Objetivo

- Contrastar la media de una variable sobre una grupo/población.
  - Ejemplo: contrastar si la nota media en matematicas de los alumnos de cierto colegio es mayor que 7.

Planteamiento formal del problema:

- Tenemos un **grupo/población** $G = \{e_{11}, e_{21}, ..., e_{N_{G,1}}\}$ con N_G elementos
- Tenemos una **muestra** g de n elementos de G

- Tenemos variable estadistica **cuantitativa** X_k medida sobre la muestra g del grupo G:

$$X_{k,g} = (x_{1k}\ ,\ x_{2k}, ...,\ x_{nk})^t \tag{1}$$

Observación: x_{ik} es el valor de X_k para el i-esimo individuo de la muestra g del grupo G sobre la que se ha medido X_k

- Desconocemos X_k medida sobre el grupo/poblacion G, a la que denotaremos como $X_{k,G}$

Observaciones:

- $X_{k,G}$ tiene la misma naturaleza que X_k , en el sentido de que ambas son variables estadisticas.
- Pero se diferencia en que $X_{k,G}$ contiene los valores de la variable X_k para los elementos de la **poblacion** G , mientras que la X_k **medida sobre la muestra** solo contiene los valores de la variable X_k para los elementos de una **muestra** g de la población G
- Los términos **poblacion** y **grupo** serán usados como sinonimos.

Los **contrastes de hipotesis** que queremos resolver son del tipo:

$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$

Donde:

μ es la media (aritmética) de $X_{k,G}$, es decir, $\mu = \overline{X}_{k,G}$

Esta información (μ) se desconoce, puesto que desconocemos $X_{k,G}$

μ_0 es un valor conocido.

T - test: Grupo/Población Normal

Supuestos

- $X_{k,G}$ tiene una distribucion normal, con media y desviación típica igual a la de la propia variable.
  - $X_{k,G} \sim N(\mu, \sigma)$

Donde: $\sigma = \sigma(X_{k,G})$

Estadistico del contraste

- El estadistico del contraste es:

$$t_{exp} = \frac{\overline{X_{k,g}} - \mu}{S(X_{k,g})/\sqrt{n}} \sim t_{n-1}$$

- El estadistico del contraste **bajo** $H_0 : \mu = \mu_0$ es:

$$t_{exp|H_0} = \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} \sim t_{n-1}$$

Donde: $S(X_{k,g}) = \frac{n}{n-1} \cdot \sigma(X_{k,g})$ es la cuasidesviación típica de $X_{k,g}$

p-valor

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

- $pvalor = P\left(\underbrace{t_{exp|H_0}}_{v.a.\ \sim\ t_{n-1}} > \underbrace{t_{exp|H_0}}_{observacion} \right) = P\left(t_{n-1} > \dfrac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} \right)$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

  - $pvalor = P\left(\underbrace{t_{exp|H_0}}_{v.a.\ \sim\ t_{n-1}} < \underbrace{t_{exp|H_0}}_{observacion} \right) = P\left(t_{n-1} < \dfrac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} \right)$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

  - $pvalor = P\left(\mid \underbrace{t_{exp|H_0}}_{v.a.\ \sim\ t_{n-1}} \mid > \underbrace{\mid t_{exp|H_0} \mid}_{observacion} \right) = P\left(\underbrace{t_{exp|H_0}}_{v.a.} > \underbrace{\mid t_{exp|H_0} \mid}_{observacion} \right) + P\left(\underbrace{t_{exp|H_0}}_{v.a.} < -\underbrace{\mid t_{exp|H_0} \mid}_{observacion} \right)$

    $\underbrace{=}_{simetria\ t} P\left(t_{n-1} > \underbrace{\mid t_{exp|H_0} \mid}_{observacion} \right) + P\left(t_{n-1} > \underbrace{\mid t_{exp|H_0} \mid}_{observacion} \right) = 2 \cdot P\left(t_{n-1} > \left| \dfrac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} \right| \right)$

Regla de decisión

Basada en el estadistico del contraste

Para un nivel de significación α :

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$Rechazar H_0 \Leftrightarrow \underbrace{t_{exp|H_0}}_{observacion} > t^{\alpha}_{n-1} \Leftrightarrow \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} > t^{\alpha}_{n-1} \Leftrightarrow \overline{X_{k,g}} > \mu_0 + t^{\alpha}_{n-1} \cdot S(X_{k,g})/\sqrt{n}$$

Donde: $P(t_{n-1} > t^{\alpha}_{n-1}) = \alpha$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$Rechazar H_0 \Leftrightarrow t_{exp|H_0} < t^{1-\alpha}_{n-1} \Leftrightarrow \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} < t^{1-\alpha}_{n-1} \Leftrightarrow \overline{X_{k,g}} < \mu_0 + t^{1-\alpha}_{n-1} \cdot S(X_{k,g})/\sqrt{n}$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$Rechazar H_0 \Leftrightarrow t_{exp|H_0} > t^{\alpha/2}_{n-1} \text{ ó } t_{exp|H_0} < t^{1-\alpha/2}_{n-1} \Leftrightarrow \overline{X_{k,g}} > \mu_0 + t^{\alpha/2}_{n-1} \cdot S(X_{k,g})/\sqrt{n} \text{ ó } \overline{X_{k,g}} < \mu_0 + t^{1-\alpha/2}_{n-1} \cdot S(X_{k,g})/\sqrt{n}$$

Observaciones:

- Las reglas de decision en los contrastes de hipotesis son conservadoras, en el sentido de que, por ejemplo, en el caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ se rechazará H_0 en favor de H_1 cuando $\overline{X_{k,g}}$ sea **suficientemente** mayor que μ_0 , no vale que sea simplemente mayor que μ_0. Este hecho se extrapola a todos los contrastes de hipotesis, y es importante tenerlo presente.

- $P(Error\ Tipo\ I) = P(Rechazar\ H_0 \mid H_0) = \alpha$ en todos los casos, veamoslo:

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$P(RH_0|H_0) = P(\underbrace{t_{exp|H_0}}_{v.a \sim t_{n-1}} > t^{\alpha}_{n-1}) = P(t_{n-1} > t^{\alpha}_{n-1}) = \alpha$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$P(RH_0|H_0) = P(\underbrace{t_{exp|H_0}}_{v.a \sim t_{n-1}} < t^{1-\alpha}_{n-1}) = P(t_{n-1} < t^{1-\alpha}_{n-1}) = \alpha$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$P(RH_0|H_0) = P(\underbrace{t_{exp|H_0}}_{v.a \sim t_{n-1}} > t^{\alpha/2}_{n-1} \text{ ó } \underbrace{t_{exp|H_0}}_{v.a \sim t_{n-1}} < t^{1-\alpha/2}_{n-1}) = P(t_{n-1} > t^{\alpha/2}_{n-1} \text{ ó } t_{n-1} < t^{1-\alpha/2}_{n-1}) = \; = P(t_{n-1} > t^{\alpha/2}_{n-1}) + P(t_{n-1}$$

Basada en el p-valor

$$Rechazar\ H_0 \Leftrightarrow pvalor < \alpha$$

Observación:

La regla de decisión basada en el p-valor se deduce de la regla de decision basada en el estadistico del contraste, veamoslo:

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$pvalor < \alpha \Leftrightarrow P(t_{n-1} > t_{exp|H0}) < \alpha \Leftrightarrow t_{exp|H0} > t^{\alpha}_{n-1} \Leftrightarrow Rechazar\ H_0$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

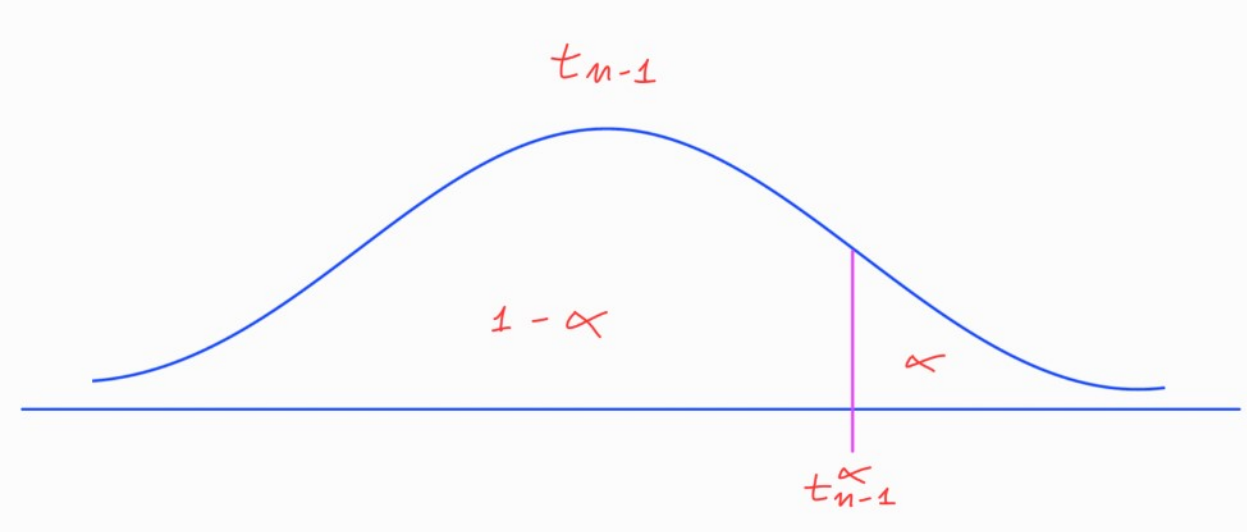
$$pvalor < \alpha \Leftrightarrow P(t_{n-1} < t_{exp|H0}) < \alpha \Leftrightarrow t_{exp|H0} < t_{n-1}^{1-\alpha} \Leftrightarrow Rechazar\ H_0$$

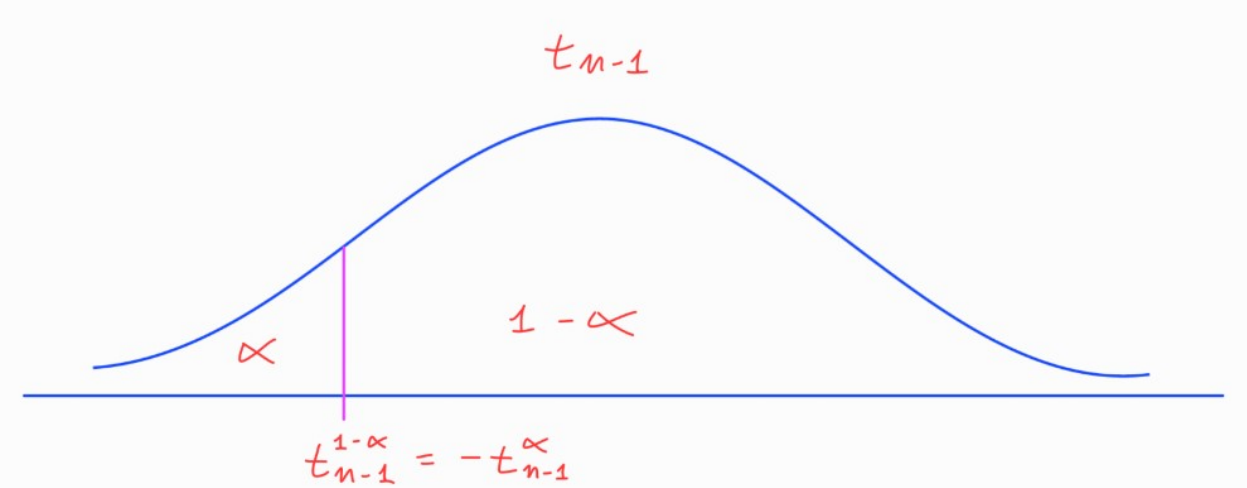
- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

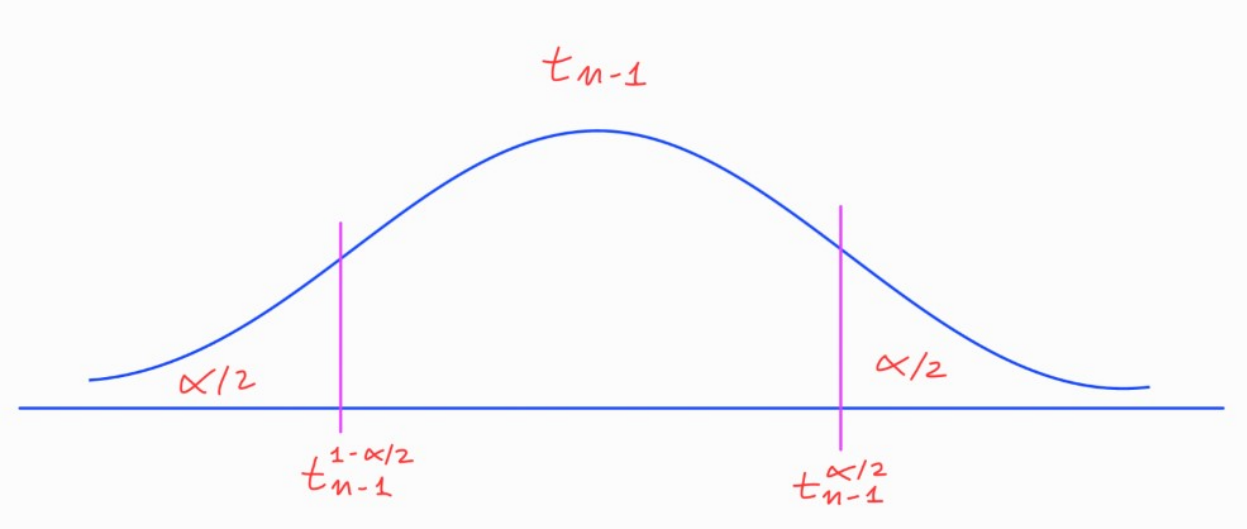
$pvalor < \alpha \Leftrightarrow 2 \cdot P(t_{n-1} < \mid t_{exp|H0} \mid) < \alpha \Leftrightarrow P(t_{n-1} < \mid t_{exp|H0} \mid) < \alpha/2 \Leftrightarrow \mid t_{exp|H0} \mid > t_{n-1}^{\alpha/2} \Leftrightarrow t_{exp|H0} > t_{n-1}^{\alpha/2}$ ò $t_{exp|H0}$

Observación:

Los razonamientos anteriores están basados en buena parte en propiedades básicas de la distribución t-student:







T-test para Grupo/Poblacion Normal en R:

Cargamos los datos con los que vamos a trabajar, un data set con precios y otros datos de multitud de productos vendidos por la empresa Mercadona:

```
library(tidyverse)
library(readr)

Mercadona_Productos <- read_csv("Mercadona_Productos.csv")

head(Mercadona_Productos)
```

```
## # A tibble: 6 x 8
##       id supermarket  category name         price reference_price reference_unit
##    <dbl> <chr>        <chr>    <chr>        <dbl>           <dbl> <chr>
## 1 248789 mercadona-es fruta    Banana        0.26            1.29 kg
## 2 248790 mercadona-es fruta    Plátano       0.34            1.99 kg
## 3 248791 mercadona-es fruta    Plátano mac~  0.58            1.95 kg
## 4 248792 mercadona-es fruta    Uva blanca ~  2.84            3.79 kg
## 5 248793 mercadona-es fruta    Uva negra s~  2.99            3.99 kg
## 6 248794 mercadona-es fruta    Uva sabores~  2.39            5.98 kg
## # ... with 1 more variable: insert_date <dttm>
```

Filtramos para quedarnos solo con las variables categoria y el precio del producto, y dentro de la categoria solo con las cervezas con y sin alcohol.

```
df<-Mercadona_Productos %>% select(category, name, reference_price, reference_unit ,insert_date) %>%
  filter(category =="cerveza" | category=="cerveza_sin_alcohol" )
```

Creamos las variables muestrales *precios_cervezas* y *precios_cervezas_sin_alcohol*, que contienen los precios de una muestra de una población de cervezas del Mercadona, con y sin alcohol, respectivamente. Ambas serán usadas en este trabajo, pero ahora nos centraremos en la primera.

```
precios_cervezas <- (df %>% filter(category=="cerveza") %>% select(reference_price))$reference_price

precios_cervezas_sin_alcohol <- (df %>% filter(category=="cerveza_sin_alcohol") %>% select(reference_pr
```

Vamos a resolver el contraste $H_0 : \mu = 1.5 \quad vs \quad H_1 : \mu \neq 1.5$, donde μ es la media de la variable *precios_cervezas* medida sobre un grupo/poblacion de cervezas del Mercadona.

Para ello tenemos que suponer que la variable *precios_cervezas* tiene una distribucion aproximadamente normal, para poder aceptar el supuesto de normalidad a nivel poblacional.

Realmente este supuesto deberia de contrastarse, usando algun procedimiento estadistico como los que expongo en el siguiente articulo https://rpubs.com/FabioScielzoOrtiz/Analisis_de_Normalidad_en_R.

Aqui no entraremos en este aspecto, nos limitaremos a aceptar el supuesto de normalidad.

Usamos la variable muestral disponible *precios_cervezas* para resolver el contraste.

Realizamos el contraste con la funcion **t.test** implementada en R:

```
t.test(x=precios_cervezas  , alternative = "two.sided", mu=1.5 , conf.level = 0.99)
```

```
##
##  One Sample t-test
##
```

```
## data:  precios_cervezas
## t = 51.728, df = 15793, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 1.5
## 99 percent confidence interval:
##  1.897172 1.938805
## sample estimates:
## mean of x
##  1.917988
```

Esta salida nos da informacion relevante, como el valor del estadisitco del contraste ($t = 51.728$) y el p-valor del contraste ($pvalor < 2.2e-16$), asi como el intervalo de confianza para μ a un nivel de confianza del 99% (puede especificarse otro).

Para un nivel de significacion $\alpha = 0.05 > pvalor \simeq 0$, se rechaza $H_0 : \mu = 1.5$ en favor de $H_1 : \mu \neq 1.5$

En general, para todo $\alpha > pvalor \simeq 0$, se rechaza $H_0 : \mu = 1.5$ en favor de $H_1 : \mu \neq 1.5$. Luego para todo nivel de significacion puede aceptarse que la media del precio de las cervezas del Mercadona es **distinta** de 1.5 €/L

Ahora mostraremos como realizar el contraste en R de manera "manual".

Cálculo del estadistico del contraste manualmente:

```
mu_0<-1.5
n<-length(precios_cervezas)

(mean(precios_cervezas)-mu_0)/(sd(precios_cervezas)/sqrt(n))
```

```
## [1] 51.72781
```

Cálculo del pvalor manualmente:

```
2*pt(abs(51.72781) , df=n-1, lower.tail = FALSE)
```

```
## [1] 0
```

Vamos a resolver ahora el contraste $H_0 : \mu = 1.5 \;\; vs \;\; H_1 : \mu < 1.5$

```
t.test(x=precios_cervezas  , alternative = "less", mu=1.5 , conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  precios_cervezas
## t = 51.728, df = 15793, p-value = 1
## alternative hypothesis: true mean is less than 1.5
```

```
## 99 percent confidence interval:
##      -Inf 1.936789
## sample estimates:
## mean of x
##  1.917988
```

Como $pvalor = 1$, para todo $\alpha \in (0,1)$ no se puede rechazar $H_0 : \mu = 1.5$ en favor de $H_1 : \mu < 1.5$, luego no se puede aceptar que el precio medio de las cervezas del Mercadona sea **menor** que $1.5£/L$

Vamos a resolver ahora el contraste $H_0 : \mu = 1.5$ vs $H_1 : \mu > 1.5$

```
t.test(x=precios_cervezas  , alternative = "greater", mu=1.5 , conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  precios_cervezas
## t = 51.728, df = 15793, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 1.5
## 99 percent confidence interval:
##  1.899188      Inf
## sample estimates:
## mean of x
##  1.917988
```

Como $pvalor < 2.2 \cdot 10^{-16}$, para todo $\alpha > 2.2 \cdot 10^{-16}$ se rechaza $H_0 : \mu = 1.5$ en favor de $H_1 : \mu > 1.5$, luego para esos niveles de significación puede aceptarse que el precio medio de las cervezas del Mercadona es **mayor** que $1.5£/L$

Z-test: Grupo/Población No necesariamente Normal

Supuestos:

- n tiene que ser grande ($n > 30$)

Estadistico del contraste

- El estadistico del contraste es:

$$z_{exp} = \frac{\overline{X_{k,g}} - \mu}{S(X_{k,g})/\sqrt{n}} \sim_{TCL} N(0,1)$$

- El estadistico del contraste **bajo** $H_0 : \mu = \mu_0$ es:

$$z_{exp|H_0} = \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} \sim_{TCL} N(0,1)$$

Donde: $S(X_{k,g}) = \frac{n}{n-1} \cdot \sigma(X_{k,g})$ es la cuasidesviación típica

p-valor

- Caso $H_0 : \mu = \mu_0$ vs $H_1 :> \mu_0$
  - $pvalor = P\left(\underbrace{z_{exp|H_0}}_{v.a. \sim N(0,1)} > \underbrace{z_{exp|H_0}}_{observacion} \right) = P\left(N(0,1) > \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} \right)$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$
  - $pvalor = P\left(\underbrace{z_{exp|H_0}}_{v.a. \sim N(0,1)} < \underbrace{z_{exp|H_0}}_{observacion} \right) = P\left(N(0,1) < \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} \right)$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$- \ pvalor = P\left(\underbrace{|\ z_{exp|H_0}\ |}_{v.a. \sim N(0.1)} \ > \ \underbrace{|\ z_{exp|H_0}\ |}_{observacion}\right) = P\left(\underbrace{z_{exp|H_0}}_{v.a.} \ > \ \underbrace{|\ z_{exp|H_0}\ |}_{observacion}\right) + P\left(\underbrace{z_{exp|H_0}}_{v.a.} \ < \ -\underbrace{|\ z_{exp|H_0}\ |}_{observacion}\right) =$$

$$= P\left(N(0,1) \ > \ \underbrace{|\ z_{exp|H_0}\ |}_{observacion}\right) + P\left(N(0,1) \ > \ \underbrace{|\ z_{exp|H_0}\ |}_{observacion}\right) = 2 \cdot P\left(N(0,1) > \left|\frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}}\right|\right)$$

Regla de decisión

Basada en el estadistico del contraste

Para un nivel de significación α :

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$Rechazar H_0 \ \Leftrightarrow \ \underbrace{z_{exp|H_0}}_{observacion} > z_\alpha \ \Leftrightarrow \ \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} > z_\alpha \ \Leftrightarrow \ \overline{X_{k,g}} > \mu_0 + z_\alpha \cdot S(X_{k,g})/\sqrt{n}$$

Donde: $P(N(0,1) > z_\alpha) = \alpha$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$Rechazar H_0 \ \Leftrightarrow \ z_{exp|H_0} < z_{1-\alpha} \ \Leftrightarrow \ \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} < z_{1-\alpha} \ \Leftrightarrow \ \overline{X_{k,g}} < \mu_0 + z_{1-\alpha} \cdot S(v)/\sqrt{n}$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$Rechazar H_0 \Leftrightarrow z_{exp|H_0} > z_{\alpha/2} \text{ ó } z_{exp|H_0} < z_{1-\alpha/2} \Leftrightarrow \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} > z_{\alpha/2} \text{ ó } \frac{\overline{X_{k,g}} - \mu_0}{S(X_{k,g})/\sqrt{n}} < z_{1-\alpha/2}$$

$$\Leftrightarrow \overline{X_{k,g}} > \mu_0 + z_{\alpha/2} \cdot S(X_{k,g})/\sqrt{n} \text{ ó } \overline{X_{k,g}} < \mu_0 + z_{1-\alpha/2} \cdot S(X_{k,g})/\sqrt{n}$$

Observaciones:

$P(Error\ Tipo\ I) = P(Rechazar\ H_0 \mid H_0) = \alpha$ en todos los casos, veamoslo:

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$P(RH_0|H_0) = P(\underbrace{z_{exp|H_0}}_{v.a\ \sim\ N(0,1)} > z_\alpha) = P(N(0,1) > z_\alpha) = \alpha$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$P(RH_0|H_0) = P(\underbrace{z_{exp|H_0}}_{v.a\ \sim\ N(0,1)} < z_{1-\alpha}) = P(N(0,1) < z_{1-\alpha}) = \alpha$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$P(RH_0|H_0) = P(\underbrace{z_{exp|H_0}}_{v.a\ \sim\ N(0,1)} > z_{\alpha/2}\ \text{ó}\ \underbrace{z_{exp|H_0}}_{v.a\ \sim\ N(0,1)} < z_{1-\alpha/2}) = P(N(0,1) > z_{\alpha/2}\ \text{ó}\ N(0,1) < z_{1-\alpha/2}) = P(N(0,1) > z_{\alpha/2})$$

Observacion:

Este procedimiento puede extrapolarse facilmente a todos los contrastes de este articulo, pero no se volverá a repetir por simplicidad.

Basada en el p-valor

$$Rechazar\ H_0 \ \Leftrightarrow\ pvalor < \alpha$$

Observación:

La regla de decisión basada en el p-valor se deduce de la regla de decision basada en el estadistico del contraste, veamoslo:

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$pvalor < \alpha \ \Leftrightarrow\ P(N(0,1) > z_{exp|H0}) < \alpha \ \Leftrightarrow\ z_{exp|H0} > z_\alpha \ \Leftrightarrow\ Rechazar\ H_0$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$pvalor < \alpha \ \Leftrightarrow\ P(N(0,1) < z_{exp|H0}) < \alpha \ \Leftrightarrow\ z_{exp|H0} < z_{1-\alpha} \ \Leftrightarrow\ Rechazar\ H_0$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$pvalor < \alpha \Leftrightarrow 2 \cdot P(N(0,1) <\mid z_{exp|H0} \mid) < \alpha \Leftrightarrow P(N(0,1) <\mid z_{exp|H0} \mid) < \alpha/2 \Leftrightarrow \mid z_{exp|H0} \mid > z_{\alpha/2} \Leftrightarrow z_{exp|H0} > z_{\alpha/2}$ ò z

Observación:

Los razonamientos anteriores están basados en buena parte en propiedades básicas de la distribución N(0,1)

Z-test en R:

Vamos a resolver el contraste $H_0 : \mu = 1.5 \quad vs \quad H_1 : \mu \neq 1.5$, donde μ es la media de la variable *precios_cervezas* medida sobre una poblacion de tipos de cervezas del Mercadona.

Como la muestra que disponemos de la variable *precio_cervezas* es suficientemente grande ($n = 15794 > 30$), **no** es necesario suponer que la variable *precios_cervezas* tiene una distribucion aproximadamente normal para aceptar el supuesto de normalidad a nivel poblacional.

Calculamos el estadistico del contraste:

```
mu_0<-1.5
n<-length(precios_cervezas)

(mean(precios_cervezas)-mu_0)/(sd(precios_cervezas)/sqrt(n))
```

```
## [1] 51.72781
```

Calculamos el pvalor:

```
2*pnorm(abs(51.72781) , mean=0, sd=1, lower.tail = FALSE)
```

```
## [1] 0
```

Vamos a resolver ahora el contraste $H_0 : \mu = 1.5 \quad vs \quad H_1 : \mu < 1.5$

Calculamos el pvalor:

```r
pnorm(51.72781 , mean=0, sd=1, lower.tail = TRUE)
```

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

Como $pvalor = 1$, para todo $\alpha \in (0, 1)$ **no** se puede rechazar $H_0 : \mu = 1.5$ en favor de $H_1 : \mu < 1.5$, puesto que los datos no aportan suficiente evidencia en favor de H_1

Vamos a resolver ahora el contraste $H_0 : \mu = 1.5$ vs $H_1 : \mu > 1.5$

Calculamos el pvalor:

```r
pnorm(51.72781 , mean=0, sd=1, lower.tail = FALSE)
```

```
## [1] 0
```

Para todo $\alpha > 0 = pvalor$ se puede rechazar $H_0 : \mu = 1.5$ en favor de $H_1 : \mu > 1.5$, puesto que los datos aportan suficiente evidencia en favor de H_1

Contraste de hipotesis para la media de una variable en dos grupos independientes

Objetivo

- Contrastar la media de una variable medida sobre dos grupos independientes.
  - Ejemplo: la nota media en matematicas de los alumnos (chicos) de cierto colegio es mayor que la de las alumnas (chicas), en ese mismo colegio.

Planteamiento formal del problema:

- Tenemos dos grupos:
  - $G_1 = \{g_{11}, g_{21}, ..., g_{N_{G1},1}\}$
  - $G_2 = \{g_{12}, g_{22}, ..., g_{N_{G2},2}\}$

- Tenemos una **muestra** g_1 de n_1 elementos de G_1
- Tenemos una **muestra** g_2 de n_2 elementos de G_2

- Tenemos una variable estadistica **cuantitativa** X_k medida sobre la muestra g_1 del grupo G_1:

$$X_{k,g_1} = (x_{g_1,1k}, x_{g_1,2k}, ..., x_{g_1,n_1k})^t \tag{2}$$

- Tenemos esa misma variable estadistica **cuantitativa** X_k pero medida sobre la muestra g_2 del grupo G_2:

$$X_{k,g_2} = (x_{g_2,1k}, x_{g_2,2k}, ..., x_{g_2,n_1k})^t \tag{3}$$

Observación: $x_{g_j,ik}$ es el valor de X_k para el i-esimo individuo de la muestra g_i del grupo G_i sobre la que se ha medido X_k

- Desconocemos X_k medida sobre el grupo G_1, a la que denotaremos como X_{k,G_1}
- Desconocemos X_k medida sobre la poblacion G_2, a la que denotaremos como X_{k,G_2}

Los **contrastes de hipotesis** que queremos resolver son del tipo:

$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 < \mu_2$

Donde:

μ_1 es la media de X_{k,G_1}

μ_2 es la media de X_{k,G_2}

Esta información (μ_1 y μ_2) se desconoce.

T-test: Dos Grupos Normales Independientes

Supuestos

- $X_{k,G_1} \sim N(\mu_1, \sigma_1)$
- $X_{k,G_2} \sim N(\mu_2, \sigma_2)$
- X_{k,G_1} y X_{k,G_2} son independientes
  - $S(X_{k,G_1}, X_{k,G_2}) = 0$

Estadistico del contraste:

- Si $\sigma_1 = \sigma_2$

El estadistico del contraste es:

$$t_{exp} = \frac{(\overline{X}_{k,g_1} - \overline{X}_{k,g_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S(X_{k,g_1})^2 + (n_2-1)S(X_{k,g_2})^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

El estadistico del contraste **bajo** $H_0 : \mu_1 = \mu_2$ es:

$$t_{exp} = \frac{(\overline{X}_{k,g_1} - \overline{X}_{k,g_2})}{\sqrt{\frac{(n_1-1)S(X_{k,g_1})^2 + (n_2-1)S(X_{k,g_2})^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

- Si $\sigma_1 \neq \sigma_2$

El estadistico del contraste es:

$$t_{exp} = \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}} \sim t_f$$

El estadistico del contraste **bajo** $H_0 : \mu_X = \mu_Y$ es:

$$t_{exp} = \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}} \sim t_f$$

Donde:

$$f = \frac{\left(\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_Y}\right)^2}{\frac{1}{n_1-1}\left(\frac{S(X_{k,g_1})^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S(X_{k,g_2})^2}{n_2}\right)^2}$$

p-valor

- Si $\sigma_1 = \sigma_2$

- Caso $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim\, t_{n_1+n_2-2}} > \underbrace{t_{exp|H_0}}_{observación} \right) = P\left(t_{n_1+n_2-2} > \frac{(\overline{X}_{k,g_1} - \overline{X}_{k,g_2})}{\sqrt{\frac{(n_1-1)S(X_{k,g_1})^2 + (n_2-1)S(X_{k,g_2})^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right)$$

- Caso $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim\, t_{n_1+n_2-2}} < \underbrace{t_{exp|H_0}}_{observación} \right) = P\left(t_{n_1+n_2-2} < \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{(n_1-1)S(X_{k,g_1})^2 + (n_2-1)S(X_{k,g_2})^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right)$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu_1 \neq \mu_2$

$$pvalor = P\left(\underbrace{|\, t_{exp|H_0} \,|}_{\sim t_{n_1+n_2-2}} > \underbrace{|\, t_{exp|H_0} \,|}_{observacion} \right) = P\left(t_{n_1+n_2-2} > |\, t_{exp|H_0} \,|\right) + P\left(t_{n_1+n_2-2} < -|\, t_{exp|H_0} \,|\right) =$$

$$= P\left(t_{n_1+n_2-2} > |\, t_{exp|H_0} \,|\right) + P\left(t_{n_1+n_2-2} > |\, t_{exp|H_0} \,|\right) =$$

$$= 2 \cdot P\left(t_{n_1+n_2-2} > \left| \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{(n_1-1)S(X_{k,g_1})^2 + (n_Y-1)S(X_{k,g_2})^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right| \right)$$

- Si $\sigma_1 \neq \sigma_2$

- Caso $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim\, t_f} > \underbrace{t_{exp|H_0}}_{observación}\right) = P\left(t_f > \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}}\right)$$

- Caso $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim\, t_{n_1+n_2-2}} < \underbrace{t_{exp|H_0}}_{observación}\right) = P\left(t_f < \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}}\right)$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu_1 \neq \mu_2$

$$pvalor = P\left(\underbrace{|\, t_{exp|H_0}\,|}_{\sim t_f} > \underbrace{|\, t_{exp|H_0}\,|}_{observacion}\right) = P\left(t_f > |\, t_{exp|H_0}\,|\right) + P\left(t_f < -|\, t_{exp|H_0}\,|\right) =$$

$$= P\left(t_f > |\, t_{exp|H_0}\,|\right) + P\left(t_f > |\, t_{exp|H_0}\,|\right) =$$

$$= 2 \cdot P\left(t_f > \left|\frac{(\overline{X_{k,g_1}} - \overline{Y_k})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}}\right|\right)$$

Regla de decisión:

Basada en el p-valor

$$Rechazar H_0 \Leftrightarrow pvalor < \alpha$$

T-test Dos Grupos Normales Independientes en R

Vamos a realizar en R el contraste $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$ donde μ_1 es el precio medio de las cervezas (con alcohol) del Mercadona y μ_2 es el precio medio de las cervezas sin alcohol del Mercadona.

Para ello tenemos que suponer quela variable *precio* medida sobre la muestra disponible de cervezas con alcohol de Mercadona tiene una distribucion aproximadamente Normal, asi como la variable *precio* medida sobre la muestra disponible de cervezas sin alcohol de Mercadona, para poder aceptar el supuesto de normalidad a nivel poblacional.

Este supuesto deberia de contrastarse, usando algun procedimiento estadistico como los que expongo en el siguiente articulo https://rpubs.com/FabioScielzoOrtiz/Analisis_de_Normalidad_en_R.

Aqui no entraremos en este aspecto, nos limitaremos a aceptar el supuesto de normalidad a nivel poblacional.

Ademas asumiremos que las varianzas de las variables *precios_cervezas* y *precios_cervezas_sin_alcohol* a nivel poblacional no son iguales. Aunque esto tambien deberia ser contrastado mediante un procedimiento adecuado.

Para resolver el contraste se usará la variable *precio* medida sobre las muestras de cervezas con y sin alcohol del Mercadona de las que disponemos, es decir, *precios_cervezas* y *precios_cervezas_sin_alcohol*. Asumiremos ademas

```
t.test(x=precios_cervezas , y=precios_cervezas_sin_alcohol , alternative = "greater", paired=FALSE, var
```

```
##
##  Welch Two Sample t-test
##
## data:  precios_cervezas and precios_cervezas_sin_alcohol
## t = 32.04, df = 3826.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  0.4925228       Inf
## sample estimates:
## mean of x mean of y
##  1.917988  1.398806
```

Como el $pvalor = 2.2e-16$, para todo α habitual (0.1, 0.05, 0.01) , al ser mayor que el pvalor, se rechaza H_0 en favor de $H_1 : \mu_1 > \mu_2$, luego puede aceptarse que el precio medio de las cervezas con alcohol del Mercadona es mayor que el de las cervezas sin alcohol.

Vamos a realizar en R el contraste $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$ donde μ_1 es el precio medio de las cervezas (con alcohol) del Mercadona y μ_2 es el precio medio de las cervezas sin alcohol del Mercadona.

Para ello se usará la variable precio medida sobre las muestras de cervezas con y sin alcohol del Mercadona de las que disponemos (*precios_cervezas* y *precios_cervezas_sin_alcohol*).

```
t.test(x=precios_cervezas , y=precios_cervezas_sin_alcohol , alternative = "less", paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  precios_cervezas and precios_cervezas_sin_alcohol
## t = 32.04, df = 3826.7, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.5458425
## sample estimates:
## mean of x mean of y
##  1.917988  1.398806
```

Como el $pvalor = 1$, para todo α habitual , al ser menor que el pvalor, no se rechaza H_0 en favor de $H_1 : \mu_1 < \mu_2$, luego no puede aceptarse que el precio medio de las cervezas con alcohol del Mercadona es menor que el de las cervezas sin alcohol.

Vamos a realizar en R el contraste $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ donde μ_1 es el precio medio de las cervezas (con alcohol) del Mercadona y μ_2 es el precio medio de las cervezas sin alcohol del Mercadona.

Para ello se usará la variable precio medida sobre las muestras de cervezas con y sin alcohol del Mercadona de las que disponemos (*precios_cervezas* y *precios_cervezas_sin_alcohol*).

```
t.test(x=precios_cervezas , y=precios_cervezas_sin_alcohol , alternative = "two.sided", paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  precios_cervezas and precios_cervezas_sin_alcohol
## t = 32.04, df = 3826.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  0.4874132 0.5509521
## sample estimates:
## mean of x mean of y
##  1.917988  1.398806
```

Como el $pvalor < 2.2e-16$, para todo α habitual , al ser mayor que el pvalor, se rechaza H_0 en favor de $H_1 : \mu_1 \neq \mu_2$, luego puede aceptarse que el precio medio de las cervezas con alcohol del Mercadona es distitno que el de las cervezas sin alcohol.

T-test: Dos Grupos No necesariamente Normales Independientes

Supuestos

- Tamaños **grandes** de las muestras g_1 y g_2 de los grupos G_1 y G_2, respectivamente
  - $n_1, n_2 > 30$

- X_{k,G_1} y X_{k,G_2} son **independientes**
  - $S(X_{k,G_1}, X_{k,G_2}) = 0$

Estadistico del contraste:

- El estadistico del contraste **bajo** $H_0 : \mu_X = \mu_Y$ es:

$$z_{exp|H_0} = \frac{(\overline{X}_{k,g_1} - \overline{X}_{k,g_2})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}} \sim_{TCL} N(0,1)$$

p-valor

- Caso $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim\ N(0,1)} > \underbrace{z_{exp|H_0}}_{observación}\right) = P\left(N(0,1) > \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}}\right)$$

- Caso $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$

$$pvalor = P\left(\underbrace{z_{exp|H_0}}_{\sim\ N(0,1)} < \underbrace{z_{exp|H_0}}_{observación}\right) = P\left(N(0,1) < \frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}}\right)$$

- Caso $H_0 : \mu = \mu_0$ vs $H_1 : \mu_1 \neq \mu_2$

$$pvalor = P\left(\underbrace{|\ z_{exp|H_0}\ |}_{\sim N(0,1)} > \underbrace{|\ z_{exp|H_0}\ |}_{observacion}\right) = P\left(N(0,1) > |\ z_{exp|H_0}\ |\right) + P\left(N(0,1) < -|\ z_{exp|H_0}\ |\right) =$$

$$= P\left(N(0,1) > |\ z_{exp|H_0}\ |\right) + P\left(N(0,1) > |\ z_{exp|H_0}\ |\right) =$$

$$= 2 \cdot P\left(N(0,1) > \left|\frac{(\overline{X_{k,g_1}} - \overline{X_{k,g_2}})}{\sqrt{\frac{S(X_{k,g_1})^2}{n_1} + \frac{S(X_{k,g_2})^2}{n_2}}}\right|\right)$$

Regla de decisión:

Basada en el p-valor

$$Rechazar H_0 \Leftrightarrow pvalor < \alpha$$

T-test Dos Grupos Independientes No necesariamente Normales en R

Vamos a realizar en R el contraste $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$ donde μ_1 es el precio medio de las cervezas (con alcohol) del Mercadona y μ_2 es el precio medio de las cervezas sin alcohol del Mercadona.

```
n1<-length(precios_cervezas)
```

```
n2<-length(precios_cervezas_sin_alcohol)
```

Como los tamaños de las muestras que tenemos de cervezas con y sin alcohol de Mercadona son suficientemente grandes ($n_1 = 15794 > 30$, $n_2 = 2194 > 30$), **no** es necesario suponer que la variable *precio* medida sobre esas muestras tiene una distribucion aproximadamente Normal. Es decir, no es necesario el supuesto de normalidad a nivel poblacional.

Para resolver el contraste se usará la variable *precio* medida sobre las muestras de cervezas con y sin alcohol del Mercadona de las que disponemos, es decir, *precios_cervezas* y *precios_cervezas_sin_alcohol*

Calculamos el estadistico del contraste en este caso:

```
(mean(precios_cervezas)-mean(precios_cervezas_sin_alcohol))/sqrt(var(precios_cervezas)/n1 + var(precios
```

```
## [1] 32.04022
```

```
pnorm(32.04022 , mean=0, sd=1, lower.tail = FALSE)
```

```
## [1] 1.502146e-225
```

Como $pvalor = 1.502146e-225$, para los α habituales , al ser mayores que el pvalor, se rechaza H_0 en favor de $H_1 : \mu_1 > \mu_2$. Luego, paea los α habituales puede aceptarse que el precio medio de las cervezas con alcohol del Mercadona es mayor que el de las cervezas sin alcohol.

Contraste de hipotesis para la media de una variable en dos grupos dependientes

Objetivo

- Contrastar la media de **dos** variables sobre una **mismo grupo** (o si se quiere, sobre **dos grupos iguales**)
  - *Ejemplo*: contrastar si la nota media en matematicas es mayor que la nota media en lengua entre los estudiantes de cierto colegio
- Contrastar la media de una variable sobre **dos grupos pareados o emparejados** (tienen los mismos individuos , pero en diferentes condiciones/circustancias)
  - *Ejemplo*: contrastar si la nota media en matematicas es menor en un grupo de estudiantes que no reciben clases de apoyo que en ese mismo grupo pero tras recibir clases de apoyo.

Planteamiento formal del problema:

- Tenemos dos grupos:
  - $G_1 = \{g_{11}, g_{21}, ..., g_{N_{G1},1}\}$
  - $G_2 = \{g_{12}, g_{22}, ..., g_{N_{G2},2}\}$
- Tenemos una **muestra** g_1 de n_1 elementos de G_1
- Tenemos una **muestra** g_2 de n_2 elementos de G_2
- Tenemos una variable estadistica **cuantitativa** X_k medida sobre la muestra g_1 del grupo G_1:

$$X_{k,g_1} = (x_{g_1,1k}, x_{g_1,2k}, ..., x_{g_1,n_1k})^t \tag{4}$$

- Tenemos esa misma variable estadistica **cuantitativa** X_k pero medida sobre la muestra g_2 del grupo G_2:

$$X_{k,g_2} = (x_{g_2,1k}, x_{g_2,2k}, ..., x_{g_2,n_1k})^t \tag{5}$$

Observación: $x_{g_j,ik}$ es el valor de X_k para el i-esimo individuo de la muestra g_i del grupo G_i sobre la que se ha medido X_k

- Desconocemos X_k medida sobre el grupo G_1, a la que denotaremos como X_{k,G_1}
- Desconocemos X_k medida sobre la poblacion G_2, a la que denotaremos como X_{k,G_2}

Hasta aqui el planteamiento del problema es el mismo que en el caso de dos grupos independientes.

- La diferencia práctica esencial es que si los grupos G_1 y G_2 son dependientes , G_1 y G_2 son el **mismo grupo** (mismos individuos) o son **grupos pareados/emparejados** (mismos individuos, distintas condiciones o circustancias), ademas $N_1 = N_2$ como consecuencia.
- En cambio si los grupos G_1 y G_2 son independientes , G_1 y G_2 son grupos distintos (distintos individuos).

Los **contrastes de hipotesis** que queremos resolver son del tipo:

$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 > \mu_2$	$H_1 : \mu_1 < \mu_2$

Donde:

μ_1 es la media (aritmética) de X_{k,G_1}

μ_2 es la media (aritmética) de X_{k,G_2}

Esta información (μ_1 y μ_2) se desconoce.

T-test: Dos Grupos Normales Dependientes

Supuestos

- $n_1 = n_2$
- $(X_{k,G_1}, X_{k,G_2}) \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix}\right)$

Donde:

$\sigma_{12} = S(X_{k,G_1}, X_{k,G_2})$ no necesariamente es 0, es decir, no son necesariamente independientes.

Distinguir en la practica una muestras de grupos dependientes de muestras de grupos independientes:

- Datos de salarios de 10 mujeres y 15 hombres ⇒ muestras de grupos *independientes.*
- Datos de calificaciones de los mismos 8 estudiantes en matemáticas y estadística ⇒ muestra de grupos *dependientes.*
- Datos del número de parados en 20 ciudades de dos paises distintos ⇒ muestras de grupos *independientes.*
- Datos del peso de 32 pacientes antes y después de un tratamiento de adelgazamiento ⇒ muestras de grupos *dependientes.*

Estadistico del contraste

Se usa la transformación $D_{G_1,G_2} = X_{k,G_1} - X_{k,G_2} \sim N(\mu_{D_{G1,G2}}, \sigma_{D_{12}})$, donde $\mu_{D_{G1,G2}} = \mu_1 - \mu_2$ es la media de D_{G_1,G_2}

Con la transformacion los contrastes equivalentes a resolver ahora serian:

$H_0 : \mu_{D_{G1,G2}} = 0$ $H_1 : \mu_{D_{G1,G2}} \neq 0$	$H_0 : \mu_{D_{G1,G2}} = 0$ $H_1 : \mu_{D_{G1,G2}} > 0$	$H_0 : \mu_{D_{G1,G2}} = 0$ $H_1 : \mu_{D_{G1,G2}} < 0\mu_2$

Con la tranformacion ahora se tiene a nivel muestral los siguientes elementos:

$$D_{g_1,g_2} = X_{k,g_1} - X_{k,g_2} = (x_{g_1,1k} - x_{g_2,1k}\ ,\ x_{g_1,2k} - x_{g_2,2k}\ ,...,\ x_{g_1,n_1k} - x_{g_2,n_2k})$$

$$\overline{D}_{g_1,g_2} = \frac{1}{n}\sum_{i=1}^{n}(x_{g_1,ik} - x_{g_2,ik})$$

Recordar que en este caso se supone $n_1 = n_2 = n$

El estadistico del contraste es:

$$t_{exp} = \frac{\overline{D}_{g_1,g_2} - \mu_{D_{G1,G2}}}{S(D_{g_1,g_2})/\sqrt{n}} \sim t_{n-1}$$

El estadistico del contraste bajo $H_0 : \mu_{D_{G1,G2}} = 0$

$$t_{exp} = \frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}} \sim t_{n-1}$$

p-valor

- Caso $H_0 : \mu_{D_{G1,G2}} = 0$ vs $H_1 : \mu_{D_{G1,G2}} > 0$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim\ t_{n-1}} > \underbrace{t_{exp|H_0}}_{observación}\right) = P\left(t_{n-1} > \frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}}\right)$$

- Caso $H_0 : \mu_{D_{G1,G2}} = 0$ vs $H_1 : \mu_{D_{G1,G2}} < 0$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim\ t_{n-1}} < \underbrace{t_{exp|H_0}}_{observación}\right) = P\left(t_{n-1} < \frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}}\right)$$

- Caso $H_0 : \mu_{D_{G1,G2}} = 0 \quad \text{vs} \quad H_1 : \mu_{D_{G1,G2}} \neq 0$

$$pvalor = P\left(\underbrace{|\, t_{exp|H_0} \,|}_{\sim t_{n-1}} \; > \; \underbrace{|\, t_{exp|H_0} \,|}_{observacion}\right) = P\left(t_{n-1} \; > |\, t_{exp|H_0} \,|\right) + P\left(t_{n-1} \; < \; -|\, t_{exp|H_0} \,|\right) =$$

$$= P\left(t_{n-1} \; > |\, t_{exp|H_0} \,|\right) + P\left(t_{n-1} \; > |\, t_{exp|H_0} \,|\right) =$$

$$= 2 \cdot P\left(t_{n-1} > \left|\frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}}\right|\right)$$

Regla de decisión:

Basada en el p-valor

$$Rechazar H_0 \Leftrightarrow pvalor < \alpha$$

T-test Dos Grupos Dependientes Normales en R

Cargamos datos de dos grupos dependientes.

Tenemos dos grupos pareados o emparejados (actuando aqui como **muestras**):

Un grupo de sujetos que se enfrentan a una prueba de esfuerzo sin usar una sustancia dopante, y el mismo grupo de sujetos que se vuelven a enfrentar a la prueba pero usando una sustancia dopante.

Se mide en ambos grupos la variable *rendimiento*, la cual mide el rendimiento de los sujetos en la prueba de esfuerzo en una escala 0-10 , donde 0 es el minimo y 10 el maximo rendimiento.

Los tamaños de los grupos son 35 sujetos.

```
rendimiento <- c(5, 4, 7, 7 , 8, 5 , 9, 3, 5, 6, 5, 7, 8,2, 8, 5, 4, 7, 7 , 8, 5 , 9, 3, 5, 6, 5, 7, 8,
                 6, 7, 8, 7,  10, 7, 9, 5, 6, 7 ,7, 7, 9, 4, 9, 6, 7, 8, 7,  10, 7, 9, 5, 6, 7 ,7, 7, 9

grupo <- c(rep("No_Dopado", 35), rep("Dopado", 35))

datos_rendimientos_grupos <- tibble(rendimiento , grupo)

head(datos_rendimientos_grupos,3)
```

```
## # A tibble: 3 x 2
##   rendimiento grupo
##         <dbl> <chr>
## 1           5 No_Dopado
## 2           4 No_Dopado
## 3           7 No_Dopado
```

Vamos a realizar en R el contraste $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$ donde μ_1 es el rendimiento medio en ese tipo de pruebas de esfuerzo si los sujetos no estan dopados y μ_2 es el rendimiento medio es ese tipo de pruebas de esfuerzo si los mismos sujetos estan dopados.

Para ello tenemos que suponer que la variable *rendimiento* medida sobre los dos grupos tiene una **distribucion normal** con media y varianza igual a la de la propia variable, como evidencia de normalidad a nivel poblacional.

Este supuesto deberia de contrastarse, usando algun procedimiento estadistico como los que expongo en el siguiente articulo https://rpubs.com/FabioScielzoOrtiz/Analisis_de_Normalidad_en_R. Aqui no entraremos en este aspecto, nos limitaremos a aceptar el supuesto de normalidad a nivel poblacional.

Para realizar el contraste se usa la informacion disponible de variable *rendimiento* medida sobre los grupos de sujetos que realizan la prueba de esfuerzo, primero sin doparse, y luego dopados, que actuan como muestras:

```
rendimiento_no_dopados <- (datos_rendimientos_grupos %>% filter(grupo=="No_Dopado") %>% select(rendimie

rendimiento_dopados <- (datos_rendimientos_grupos %>% filter(grupo=="Dopado") %>% select(rendimiento))$
```

Realizamos el contraste en R:

```
t.test(x=rendimiento_no_dopados , y=rendimiento_dopados , alternative = "greater", paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  rendimiento_no_dopados and rendimiento_dopados
## t = -8.4737, df = 34, p-value = 1
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  -1.610825       Inf
## sample estimates:
## mean difference
##       -1.342857
```

Como el *pvalor* $= 1$, para todo α habitual (0.1, 0.05, 0.01) , al ser menor que el pvalor, no se rechaza H_0 en favor de $H_1 : \mu_1 > \mu_2$, luego no puede aceptarse el rendimiento medio de los sujetos cuando no se dopan sea superior a cuando se dopan.

Vamos a realizar en R el contraste $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$ donde μ_1 es el rendimiento medio en ese tipo de pruebas de esfuerzo si los sujetos no estan dopados y μ_2 es el rendimiento medio en ese tipo de pruebas de esfuerzo si los mismos sujetos estan dopados.

Para ello se usa la informacion disponible de variable *rendimiento* medida en los grupos que actuan como muestras:

```
t.test(x=rendimiento_no_dopados , y=rendimiento_dopados , alternative = "less", paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  rendimiento_no_dopados and rendimiento_dopados
## t = -8.4737, df = 34, p-value = 3.383e-10
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##       -Inf -1.074889
## sample estimates:
## mean difference
##       -1.342857
```

Como el $pvalor = 3.569e-05$, para todo $\alpha > 3.569e-05$, como el habitual 0.05, al ser mayor que el pvalor, se rechaza H_0 en favor de $H_1 : \mu_1 < \mu_2$, luego para $\alpha > 0.034$ puede aceptarse que el rendimiento medio de los sujetos cuando se dopan es superior a cuando no se dopan.

Vamos a realizar en R el contraste $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ donde μ_1 es el rendimiento medio en ese tipo de pruebas de esfuerzo si los sujetos no estan dopados y μ_2 es el rendimiento medio en ese tipo de pruebas de esfuerzo si los mismos sujetos estan dopados.

Para ello se usa la informacion disponible de variable *rendimiento* medida en los grupos que actuan como muestras:

```
t.test(x=rendimiento_no_dopados , y=rendimiento_dopados , alternative = "two.sided", paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  rendimiento_no_dopados and rendimiento_dopados
## t = -8.4737, df = 34, p-value = 6.766e-10
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.664915 -1.020799
## sample estimates:
## mean difference
##       -1.342857
```

Como el $pvalor = 7.138e-05$, para todo $\alpha > 7.138e-05$, como el habitual 0.05 , al ser mayor que el pvalor, se puede rechazar H_0 en favor de $H_1 : \mu_1 \neq \mu_2$, luego puede aceptarse que el rendimiento medio de los sujetos cuando se dopan es diferente a cuando no se dopan.

T-test: Dos Grupos Dependientes No necesariamente Normales

Supuestos

- Tamaños grandes de las muestras g_1 y g_2 de los grupos dependientes G_1 y G_2, respectivamente.
  - $n_1 = n_2 = n > 30$

Estadistico del contraste:

El estadistico del contraste bajo $H_0 : \mu_{D_{G1,G2}} = 0$

$$z_{exp|H_0} = \frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}} \sim_{TCL} N(0,1)$$

p-valor

- Caso $H_0 : \mu_{D_{G1,G2}} = 0$ vs $H_1 : \mu_{D_{G1,G2}} > 0$

$$pvalor = P\left(\underbrace{t_{exp|H_0}}_{\sim N(0,1)} > \underbrace{z_{exp|H_0}}_{observación}\right) = P\left(N(0,1) > \frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}}\right)$$

- Caso $H_0 : \mu_{D_{G1,G2}} = 0$ vs $H_1 : \mu_{D_{G1,G2}} < 0$

$$pvalor = P\left(\underbrace{z_{exp|H_0}}_{\sim N(0,1)} < \underbrace{z_{exp|H_0}}_{observación}\right) = P\left(N(0,1) < \frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}}\right)$$

- Caso $H_0 : \mu_{D_{G1,G2}} = 0 \quad \text{vs} \quad H_1 : \mu_{D_{G1,G2}} \neq 0$

$$pvalor = P\left(\underbrace{|\, z_{exp|H_0} \,|}_{\sim N(0,1)} \; > \; \underbrace{|\, t_{exp|H_0} \,|}_{observacion}\right) = P\left(N(0,1) \; > |\, z_{exp|H_0} \,|\right) + P\left(N(0,1) \; < \; -|\, t_{exp|H_0} \,|\right) =$$

$$= P\left(N(0,1) \; > |\, z_{exp|H_0} \,|\right) + P\left(N(0,1) \; > |\, z_{exp|H_0} \,|\right) =$$

$$= 2 \cdot P\left(N(0,1) > \left|\frac{\overline{D}_{g_1,g_2}}{S(D_{g_1,g_2})/\sqrt{n}}\right|\right)$$

Regla de decisión:

Basada en el p-valor

$$Rechazar H_0 \Leftrightarrow pvalor < \alpha$$

T-test Dos Grupos Dependientes No necesariamente Normales en R

Vamos a realizar en R el contraste $H_0 : \mu_{D_{G1,G2}} = \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_{D_{G1,G2}} = \mu_1 - \mu_2 > 0$ donde μ_1 es el rendimiento medio en ese tipo de pruebas de esfuerzo si los sujetos no estan dopados y μ_2 es el rendimiento medio es ese tipo de pruebas de esfuerzo si los mismos sujetos estan dopados.

Como los tamaños de los grupos dependientes son grandes $n_1 = n_2 = 35 > 30$, ahora **no** es necesario suponer la normalidad de la variable *rendimiento* medida sobre los dos grupos, como evidencia de normalidad a nivel poblacional.

Realizamos el contraste en R.

Calculamos el estadistico del contraste:

```r
D<- rendimiento_no_dopados - rendimiento_dopados
D_media <- mean(D)
D_sd <- sd(D)
n<-length(D)

D_media/(D_sd/sqrt(n))
```

```
## [1] -8.473669
```

Calculamos el pvalor del contraste:

```r
pnorm(-8.473669, mean=0, sd=1, lower.tail = FALSE)
```

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

Como $pvalor = 1$, para cualquier α habitual, no se puede rechazar H_0 en favor de $H_1 : \mu_{D_{G1,G2}} = \mu_1 - \mu_2 > 0$. Luego, no puede aceptarse que el rendimiento medio en ese tipo de pruebas de esfuerzo cuando los sujetos no se dopan sea mayor que cuando se dopan.

Contraste de hipotesis para la media de una variable en multiples grupos independientes.

Objetivo

- Contrastar la media de una variable medida sobre de multiples grupos.

Planteamiento formal del problema:

- Tenemos h grupos:

$$G_1 = \{g_{11}, g_{21}, ..., g_{N_{G1},1}\}$$
$$G_2 = \{g_{12}, g_{22}, ..., g_{N_{G2},2}\}$$
$$...$$
$$G_h = \{g_{1h}, g_{2h}, ..., g_{N_{G2},h}\}$$

- Tenemos una **muestra** g_1 de n_1 elementos de G_1
- Tenemos una **muestra** g_2 de n_2 elementos de G_2

...

- Tenemos una **muestra** g_h de n_h elementos de G_h

- Tenemos una variable estadistica **cuantitativa** X_k medida sobre las h muestras $g_1, g_2, ..., g_h$ de los grupos $G_1, G_2, ..., G_h$:

$$X_{k,g_j} = (x_{g_j,1k}, x_{g_j,2k}, ..., x_{g_j,n_jk})^t \tag{6}$$

para $j = 1, 2, ..., h$

Observación: $x_{g_j,ik}$ es el valor de X_k para el i-esimo individuo de la muestra g_j del grupo G_i sobre la que se ha medido X_k

- Desconocemos X_k medida sobre el grupo G_j, a la que denotaremos como X_{k,G_j} , para $j = 1, 2, ..., h$

Los **contrastes de hipotesis** que queremos resolver son del tipo:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_h$$
$$H_1 : \mu_j \neq \mu_r \ , \ \exists \ j \neq r = 1, ..., h$$

Donde:

μ_j es la media de X_{k,G_j}

Esta información (μ_1,μ_2,...,μ_h) se desconoce.

ANOVA

Supuestos

- $X_{k,G_j} \sim N(\mu_j, \sigma)$, para $j = 1, ..., h$
- Normalidad y Homocedasticidad (igualdad de varianzas) a nivel poblacional
- Independiencia entre poblaciones/grupos

Estadistico del contraste

Para calcular el estadistico del contraste primero se necesita calcular los siguientes parametros:

- Media de la variable X_{k,g_j}:

$$\overline{X}_{k,g_j} = \frac{1}{n_h} \cdot \sum_{i=1}^{n_h} x_{g_j,ik}$$

para $j = 1, ..., h$

- Media total:

$$\overline{X}_k = \frac{1}{n_1 + n_2 + ... + n_h} \cdot \sum_{j=1}^{h} \sum_{i=1}^{n_j} x_{g_j,ik} = = \frac{1}{\sum_{j=1}^{h} n_j} \cdot \left(\sum_{i=1}^{n_1} x_{g_1,ik} + \sum_{i=1}^{n_2} x_{g_2,ik} + ... + \sum_{i=1}^{n_3} x_{g_3,ik} \right)$$

- Suma de Cuadrados Total (TSS):

$$TSS = \sum_{j=1}^{h} \sum_{i=1}^{n_j} \left(x_{g_j,ik} - \overline{X}_k \right)^2$$

- Suma de Cuadrados explixado por el Factor (FSS):

$$FSS = \sum_{j=1}^{h} \sum_{i=1}^{n_j} \left(\overline{X}_{k,g_j} - \overline{X}_k \right)^2$$

- Suma de Cuadrados Residual (RSS):

$$RSS = \sum_{j=1}^{h} \sum_{i=1}^{n_j} \left(x_{g_j,ik} - \overline{X}_{k,g_j} \right)^2$$

Se cumple que: $TSS = FSS + RSS$

El **estadistico del contraste** bajo H_0 es :

$$F_{exp|H_0} = \frac{FSS/(h-1)}{RSS/(n-h)} \sim F_{h-1,n-h}$$

Donde: $n = n_1 + n_2 + ... + n_h$

p-valor

El pvalor del contraste es:

$$pvalor = P(F_{h-1,n-h} > \frac{FSS/(h-1)}{RSS/(n-h)})$$

Regla de decisión:

Basada en el p-valor

$$Rechazar H_0 \Leftrightarrow pvalor < \alpha$$

ANOVA para analizar la influencia de un factor en una variable

El metodo ANOVA puede ser empleado para analizar la influencia de un factor (variable categorica) en una variable cuantitativa.

Para ello los distintos grupos a considerar deben formarse/definirse en base a una variable categorica. De modo que cada grupo tiene asignada una categoria de la variable categorica y se diferencia entre sí a traves de ella.

Por ejemplo: la variable categorica podria ser una serie de colores de un movil, cada grupo tendria asignado un color. Y lo que se quiere analizar podria ser la influencia del color del movil en la disposición de compra de los sujetos.

Tamaño del efecto

Se define el tamaño del efecto del factor en la variable cuantitativa como:

$$\eta^2 = \frac{FSS}{TSS}$$

Criterio:

- Si $\eta^2 \in [0, 0.03] \Rightarrow$ el efecto del factor sobre la variable es pequeño (poco significativo).
- Si $\eta^2 \in (0.03, 0.14] \Rightarrow$ el efecto del factor sobre la variable es moderado (moderadamente significativo).
- Si $\eta^2 > 0.14 \Rightarrow$ el efecto del factor sobre la variable es grande (significativo).

Para entender el coeficiente η^2 es necesario entender los elementos FSS, RSS y TSS:

- η^2 es la proporcion de variabilidad de la variable (cuantitativa) de interes explicada por el factor sobre su variabilidad total
- FSS mide la variabilidad de la variable de interes entre los grupos, es decir, la distancia de las medias de la variable en cada grupo respecto de la media global.

  - Si el valor de FSS es alto, significa que existe al menos un grupo en el que la media de la variable de interes es notablemente diferente de la media global de los grupos, lo que indica un comportamiento diferente de la variable en ese (o esos) grupo respecto de la media global de los grupos.
  - Si el valor de FSS es pequeño, significa que no existe ningun grupo en el que la media de la variable de interes sea notablemente diferente de la media global, lo que indica un comportamiento similar de la variable en todos los grupos.
  - Por ello FSS se considera una medida de la cantidad de variabilidad de la variable de interes que es explicada por el factor. Cuanto mayor sea, mas variabilidad es explcada por el factor, y viceversa.

- TSS mide la variabilidad total de la variable de interes, es decir, la distancia de las observaciones de la variable (sin atender al grupo) resepecto de la media global.
- RSS mide la variabilidad de la variable de interes dentro de cada grupo, es decir, la distancia de las observaciones de la variable en cada grupo respecto de la media del grupo.

Test HSD de Tukey

Cuando se ha rechazado la hipótesis de igualdad de medias con el test ANOVA, el interés está en averiguar cuál o cuáles pares de medias son diferentes entre sí.

Se podría pensar en realizar un t-test de igualdad de medias para cada par de variables. Pero el problema es que si realizamos múltiples contrastes de hipótesis se incrementa la probabilidad de cometer un error de tipo I.

La solución es usar el test HSD (Honestly-significant-difference) de Tukey que es un test de comparaciones múltiples que contrasta simultáneamente para todos los pares (i; j) y que tiene en cuenta a todos los grupos para la realización del test (no solo a los dos grupos que entran en juego)

El contraste que se lleva a cabo con el test de Tukey es:

$$H_0 : \mu_r = \mu_j$$
$$H_1 : \mu_r \neq \mu_j$$

$\forall\ r \neq j = 1, ..., h$

Estadistico del contraste:

Para $H_0 : \mu_r = \mu_j \;\; vs \;\; H_1 : \mu_r \neq \mu_j$

$$q_{exp|H_0} = \frac{\overline{X}_{k,g_r} - \overline{X}_{k,g_j}}{\frac{RSS}{n-h}/\sqrt{2} \cdot \sqrt{1/n_r + 1/n_j}} \sim q_{h,n-h}$$

p-valor

Para $H_0 : \mu_r = \mu_j \;\; vs \;\; H_1 : \mu_r \neq \mu_j$

$$pvalor = P\left(q_{h,n-h} > \frac{\overline{X}_{k,g_r} - \overline{X}_{k,g_j}}{\frac{RSS}{n-h}/\sqrt{2} \cdot \sqrt{1/n_r + 1/n_j}}\right)$$

Regla de decisión:

$$Rechazar H_0 \Leftrightarrow pvalor < \alpha$$

ANOVA y Test de Tukey en R:

Cargamos el data-set con el que expondremos la aplicacion del ANOVA en R.

Imaginemos que Apple quiere analizar si el color de su nuevo Iphone influye en la decision de compra de los clientes potenciales.

Tenemos 5 grupos independientes de individuos que tienen la intencion de comprar el nuevo Iphone, pero a cada grupo se le ofrece un color distinto (negro, blanco, azul, rojo, morado), y se les pide que indiquen en una escala del 1 al 10 su intencion de compra el nuevo Iphone pero con el color establecido.

El tamaño de los grupos es 10, 15, 10, 20 y 13 respectivamente.

Los datos disponibles son los siguientes:

```r
library(tidyverse)

grupos<- c(rep("negro", 10), rep("blanco", 15), rep("azul", 10), rep("rojo",13), rep("morado", 10))

intencion_compra_Iphone <-
  c(8,9,7.5,8,7,6.5,8,9,8,7.7,
    8,7,6,9,7,8.5,7.5,9,8.5,7,8.3,6.7,7.5,8,9,
    7,6,8,7.5,8,8.5,9,7,6.5,7.5,
    6,6.5,7,7.5,6,7,5.5,6,7.2,6,6.5,7.5,8,
    8,9,8.5,8,7,7,8.5,9,7.5,9)

Datos_Apple <- tibble(intencion_compra_Iphone, grupos)
```

Realizamos un ANOVA para contrastar si algun par de medias son significativamente diferentes entre si, y por tanto se puede rechazar H_0 en favor de H_1:

```r
ANOVA_resumen <- summary( aov(Datos_Apple$intencion_compra_Iphone ~ Datos_Apple$grupos) )

ANOVA_resumen
```

```
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## Datos_Apple$grupos  4  15.65   3.914   5.532 0.00085 ***
## Residuals          53  37.49   0.707
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La salida nos da informacion relevante para el ANOVA como:

$pvalor = 0.00085$

$RSS = 37.49$, $FSS = 15.65$, $TSS = RSS + FSS = 37.49 + 15.65$

Para cualquier α habitual, al ser mayor que el pvalor, se rechaza $H_0 : \mu_1 = \mu_2 = ... = \mu_5$ en favor de que algun par de medias es diferente (H_1)

Para hacernos una idea del por que del resultado del contraste, y de que pares de medias son las que difieren mas notablemente:

```r
Datos_Apple %>% group_by(grupos) %>% summarise(mean(intencion_compra_Iphone))
```

```
## # A tibble: 5 x 2
##   grupos `mean(intencion_compra_Iphone)`
##   <chr>                            <dbl>
## 1 azul                              7.5
## 2 blanco                            7.8
## 3 morado                            8.15
## 4 negro                             7.87
## 5 rojo                              6.67
```

Realizamos el test de Tukey para contrastar que pares de medias son significativamente diferentes:

```r
anova <- aov(Datos_Apple$intencion_compra_Iphone ~ Datos_Apple$grupos)

TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Datos_Apple$intencion_compra_Iphone ~ Datos_Apple$grupos)
##
## $`Datos_Apple$grupos`
##                     diff        lwr        upr     p adj
## blanco-azul    0.3000000 -0.6696496  1.2696496 0.9052729
## morado-azul    0.6500000 -0.4121980  1.7121980 0.4260225
## negro-azul     0.3700000 -0.6921980  1.4321980 0.8614555
## rojo-azul     -0.8307692 -1.8298088  0.1682704 0.1460636
## morado-blanco  0.3500000 -0.6196496  1.3196496 0.8453541
## negro-blanco   0.0700000 -0.8996496  1.0396496 0.9996023
## rojo-blanco   -1.1307692 -2.0307893 -0.2307492 0.0070568
## negro-morado  -0.2800000 -1.3421980  0.7821980 0.9449235
## rojo-morado   -1.4807692 -2.4798088 -0.4817296 0.0009846
## rojo-negro    -1.2007692 -2.1998088 -0.2017296 0.0110000
```

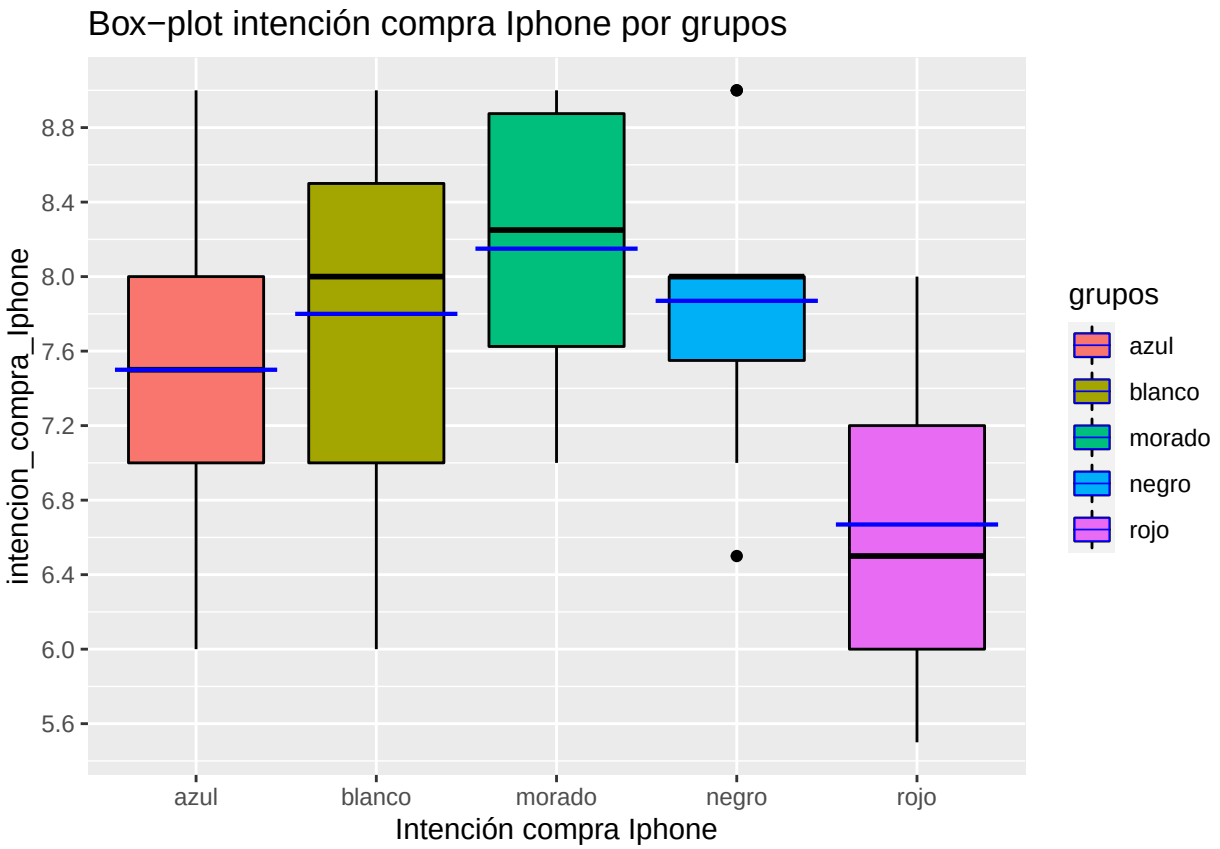
Teniendo en cuenta los pvalores obtenidos, para los α habituales, se aprecian diferencias significativas entre las medias de la variable *intencion de compra* para los grupos asociados al Iphone de color rojo y morado, asi como rojo y negro, y tambien rojo y blanco.

Luego, para los α habituales, puede aceptarse que $\mu_{rojo} \neq \mu_{morado}$, $\mu_{rojo} \neq \mu_{negro}$ y $\mu_{rojo} \neq \mu_{blanco}$.

Hacemos un grafico box-plot para evidenciar graficamente estas diferencias, en el que la linea azul marca la media de intención de compra del Iphone según el grupo-color:

```
ggplot(data=as.data.frame(Datos_Apple), aes( x = grupos, y= intencion_compra_Iphone , fill=grupos)) +
  geom_boxplot(  color = 'black') +
  stat_summary(fun=mean, geom="crossbar", shape=2, size=0.3, color="blue")+
  xlab("Intención compra Iphone") +
  ggtitle("Box-plot intención compra Iphone por grupos")+
  scale_y_continuous(n.breaks = 12)
```

```
## Warning: Ignoring unknown parameters: shape
```



Ahora vamos a analizar el tamaño del efecto del factor (color del Iphone) sobre la variable de interes (intencion de compra sobre el Iphone). Para ello calculamos el coeficiente η^2 :

```
FSS<- 15.65
RSS<- 37.49
TSS<-FSS+RSS

(eta_2 <- FSS/TSS)
```

```
## [1] 0.2945051
```

El tamaño del efecto es grande, al ser $\eta^2 > 14$

Se puede concluir que hay un **gran** efecto del color del Iphone (factor) sobre la intencion de compra del propio Iphone (variable cuantitativa de interes). Es decir, el color del Iphone produce un efecto importante (**influye significativamente**) en la intencion de compra de los clientes potenciales.

Una forma autamatizada de calcular η^2 en R es la siguiente:

```
library(lsr)

etaSquared(anova)
```

```
##                         eta.sq eta.sq.part
## Datos_Apple$grupos 0.2945386   0.2945386
```