

Unidad 6: Inferencia estadística

[Descargar este documento en PDF](#)

Introducción

La **estadística inferencial** es la rama de la estadística que permite formular conclusiones sobre una población a partir del análisis de una muestra. Se apoya en el cálculo de probabilidades, que proporciona el marco teórico para modelar fenómenos aleatorios y generalizar los resultados muestrales a toda la población. Dado que las inferencias basadas en muestras están sujetas a incertidumbre, es fundamental expresarlas siempre en términos **probabilísticos**.

Las dos actividades principales en este proceso son:

- **Estimación de parámetros:** consiste en calcular, a partir de los datos muestrales, valores que aproximen parámetros desconocidos de la población.

– *Ejemplo: Estimar la prevalencia de una enfermedad X en la población.*

- **Pruebas de hipótesis:** implica evaluar con base estadística afirmaciones acerca de uno o más parámetros.

Ejemplos:

– *¿La prevalencia de la enfermedad X en Argentina es menor que en Uruguay?*

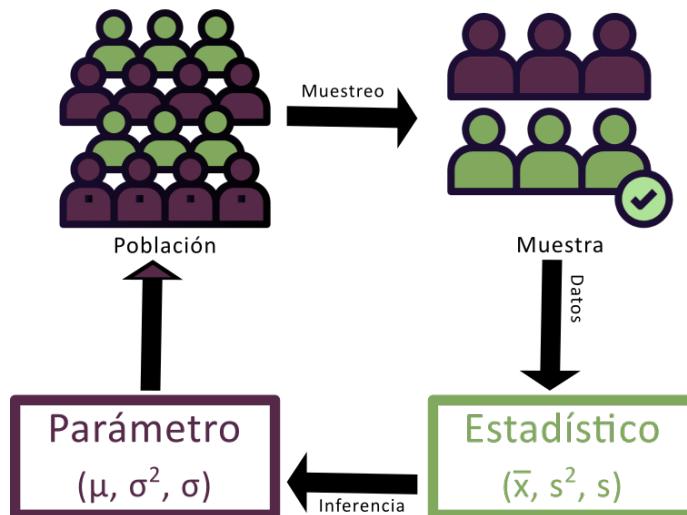
– *¿La prevalencia de la enfermedad X en Argentina en 2010 fue menor que en el año 2000?*

Al trabajar con muestras, entra en juego el concepto de **teoría del muestreo**, que, si bien no abordaremos en profundidad aquí, es clave para comprender cómo se relacionan los valores observados en la muestra con los de la población.

Esta teoría estudia la relación entre la distribución de una variable en la población y el comportamiento de dicha variable en muestras aleatorias extraídas de ella. A las medidas obtenidas a partir de la muestra se las denomina **estadísticos muestrales** o simplemente *estadísticos*, mientras que sus contrapartes en la población se denominan **parámetros**.

Por ejemplo, supongamos que queremos conocer el valor medio de colesterol total de la población de Mar del Plata y tomamos una muestra de tamaño n .

- La media poblacional del colesterol total se representa con la letra griega μ y corresponde al **parámetro**.
- La media muestral, que se obtiene a partir de los datos de la muestra, se representa como \bar{x} y es un **estimador** o estadístico muestral.



La **distribución muestral de un estadístico** es la distribución de todos los valores posibles que ese estadístico puede tomar al calcularse en muestras aleatorias del mismo tamaño extraídas de una misma población. Este concepto es central en la inferencia estadística, ya que permite cuantificar la incertidumbre asociada a las estimaciones.

Como construir una distribución muestral puede resultar muy laborioso cuando la población es grande, y directamente imposible si es infinita, se suelen utilizar aproximaciones basadas en la toma de un gran número de muestras aleatorias del mismo tamaño.

Intervalos de Confianza (IC)

Una forma eficaz de abordar la inferencia estadística es a través de los **intervalos de confianza (IC)**, ya que, aunque son procedimientos inferenciales, están estrechamente vinculados con la estadística descriptiva.

Supongamos que queremos estimar la **media de colesterol** de la población de Mar del Plata. Sería inviable medir el colesterol de cada habitante, por lo que optamos por tomar una muestra de, por ejemplo, 100, 200 o 300 individuos (más adelante veremos cómo determinar el tamaño adecuado de la muestra). Debemos recordar que diferentes muestras producirán en general medias diferentes. Existe, por tanto, un grado de incertidumbre asociado. Si hiciéramos una estimación puntual, obtendríamos un solo valor, pero sin información sobre su variabilidad.

No sabríamos qué tan cerca o lejos está nuestra estimación (\bar{x}) de la verdadera media poblacional (μ).

El **intervalo de confianza** proporciona un rango de valores dentro del cual se espera que se encuentre el valor verdadero del parámetro poblacional, con un cierto nivel de confianza. A diferencia de la estimación puntual que proporciona un único valor numérico, el intervalo consta de dos valores entre los cuales se supone está contenido el parámetro estimado. Entonces, el intervalo de confianza puede expresarse como:

$$IC = estimador puntual \pm (\text{coeficiente de confiabilidad}) * (\text{error standar})$$

donde:

- **Estimador puntual:**

- Para la media poblacional (μ), se toma la media muestral(\bar{x}).
- Para una proporción de la población (p), se toma la proporción muestral (\hat{p}).

- **Coeficiente de confiabilidad:** Se relaciona con el nivel de confianza deseado (por ejemplo, 90%, 95% o 99%), y se expresa como $1 - \alpha$, es decir la probabilidad de que el parámetro se encuentre dentro del IC. Recordemos que el nivel de significancia (α) es la probabilidad de que el parámetro no se halle dentro del IC y es un valor generalmente pequeño (por ejemplo, 0.1, 0.05 o 0.01) expresado como probabilidad o porcentaje (por ejemplo, 10%, 5% o 1%).

- **Error estándar (SE):** Representa la variabilidad de la distribución muestral. Por ejemplo, para la media el error estándar se calcula como la raíz cuadrada de la varianza de la distribución muestral:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Donde σ es la desviación estándar poblacional y n el tamaño de la muestra. Si se estima un IC para una proporción, el error estándar es:

$$SE = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

El proceso se fundamenta en el Teorema del Límite Central (TCL), que establece que, para muestras suficientemente grandes, la distribución de \bar{x} es aproximadamente normal, con media μ y varianza σ^2/n . Así, la variable tipificada:

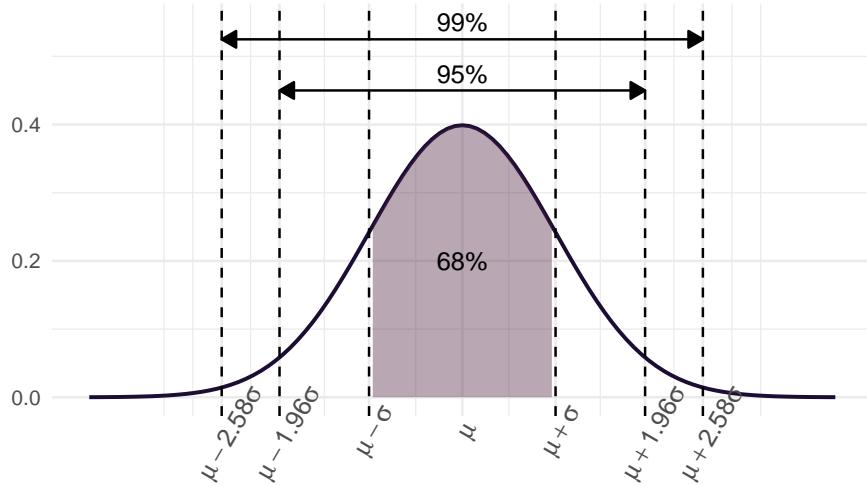
$$Z = \frac{\bar{x} - \mu}{\sigma}$$

sigue una distribución normal estándar (media 0 y desviación estándar 1), lo que permite calcular probabilidades y construir el IC.

En cualquier distribución normal:

- Entre $\mu \pm \sigma$ se encuentra el 68% de los datos.
- Entre $\mu \pm 2\sigma$ se encuentra el 95%.
- Entre $\mu \pm 3\sigma$ se encuentra el 99%.

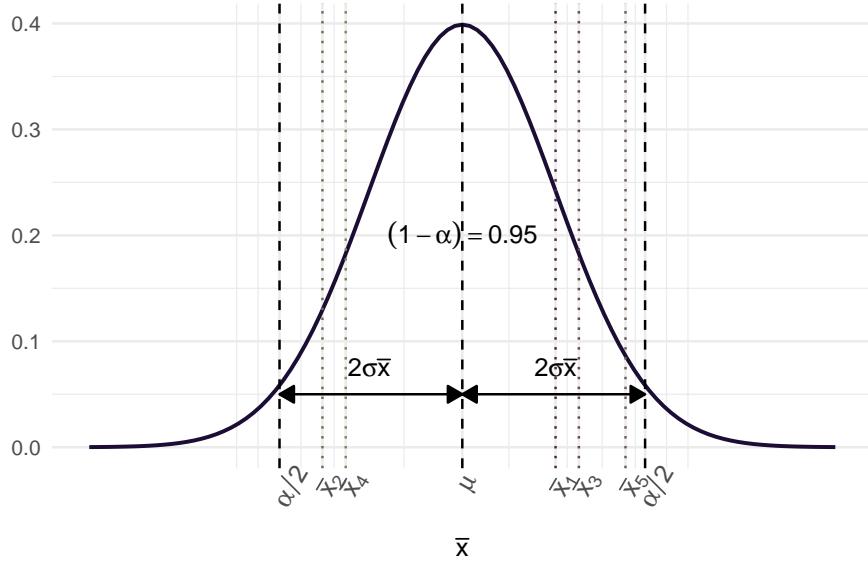
El siguiente gráfico ilustra lo explicado anteriormente:



Sabemos que, independientemente de la localización de los valores, aproximadamente el **95%** de los valores posibles de \bar{x} en la distribución muestral estarán a menos de dos desviaciones estándar de la media μ . Es decir, el intervalo $\mu \pm 2\sigma$ contendrá el **95%** de los valores posibles de \bar{x} .

Supongamos que formamos intervalos a partir de todos los posibles valores de \bar{x} , calculados a partir de todas las muestras posibles de tamaño n tomadas de la población de interés. Esto generará una gran cantidad de intervalos de la forma $\mu \pm 2\sigma$, todos con la misma amplitud, centrados en torno a una μ desconocida.

Aproximadamente el **95%** de estos intervalos tendrán sus centros dentro del intervalo $\mu \pm 2\sigma$. Cada uno de estos intervalos, que se encuentran dentro de $\mu \pm 2\sigma$, puede contener el valor verdadero de μ .



Finalmente, y basándonos en las propiedades de la distribución Normal, se puede deducir la expresión del IC:

$$P(Z_{\alpha/2} < Z_{1-\alpha/2}) = 1 - \alpha$$

Reemplazando Z :

$$P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < Z_{(1-\alpha/2)}\right) = 1 - \alpha$$

Reordenando, la expresión del IC para la media queda:

$$\bar{x} - Z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} < \mu < \bar{x} + Z_{\alpha/2} \frac{\sigma^2}{n}$$

¿Cómo se interpreta un IC?

Si hubiésemos tomado múltiples muestras del mismo tamaño de la población, en al menos $100 * (1 - \alpha)\%$ de las ocasiones el intervalo calculado contendría el parámetro poblacional real. Es decir, un IC al 95% implica que, a largo plazo, el 95% de los intervalos obtenidos a partir de muestras repetidas incluirán el valor verdadero del parámetro.

El producto del coeficiente de confiabilidad y el error estándar se denomina **precisión de la estimación** y es el componente responsable de la amplitud del IC. Recordemos que la fórmula general para construir un intervalo de confianza era:

$$IC = estimador puntual \pm (\text{coeficiente de confiabilidad}) * (\text{error estandar})$$

Para el caso de la media:

- **Aumento de la confiabilidad:** Si se incrementa el nivel de confianza, el coeficiente (por ejemplo, pasando de 1.96 a un valor mayor) aumenta, lo que a su vez incrementa la amplitud del IC.
- **Reducción del error estándar:** Si se fija la confiabilidad (por ejemplo, al 95%), para disminuir la amplitud del IC es necesario reducir el error estándar. Dado que el error estándar de la media es:

$$SE = \frac{\sigma}{\sqrt{n}}$$

y considerando que σ es constante, la única forma de disminuir el error estándar es aumentando el tamaño muestral (n).

Surge entonces la pregunta: **¿qué tan grande debe ser n ?**

La respuesta dependerá de σ , del nivel de significación (α) y de la amplitud deseada para el IC. La relación es:

$$\text{Amplitud} = Z \frac{\sigma}{\sqrt{n}} \implies n = \frac{Z^2 \sigma^2}{\text{Amplitud}^2} \quad (Z = 1.96 \text{ si } \alpha = 0.05)$$

(En la práctica, σ generalmente no se conoce, así que se usa su estimación muestral).

La expresión del error estándar varía según el parámetro a estimar. Hemos visto el caso de la media; si lo que se desea es calcular un IC para una proporción, recordemos que, para muestras grandes, la distribución de las proporciones de la muestra es aproximadamente normal de acuerdo con el TCL. En este caso:

- La media de la distribución es la proporción real p
- La varianza es $p(1 - p)/n$, lo que nos lleva a que el error estándar es:

$$SE = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

y el IC para la proporción se expresa como:

$$\hat{p} - Z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + Z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Dado que un intervalo de confianza implica una declaración probabilística, su cálculo se fundamenta en las distribuciones muestrales de los estimadores y en el correspondiente error estándar. Aunque las fórmulas pueden parecer complejas, los paquetes estadísticos (como R) realizan estos cálculos automáticamente. Lo fundamental es comprender en qué depende la amplitud del IC (nivel de confianza, error estándar y tamaño muestral) y cómo cada uno de estos componentes influye en la precisión de la estimación.

Para profundizar y visualizar simulaciones sobre estos conceptos, pueden explorar recursos interactivos como:

Viendo la teoría: Una introducción visual a probabilidad y estadística

Normalidad y homocedasticidad

Las características fundamentales a la hora de decidir si utilizaremos métodos paramétricos o no paramétricos para la inferencia estadística, es que los datos se ajusten a una distribución normal y conocer si tienen una dispersión homogénea o heterogénea.

Normalidad

Determinar que una distribución es aproximadamente normal nos permite decidirnos por test de comparaciones paramétricos.

Existen tres enfoques que debemos analizar simultáneamente:

- Métodos gráficos
- Métodos analíticos
- Pruebas de bondad de ajuste

Métodos gráficos

El gráfico por excelencia para evaluar normalidad es el Q-Q Plot que consiste en comparar los cuantiles de la distribución observada con los cuantiles teóricos de una distribución normal con la misma media y desviación estándar que los datos.

Cuanto más se aproximen los datos a una normal, más alineados están los puntos entorno a la recta.

En el lenguaje R hay varios paquetes que tienen funciones para construirlos:

```
library(dlookr)
library(ggpubr)
library(moments)
library(nortest)
library(car)
library(tidyverse)
```

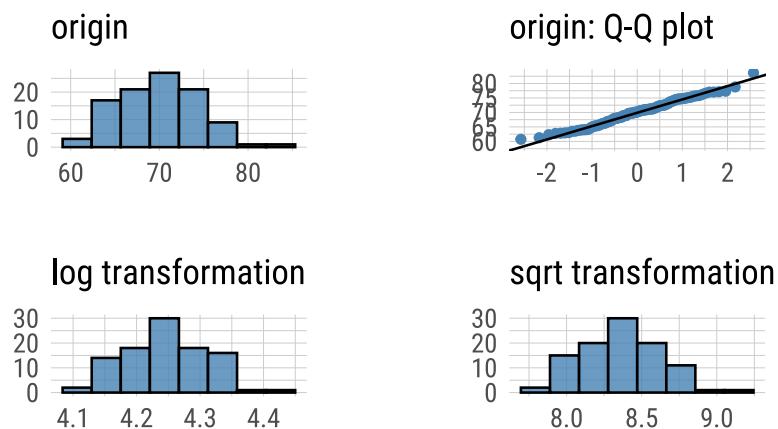
Cargamos datos de ejemplo:

```
datos <- read_csv2("datos/datos_normalidad.csv")
```

Evaluación gráfica de normalidad usando la función `plot_normality()` del paquete `dlookr`:

```
datos |>  
  plot_normality(peso)
```

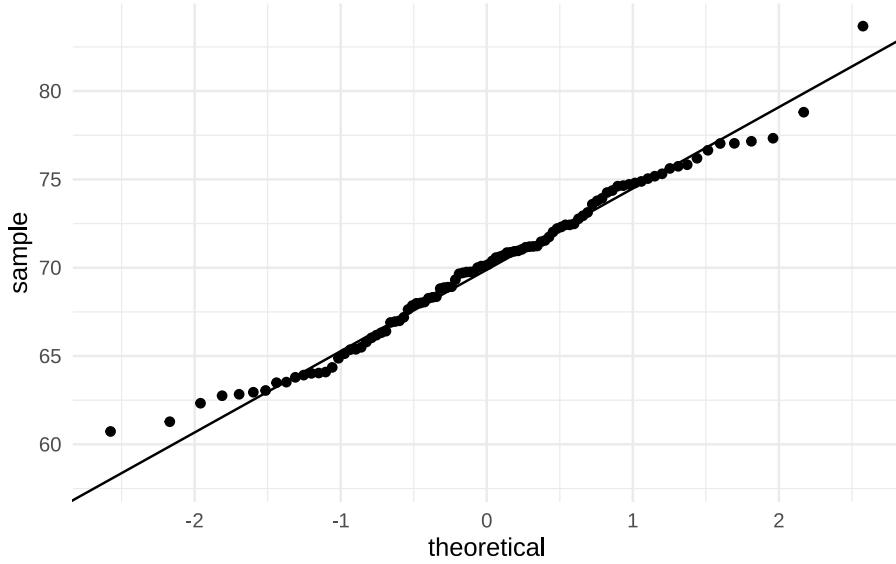
Normality Diagnosis Plot (peso)



A simple vista observamos que los puntos de la variable `peso` se *ajustan bastante bien a la recta*.

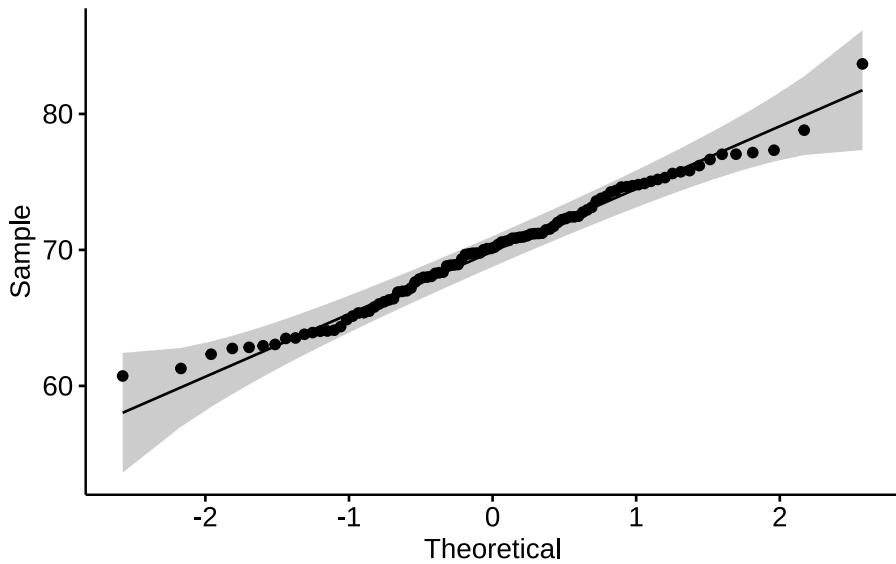
También podemos generar el qqplot usando las funciones `geom_qq_line()` y `geom_qq()` de `ggplot2`:

```
datos |>  
  ggplot(mapping = aes(sample = peso)) +  
    geom_qq_line() +  
    geom_qq() +  
    theme_minimal()
```



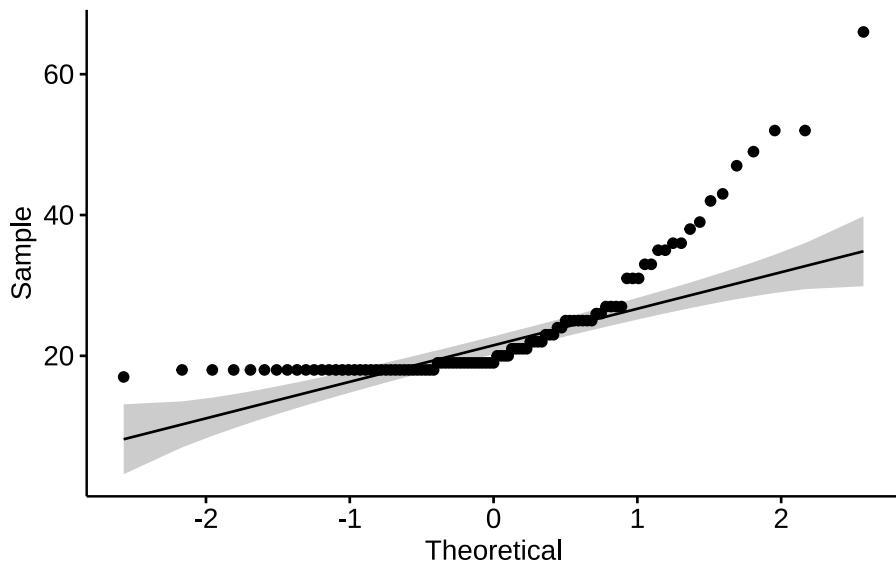
La función `ggqqplot()` del paquete `ggpubr` [@ggpubr] nos permite agregar intervalos de confianza (zona gris alrededor de la recta) que nos orienta mejor sobre “donde caen” los puntos de la variable analizada:

```
datos$peso |>
  ggqqplot()
```



Un ejemplo donde la variable parece no cumplir con el supuesto de normalidad en estos datos de prueba es `edad`:

```
datos$edad |>  
ggqqplot()
```



Métodos analíticos

Medidas de forma

Existen dos medidas de forma útiles que podemos calcular mediante funciones de R.

- La curtosis (kurtosis)
- La asimetría (skewness)

La curtosis mide el grado de agudeza o achatamiento de una distribución con relación a la distribución normal.

- < 0 Distribución platicúrtica (apuntamiento negativo): baja concentración de valores
- > 0 Distribución leptocúrtica (apuntamiento positivo): gran concentración de valores
- $= 0$ Distribución mesocúrtica (apuntamiento normal): concentración como en la distribución normal.

El paquete `moments` [@moments] posee algunas funciones interesantes para analizar medidas de forma, como el estimador de Pearson para curtosis:

```

datos |>
  summarise(kurtosis_edad = kurtosis(edad, na.rm = T),
             kurtosis_peso = kurtosis(peso, na.rm = T))

# A tibble: 1 x 2
  kurtosis_edad kurtosis_peso
  <dbl>          <dbl>
1        8.05       2.72

```

En los dos casos estamos frente a una *distribución leptocúrtica* pero de magnitudes bien diferentes. Muy alta en el caso de la variable `edad` (8,0) y mucho menor para la variable `peso` (2,7).

El índice de asimetría es un indicador que permite establecer el *grado de asimetría* que presenta una distribución. Los valores menores que 0 indican distribución asimétrica negativa; los mayores a 0: distribución asimétrica positiva y cuando sea 0, o muy próximo a 0, distribución simétrica:

```

datos |>
  summarise(asimetria_edad = skewness(edad, na.rm = T),
             asimetria_peso = skewness(peso, na.rm = T))

# A tibble: 1 x 2
  asimetria_edad asimetria_peso
  <dbl>          <dbl>
1        2.19       0.122

```

Los valores obtenidos con la función `skewness()` del paquete `moments` nos informan que la distribución de la edad tienen una asimetría positiva (2,2) y que los valores de peso se distribuyen bastante simétricos (0,1).

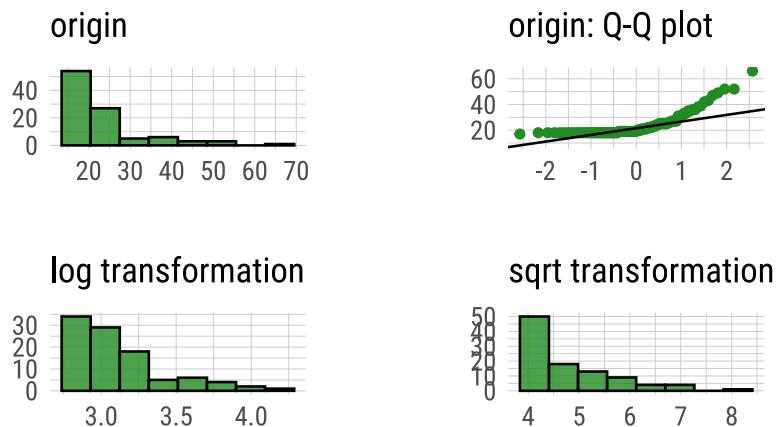
Estas características de las distribuciones también se pueden ver mediante histogramas o gráficos de densidad:

```

datos |>
  plot_normality(edad, col = "forestgreen")

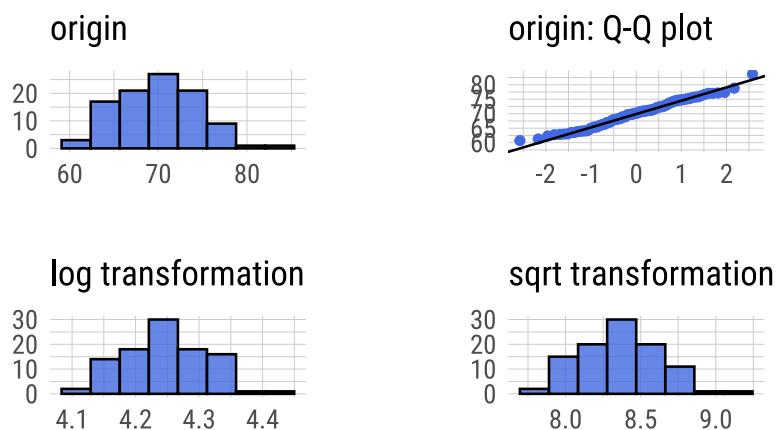
```

Normality Diagnosis Plot (edad)



```
datos |>
  plot_normality(peso, col = "royalblue")
```

Normality Diagnosis Plot (peso)



Los histogramas que se acerquen a la clásica “campana de Gauss” tendrán curtosis y asimetrías alrededor del valor cero.

Pruebas de bondad de ajuste

Una prueba de bondad de ajuste permite testear la hipótesis de que una variable aleatoria sigue cierta distribución de probabilidad y se utiliza en situaciones donde se requiere comparar una distribución observada con una teórica o hipotética.

El mecanismo es idéntico a cualquier test de hipótesis salvo que aquí esperamos no descartar la hipótesis nula de igualdad, por lo que obtener valores p de probabilidad mayores a 0,05 es signo de que la distribución de la variable analizada se ajusta.

A continuación, presentaremos los test de hipótesis más utilizados para analizar normalidad.

Test de Shapiro-Wilk

Lleva el nombre de sus [autores](#) (Samuel Shapiro y Martin Wilk) y es usado preferentemente para muestras de hasta 50 observaciones.

La función se encuentra desarrollada en el paquete **stats** y se llama **shapiro.test()**:

```
shapiro.test(datos$edad)
```

```
Shapiro-Wilk normality test

data: datos$edad
W = 0.69517, p-value = 4.912e-13
shapiro.test(datos$peso)
```

```
Shapiro-Wilk normality test

data: datos$peso
W = 0.98615, p-value = 0.383
```

Interpretación: Siendo la hipótesis nula que la población está distribuida normalmente, si el p-valor es menor a α (nivel de significancia, convencionalmente un 0,05) entonces la hipótesis nula es rechazada (se concluye que los datos no provienen de una distribución normal). Si el p-valor es mayor a α , se concluye que no se puede rechazar dicha hipótesis.

En función de esta interpretación (que es común a todos los test de hipótesis de normalidad), podemos decir que la distribución de la variable **edad** *no se ajusta a la normal* y no podemos rechazar que la distribución de la variable **peso** se ajuste.

Test de Kolmogorov-Smirnov

El test de **Kolmogorov-Smirnov** permite estudiar si una muestra procede de una población con una determinada distribución que no está limitado únicamente a la distribución normal.

El test asume que se conoce la media y varianza poblacional, lo que en la mayoría de los casos no es posible. Para resolver este problema, se realizó una modificación conocida como test Lilliefors.

Test de Lilliefors

El test de **Lilliefors** asume que la media y varianza son desconocidas y está especialmente desarrollado para contrastar la normalidad.

Es la alternativa al test de Shapiro-Wilk cuando el número de observaciones es mayor de 50.

La función `lillie.test()` del paquete `nortest` [@nortest] permite aplicarlo:

```
lillie.test(datos$edad)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: datos$edad  
D = 0.24892, p-value < 2.2e-16
```

```
lillie.test(datos$peso)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: datos$peso  
D = 0.049534, p-value = 0.7905
```

Los resultados son coincidentes con los obtenidos anteriormente.

Test de D'agostino

Esta prueba se basa en las transformaciones de la curtosis y la asimetría de la muestra, y solo tiene poder frente a las alternativas de que la distribución sea sesgada.

El paquete `moments` la tiene implementada en `agostino.test()`:

```
agostino.test(datos$edad)
```

```
D'Agostino skewness test
```

```
data: datos$edad  
skew = 2.1947, z = 6.3465, p-value = 2.203e-10
```

```
alternative hypothesis: data have a skewness  
agostino.test(datos$peso)
```

```
D'Agostino skewness test  
  
data: datos$peso  
skew = 0.12160, z = 0.52683, p-value = 0.5983  
alternative hypothesis: data have a skewness
```

Los resultados coinciden con la observación de asimetría que efectuamos con los métodos analíticos, confirmando que la variable **edad** *no se ajusta* a una curva simétrica y la variable **peso** si lo hace.

Cuando estos test se emplean con la finalidad de verificar las condiciones de métodos paramétricos es importante tener en cuenta que, al tratarse de valores probabilidad, cuanto mayor sea el tamaño de la muestra más poder estadístico tienen y más fácil es encontrar evidencias en contra de la hipótesis nula de normalidad.

Por otra parte, cuanto mayor sea el tamaño de la muestra, menos sensibles son los métodos paramétricos a la falta de normalidad. Por esta razón, es importante no basar las conclusiones únicamente en los resultados de los test, sino también considerar los otros métodos (gráfico y analítico) y no olvidar el tamaño de la muestra.

Homocedasticidad

La homogeneidad de varianzas es un supuesto que considera constante la varianza en los distintos grupos que queremos comparar.

Esta homogeneidad es condición necesaria antes de aplicar algunos test de hipótesis de comparaciones o bien para aplicar correcciones mediante los argumentos de las funciones de R.

Existen diferentes test de bondad de ajuste que permiten evaluar la distribución de la varianza. Todos ellos consideran como H_0 que la varianza es igual entre los grupos y como H_1 que no lo es.

La diferencia entre ellos es el estadístico de centralidad que utilizan:

- Media de la varianza: son los más potentes pero se aplican en distribuciones que se aproximan a la normal.
- Mediana de la varianza: son menos potentes pero consiguen mejores resultados en distribuciones asimétricas.

F-test

Este test es un contraste de la razón de varianzas, mediante el estadístico F que sigue una distribución F-Snedecor.

Se utiliza cuando las distribuciones se aproximan a la “normal” y en R base se la encuentra en la función `var.test()` que permite utilizar la sintaxis de fórmula:

```
variable_cuantitativa ~ variable_categórica_grupos
```

Por ejemplo:

```
var.test(formula = peso ~ sexo, data = datos)
```

```
F test to compare two variances

data: peso by sexo
F = 0.67914, num df = 50, denom df = 48, p-value = 0.1781
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.384712 1.194943
sample estimates:
ratio of variances
 0.6791414
```

Comparamos las varianzas de la variable peso entre el grupo de mujeres y hombres. El valor p del test indica que no podemos descartar la igualdad de varianzas entre los grupos (H_0) o lo que es lo mismo el test no encuentra diferencias significativas entre las varianzas de los dos grupos.

Test de Bartlett

Este test se puede utilizar como alternativa al F-test, sobre todo porque nos permite aplicarlo cuando tenemos más de 2 grupos de comparación. Al igual que el anterior es sensible a las desviaciones de la normalidad.

La función en R base es `bartlett.test()` y también se pueden usar argumentos tipo fórmula:

```
bartlett.test(formula = peso ~ sexo, data = datos)
```

```
Bartlett test of homogeneity of variances

data: peso by sexo
Bartlett's K-squared = 1.8082, df = 1, p-value = 0.1787
```

El resultado es coincidente con el mostrado por `var.test()`. No se encuentran diferencias significativas entre las varianzas de los pesos en los dos grupos (Mujer - Varón)

Test de Levene

El test de Levene sirve para comparar la varianza de 2 o más grupos pero además permite elegir distintos estadísticos de tendencia central. Por lo tanto, la podemos adaptar a distribuciones alejadas de la normalidad seleccionando por ejemplo la mediana.

La función `leveneTest()` se encuentra disponible en el paquete `car`. La vemos aplicada sobre `peso` para los diferentes grupos de `sexo` y utilizando la `media` como estadístico de centralidad, dado que la distribución de `peso` se aproxima a la normal.

```
leveneTest(y = peso ~ sexo, data = datos, center = "mean")
```

```
Levene's Test for Homogeneity of Variance (center = "mean")
  Df F value Pr(>F)
group  1      2 0.1605
      98
```

La conclusión es la misma que la encontrada anteriormente.

Ahora vamos aplicarla sobre la variable `edad`, de la que habíamos descartado “normalidad”. Lo hacemos usando el argumento `center` con `"mean"` (media) y con `"median"` (mediana).

```
leveneTest(y = edad ~ sexo, data = datos, center = "mean")
```

```
Levene's Test for Homogeneity of Variance (center = "mean")
  Df F value Pr(>F)
group  1  4.6784  0.033 *
      97
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(y = edad ~ sexo, data = datos, center = "median")
```

```
Levene's Test for Homogeneity of Variance (center = "median")
  Df F value Pr(>F)
group  1  2.4413  0.1214
      97
```

Los resultados son diferentes. Mientras con el centrado en la media nos da un p valor significativo menor a 0,05 con el centrado en la mediana no nos permite descartar homocedasticidad.

Observamos aquí las distorsiones sobre la media y las formas paramétricas que devienen de distribuciones asimétricas y alejadas de la curva normal. El código correcto para este caso (variable `edad`) es usar el centrado en la mediana (`center = "median"`).

Test de hipótesis

Aunque los estudios de corte transversal no se diseñan originalmente con grupos de comparación, en aquellos de carácter más analítico es frecuente establecer comparaciones. Por ejemplo, pueden surgir preguntas como:

¿La prevalencia de la enfermedad es mayor en mujeres, en determinados grupos etarios o en una provincia específica?

Con esto en mente, revisaremos las herramientas que permiten comparar grupos mediante test o contrastes de hipótesis. El propósito de estos test es ofrecer al investigador una herramienta para tomar decisiones sobre la población a partir de la información obtenida en una muestra.

Antes de adentrarnos en la parte estadística, es importante distinguir entre dos tipos de hipótesis:

- **Hipótesis de investigación:** Es la conjeta o suposición que motiva la investigación.
- **Hipótesis estadística:** Es aquella que puede ser evaluada mediante técnicas estadísticas apropiadas.

En este texto nos centraremos en aclarar aspectos relativos a las hipótesis estadísticas, asumiendo que las hipótesis de investigación ya han sido discutidas previamente por los investigadores. Describiremos brevemente el razonamiento subyacente a estos test.

Los contrastes de hipótesis parten de una **hipótesis nula**, la cual afirma que los dos grupos comparados son iguales o, en otras palabras, que las diferencias observadas se deben únicamente al azar. Como ya se ha mencionado, la variabilidad intrínseca de cualquier muestra impide que la diferencia entre grupos sea exactamente cero.

El método estadístico nos permite cuantificar la diferencia entre grupos asumiendo que, si repitiésemos el experimento infinitas veces obtenemos todas las posibles muestras del tamaño indicado a partir de nuestras poblaciones (*distribución muestral*), las diferencias entre grupos “iguales” se distribuirían conforme a una curva teórica. Basándonos en las propiedades de esta distribución, podemos determinar un valor límite que comprende, por ejemplo, el 95% o el 99% de las diferencias esperadas. Si la diferencia observada entre las muestras supera este valor límite, se considera excesiva para ser atribuible al azar y, por tanto, se rechaza la hipótesis nula. Por el contrario, si la diferencia cae dentro del área del 95%, se concluye que la diferencia encontrada podría atribuirse al azar, y no habría evidencia que nos permita rechazar la hipótesis nula. En estos casos decimos que los grupos “no son diferentes” pero no “son iguales”, ya que la variabilidad inherente impide probar una igualdad exacta.

Los contrastes de hipótesis se realizan generalmente bajo las siguientes condiciones:

- Se asume a priori que la ley de distribución de la población es conocida.
- Se extrae una muestra aleatoria de dicha población.

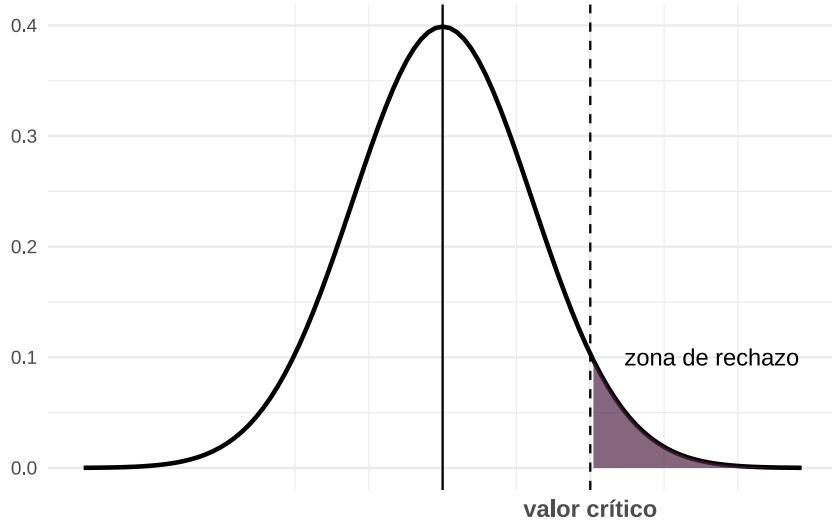
El conjunto de estas técnicas de inferencia se denomina **técnicas paramétricas**. Sin embargo, existen otros métodos, denominados **técnicas no paramétricas** o contrastes de distribuciones libres, que no requieren estimar parámetros ni suponer una ley de probabilidad específica para la población. Algunos de estos métodos serán desarrollados más adelante en este módulo.

Es importante señalar que los contrastes de hipótesis se utilizan no solo en estudios transversales, sino con mayor frecuencia en diseños en los que *a priori* existen grupos de comparación, como en estudios de casos y controles, cohortes, ensayos clínicos, entre otros. Desarrollaremos aquí la teoría que los sustenta y los utilizaremos a lo largo de todo el curso.

Estructura del test de hipótesis

Los componentes básicos de cualquier test de contraste de hipótesis son:

- **Hipótesis nula (H_0):** Afirma que no existe diferencia entre los grupos que se comparan, es decir, que las variaciones observadas se deben únicamente al azar.
- **Hipótesis alternativa (H_1):** Es la conjectura o suposición que plantea el investigador, estableciendo que sí existe una diferencia entre los grupos. Generalmente es complementaria de la H_0 .
- **Estadístico de prueba:** Es el valor calculado a partir de los datos muestrales que se utiliza para tomar la decisión sobre la H_0 . Cada tipo de problema tiene un estadístico adecuado, cuya magnitud, al compararse con su distribución muestral (por ejemplo, la distribución normal estándar en el caso del estadístico Z), permite determinar si las diferencias observadas son atribuibles al azar.
- **Valor crítico o Región crítica:** La región crítica se establece en función del nivel de significación (α) y consiste en el conjunto de valores extremos del estadístico de prueba que, de ser alcanzados, llevarían a rechazar la H_0 . Todos los posibles valores del estadístico se ubican en el eje horizontal de la gráfica de su distribución, y la región crítica delimita aquellos valores que son muy poco probables bajo la hipótesis nula.



La regla de decisión es la siguiente:

- Si el valor del estadístico de prueba calculado a partir de la muestra cae en la región crítica, se rechaza la H_0 y se concluye que las diferencias observadas son **estadísticamente significativas**.
- Si el valor no cae en la región crítica, no se rechaza la H_0 ; esto indica que las diferencias entre lo observado y lo esperado pueden explicarse por el azar. Es decir, **no son estadísticamente significativas**.
- **Nivel de significación (α)**: Es la probabilidad de cometer un error tipo I, es decir, rechazar H_0 cuando es verdadera. Este valor, elegido por el investigador (comúnmente 5% o 1%), determina el límite entre la región de no rechazo y la región crítica.
- **Valor p** : Es una medida de qué tan probable son los resultados de la muestra, considerando que H_0 sea verdadera.
 - Un valor p muy pequeño indica que es muy poco probable obtener los resultados observados para el estadístico muestral si H_0 fuese cierta, por lo que debemos rechazarla.
 - Esto significa que si el valor $p \leq \alpha$, es posible rechazar H_0 ; mientras que si $p > \alpha$, no es posible rechazar la hipótesis nula.

Por ejemplo, si en un test de contraste de dos proporciones se obtiene un estadístico $Z = 3,034$ y un valor $p = 0.0024$, esto significa que la probabilidad de obtener un valor de Z de 3.034 o mayor, suponiendo que H_0 es cierta, es del 0.24%. Dado que este valor es mucho menor que un nivel de significación del 5% (o incluso del 1%), se rechaza H_0 y se concluye que la diferencia observada es estadísticamente significativa. En otras palabras, si bajo un supuesto dado, la probabilidad de un suceso observado particular

es excepcionalmente pequeña, concluimos que el supuesto probablemente es incorrecto (*regla del suceso infrecuente*).

Tipos de contrastes

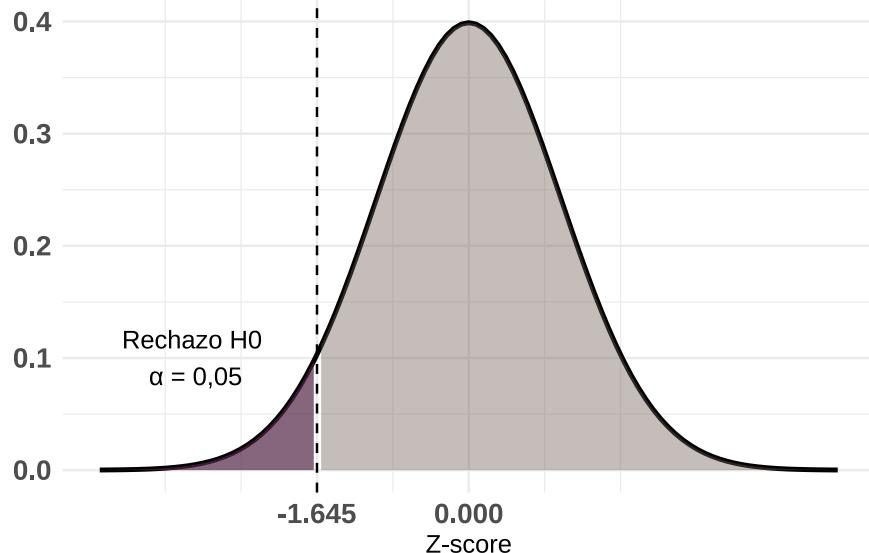
Los contrastes de hipótesis se clasifican según la forma de la hipótesis alternativa (H_1). Esta clasificación determina si la prueba es **unilateral** de cola izquierda o derecha) o **bilateral** (de dos colas). Siendo las colas de la distribución las regiones extremas limitadas por los valores críticos.

Test de cola izquierda

La hipótesis alternativa plantea que la media del primer grupo es significativamente **menor** que la del segundo:

$$H_1 : \mu_1 < \mu_2$$

La región crítica se encuentra en el extremo izquierdo de la distribución. Todo el área crítica tiene un tamaño α con un valor crítico de $-1,645$.

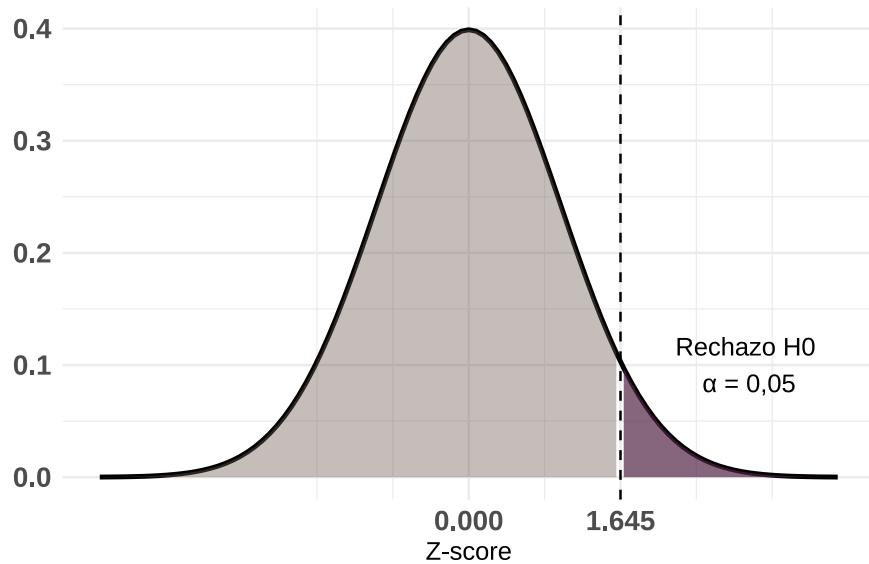


Test de cola derecha

La hipótesis alternativa establece que la media del primer grupo es significativamente **mayor** que la del segundo:

$$H_1 : \mu_1 > \mu_2$$

La región crítica se concentra en el extremo derecho de la distribución y toda el área crítica tiene un tamaño α con un valor crítico de 1,645.

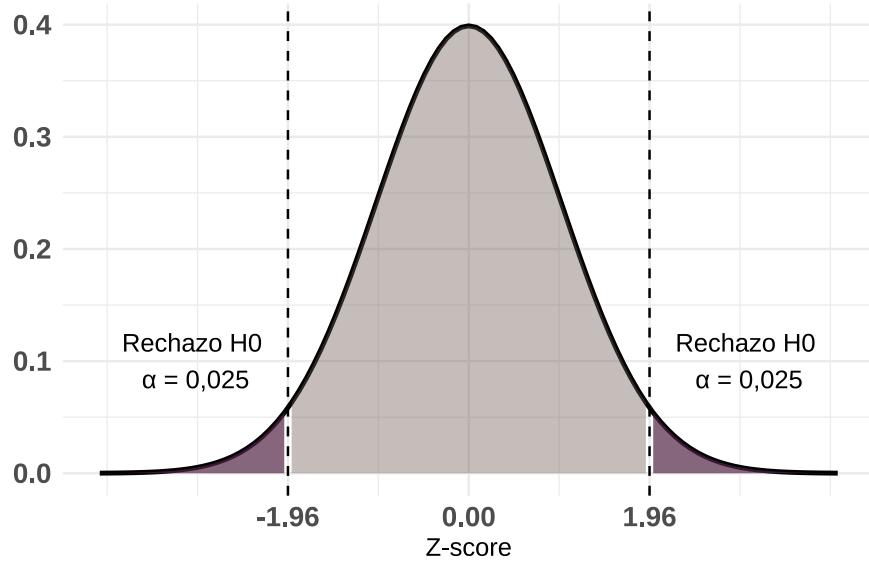


Pruebas bilaterales

La hipótesis alternativa afirma que existen diferencias entre los grupos, **sin especificar la dirección:**

$$H_1 : \mu_1 \neq \mu_2$$

La región crítica se divide entre ambos extremos de la distribución, con valores críticos de $\pm 1,96$. El nivel de significación total (α) se reparte en partes iguales entre las dos colas ($\alpha/2$ en cada una), lo que implica un 2,5% de probabilidad en cada cola si H_0 es verdadera.



Errores

Al utilizar el razonamiento de los contrastes de hipótesis, existen dos tipos principales de errores que podemos cometer:

- **Error tipo I (α):** Ocurre cuando el investigador rechaza la hipótesis nula (H_0) siendo esta verdadera en la población y se concluye erróneamente que existe una diferencia cuando en realidad no la hay. Se suele elegir un valor pequeño de α (0.01, 0.05 y 0.10) para hacer que la probabilidad de rechazo de H_0 sea pequeña.
- **Error tipo II (β):** Ocurre cuando el investigador no rechaza la H_0 siendo esta falsa en la población, es decir, se falla en detectar una diferencia real. Generalmente β es mayor que α , pero su valor real se desconoce en la práctica.

Es importante notar que, una vez que se realiza el procedimiento de prueba, no es posible saber con certeza si se ha cometido alguno de estos errores, ya que se desconoce el verdadero estado de la realidad. Sin embargo, al fijar un α pequeño, se busca asegurar que, en caso de rechazar la H_0 , la probabilidad de haber cometido un error Tipo I sea baja.

En resumen, al interpretar los resultados de un test de hipótesis:

- **Si se rechaza la H_0 :** se asume que la probabilidad de haber cometido un error Tipo I es baja (debido al valor pequeño de α).
- **Si no se rechaza la H_0 :** se desconoce el riesgo real de un error Tipo II, pero es importante tener en cuenta que, en muchas situaciones, este riesgo

	No rechazar H_0	Rechazar H_0
H_0 es cierta	Correcto (1- α)	Error tipo I (α)
H_0 es falsa	Error tipo II (β)	Correcto (1- β)

es mayor que el de un error Tipo I.

La tabla que se presenta a continuación resume las posibles situaciones a las que nos enfrentamos con los test de hipótesis:

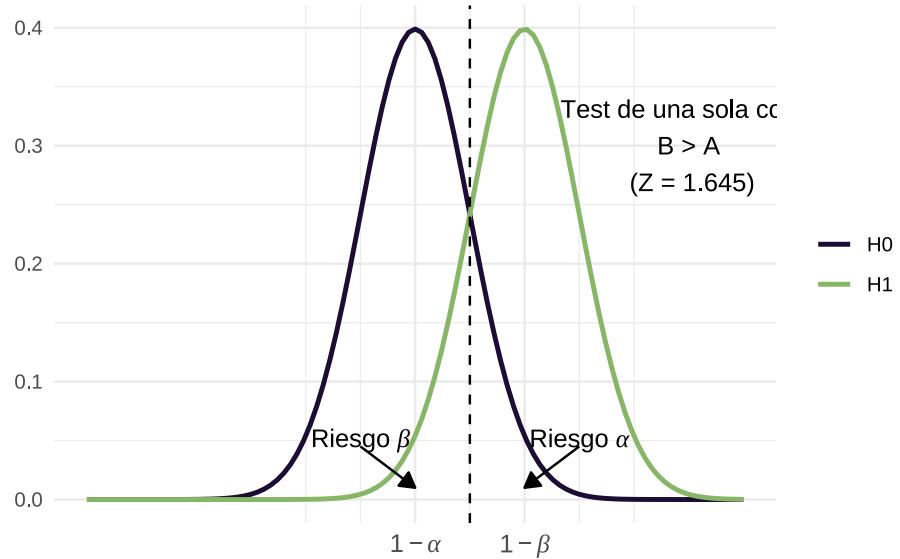
Para quienes están familiarizados con el ámbito del diagnóstico, existe una clara analogía entre los falsos positivos y falsos negativos en las pruebas diagnósticas y, respectivamente, el error Tipo I y el error Tipo II en los contrastes de hipótesis.

Hemos discutido el significado de α (error Tipo I); ahora veamos qué implica β . Recordemos que el error Tipo II es análogo a los falsos negativos de las pruebas diagnósticas: es la probabilidad de no detectar una diferencia cuando, en realidad, ésta existe. En otras palabras, β es la probabilidad de no rechazar la hipótesis nula (H_0) siendo esta falsa.

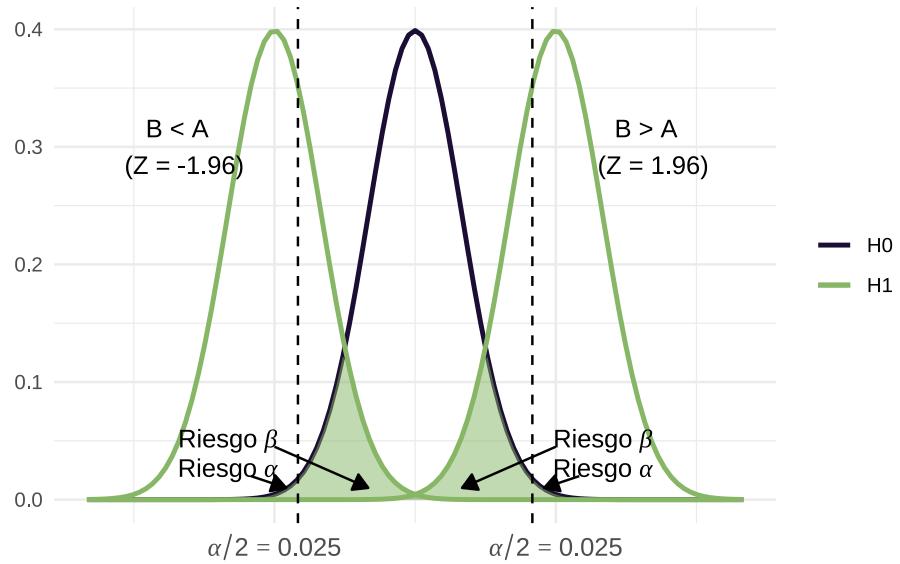
A diferencia de α que se fija en un único valor y es determinado por el investigador, β varía según el valor real del parámetro en estudio. Por ejemplo, si consideramos la hipótesis nula $H_0 : \mu_1 - \mu_2 = 0$, habrá un valor de β para cada posible diferencia entre μ_1 y μ_2 cuando el valor real no sea cero. La probabilidad de detectar correctamente una diferencia real - es decir, de obtener un resultado estadísticamente significativo cuando la diferencia existe- se denomina **potencia estadística** y se expresa como $1 - \beta$.

Los contrastes de hipótesis no son exclusivos de los estudios transversales; por el contrario, su uso es más común en estudios analíticos que involucran grupos de comparación, tales como estudios de casos y controles, cohortes o ensayos clínicos. Por ejemplo, en un ensayo clínico que evalúa dos tratamientos, la potencia estadística refleja la capacidad del estudio para identificar un efecto real del tratamiento.

Es deseable que la potencia del estudio sea lo mayor posible, ya que esto incrementa la probabilidad de detectar diferencias verdaderas. Sin embargo, no es posible minimizar ambos errores simultáneamente ya que al disminuir α (es decir, al ser más exigentes para rechazar la H_0), β tiende a aumentar, y viceversa. En los contrastes, la hipótesis privilegiada es H_0 que solo será rechazada cuando la evidencia de su falsedad supere el umbral del $1 - \alpha$. Esto significa que, a menos que la evidencia en contra de H_0 sea muy significativa, se opta por no rechazarla. Lo ideal a la hora de definir un test es encontrar un compromiso entre β y α .



Mientras que para tests bilaterales:



Finalmente, podemos decir que la potencia estadística ofrece un segundo mecanismo de seguridad en un contraste de hipótesis. Es como contar con una protección adicional en la toma de decisiones: si solo dispusiéramos del nivel de significación (α), tendríamos menos garantías. Al incorporar la potencia ($1 - \beta$) agregamos un segundo control. Por ello, en un contraste no basta con tener un

valor p pequeño; también se necesita una potencia alta, que en la práctica suele fijarse en un 80% (es decir, $1 - \beta = 0.8$).

Por otro lado, cuando existe una diferencia real —o un efecto real de una terapia, o una verdadera diferencia entre dos fármacos—, la **magnitud** de ese efecto influye en la facilidad para detectarlo. Los efectos grandes son más fáciles de identificar que los pequeños. Para estimar la potencia de una prueba, debemos especificar el **efecto mínimo** que valga la pena identificar.

La potencia de un test estadístico depende de tres factores que actúan de manera interrelacionada:

- El **riesgo de error** que se tolerará al rechazar la hipótesis de ausencia de efecto o diferencia (α).
- La **dimensión de la diferencia** que se desea identificar en relación con la variabilidad en las poblaciones.
- El **tamaño de la muestra**.

Del mismo modo que en un problema de estimación se necesita una idea de la magnitud a estimar y del error aceptable para definir el tamaño de la muestra, en un contraste de hipótesis se requiere conocer el **tamaño del efecto** que se quiere detectar. Así, el tamaño muestral se determina en función del nivel de confianza y de la potencia de la prueba, además de otros aspectos relacionados con el diseño y la prueba estadística elegida.

En epidemiología, una de las situaciones más frecuentes al diseñar un estudio es el cálculo del tamaño muestral para un **nivel de confianza del 95%** y una **potencia del 80%**, que, como se mencionó, es un nivel de potencia alto y que además permite manejar un α relativamente bajo. El lenguaje **R** cuenta con diversas funciones para **calcular y visualizar** la relación entre tamaño muestral, potencia, tamaño del efecto y nivel de confianza, facilitando el diseño de estudios con un adecuado balance entre la probabilidad de detectar diferencias reales y la de controlar errores estadísticos.

Por ejemplo, la función `pwr.t.test()` del paquete `pwr` [@pwr], calcula la potencia para pruebas t de Student de medias (para una muestra, dos muestras y muestras pareadas), basado en el tamaño de la muestra, el nivel de confianza y el tamaño de efecto:

```
pwr.t.test(n, d, sig.level = 0.05, power, type, alternative)
```

Donde:

- **n**: Número de observaciones para cada grupo.
- **d**: tamaño de efecto (d de Cohen) - diferencia (estandarizada) entre grupos.

Nota: La d de Cohen representa las desviaciones estándar que separan dos o más grupos. Por ejemplo: $d_{Cohen} = 0.5$ representa que la diferencia entre grupo experimental y muestral es de media

desviación estándar. Cohen sugirió (provisoriamente) que 0.2 es un tamaño de efecto pequeño, 0.5 es mediano y 0.8 es grande,

- `sig.level`: nivel de significación (probabilidad del error de tipo I).
- `power`: potencia del test (1 menos la probabilidad del error tipo II).
- `type`: tipo de test ("two.sample", "one.sample", "paired").
- `alternative`: palabra que especifica la hipótesis alternativa, debe ser "two.sided" (predeterminado), "greater" or "less".

La función se ejecuta incorporando todos los argumentos obligatorios (`d`, `n`, `power` y `sig.level`) menos el que se quiere calcular. En ese caso se iguala a `NULL` o se omite.

Supongamos que queremos conocer el tamaño de la muestra para detectar diferencias en la media de la hemoglobina glicosilada entre dos grupos de pacientes con tratamientos de control de la diabetes distintos. Aceptamos un nivel de efecto convencional de una pequeña desviación ($d_{Cohen} = 0.2$), una potencia del 80% y una significación habitual de 0,05.

Cargamos el paquete requerido:

```
library(pwr)
```

Calculamos el tamaño de muestra:

```
pwr.t.test(d = 0.2,
            power = 0.8,
            sig.level = 0.05,
            type = "two.sample",
            alternative = "two.sided")
```

```
Two-sample t test power calculation

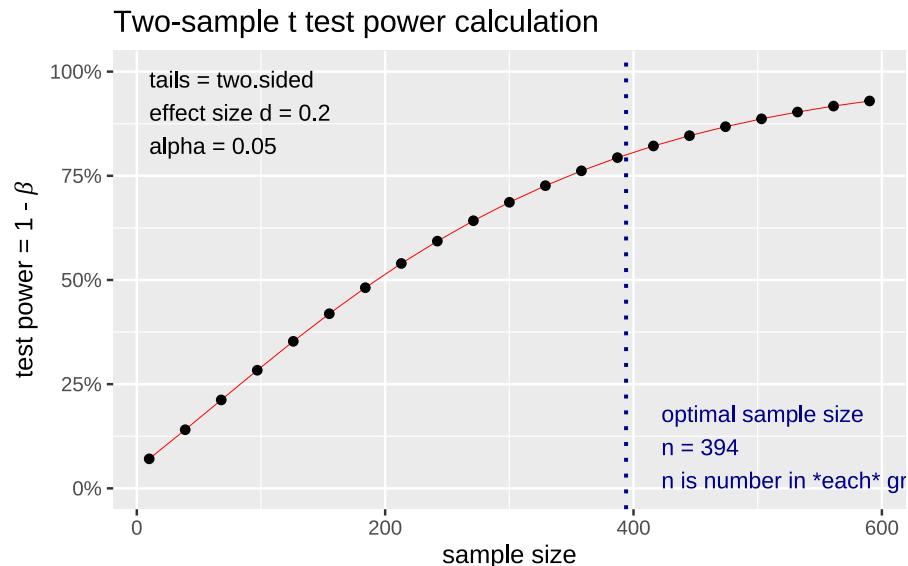
n = 393.4057
d = 0.2
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: `n` is number in *each* group

También podemos graficar la salida:

```
pwr.t.test(d = 0.2,
            power = 0.8,
            sig.level = 0.05,
            type = "two.sample",
```

```
alternative = "two.sided") |>
plot()
```



Observamos que, si por ejemplo tomásemos una muestra de 300 individuos por grupo, la potencia del estudio con ese tamaño de efecto y nivel de significación del 0,05 sería aproximadamente de 68%.

¿Qué test se debe aplicar en cada caso?

Hemos discutido la mecánica general de los test de hipótesis. Ahora, nos centraremos en orientarnos sobre qué test aplicar en cada situación. Aunque existe un desarrollo teórico detrás de cada caso, aquí nos quedaremos con ciertas reglas prácticas.

Para sistematizar los contrastes de hipótesis, es útil responder dos preguntas fundamentales:

- **¿Qué tipo de variable dependiente tengo?**

La variable resultado puede ser cuantitativa, cualitativa, de tiempo hasta un evento, etc.

- **¿Qué se está comparando?**

Esto se traduce en evaluar el tipo de experimento: ¿se comparan dos grupos o más? ¿Son muestras independientes o relacionadas (por ejemplo, medidas en el mismo grupo antes y después de una intervención)?

Variable resultado cuantitativa

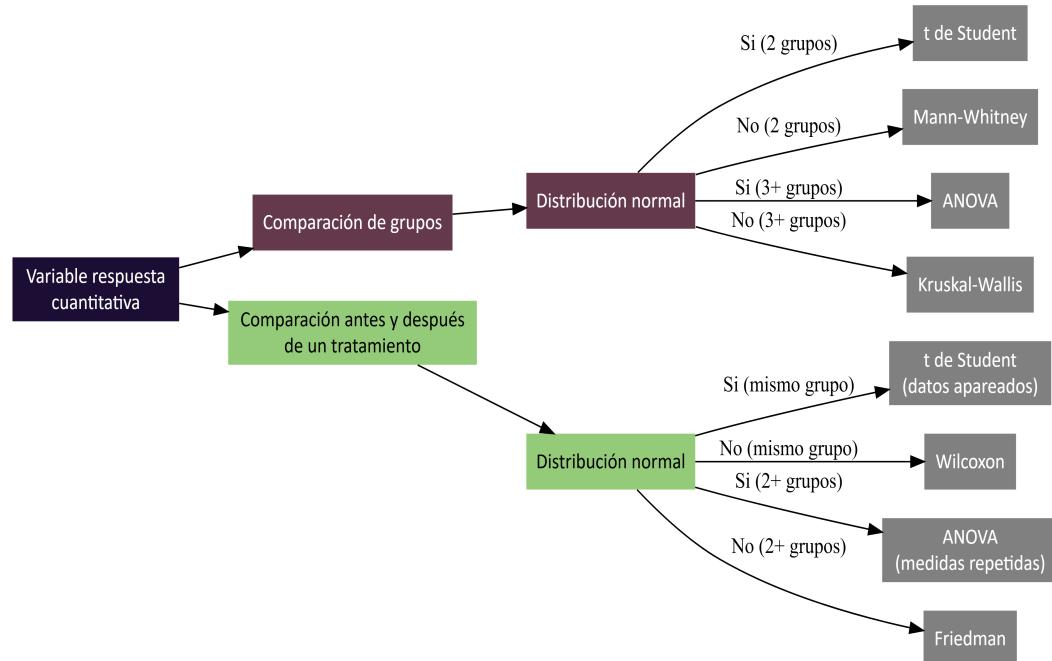
Cuando la variable de interés es cuantitativa, la comparación de grupos generalmente se traduce en comparar las medias de dichos grupos. En este contexto, se deben responder una tercera pregunta y es si dicha variable se *distribuye normalmente*, porque si no fuera así, debemos recurrir a los contrastes *no paramétricos*.

Por ejemplo, para comparar dos grupos se plantearía:

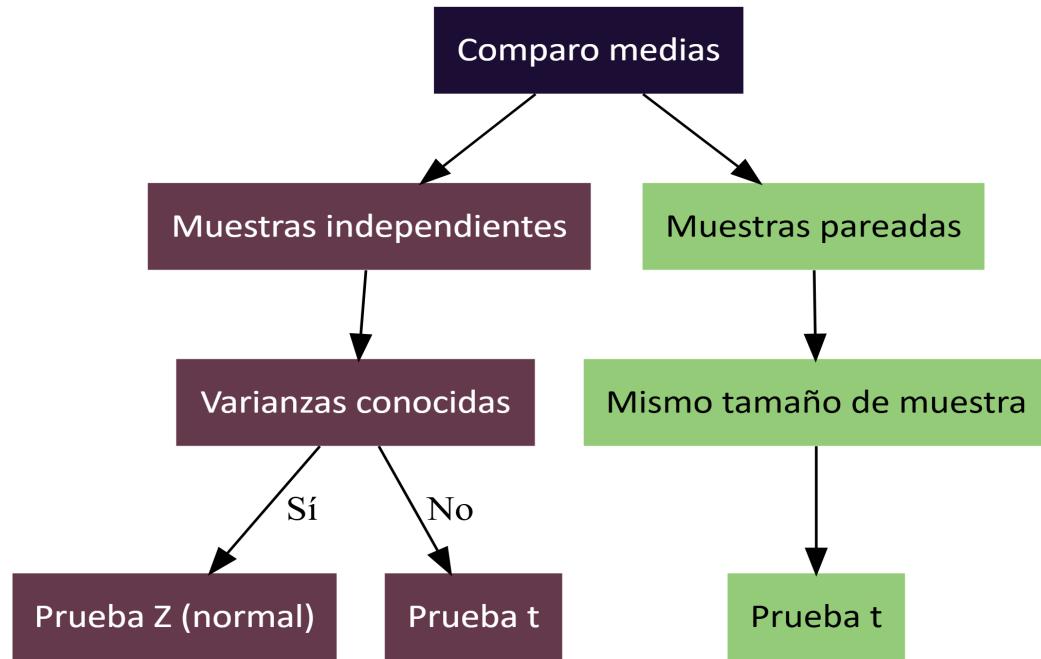
- $H_0 : \mu_1 = \mu_2$ (o su equivalente $\mu_1 - \mu_2 = 0$)
- $H_1 : \mu_1 \neq \mu_2$ (contraste bilateral) o bien $\mu_1 < \mu_2$ o $\mu_1 > \mu_2$ (contrastos unilaterales).

Se calcula el estadístico de prueba (por ejemplo, t de Student o Z , según el tamaño de la muestra y si se conoce la varianza poblacional), se determina la región de rechazo y, finalmente, si $p < \alpha$ rechazo H_0 y si $p > \alpha$ no rechazo H_0 .

A continuación presentamos un esquema que sirve de guía para potenciales situaciones:



Hay algo más que los estadísticos nos “obligan” a considerar en el caso que compare medias independientes y se haya verificado la normalidad de la distribución, y se trata de considerar si las varianzas de las poblaciones que comparo son iguales o diferentes. El algoritmo de resolución es el siguiente:

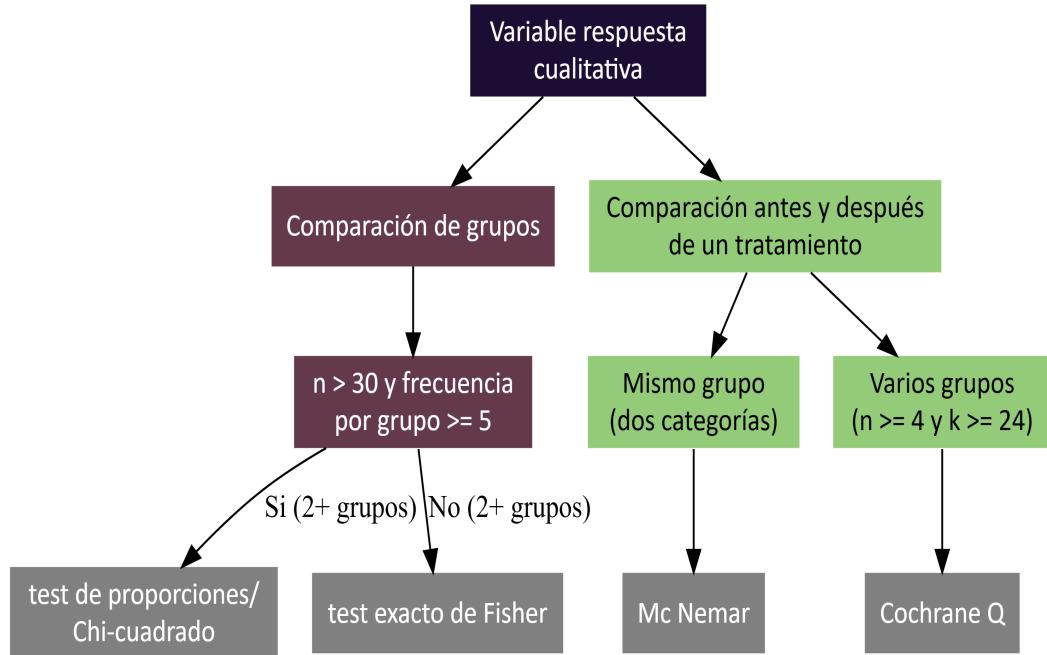


Cuando trabajamos con muestras independientes, a menudo nos encontramos en el escenario de *varianzas desconocidas*. Esto nos obliga a realizar, de manera previa, un test de homogeneidad de varianzas para determinar si se puede asumir que son iguales o si, por el contrario, difieren. La razón es que, aunque se utiliza el estadístico *t* de Student para comparar las medias, su cálculo y la forma en que se distribuye varían en función de si las varianzas de los grupos son iguales o no.

Variable resultado cualitativa

Cuando la variable dependiente es cualitativa (categórica), como por ejemplo Enfermo (Sí/No) o Expuesto (Sí/No), la comparación de grupos no se basa en medias, sino en **proporciones**. Para ello, se utilizan pruebas estadísticas específicas según el número de grupos y el diseño del estudio.

A continuación, presentamos una guía análoga a la anterior:



ANOVA

El análisis de varianza (ANOVA) es una extensión del modelo lineal general que se utiliza para comparar las medias de una variable dependiente continua (Y) entre diferentes niveles de una variable explicativa categórica (X), que debe tener al menos tres niveles.

La hipótesis nula (H_0) del test estadístico establece que las medias de la variable dependiente son iguales en todos los grupos, mientras que la hipótesis alternativa (H_1) plantea que al menos dos medias difieren significativamente:

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_i$
- $H_1: \text{al menos una } \mu_i \neq \mu_j$

Por lo tanto, el ANOVA permite comparar múltiples medias, pero lo hace analizando la variabilidad entre y dentro de los grupos.

La variabilidad total se descompone en dos componentes:

- **Intervarianza (SSB):** Variabilidad entre los grupos.
- **Intravarianza (SSE):** Variabilidad dentro de los grupos.

El estadístico F del ANOVA, que sigue una distribución F de Fisher-Snedecor, compara estas dos fuentes de variabilidad:

$$F = \frac{SSB/(k-1)}{SSE/(n-k)}$$

donde:

- k : el número de grupos.
- n : número total de observaciones.

Si se cumple H_0 , el estadístico F tiende a 1, ya que las varianzas entre y dentro de los grupos son similares. Si las medias difieren significativamente, la intervarianza será mayor que la intravarianza, resultando en valores de F superiores a 1.

ANOVA de un factor o de una vía

El ANOVA de una vía puede considerarse como una extensión de los *t-test* independientes para comparar más de dos grupos de un factor. En este contexto, factor se refiere a la variable categórica que define los grupos.

Para que los resultados del ANOVA sean válidos, deben cumplirse los siguientes supuestos:

- **Aleatoriedad:** Las observaciones deben ser aleatorias.
- **Independencia:** Las observaciones entre grupos deben ser independientes.
- **Variable dependiente:** Cuantitativa continua.
- **Variable explicativa:** Categórica con más de dos niveles.
- **Normalidad:** La distribución de la variable dependiente dentro de cada grupo debe ser normal.
- **Homocedasticidad:** Las varianzas de los grupos deben ser homogéneas.

Funciones de R

Podemos realizar un test ANOVA de 1 factor o 1 vía mediante la función `aov()` de R `base`. Utiliza sintaxis fórmula del tipo `varcuantि ~ factor` y su resultado es un objeto lista (modelo “aov-lm”).

Siempre conviene que asignemos los resultados a objetos que luego nos servirán para aplicar otras funciones como `summary()` para obtener los resultados completos, incluido el p-valor.

También podemos usar la función `lm()` del paquete `stats` que ajusta modelos lineales y luego la función `anova()` sobre el objeto para obtener la significancia del test F.

Si los resultados del test F nos muestra un $p - valor$ menor a 0,05 el análisis sugiere que al menos dos grupos son diferentes entre sí.

Comparaciones múltiples

Una vez que comprobamos que existen diferencias significativas entre grupos, nos interesa saber cuáles grupos son diferentes entre sí. Para ello, existen distintos algoritmos de comparaciones múltiples con sus respectivas correcciones, como el test de *Diferencia Honestamente Significativa de Tukey*, también llamado Tukey HSD o **test de Tukey**. Esta prueba se aplica para grupos equilibrados (mismo tamaño) y varianzas similares (homocedásticas). Es una prueba conservadora, dado que mantiene bajo el error de tipo I, sacrificando la capacidad de detectar diferencias existentes.

Si las varianzas son homocedásticas pero los grupos difieren en tamaño, podemos usar el test de Tukey si tenemos que comparar entre varios grupos, o la **corrección de Bonferroni** para grupos más reducidos.

El paquete `emmeans` es una herramienta muy versátil y poderosa para realizar comparaciones múltiples. Comenzaremos utilizando la función `emmeans()` con el argumento `specs = a` la variable de grupo/factor para crear un objeto que contenga las medias marginales por grupo:

```
comp <- emmeans(anova, specs = "grupo")
```

Para realizar las comparaciones múltiples mediante test de Tukey usamos el comando `contrast()` con el argumento `method = "pairwise"`:

```
contrast(comp, method = "pairwise")
```