Laboratorio de Estimación y Contrastes de Hipótesis

Carmen Lancho - Isaac Martín - Víctor Aceña

Grado en Ciencia e Ingeniería de Datos - Inferencia Estadística - Curso 2024/2025

Índice

Objetivo

El objetivo principal de este laboratorio es aplicar técnicas de **estimación estadística** y realizar **contrastes de hipótesis** utilizando R. Al finalizar este laboratorio, deberías ser capaz de:

1. Calcular estimaciones puntuales e intervalos de confianza para medias y proporciones.
2. Realizar contrastes de hipótesis para medias y proporciones.
3. Interpretar los resultados estadísticos obtenidos.
4. Visualizar datos y resultados para apoyar tus conclusiones.

1. Introducción

En esta sección, introduciremos los conceptos básicos de **estimación** y **contraste de hipótesis**.

- **Estimación Estadística**: Proceso de inferir el valor de un parámetro poblacional a partir de datos muestrales.

- **Estimación Puntual**: Un único valor estimado del parámetro.
- **Estimación por Intervalo**: Un rango de valores dentro del cual se espera que se encuentre el parámetro con cierto nivel de confianza.

- **Contraste de Hipótesis**: Procedimiento para decidir si los datos muestrales proporcionan suficiente evidencia para rechazar una hipótesis sobre un parámetro poblacional.

2. Estimación Puntual y por Intervalos

Existen diferentes métodos de estimación que podemos utilizar dependiendo de la situación y la naturaleza de los datos. A continuación, se describen brevemente los métodos más comunes:

2.1. Estimación de la Media

El objetivo es estimar la media poblacional μ utilizando la media muestral $\bar{X}$. La **media muestral** es un estimador puntual de la media poblacional y se calcula como:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

2.1.0.1. Intervalo de Confianza para la Media

- **Cuando σ es conocido**:

$$IC = \bar{X} \pm Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right)$$

- **Cuando σ es desconocido** (más común):

$$IC = \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right)$$

Donde:

- $\bar{X}$: Media muestral.
- s: Desviación estándar muestral.
- n: Tamaño de la muestra.
- $Z_{\frac{\alpha}{2}}$: Valor crítico de la distribución normal.
- $t_{\frac{\alpha}{2}, n-1}$: Valor crítico de la distribución t de Student con $n-1$ grados de libertad.

Estimación de la Media de la Estatura de una Población

Ejemplo: Supongamos que deseamos conocer la estatura media de una población de adultos, pero no podemos medir a cada individuo en la población debido a limitaciones de tiempo y recursos. En su lugar, tomamos una muestra aleatoria de 30 personas y registramos sus estaturas en centímetros. Utilizaremos esta muestra para hacer una estimación puntual de la media poblacional y construir un intervalo de confianza que nos permita inferir, con cierto nivel de confianza, el rango en el que se encuentra la verdadera media de estatura en la población.

```r
set.seed(123)
estaturas <- rnorm(30, mean = 170, sd = 10)
```

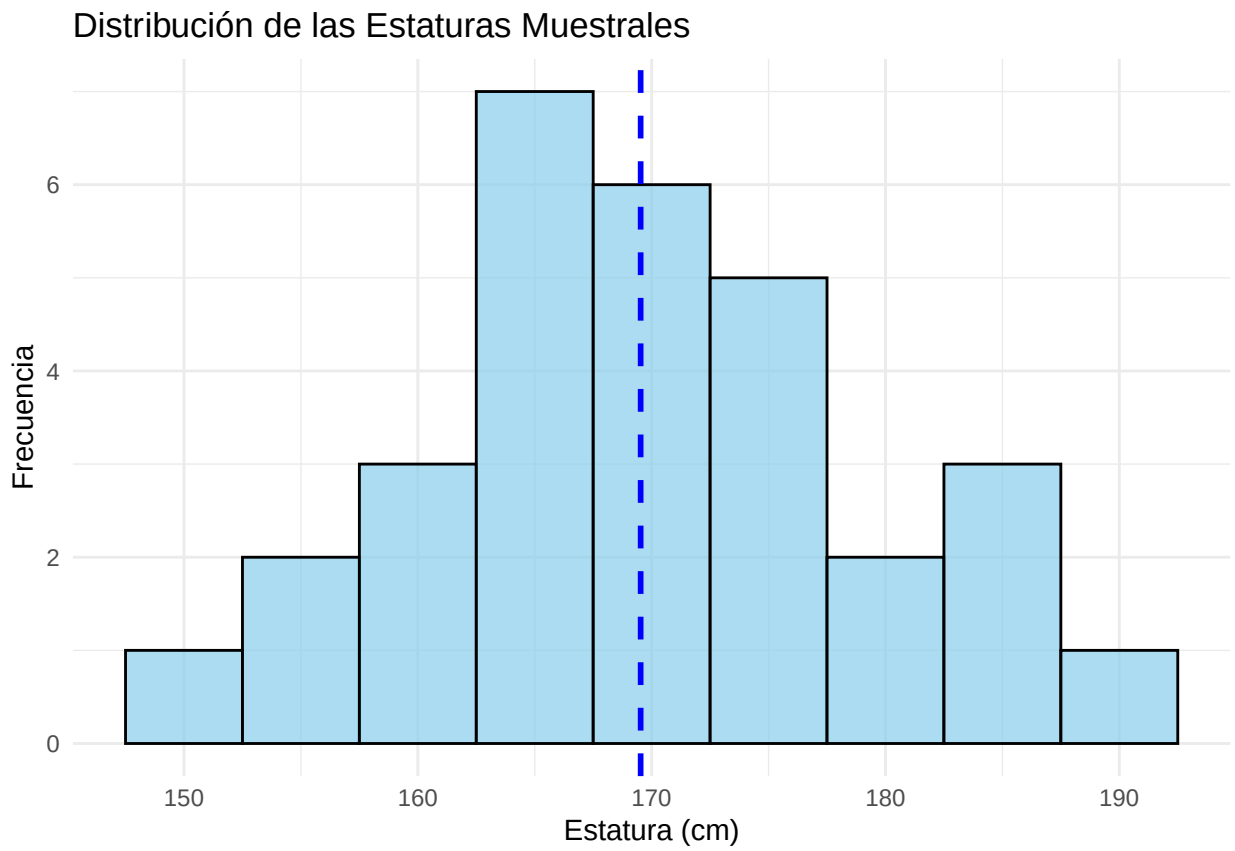
Primero, calculamos la media muestral como estimación puntual de la media poblacional. Este valor representa la mejor estimación del promedio de la población, basada en nuestra muestra.

```r
media_muestral <- mean(estaturas)
cat("La media muestral es:", round(media_muestral, 2), "cm")
```

```
## La media muestral es: 169.53 cm
```

A continuación, visualizamos la distribución de las estaturas muestrales para observar su forma y el estimador de la media que acabamos de calcular.

```r
ggplot(data.frame(estaturas), aes(x = estaturas)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = media_muestral), color = "blue", linetype = "dashed", size = 1) +
  labs(title = "Distribución de las Estaturas Muestrales",
       x = "Estatura (cm)",
       y = "Frecuencia") +
  theme_minimal()
```



Como no conocemos la desviación estándar de la población, utilizamos la distribución t de Student para construir un intervalo de confianza. Este intervalo nos proporciona un rango de valores dentro del cual esperamos que se ubique la media poblacional, con un nivel de confianza específico.

```r
# Definimos una función llamada `calcular_ic` que calcula el intervalo de confianza
# para un conjunto de datos (en este caso, las estaturas) y un nivel de confianza específico.
# Por defecto, usa un nivel de confianza del 95% (alpha = 0.05).
calcular_ic <- function(datos, alpha = 0.05) {

  # Tamaño de la muestra: `n` representa el número de datos en la muestra.
  # Esto es importante para calcular el error estándar.
  n <- length(datos)

  # Calculamos la media muestral de `datos`, que representa la estimación puntual
  # de la media poblacional.
  media <- mean(datos)

  # Calculamos la desviación estándar muestral `s`, que mide la dispersión de los datos.
  # Como no conocemos la desviación estándar poblacional, utilizamos la de la muestra.
  s <- sd(datos)

  # Calculamos el valor crítico de la t de Student `t_critico`.
  # Este valor depende del nivel de confianza (`alpha`) y de los grados de libertad (`df = n - 1`).
  # Nos indica cuántas desviaciones estándar debemos alejarnos de la media para cubrir
  # el intervalo de confianza deseado.
  t_critico <- qt(1 - alpha/2, df = n - 1)

  # Calculamos el error estándar de la media `error_estandar`, que nos dice
  # cuánto varía la media muestral en relación con la media poblacional.
  # Se calcula dividiendo la desviación estándar muestral `s` entre la raíz cuadrada del tamaño de la m
  error_estandar <- s / sqrt(n)

  # Calculamos el margen de error `ME`, que determina cuánto sumamos y restamos
  # a la media para crear el intervalo de confianza. Es el producto del valor crítico
  # y el error estándar.
  ME <- t_critico * error_estandar

  # Retornamos el intervalo de confianza, que va desde `media - ME` hasta `media + ME`.
  # Esto nos da un rango de valores dentro del cual, con un 95% de confianza, se encuentra la media pob
  c(media - ME, media + ME)
}

# Llamamos a la función `calcular_ic` con nuestros datos de estatura y obtenemos el intervalo de confia
IC <- calcular_ic(estaturas)

# Mostramos el resultado en pantalla, redondeando cada límite del intervalo a dos decimales.
cat("Intervalo de confianza al 95%: [", round(IC[1], 2), ",", round(IC[2], 2), "] cm")
```

```
## Intervalo de confianza al 95%: [ 165.87 , 173.19 ] cm
```

Interpretación

Con un 95 % de confianza, estimamos que la estatura media de la población se encuentra dentro del intervalo calculado. Esto significa que, si repitiéramos este procedimiento muchas veces con diferentes muestras, aproximadamente el 95 % de las veces el intervalo de confianza incluiría la verdadera media poblacional.

2.2. Cálculo del Tamaño de Muestra Necesario para una Estimación

Cuando queremos obtener un intervalo de confianza con un margen de error específico, es importante determinar de antemano cuántas observaciones necesitamos. Esto nos ayuda a optimizar la recolección de datos y asegurar que nuestra muestra sea representativa y precisa.

2.2.1. Teoría

Para calcular el tamaño de muestra necesario, utilizamos la fórmula:

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{ME} \right)^2$$

Donde:

- `n`: Tamaño de la muestra necesario.
- `Z_{\alpha/2}`: Valor crítico de la distribución normal para el nivel de confianza deseado.
- `\sigma`: Desviación estándar poblacional (o una estimación).
- `ME`: Margen de error que deseamos obtener en el intervalo de confianza.

> **Ejemplo**: Supongamos que queremos estimar la estatura media de una población de adultos con un intervalo de confianza del 95 % y un margen de error de 2 cm. Sabemos, por estudios previos, que la desviación estándar de estatura en esta población es de aproximadamente 10 cm.

Podemos definir una función para calcular el tamaño de muestra necesario dado un margen de error específico:

```
calcular_tamaño_muestra <- function(sigma, ME, alpha = 0.05) {
  # Valor crítico Z para el nivel de confianza deseado
  Z_critico <- qnorm(1 - alpha / 2)

  # Cálculo del tamaño de muestra necesario
  n <- (Z_critico * sigma / ME)^2

  # Redondeamos al número entero superior
  ceiling(n)
}

# Parámetros de ejemplo
sigma <- 10     # Desviación estándar de la población
ME <- 2         # Margen de error deseado

# Llamada a la función
tamaño_muestra <- calcular_tamaño_muestra(sigma, ME)
cat("Tamaño de muestra necesario:", tamaño_muestra)
```

```
## Tamaño de muestra necesario: 97
```

Interpretación

Con estos parámetros, necesitamos al menos 97 observaciones para asegurarnos de que nuestro intervalo de confianza tenga un margen de error de 2 cm con un nivel de confianza del 95 %. Esta información es útil para planificar la recolección de datos de forma efectiva y asegurar la precisión de nuestra estimación.

2.3. Estimación de una Proporción

A veces, estamos interesados en conocer la proporción de individuos en una población que posee cierta característica. En este caso, utilizaremos la **proporción muestral** como una estimación de la **proporción poblacional**.

La proporción muestral, denotada como $\hat{p}$, es un estimador puntual de la proporción poblacional p. Para calcular un **intervalo de confianza** para la proporción, utilizamos la siguiente fórmula:

$$IC = \hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Donde:

- $\hat{p}$: Proporción muestral, calculada como el número de éxitos (o individuos con la característica de interés) dividido por el tamaño de la muestra n.
- $Z_{\frac{\alpha}{2}}$: Valor crítico de la distribución normal para el nivel de confianza deseado.
- n: Tamaño de la muestra.

> **Ejemplo**: Supongamos que queremos estimar la proporción de personas en una ciudad que practica regularmente algún deporte. Tomamos una muestra de 200 personas y encontramos que 120 de ellas practican deporte con regularidad. Calcularemos la proporción muestral y construimos un intervalo de confianza del 95 % para esta proporción.

Empezaremos definiendo los datos de nuestro ejemplo y luego calcularemos la proporción muestral.

```r
# Definimos los datos iniciales del ejemplo
datos_deporte <- data.frame(
  muestra = 1:200,
  practica_deporte = c(rep(1, 120), rep(0, 80)) # 1 si practica deporte, 0 si no
)

# Cálculo de la proporción muestral
p_muestral <- mean(datos_deporte$practica_deporte)
cat("La proporción muestral es:", round(p_muestral, 3))
```

```
## La proporción muestral es: 0.6
```

A continuación, utilizaremos la proporción muestral calculada para construir un intervalo de confianza para la proporción poblacional, lo que nos permitirá estimar el rango en el que se encuentra la verdadera proporción de personas que practican deporte regularmente en la población.

2.3.1. Intervalo de Confianza para una Proporción

El intervalo de confianza para una proporción se calcula utilizando la fórmula:

$$IC = \hat{p} \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Donde:

- $\hat{p}$: Proporción muestral.
- $Z_{\frac{\alpha}{2}}$: Valor crítico de la distribución normal para el nivel de confianza deseado.
- n: Tamaño de la muestra.

En este ejemplo, utilizaremos un nivel de confianza del 95 %, por lo que el valor crítico $Z_{\frac{\alpha}{2}}$ será aproximadamente 1.96.

```
# Parámetros iniciales
p_muestral <- mean(datos_deporte$practica_deporte)  # Proporción muestral de personas que practican dep
n <- length(datos_deporte$practica_deporte)         # Tamaño de la muestra
alpha <- 0.05                                       # Nivel de significancia para un IC del 95%
z_critico <- qnorm(1 - alpha / 2)                   # Valor crítico Z para un IC del 95%

# Cálculo del error estándar
error_estandar <- sqrt((p_muestral * (1 - p_muestral)) / n)

# Cálculo del margen de error
ME <- z_critico * error_estandar

# Cálculo del intervalo de confianza
IC_inferior <- p_muestral - ME
IC_superior <- p_muestral + ME

# Mostrar el intervalo de confianza
cat("Intervalo de confianza al 95% para la proporción poblacional: [", round(IC_inferior, 3), ",", roun
```

```
## Intervalo de confianza al 95% para la proporción poblacional: [ 0.532 , 0.668 ]
```

Interpretación

Con un 95 % de confianza, estimamos que la proporción de personas en la población que practican deporte regularmente se encuentra dentro del intervalo calculado. Esto significa que, si tomáramos muchas muestras y construyéramos un intervalo de confianza para cada una, aproximadamente el 95 % de esos intervalos contendrían la verdadera proporción poblacional.

Este enfoque es particularmente útil en estudios donde queremos hacer inferencias sobre un comportamiento o característica en una población, como en este caso la práctica regular de deporte. A partir de este intervalo, podemos tener una idea más precisa y confiable del porcentaje de personas que practican deporte en la ciudad sin haber necesitado encuestar a toda la población.

2.4. Diferencia de Proporciones

En este ejemplo, queremos comparar la proporción de personas que practican deporte regularmente en dos ciudades: **Ciudad A** y **Ciudad B**. Tomamos muestras independientes en cada ciudad y calculamos la proporción de personas que practican deporte. Utilizaremos estas proporciones muestrales para construir un intervalo de confianza que nos permita inferir si existe una diferencia significativa entre las dos proporciones poblacionales.

Para construir un intervalo de confianza para la diferencia entre dos proporciones poblacionales, utilizamos la siguiente fórmula:

$$IC = (p_1 - p_2) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

donde:

- p_1 y p_2: Proporciones muestrales de personas que practican deporte en Ciudad A y Ciudad B, respectivamente.
- n_1 y n_2: Tamaños de las muestras en Ciudad A y Ciudad B, respectivamente.
- $Z_{\frac{\alpha}{2}}$: Valor crítico de la distribución normal para el nivel de confianza deseado (por ejemplo, 1.96 para un IC del 95 %).

Esta fórmula permite estimar el rango en el que se encuentra la diferencia real entre las dos proporciones poblacionales con un nivel de confianza especificado.

> **Ejemplo**: Supongamos que en una muestra de 300 personas en **Ciudad A**, 180 personas practican deporte regularmente, mientras que en una muestra de 250 personas en **Ciudad B**, 130 personas practican deporte regularmente. Queremos calcular el intervalo de confianza al 95 % para la diferencia de proporciones entre las dos ciudades.

```
# Datos del ejemplo
n1 <- 300          # Tamaño de la muestra en Ciudad A
exitos1 <- 180     # Número de personas que practican deporte en Ciudad A
n2 <- 250          # Tamaño de la muestra en Ciudad B
exitos2 <- 130     # Número de personas que practican deporte en Ciudad B

# Cálculo de las proporciones muestrales
p1_muestral <- exitos1 / n1
p2_muestral <- exitos2 / n2

# Mostrar las proporciones muestrales
cat("Proporción muestral en Ciudad A:", round(p1_muestral, 3), "\n")
```

```
## Proporción muestral en Ciudad A: 0.6
```

```
cat("Proporción muestral en Ciudad B:", round(p2_muestral, 3), "\n")
```

```
## Proporción muestral en Ciudad B: 0.52
```

2.4.1. Intervalo de Confianza para la Diferencia de Proporciones

En este ejemplo, queremos estimar la diferencia en la proporción de personas que practican deporte regularmente entre dos ciudades, Ciudad A y Ciudad B. Con los datos obtenidos de cada muestra, calcularemos un intervalo de confianza para esta diferencia de proporciones con un nivel de confianza del 95 %.

Para ello, utilizaremos la siguiente fórmula de intervalo de confianza para la diferencia de proporciones:

$$IC = (p_1 - p_2) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Donde:

- p_1 y p_2 son las proporciones muestrales de Ciudad A y Ciudad B, respectivamente.
- n_1 y n_2 son los tamaños de las muestras en Ciudad A y Ciudad B.
- $Z_{\frac{\alpha}{2}}$ es el valor crítico de la distribución normal para el nivel de confianza deseado.

```r
# Nivel de significancia y valor crítico Z para un IC del 95%
alpha <- 0.05
z_critico <- qnorm(1 - alpha / 2)

# Cálculo del error estándar para la diferencia de proporciones
error_estandar <- sqrt((p1_muestral * (1 - p1_muestral) / n1) + (p2_muestral * (1 - p2_muestral) / n2))

# Cálculo del margen de error
ME <- z_critico * error_estandar

# Cálculo del intervalo de confianza para la diferencia de proporciones
IC_inferior <- (p1_muestral - p2_muestral) - ME
IC_superior <- (p1_muestral - p2_muestral) + ME

# Mostrar el intervalo de confianza
cat("Intervalo de confianza al 95% para la diferencia de proporciones: [",
    round(IC_inferior, 3), ",", round(IC_superior, 3), "]")
```

```
## Intervalo de confianza al 95% para la diferencia de proporciones: [ -0.003 , 0.163 ]
```

Interpretación

El intervalo de confianza calculado para la diferencia de proporciones nos indica que, con un 95 % de confianza, la diferencia en la proporción de personas que practican deporte regularmente entre Ciudad A y Ciudad B se encuentra dentro del rango obtenido. Si el intervalo incluye el cero, esto sugiere que no hay una diferencia significativa entre las proporciones de las dos ciudades en cuanto a la práctica deportiva. Si el intervalo no contiene el cero, podemos inferir que existe una diferencia significativa en la proporción de personas que practican deporte regularmente entre ambas ciudades.

Este análisis es útil para comparar la prevalencia de una característica o comportamiento entre dos grupos o poblaciones y es ampliamente aplicable en estudios de encuestas, marketing, salud pública, y otros campos de investigación.

2.5. Contraste de Igualdad de Medias

Cuando queremos comparar las medias de dos grupos diferentes para ver si existe una diferencia significativa entre ellos, utilizamos un **contraste de igualdad de medias**. Este contraste de hipótesis nos permite evaluar si la diferencia observada entre las medias muestrales es suficientemente grande como para inferir que existe una diferencia real en las medias poblacionales.

En un contraste de igualdad de medias, las hipótesis se plantean de la siguiente manera:

- **Hipótesis Nula (H_0)**: No hay diferencia en las medias poblacionales, es decir, $\mu_1 = \mu_2$.
- **Hipótesis Alternativa (H_1)**: Existe una diferencia en las medias poblacionales, es decir, $\mu_1 \neq \mu_2$.

Para realizar este contraste, utilizamos el **test t de Student** para dos muestras independientes. Los resultados del contraste incluyen:

- **p-valor**: Nos indica si la diferencia es estadísticamente significativa. Un p-valor bajo (típicamente menor que 0.05) sugiere que podemos rechazar la hipótesis nula y concluir que existe una diferencia significativa entre las medias.

- **Intervalo de Confianza (IC)**: Nos permite observar si el intervalo incluye el valor cero. Si el intervalo de confianza de la diferencia de medias no contiene el cero, esto refuerza la conclusión de que hay una diferencia significativa entre las medias de los grupos.

Un **intervalo de confianza que no incluye el cero** y un **p-valor menor que 0.05** sugieren una diferencia significativa entre las medias de los dos grupos.

> **Ejemplo**: Supongamos que estamos interesados en comparar los niveles de pH de agua de dos fuentes diferentes (orígenes) en un laboratorio. Para este propósito, utilizamos una muestra de 100 mediciones de pH para cada origen y aplicamos un contraste de igualdad de medias para determinar si existe una diferencia significativa en el nivel de pH entre ambas fuentes.

En primer lugar, cargamos y preparamos los datos:

```
# Cargamos y preparamos los datos del laboratorio
set.seed(1)
library(tidyverse)
library(readxl)

lab <- read_excel("../../data/lab.xlsx") |>
  mutate(fecha = as.Date(fecha)) |>
  slice_sample(n = 100)
```

```
## Error: ‘path‘ does not exist: ’../../data/lab.xlsx’
```

A continuación, observamos un resumen descriptivo de las mediciones de pH por grupo de origen. Esta descripción nos permite tener una visión general de los datos y prepararnos para la interpretación de los resultados en el contraste de hipótesis.

```
# Generar un resumen estadístico de la variable ph según el grupo de origen
with(lab, stby(ph, origen, descr, stats = "common", transpose = TRUE))
```

```
## Error in eval(expr, envir, enclos): objeto ’lab’ no encontrado
```

Para confirmar si existe una diferencia significativa en los niveles de pH entre los grupos, realizaremos un **contraste de hipótesis para la igualdad de medias**. Este contraste es útil cuando queremos determinar si la media de una variable continua difiere significativamente entre dos grupos.

2.5.1. Hipótesis del contraste

En este caso, formulamos las siguientes hipótesis:

- **Hipótesis nula** (H_0): La media de pH es igual en ambos grupos de origen, es decir, $\mu_1 = \mu_2$.
- **Hipótesis alternativa** (H_a): La media de pH es diferente entre los grupos de origen, es decir, $\mu_1 \neq \mu_2$.

2.5.2. Procedimiento del contraste

Utilizaremos el **t-test** de Student para comparar las medias entre ambos grupos. Este método es adecuado cuando la variable sigue una distribución normal y las varianzas de ambos grupos son iguales (supuesto de homogeneidad de varianzas).

Implementación

En este análisis, utilizamos la función `t.test()` de R para realizar un **contraste de hipótesis de igualdad de medias** entre los grupos "Norte" y "Sur". A continuación, desglosamos su propósito y uso:

- `ph ~ origen` define la variable dependiente (`ph`, que representa el nivel de pH) y la variable independiente (`origen`, que indica el grupo, "Norte" o "Sur"). Esto indica a R que compare las medias de `ph` entre los dos grupos definidos en `origen`.
- `data = lab` indica que los datos de pH y de origen se encuentran en el conjunto de datos `lab`.
- `var.equal = TRUE` asume que las varianzas de ambos grupos son iguales, lo cual permite realizar un test t de Student para muestras independientes, adecuado cuando los grupos tienen varianzas similares.

El uso de `t.test()` es fundamental para **determinar si existe una diferencia significativa entre las medias** de dos grupos en una variable cuantitativa (en este caso, el nivel de pH entre dos orígenes de agua).

Realizaremos este contraste con un nivel de significancia del 5 % ($\alpha = 0{,}05$), y evaluaremos el **p-valor** y el **intervalo de confianza** para interpretar los resultados.

```
# Realizamos el t-test para comparar las medias de pH entre los grupos de origen
t.test(ph ~ origen, data = lab, var.equal = TRUE)
```

```
## Error in eval(m$data, parent.frame()): objeto 'lab' no encontrado
```

Interpretación de los Resultados

Tras ejecutar el test de hipótesis, interpretamos los resultados obtenidos:

1. **Valor p**: El valor p reportado es **3.862e-06**, que es mucho menor al nivel de significancia comúnmente utilizado (0.05). Esto sugiere que existe suficiente evidencia para rechazar la hipótesis nula de que las medias de los grupos "Norte" y "Sur" son iguales. En otras palabras, hay una diferencia estadísticamente significativa en los niveles de pH entre los dos grupos de origen.
2. **Intervalo de Confianza (IC)**: El intervalo de confianza al 95 % para la diferencia en las medias de pH entre el grupo "Norte" y el grupo "Sur" es **[0.0341, 0.0805]**. Dado que este intervalo no incluye el valor cero, esto respalda la conclusión de que existe una diferencia significativa entre las medias de los dos grupos. La diferencia de medias es positiva, lo que indica que, en promedio, el pH del grupo "Norte" es ligeramente mayor que el del grupo "Sur".
3. **Estimaciones de las Medias**: La media de pH en el grupo "Norte" es **6.6614** y en el grupo "Sur" es **6.6041**. Esto indica que, aunque la diferencia es pequeña, el pH promedio en el grupo "Norte" es mayor.

Con un 95 % de confianza, podemos afirmar que existe una diferencia significativa en el nivel de pH entre el grupo "Norte" y el grupo "Sur". Esta diferencia, aunque pequeña, es estadísticamente significativa, y el intervalo de confianza nos indica que la verdadera diferencia en las medias de los niveles de pH entre los dos grupos probablemente se encuentra entre 0.034 y 0.080 unidades de pH.

Para visualizar esta diferencia y entender mejor la distribución de los datos entre los grupos de origen, generaremos un gráfico comparativo utilizando la librería `ggstatsplot`, que mostrará las medias de cada grupo y sus respectivas distribuciones.

```r
# Gráfico de comparación de medias entre los grupos de origen
lab |> ggbetweenstats(x = origen, y = ph)
```

```
## Error in eval(expr, envir, enclos): objeto 'lab' no encontrado
```

Este gráfico visualiza las diferencias entre los grupos en términos de sus distribuciones y medias, permitiendo una interpretación más intuitiva de la posible diferencia en pH entre los orígenes.

Interpretación del Gráfico

El gráfico muestra visualmente si hay una superposición significativa entre las distribuciones de los grupos, junto con las medias y los intervalos de confianza. Si las distribuciones no se superponen o hay poca superposición, y las medias son claramente distintas, esto refuerza la conclusión del test de hipótesis de que existe una diferencia significativa entre los grupos en el pH medido.

2.6. Potencia Estadística

La **potencia estadística** es la probabilidad de rechazar correctamente la hipótesis nula (H_0) cuando es falsa. Matemáticamente, se define como:

$$\text{Potencia} = P(\text{Rechazar } H_0 \mid H_0 \text{ es falsa})$$

Una potencia alta (generalmente considerada como 0.8 o más) significa que el test es sensible para detectar diferencias reales. El cálculo de la potencia es fundamental en el diseño experimental, ya que ayuda a determinar el tamaño de muestra necesario y a minimizar los errores de Tipo II (β), que ocurren cuando se falla en rechazar la hipótesis nula cuando en realidad es falsa.

2.6.1. Factores que Afectan la Potencia del Test

La potencia de un test depende de varios factores, que se pueden expresar de la siguiente manera:

1. **Nivel de significancia (α)**:
   - Es la probabilidad de cometer un error de Tipo I, que se define como la probabilidad de rechazar la hipótesis nula cuando es verdadera. En términos matemáticos:

$$\alpha = P(\text{Rechazar } H_0 \mid H_0 \text{ es verdadera})$$

   - Un nivel de significancia más alto (como 0.10 en lugar de 0.05) incrementa la potencia, ya que aumenta la probabilidad de rechazar H_0. Sin embargo, esto también aumenta el riesgo de cometer un error Tipo I.
2. **Tamaño del efecto (d)**:
   - Representa la magnitud de la diferencia que queremos detectar, calculada como la diferencia entre las medias poblacionales en unidades de desviación estándar:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

   - Un tamaño del efecto más grande resulta en una mayor potencia, ya que es más fácil detectar diferencias significativas cuando las diferencias entre los grupos son grandes.
3. **Tamaño de la muestra (n)**:

- El tamaño de la muestra se refiere al número de observaciones en el estudio. A medida que aumentamos n, la potencia del test también aumenta. Esto se debe a que muestras más grandes reducen la variabilidad de la estimación de la media y proporcionan una estimación más precisa de la diferencia entre las poblaciones:

$$\text{Error estándar} = \frac{\sigma}{\sqrt{n}}$$

- Con un error estándar menor, es más probable que una diferencia observada sea considerada significativa.

4. **Variabilidad de los datos (σ)**:
   - Representa la desviación estándar de la población. Una mayor variabilidad en los datos reduce la potencia, ya que las diferencias son más difíciles de detectar en datos más dispersos. La relación se puede expresar como:

$$\text{Potencia} \propto \frac{1}{\sigma}$$

   - Esto significa que si aumentamos la variabilidad de los datos (un σ más grande), la potencia disminuye, haciendo más difícil detectar diferencias significativas.

La relación entre estos factores y la potencia estadística es esencial para diseñar experimentos eficaces. Es crucial encontrar un equilibrio entre el tamaño de la muestra, el nivel de significancia, el tamaño del efecto y la variabilidad de los datos para asegurar que el estudio tenga una potencia adecuada para detectar diferencias significativas, minimizando así el riesgo de cometer errores Tipo II.

Ejemplo:Vamos a utilizar el paquete `pwr` para calcular la potencia de un test t para dos muestras independientes. Supongamos que queremos detectar una diferencia en la media de dos grupos con un tamaño del efecto medio.

1. Definir los Parámetros del Estudio

- **Nivel de significancia (α)**: 0.05
- **Tamaño del efecto (d)**: Utilizamos el estadístico de Cohen para el tamaño del efecto.
  - Pequeño: 0.2
  - Medio: 0.5
  - Grande: 0.8
- **Tamaño de la muestra (n)**: Número de observaciones en cada grupo.

2. Calcular la Potencia

Supongamos que tenemos 50 observaciones en cada grupo y queremos detectar un tamaño del efecto medio.

```r
# Parámetros
n <- 50            # Tamaño de muestra por grupo
d <- 0.5           # Tamaño del efecto medio
alpha <- 0.05      # Nivel de significancia

# Cálculo de la potencia
resultado_potencia <- pwr.t.test(n = n, d = d, sig.level = alpha,
                                 type = "two.sample", alternative = "two.sided")

# Mostrar los resultados
print(resultado_potencia)
```

```
##
##      Two-sample t test power calculation
##
##               n = 50
##               d = 0.5
##       sig.level = 0.05
##           power = 0.6968934
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Interpretación

La potencia del test realizado es de aproximadamente **69.7 %**, lo que indica una **probabilidad moderada** de detectar una diferencia real si esta existe. Para aumentar la confiabilidad de los resultados y reducir la posibilidad de no detectar diferencias significativas, se recomienda aumentar el tamaño de la muestra.

> **Ejemplo**: Ahora, supongamos que queremos determinar el tamaño de muestra necesario para alcanzar una potencia del 80 %.

```r
# Parámetros
d <- 0.5           # Tamaño del efecto medio
power <- 0.8       # Potencia deseada
alpha <- 0.05      # Nivel de significancia

# Cálculo del tamaño de muestra
resultado_tamano <- pwr.t.test(power = power, d = d, sig.level = alpha,
                               type = "two.sample", alternative = "two.sided")

# Mostrar los resultados
print(resultado_tamano)
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Interpretación

- **n = 63.76561**: Se requieren aproximadamente **64 observaciones por grupo** para alcanzar una potencia del **80 %** con un tamaño de efecto medio.
- **d = 0.5**: Tamaño del efecto medio.
- **sig.level = 0.05**: Nivel de significancia del **5 %**.
- **power = 0.8**: Potencia del **80 %**.
- **alternative = two.sided**: Contraste bilateral.

En este caso, determinar que se necesitan aproximadamente **64 observaciones por grupo** garantiza que el test t de Student tendrá una potencia del **80 %** para detectar una diferencia media con un nivel de significancia del **5 %**. Esto asegura que el estudio es suficientemente sensible para identificar diferencias reales, minimizando la probabilidad de errores Tipo II y aumentando la confiabilidad de los resultados obtenidos.

2.7. Medidas del Tamaño del Efecto

Además de la significancia estadística proporcionada por el p-valor, es importante evaluar la **magnitud del efecto** para entender la relevancia práctica de los resultados. El **d de Cohen** es una medida del tamaño del efecto que expresa la diferencia entre dos medias en términos de desviaciones estándar, permitiendo comparar la magnitud de la diferencia independientemente de las unidades de medida.

2.7.1. Cálculo del d de Cohen

El **d de Cohen** se calcula utilizando la siguiente fórmula:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_{\text{pooled}}}$$

Donde:

- $\bar{X}_1$ y $\bar{X}_2$ son las medias muestrales de los grupos 1 y 2, respectivamente.
- s_{pooled} es la desviación estándar combinada (desviación estándar agrupada), calculada como:

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Donde:

- s_1^2 y s_2^2 son las varianzas muestrales de los grupos 1 y 2, respectivamente.
- n_1 y n_2 son los tamaños de muestra de los grupos 1 y 2.

Esta fórmula combina las desviaciones estándar de ambos grupos, asumiendo que las varianzas son homogéneas.

2.7.2. Interpretación del d de Cohen

Los valores del **d de Cohen** se interpretan generalmente de la siguiente manera:

- **Pequeño**: $d = 0{,}2$
- **Medio**: $d = 0{,}5$
- **Grande**: $d = 0{,}8$ o mayor

Un valor mayor de d indica una diferencia más grande entre los grupos en términos de desviaciones estándar.

2.7.3. Implementación en R

Para calcular el **d de Cohen** en R, utilizamos el paquete `effsize`.

- `library(effsize)`:
  - **Descripción**: Carga el paquete `effsize` para poder utilizar la función `cohen.d()`.
- `cohen.d(ph ~ origen, data = lab, pooled = TRUE, hedges.correction = FALSE)`: Calcula el **d de Cohen** para comparar las medias de `ph` entre los niveles de `origen` en el conjunto de datos `lab`.
  - **Argumentos**:
    - `ph ~ origen`: Fórmula que indica que queremos comparar la variable `ph` entre los niveles de `origen`.
    - `data = lab`: Especifica el conjunto de datos que contiene las variables.
    - `pooled = TRUE`: Utiliza la desviación estándar combinada de ambos grupos, asumiendo homogeneidad de varianzas.
    - `hedges.correction = FALSE`: Indica que no se aplicará la corrección de Hedges; si se desea una estimación más precisa en muestras pequeñas, se puede establecer como `TRUE`.

Ejemplo: Vamos a calcular el tamaño del efecto utilizando el **d de Cohen** para comparar los niveles de pH entre los grupos "Norte" y "Sur" en nuestro conjunto de datos `lab`. Esto nos permitirá cuantificar la magnitud de la diferencia observada entre los dos grupos y entender su relevancia práctica.

```
# Calcular el d de Cohen para la variable ph entre los grupos de origen
cohen_d <- cohen.d(ph ~ origen, data = lab, pooled = TRUE, hedges.correction = FALSE)
```

```
## Error in eval(expr, envir, enclos): objeto 'lab' no encontrado
```

```
print(cohen_d)
```

```
## Error in eval(expr, envir, enclos): objeto 'cohen_d' no encontrado
```

Interpretación

- **Valor del d de Cohen (`d estimate`):** El valor obtenido es **1.078852**, lo que se considera un **tamaño del efecto grande** según las convenciones establecidas. Este valor indica que la diferencia entre las medias de los grupos "Norte" y "Sur" es de aproximadamente **1.08 desviaciones estándar**, lo que representa una diferencia sustancial entre los grupos.
- **Intervalo de Confianza del 95 %:** Con un 95 % de confianza, el verdadero tamaño del efecto en la población se encuentra entre **0.616** y **1.542**. Dado que todo el intervalo está por encima de 0, esto refuerza la conclusión de que existe una diferencia significativa y positiva entre los grupos.

Relevancia Práctica:

- Un **tamaño del efecto grande** sugiere que la diferencia observada no solo es estadísticamente significativa, sino también relevante en la práctica.
- Esto implica que el origen del agua ("Norte" vs. "Sur") tiene un impacto considerable en los niveles de pH medidos.

Conclusión:

El **d de Cohen** de **1.078852** indica una diferencia grande y significativa entre los grupos "Norte" y "Sur" en términos de pH. Este hallazgo tiene implicaciones prácticas importantes, ya que sugiere que el origen del agua afecta significativamente su nivel de pH. Los resultados respaldan la necesidad de considerar el origen del agua en análisis futuros y posibles intervenciones o regulaciones relacionadas con la calidad del agua.

3. Conclusiones

A lo largo de este laboratorio, hemos aplicado técnicas de **estimación estadística** y **contraste de hipótesis** para analizar diferencias en **medias** y **proporciones**. En particular, se destacan los siguientes puntos:

- **Visualización de Datos**:
  - Utilizamos herramientas gráficas para **visualizar las distribuciones** de los datos y las diferencias entre grupos, lo que facilita una interpretación más intuitiva y efectiva de los resultados.
  - Las visualizaciones permiten identificar patrones, tendencias y posibles anomalías en los datos que podrían influir en el análisis estadístico.
- **Significancia Estadística**:
  - Aprendimos a determinar si las diferencias observadas entre grupos son estadísticamente significativas utilizando **p-valores** e **intervalos de confianza**.
  - La interpretación correcta de estos resultados nos permite evaluar la evidencia contra la hipótesis nula de manera objetiva.
- **Verificación de Supuestos Estadísticos**:
  - Exploramos la importancia de verificar los **supuestos necesarios** para la aplicación de pruebas paramétricas, como la **normalidad de los datos** y la **homogeneidad de varianzas**.
  - La correcta verificación de estos supuestos asegura la validez y confiabilidad de los análisis realizados.
- **Cálculo de la Potencia Estadística**:
  - Entendimos cómo la **potencia estadística** influye en la capacidad de un test para detectar diferencias reales cuando existen.
  - Aprendimos a calcular el **tamaño de muestra necesario** para alcanzar una potencia deseada, optimizando así el diseño experimental y reduciendo la probabilidad de errores de Tipo II.
- **Tamaño del Efecto**:
  - Incorporamos el cálculo y la interpretación del **d de Cohen**, una medida que cuantifica la magnitud de la diferencia entre grupos en términos de desviaciones estándar.
  - Esta medida complementa la significancia estadística al proporcionar información sobre la relevancia práctica de los resultados obtenidos.

Nota: Algunos de los ejemplos de este laboratorio utilizan datos del libro *Introducción al software estadístico R* de López Cano E. (2024), disponible en https://www.lcano.com/b/iser/_book/.