



# 01. Estadística Inferencial

En el encuentro anterior, vieron que los **objetivos** principales del **análisis exploratorio de datos (EDA)** son:

- Sugerir hipótesis acerca de las causas de los fenómenos observados
- Guiar en la selección de herramientas y técnicas estadísticas apropiadas
- Evaluar los supuestos modelos en los que se basa el análisis estadístico
- Proporcionar una base para una recolección de datos futura

Esta exploración implica una mezcla de métodos analíticos y visuales de análisis. Otros métodos estadísticos se utilizan a veces para complementar EDA, pero su principal objetivo es facilitar la comprensión antes de sumergirse en el modelado estadístico formal.

Incluso si pensamos que ya sabemos qué tipo de análisis necesitamos llevar a cabo, siempre es una buena idea **explorar el conjunto de datos antes de sumergirse en el análisis**.

En este encuentro hablaremos de estadística inferencial.

La **inferencia estadística** es el conjunto de métodos que permiten inducir, a través de una muestra estadística, el comportamiento de una determinada población. Estudia cómo, a través de la aplicación de dichos métodos sobre los datos de una muestra, se pueden **extraer conclusiones** sobre los parámetros de la población de datos. De la misma manera estudia también el grado de fiabilidad de los resultados extraídos del estudio.

Entonces, la inferencia estadística, nunca nos ofrecerá seguridad absoluta, sí nos ofrecerá una respuesta probabilística. Esto es importante: la estadística no decide; sólo ofrece elementos para que el investigador o el lector decidan.

La estimación estadística tiene por objetivo aproximar o predecir el valor de un parámetro de una población a partir de datos proporcionados por una muestra (estadísticos). Por ejemplo, la media muestral  $\bar{x}$  (estadístico) es usada para estimar la media poblacional  $\mu$  (parámetro). Esta estimación la podemos realizar de manera puntual o por intervalos.

**Repasemos tres conceptos:**

**Inferencia:** Inferir significa, extraer juicios o conclusiones a partir de ciertos supuestos, sean estos generales o particulares.

**Población:** Una población de datos, es el conjunto total de datos que existen sobre un variable.

**Muestra estadística:** Una muestra es una parte de la población de datos.

Normalmente, en estadística se trabaja con muestras debido a la gran cantidad de datos que tiene una población. Por ejemplo, si queremos sacar conclusiones sobre las elecciones presidenciales, esto es, inferir los resultados de las elecciones, es imposible encuestar a toda la población del país. Para resolver este problema se escoge una muestra variada y representativa, gracias a la cual se puedan extraer una estimación del resultado final.

La elección de la muestra adecuada depende de la técnica de muestreo. Métodos y técnicas de la inferencia estadística:

**Métodos de estimación de parámetros:** Se encarga de asignar un valor al parámetro o al conjunto de parámetros que caracterizan el campo sujeto a estudio. Claro que al ser una estimación existe cierto error. Para obtener estimaciones adaptadas a esa realidad, se crean intervalos de confianza. Para hacer la estimación de parámetros descriptivos podemos hacer un boxplot.

### / Estimación puntual

Una forma natural de estimar muchos parámetros poblacionales consiste en utilizar el estadístico muestral correspondiente. Así, la media muestral ( $\bar{x}$ ) es un estimador puntual de la media poblacional ( $\mu$ ) y la proporción de casos de una enfermedad en la muestra es un estimador puntual de la probabilidad de tener la enfermedad en la población. No obstante, para un determinado parámetro poblacional, pueden contemplarse distintos estimadores alternativos. Algunos estimadores de la media poblacional distintos de la media muestral podrían ser, por ejemplo, la mediana, la media del 50% central de la muestra o la media de los valores máximo y mínimo

### / Estimación por intervalo

Como ya comentamos, las estimaciones puntuales obtenidas a partir de una muestra diferirán del parámetro poblacional y, en consecuencia, quedará un margen de incertidumbre que se expresa en términos del **error estándar del estimador**. Así, resulta natural la pretensión de disponer de una medida del parámetro poblacional que incorpore tanto la estimación puntual como su error estándar. Esta medida es el **intervalo de confianza**, que facilita un rango de valores dentro del cual se encontrará el verdadero valor del parámetro poblacional con un cierto grado de confianza, es decir, el rango de valores en que se encuentra el parámetro, cuyos límites confidenciales son los valores LI (Límite inferior) y LS (Límite superior).

Se denomina **nivel de confianza** ( $1 - \alpha$ ) a la probabilidad de que el intervalo de confianza incluya el verdadero valor del parámetro, es decir, que refleje la confianza en la construcción del intervalo.

Cuando se trabaja con datos se busca intentar alcanzar una decisión, para lo cual es útil hacer hipótesis (o conjeturas) sobre la población aplicada.



Pero... ¿Qué es una hipótesis?

## 02. ¿Qué es una hipótesis?

### Hipótesis estadística:

Son, en general, enunciados acerca de las distribuciones de probabilidad de las poblaciones. Tales hipótesis, que pueden ser o no ciertas.

Las hipótesis estadísticas se pueden contrastar con la información extraída de las muestras y tanto si se aceptan como si se rechazan se puede cometer un error.

El procedimiento de toma de decisión sobre la hipótesis se conoce como **prueba o test de hipótesis**. Su objetivo es comprobar si una estimación corresponde con los valores poblacionales.

Como se emplean distribuciones de probabilidad para representar poblaciones, también podemos decir que una hipótesis estadística es una proposición sobre la distribución de probabilidad de una variable aleatoria, donde la hipótesis involucra a uno más parámetros de esta distribución.

Por lo tanto, una Prueba de Hipótesis es un procedimiento basado en evidencia muestral (estadístico) y en la teoría de probabilidad (distribución muestral del estadístico) para determinar si una hipótesis es razonable y no debe rechazarse, o si es irrazonable y debe ser rechazada.

En todo contraste de hipótesis existen dos supuestos. La **hipótesis nula ( $H_0$ )** que recoge la idea de que un valor tiene un valor predeterminado.

Por ejemplo, si queremos decidir si una moneda está trucada, formulamos la hipótesis de que la moneda es buena (o sea  $p = 0,5$ , donde  $p$  es la probabilidad de al arrojarla salga cara).

Analógicamente, si deseamos decidir si un procedimiento es mejor que otro, formulamos la hipótesis de que no hay diferencia entre ellos (o sea. que cualquier diferencia observada se debe simplemente a fluctuaciones en el muestreo de la misma población).

Para todo tipo de investigación en la que tenemos dos o más grupos, se establecerá una hipótesis nula, la cual es aquella que nos dice que no existen diferencias significativas entre los grupos.

Una hipótesis nula es importante por varias razones:

Es una hipótesis que se acepta o se rechaza según el resultado de la investigación.



El hecho de contar con una hipótesis nula ayuda a determinar si existe una diferencia entre los grupos, si esta diferencia es significativa, y si no se debió al azar.

No toda investigación precisa de formular hipótesis nula. Recordemos que la hipótesis nula es aquella por la cual indicamos que la información a obtener es contraria a la hipótesis de trabajo.

Toda hipótesis que difiere de una dada se llamará una **hipótesis alternativa (H1)**.

Por ejemplo: si una hipótesis es  $p = 0,5$ , hipótesis alternativa podrían ser  $p = 0,7$ ,  $p = 0,5$  ó  $p > 0,5$ . Si se rechaza la hipótesis nula ( $H_0$ ), entonces se acepta la hipótesis alternativa ( $H_1$ ).

### Pasos para realizar un contraste de hipótesis:

1. Establecer la hipótesis nula en términos de igualdad

$$H_0 : \mu = X$$

2. Establecer la hipótesis alternativa, que puede hacerse de tres maneras, dependiendo del interés del investigador

$$H_1 : \mu \neq X$$

$$H_1 : \mu > X$$

$$H_1 : \mu < X$$

En el primer caso se habla de contraste bilateral o de dos colas, y en los otros dos de lateral (derecho en el 2º caso, o izquierdo en el 3º) o una cola. ¿Qué significa esto?

Existen **zonas críticas** o también llamadas zonas de rechazo de la hipótesis nula.

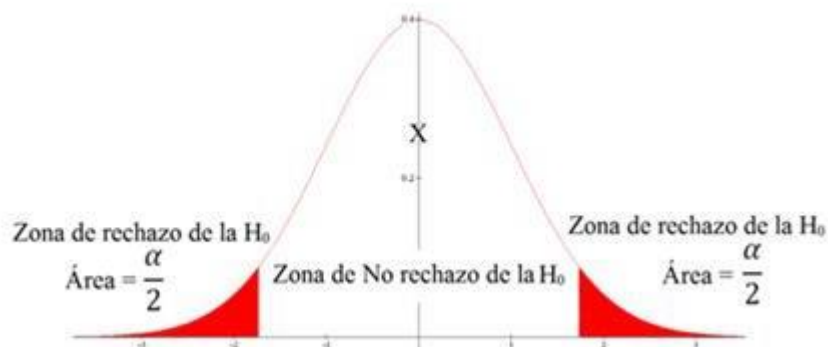
**Región de rechazo** es la región formada por los valores que se alejan del valor aceptado bajo la  $H_0$ .

**Región de aceptación** es la región formada por los valores que no se alejan tanto del valor aceptado bajo la  $H_0$ .

Un **valor crítico** son valores del estadístico de contraste que delimitan la región del rechazo.



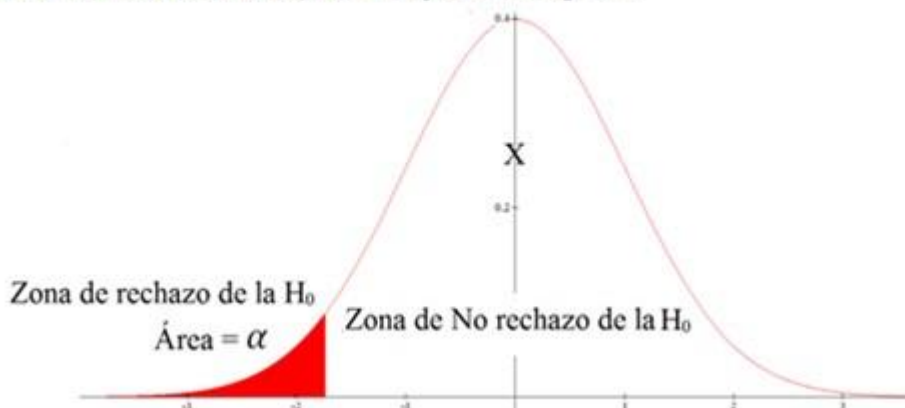
1) Prueba Bilateral o a dos colas:  $H_0: \mu = X; H_1: \mu \neq X$



2) Prueba Unilateral con cola hacia la derecha:  $H_0: \mu \leq X; H_1: \mu > X$



3) Prueba Unilateral con cola hacia la izquierda:  $H_0: \mu \geq X; H_1: \mu < X$



3. Elegir un nivel de significación: nivel crítico para. El **nivel de significación** es la probabilidad de que una muestra genere el valor del estadístico de contraste que esté dentro de la región de rechazo.



4. Elegir un estadístico de contraste: estadístico cuya distribución muestral se conozca en  $H_0$  y que esté relacionado con el parámetro, y establecer, en base a dicha distribución, la **región crítica**: región en la que el estadístico tiene una probabilidad menor que  $\alpha$  si  $H_0$  fuera cierta y, en consecuencia, si el estadístico cayera en la misma, se rechazaría  $H_0$ .

Obsérvese que, de esta manera, se está más seguro cuando se rechaza una hipótesis que cuando no. Por eso se fija como  $H_0$  lo que se quiere rechazar. Cuando no se rechaza, no se ha demostrado nada, simplemente no se ha podido rechazar. Por otro lado, la decisión se toma en base a la distribución muestral en  $H_0$ , por eso es necesario que tenga la igualdad.

5. Calcular el estadístico para una muestra aleatoria y compararlo con la región crítica, o equivalentemente, calcular el **"valor p"** del estadístico (probabilidad de obtener ese valor, u otro más alejado de la  $H_0$ , si  $H_0$  fuera cierta) y compararlo con  $\alpha$ .

Ninguna prueba de hipótesis es 100% cierta. Dado que la prueba se basa en probabilidades, siempre existe la posibilidad de llegar a una conclusión incorrecta.

Cuando realizamos una prueba de hipótesis, podemos cometer dos **tipos de error**: tipo I y tipo II.

Los riesgos de estos dos errores están inversamente relacionados y se determinan según el nivel de significancia y la potencia de la prueba. Por lo tanto, debemos determinar qué error tiene consecuencias más graves para la situación antes de definir los riesgos.

**Error de tipo I** se define como el rechazo de la hipótesis nula cuando esta es, en realidad, cierta. Se le conoce también como **falso positivo o error de tipo alfa**, siendo  $\alpha$  la probabilidad de cometer un error de tipo I, que es el nivel de significancia que establecemos para nuestra prueba de hipótesis.

Al valor  $1 - \alpha$  se lo denomina **nivel de confianza**, es decir, probabilidad de que una muestra genere el valor del estadístico de contraste que esté fuera de la región de rechazo (dentro de la región de aceptación).

Un  $\alpha$  de 0.05 indica que estamos dispuestos a aceptar una probabilidad de 5% de estar equivocado al rechazar la hipótesis nula.

**Error de tipo II** se produce cuando la hipótesis nula es falsa y no la rechazamos. La probabilidad de cometer un error de tipo II es  $\beta$ , que depende de la potencia de la prueba.

La probabilidad de rechazar la hipótesis nula cuando es falsa es igual a  $1 - \beta$ . Este valor es la **potencia de la prueba**.



Decisión sobre $H_0$	Se acepta $H_0$	Se rechaza $H_0$
$H_0$ es verdadera	Decisión correcta Probabilidad = $1 - \alpha$ $1 - \alpha$ es el grado de confianza	Error Tipo I Probabilidad = $\alpha$ $\alpha$ es el nivel de significancia
$H_0$ es falsa	Error Tipo II Probabilidad = $\beta$	Decisión correcta Probabilidad = $1 - \beta$ $1 - \beta$ es la potencia

Por ejemplo, se realiza un test para saber si los sujetos de una muestra padecen una patología determinada.

$H_0$ : no padece la enfermedad

$H_1$ : padece la enfermedad

El **error  $\alpha$**  me da la probabilidad de que el test de positivo y que el paciente realmente no tenga la enfermedad. Es lo que llamamos **falsos positivos**.

El **error  $\beta$**  me da la probabilidad de que el test de negativo y que el paciente realmente esté enfermo. Es lo que llamamos **falsos negativos**.

Al contrastar una cierta hipótesis, la máxima probabilidad con la que estamos dispuesto a correr el riesgo de cometerán error de tipo I, se llama **nivel de significación (p-valor)**, y lo definimos como el mínimo nivel de significancia bajo la cual  $H_0$  es rechazada.

Esta probabilidad, se suele especificar antes de tomar la muestra, de manera que los resultados obtenidos no influyan en nuestra elección.

En la práctica, es frecuente un nivel de significación de 0,05 ó 0,01. Si por ejemplo se escoge el nivel de significación 0,05 (ó 5%) al diseñar una regla de decisión, entonces hay unas cinco (5) oportunidades entre 100 de rechazar la hipótesis cuando debiera haberse aceptado; es decir, tenemos un 95% de confianza de que hemos adoptado la decisión correcta. En tal caso decimos que la hipótesis ha sido rechazada al nivel de significación 0,05, lo cual quiere decir que tal hipótesis tiene una probabilidad 0,05 de ser falsa.

Entonces podemos decir que, según el nivel de significación que hemos preestablecido (habitualmente 95%):

- $p > 0,05$  no podemos **rechazar la hipótesis nula** (no decimos que sea cierta, sino que no podemos rechazarla).
- $p < 0,05$  rechazamos la hipótesis nula, **aceptamos la hipótesis alternativa**.



Algunos sinónimos que se suelen usar para un resultado estadísticamente significativo son:

- Rechazo la hipótesis nula con  $p < 0.05$
- Aceptamos la hipótesis alternativa.
- Existe evidencia suficiente para dudar de la hipótesis nula.
- El resultado observado no es compatible con la hipótesis nula.
- Es improbable que el resultado observado sea debido únicamente al azar.
- La variabilidad debida al muestreo no es suficiente para explicar el resultado observado.

La media y la varianza poblacional son parámetros que representan la tendencia central y dispersión de la distribución subyacente de una variable aleatoria. Estos parámetros son típicamente desconocidos y, en consecuencia, han de ser estimados a partir de los valores observados de dicha variable en una muestra.

Un paso importante en el contraste de hipótesis es determinar si los **datos** son **independientes** (desapareados) o **dependientes** (apareados) ya que los test estadísticos que se aplicarán serán diferentes en cada caso. Como regla práctica general, aunque siempre sometemos a discusión, decimos que si una misma variable fue medida dos o más veces sobre la misma unidad experimental los datos son dependientes o apareados. Contrariamente si una misma variable fue medida en unidades experimentales distintas los datos son no apareados o independientes.

Además, como vieron en el encuentro anterior, los datos de una muestra pueden tener una distribución de probabilidad con una función conocida o no. Lo más común dentro de las funciones conocidas es la distribución normal. Por lo tanto, es común que nos fijemos si una muestra de datos tiene o no distribución normal, lo cual es importante para la elección del test estadístico a aplicar para contrastar las hipótesis. Y esto lo podemos hacer, mediante pruebas gráficas o analíticas.

## 03. Prueba de medias para una muestra

Se utiliza para probar una afirmación con respecto a una media de una población única.

Si se conoce la desviación estándar de la población ( $\sigma$ ), la distribución de muestreo adecuada es la **distribución normal**. Si la población que se muestrea es normal, la distribución de muestreo será normal en el caso de todos los tamaños de muestra, y el valor estadístico de prueba que se utiliza es:



$$Z_{prueba} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Si la población no es normal, o si no conocemos su forma, la ecuación anterior solo se usa para tamaños de muestra mayores o iguales a 30.

Si no se conoce la desviación estándar de la población ( $\sigma$ ), el estadístico de prueba que se utiliza es:

$$t_{prueba} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

La **distribución t de Student** es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño.

Es una distribución simétrica alrededor de 0 y de aspecto parecido al de una distribución normal estandarizada, aunque menos apuntada en el centro y con más probabilidad en los extremos. Los grados de libertad de una distribución t de Student determinan su dispersión: al aumentar los grados de libertad, disminuye la variabilidad y la distribución t de Student se aproxima a una distribución normal estandarizada.

El método más extendido para el cálculo de intervalos de confianza se basa en las propiedades de la distribución muestral del estimador. Por el teorema central del límite sabemos que, para cualquier variable aleatoria con media  $\mu$  y varianza  $\sigma^2$ , la distribución de las medias muestrales  $\bar{X}$  es aproximadamente normal con la media  $\mu$  y la varianza  $\sigma^2$  si el tamaño muestral es suficientemente grande.

$$\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma^2}{n}\right)$$

o de forma equivalente aplicando la estandarización de una distribución normal

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightsquigarrow N(0, 1).$$



Se considera práctico utilizar la distribución t solamente cuando se requiera que el tamaño de la muestra sea menor de 30, ya que para muestras más grandes los valores t y z son aproximadamente iguales, y es posible emplear la distribución normal en lugar de la distribución.

En el Lenguaje R, para realizar este tipo de prueba para la media de una población normal, se puede usar la función `t.test()` que tiene la siguiente estructura.

```
t.test(x, y = NULL,  
       alternative = c("two.sided", "less", "greater"),  
       mu = 0, paired = FALSE, var.equal = FALSE,  
       conf.level = 0.95, ...)
```

Donde:

x: vector numérico con los datos.

alternative: tipo de hipótesis alterna. Los valores disponibles son "two.sided" cuando la hipótesis alterna es  $\neq$ , "less" para el caso  $<$  y "greater" para  $>$

mu: valor de referencia de la prueba.

conf.level: nivel de confianza para reportar el intervalo de confianza asociado (opcional).

Veamos un ejemplo:

Queremos verificar si el proceso de llenado de bolsas de café con 500 gramos, en una fábrica cumple con las especificaciones de la envoltura, asumiendo una distribución normal con un nivel de significancia de 5%. Para verificar se toman aleatoriamente muestras de tamaño diez cada cuatro horas. Una muestra de bolsas está compuesta por las siguientes observaciones: 502, 501, 497, 491, 496, 501, 502, 500, 489, 490.

```
>contenido <- c(510, 492, 494, 498, 492,  
                496, 502, 491, 507, 496)
```

Planteamos el test de Hipótesis

$H_0: \mu = 500\text{gr}$



$H_a: \mu \neq 500\text{gr}$

```
>t.test(contenido, alternative='two.sided',
        conf.level=0.95, mu=500)
One Sample t-test

data:  contenido
t = -1.0629, df = 9, p-value = 0.3155
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 493.1176 502.4824
sample estimates:
mean of x
 497.8
```

Como el p-valor es 30% y mayor que el nivel de significancia 5%, no se rechaza la hipótesis nula, es decir, las evidencias no son suficientes para afirmar que el proceso de llenando no está cumpliendo con lo impreso en la envoltura.

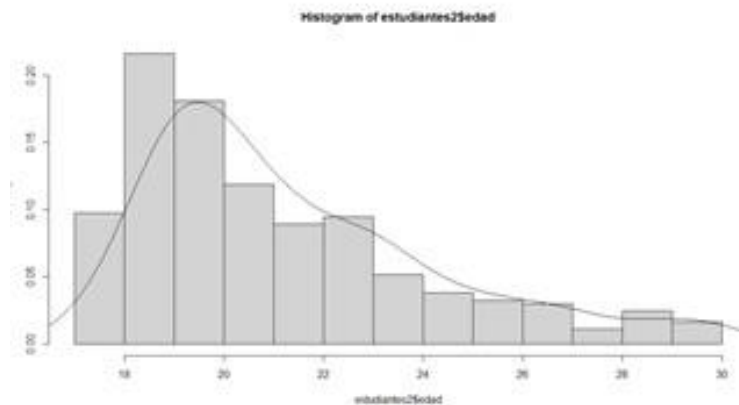
Más ejemplos:

Carguemos la table “estudiantes2.xlsx” que contiene los datos de datos de edad, sexo, peso, altura y actividad física de 370 estudiantes.

```
>estudiantes2 <- read_excel("estudiantes2.xlsx")
```

La gráfica del histograma y la superposición de la función **density()** puede darnos una idea de la distribución de probabilidad de la variable edad.

```
>hist(estudiantes2$edad, freq=F)
>lines(density(estudiantes2$edad))
```



Podemos observar un desplazamiento hacia la izquierda de la densidad de probabilidad, lo que podría estar indicando que no tienen una distribución normal.

La prueba de Shapiro Wilk nos permite tomar una decisión sobre la distribución de probabilidad y es recomendada para variables continuas con gran número de datos. Si el número de datos es menor a 50, esta prueba es recomendada por sobre la prueba de Kolmogorov Smirnov.

```
>shapiro.test(estudiantes2$edad)
Shapiro-Wilk normality test

data:  estudiantes2$edad
W = 0.89362, p-value = 2.23e-15
```

La hipótesis nula es que la distribución analizada tiene distribución normal. Como p-value es menor a 0.01, no tiene distribución normal, lo cual concuerda con el histograma.

Veamos otro ejemplo.

Cargamos el data frame que contiene los resultados obtenidos al realizar una encuesta a 400 estudiantes universitarios y lo asignamos al objeto llamado estudiantes. Suponemos que la población es normal.

```
> estudiantes <- read.csv2("estudiantes.csv", header = TRUE,
                           sep = ",", dec = ".")
```



Consideremos solamente las observaciones que van desde la columna 2 hasta la 35. Luego, con estas observaciones, definiremos un nuevo data frame llamado "datos2a35", al cual le verificaremos su tamaño, las variables y su estructura.

Para este tipo de pruebas no hay una función de R que haga los cálculos, por esta razón uno mismo debe escribir unas líneas de código para obtener los resultados deseados.

```
>datos2a35 <- estudiantes[2:35,] # nuevo data frame
>n <- nrow(datos2a35); n         # N° observaciones (tamaño muestral)
>dim (datos2a35)                 # N° observaciones y N°de variables
>str(datos2a35)                  # Estructura
>names(datos2a35)                # Variable
```

Definamos el objeto "P3" como las calificaciones del tercer parcial.

```
>P3 <- as.numeric(datos2a35$P3) # La variable P3 en "datos2a35"
```

Calculemos la media, la varianza y desviación estándar de P3.

```
>xbarra <- mean(P3)              # Media muestral
>xbarra
>v <- var(P3)                    # Varianza (muestral)
>v
>s <- sd(P3)                     # Desviación estándar (muestral)
>s
```

Para que sea más fácil la visualización, resumimos las tres medidas anteriores en una tabla. Dentro de "datos2a35" y utilizando el método de la región crítica determinemos con un nivel del 5%, si el promedio poblacional de las calificaciones del tercer parcial es igual a 3.5.



```
>datos2a35 %>% summarise(Media=mean(as.numeric(P3)),  
                             Varianza=var(P3), Desviacion=sd(P3)) %>%  
as.data.frame()
```

Escribamos un resumen del enunciado del problema, verifiquemos los supuestos, y por último concluyamos, cuál es el valor de prueba, el valor crítico, la región crítica e interpretemos.

¿Qué nos dice el enunciado?

- Unidades experimentales: Los estudiantes.
- Población: Las calificaciones del tercer parcial.
- Estadístico: la media muestral de las calificaciones del tercer parcial.
- Parámetro: la media poblacional de las calificaciones del tercer parcial.
- Tamaño muestral:  $n=34$ .
- Tamaño poblacional:  $N$  es desconocido.
- Grado de confianza: 95%.
- Nivel de significancia: 5%.
- Tipo de prueba: Prueba de bilateral o de dos colas.

Con éstos datos, planteamos nuestro test de hipótesis:

Hipótesis nula	$H_0: \mu=3.5$
Hipótesis alternativa	$H_1: \mu \neq 3.5$

Verificamos los supuestos:

- La forma de la población es normal.
- La varianza poblacional es desconocida.
- El tamaño muestral es grande ( $n > 30$ ).

Entonces, podemos concluir que la distribución muestral de la media muestral es normal. La fórmula a utilizar es:

$$Z_{prueba} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

A continuación calculamos el valor de prueba.

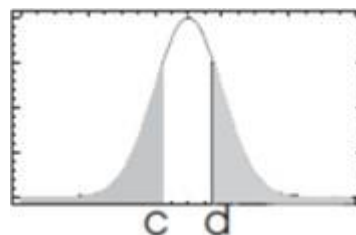


```
>mu <- 3.5      #) Ver valor numérico en las hipótesis
>media <- xbarra
>desviacion <- s
>muestra <- n
>ES <- s/sqrt(n)      #) Error estándar (= SD del estadístico)
>Z <- (xbarra - mu)/ES  #) Valor de prueba
>Z
[1] 0.9746017
```

La región crítica es:

$H_0: \mu=3.5$      $H_1: \mu \neq 3.5$

$c = -Z_{\alpha/2}$  y  $d = Z_{\alpha/2}$



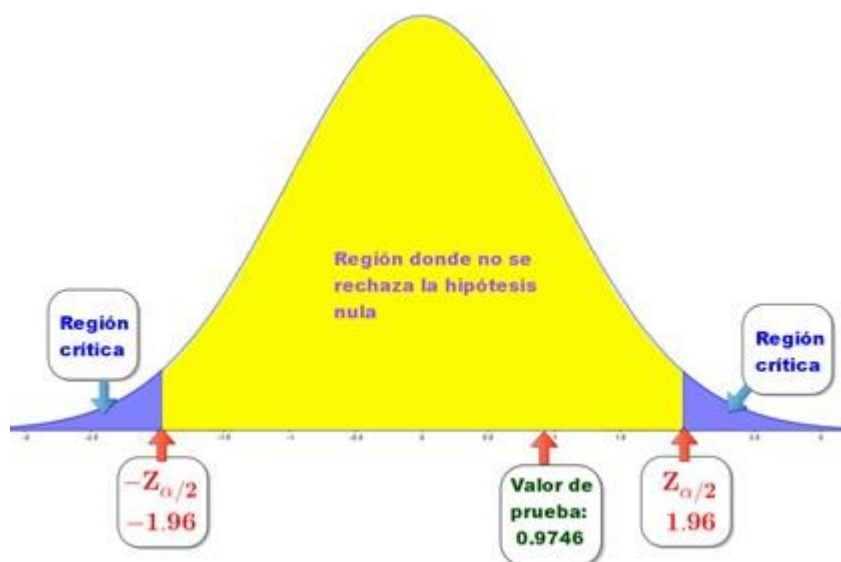
El valor crítico (o valor cuantil de la distribución normal)  $Z_{\alpha/2}$  es:

```
>alfa <- 0.05
>Critico <- qnorm(1- (alfa/2))  #) Valor crítico
>Critico
[1] 1.959964
```

Es decir,  $Z_{\alpha/2} = 1.95996$ .

Con lo cual podemos observar que el valor de prueba  $Z = 0.9746$  no cae en la región crítica. Es decir, se cumple:

$$-Z_{\alpha/2} < Z < Z_{\alpha/2}$$



```
Z < -Critico | Z > Critico    #) Región crítica
[1] FALSE
```

Por lo tanto, no rechazamos  $H_0: \mu=3.5$ , con lo cual, con una confianza del 95%, podemos afirmar que el promedio poblacional de las calificaciones del tercer parcial de todos los estudiantes es 3.5.

Nos falta calcular el p-valor. Para eso sabemos que tenemos una prueba a dos colas.

Con  $X = Z$  y  $x = 0.9746$  (el valor de prueba), el P-valor es:

$$P\text{-valor} = 2 P(Z \geq |0.9746|) = 2 (0.1649) = 0.3298$$

```
>2*(1-pnorm(Z)) # calculo del p-valor
[1] 0.3297579
```

Recordemos la regla de decisión:



Se rechaza  $H_0$  cuando  $p\text{-valor} \leq \alpha$ . No

se rechaza  $H_0$  cuando  $p\text{-valor} > \alpha$ .

Como el p-valor es mayor que 0.05, no se rechaza  $H_0$  al nivel el 5%. Es decir,  $\mu=3.5$ .

Por consiguiente, con una confianza del 95%, podemos afirmar que el promedio poblacional de las calificaciones del tercer parcial de todos los estudiantes es 3.5.

Otro ejemplo...

Una compañía afirma que los automóviles recorren en promedio más de 20000 kilómetros por año pero en realidad creemos que el promedio es menor. Para probar tal afirmación se pide a una muestra de 100 propietarios de automóviles seleccionada de manera aleatoria que lleven un registro de los kilómetros que recorren.

¿Estarías de acuerdo con la afirmación si la muestra aleatoria indicara un promedio de 19500 kilómetros y una desviación estándar de 3900 kilómetros? Utiliza un valor P en tu conclusión y usa una significancia del 3%.

Planteamos el test de Hipótesis

$H_0: \mu \geq 20000$  Km

$H_a: \mu < 20000$  Km

```
>xbarra <- 19500          # Datos del problema
>desvia <- 3900
>n <- 100
>mu <- 20000              # Media de referencia
>est <- (xbarra - mu) / (desvia / sqrt(n))
>est                      # Para obtener el valor del estadístico
[1] -1.282051
>pnorm(est)               # Para obtener el valor-P
[1] 0.09991233
```

Como el p-valor es mayor que el nivel de significancia 3%, no hay evidencias suficientes para pensar que ha disminuido el recorrido anual promedio de los autos.

## 04. Prueba de medias para dos muestras

Las pruebas de dos muestras se utilizan para decidir si las medias de dos poblaciones son iguales.

El estadístico del test dependerá de la estructura de los conjuntos de datos. En particular es importante establecer si los datos corresponden a muestras independientes o apareadas.

Se requieren **dos muestras independientes**, es decir donde las observaciones de una muestra no están relacionadas o emparejadas con las observaciones de la otra muestra.

Con frecuencia se utilizan pruebas de dos muestras para comparar dos métodos de enseñanza, dos marcas, dos ciudades, dos distritos escolares y otras cosas semejantes.

Por ejemplo, una compañía está experimentando con dos diferentes mezclas de pintura, para ver si se puede modificar el tiempo de secado de una pintura para uso doméstico. Cada mezcla es probada un determinado número de veces, y comparados posteriormente los tiempos medios de secado de las dos muestras. Una parece ser superior, ya que su tiempo medio de secado (muestra) es 30 minutos menor que el de la otra muestra.

Pero... ¿son realmente diferentes los tiempos medios de secado de las dos pinturas, o esta diferencia muestral es nada más la variación aleatoria que se espera, aun cuando las dos fórmulas presentan idénticos tiempos medios de secado? Una vez más, las diferencias casuales se deben distinguir de las diferencias reales.

La hipótesis nula puede establecer que las dos poblaciones tienen medias iguales:

$$H_0 : \mu_1 = \mu_2$$

La hipótesis alternativa, puede ser alguna de las siguientes:

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

$$H_1 : \mu_1 > \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

Cuando conocemos las desviaciones estándar de la población y el estadístico de prueba que se utiliza es el siguiente:



$$Z_{prueba} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Suponemos que, el valor real de Z, cuando  $H_0$  es verdadero, está distribuido normalmente con una media de 0 y una desviación estándar de 1, (distribución normal estandarizada), para casos en los que la suma  $n_1+n_2$  es mayor o igual a 30.

Cuando no se conocen la desviación estándar de la población y  $n_1+n_2$  es menor que 30, el estadístico de prueba que se utiliza es:

$$t_{prueba} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Podemos concluir que, un test de hipótesis para dos muestras es similar en muchos aspectos al test para una muestra.

- Se especifica una hipótesis nula, en la mayoría de los casos se propone que las medias de las dos poblaciones son iguales y se establece la hipótesis alternativa (uni o bilateral).
- Se especifica un nivel de significación  $\alpha$ .
- Se calcula el p-valor: la probabilidad de obtener datos cuyas medias muestrales difieren tanto o más que la diferencia observada cuando  $H_0$  es verdadera. Si esta probabilidad es pequeña (menor que  $\alpha$ ) se rechaza  $H_0$  y se concluye que la diferencia observada no es atribuible al azar y las medias de las dos poblaciones son diferentes.

Las **muestras apareadas** se obtienen cuando se realizan comparaciones sobre una misma unidad experimental:

- se determina en la misma unidad la concentración de una sustancia con dos métodos diferentes.
- se estudia un mismo individuo antes y después de un tratamiento.



## 05. Prueba de hipótesis para la varianza de una población normal

El test para la igualdad de varianzas poblacionales se basa en la comparación de las varianzas muestrales  $S_1^2$  y  $S_2^2$

Para realizar este tipo de prueba, R dispone de la función **var.test** la cual tiene la siguiente estructura:

```
var.test(x, alternative = "two.sided",  
         null.value = 1, conf.level = 0.95)
```

Donde:

x: vector numérico con los datos.

alternative: tipo de hipótesis alterna. Los valores disponibles son "two.sided" cuando la alterna es  $\neq$ , "less" para el caso  $<$  y "greater" para  $>$

null.value: valor de referencia de la prueba.

conf.level: nivel de confianza para reportar el intervalo de confianza asociado (opcional).

Veamos un ejemplo.

Se realiza un estudio para comparar dos tratamientos que se aplicarán a frijoles crudos con el objetivo de reducir el tiempo de cocción. El tratamiento T1 es a base de bicarbonato de sodio, el T2 es a base de cloruro de sodio o sal común. La variable respuesta es el tiempo de cocción en minutos. ¿Son las varianzas de los tiempos iguales o diferentes? Usar  $\alpha = 0.05$ , sabiendo que

T1: 76, 85, 74, 78, 82, 75, 82.

T2: 57, 67, 55, 64, 61, 63, 63.

Nos interesa probar si las varianzas poblacionales son iguales o no, por esta razón el cociente de  $\sigma^2_{T1}/\sigma^2_{T2}$  se iguala al valor de 1 que será el valor de referencia de la prueba.



$$H_0: \sigma^2_{T1} / \sigma^2_{T2} = 1$$

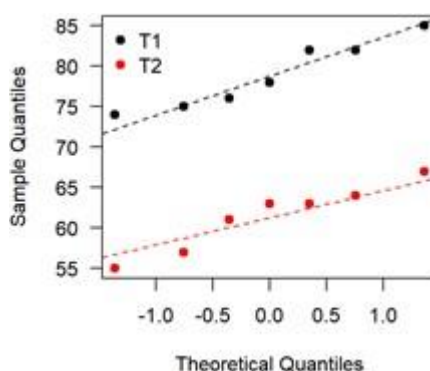
$$H_a: \sigma^2_{T1} / \sigma^2_{T2} \neq 1$$

```
>T1 <- c(76, 85, 74, 78, 82, 75, 82)
```

```
>T2 <- c(57, 67, 55, 64, 61, 63, 63)
```

Primero se debe explorar si las muestras provienen de una población normal

```
>q1 <- qqnorm(T1, plot.it=FALSE)
>q2 <- qqnorm(T2, plot.it=FALSE)
>plot(range(q1$x, q2$x), range(q1$y, q2$y), type="n", las=1,
      xlab='cuantiles teóricos', ylab='cuantiles de muestra')
>points(q1, pch=19)
>points(q2, col="red", pch=19)
>qqline(T1, lty='dashed')
>qqline(T2, col="red", lty="dashed")
>legend('topleft', legend=c('T1', 'T2'), bty='n',
      col=c('black', 'red'), pch=19)
```





Se observa que los puntos están bastante alineados lo cual nos lleva a pensar que las muestras si provienen de una población normal.

```
>var.test(T1, T2, null.value=1, alternative="two.sided",
          conf.level=0.95)

F test to compare two variances

data:  T1 and T2
F = 1.011, num df = 6, denom df = 6, p-value = 0.9897
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1737219 5.8838861
sample estimates:
ratio of variances
      1.011019
```

Como el p-valor es 0.9897, muy superior al nivel  $\alpha$  de significancia 5%, se puede concluir que las varianzas son similares.

Más ejemplos...

El arsénico en agua potable es un posible riesgo para la salud. Un artículo reciente reportó concentraciones de arsénico en agua potable en partes por billón (ppb) para diez comunidades urbanas y diez comunidades rurales. Los datos son los siguientes:

Urbana: 3, 7, 25, 10, 15, 6, 12, 25, 15, 7

Rural: 48, 44, 40, 38, 33, 21, 20, 12, 1, 18

Planteamos las hipótesis:

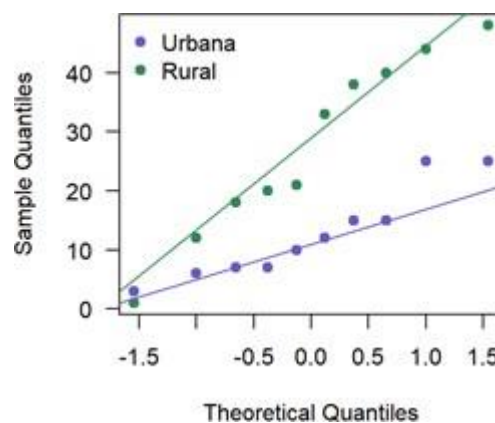
$$\begin{aligned} H_0: \sigma^2_{urb} / \sigma^2_{rur} \\ &= 1 \quad H_a: \sigma^2_{urb} / \\ &\quad \sigma^2_{rur} \neq 1 \end{aligned}$$



```
>urb <- c(3, 7, 25, 10, 15, 6, 12, 25, 15, 7)
>rur <- c(48, 44, 40, 38, 33, 21, 20, 12, 1, 18)
```

Exploremos si las muestras provienen de una población normal

```
>q1 <- qqnorm(urb, plot.it=FALSE)
>q2 <- qqnorm(rur, plot.it=FALSE)
>plot(range(q1$x, q2$x), range(q1$y, q2$y), type="n", las=1,
      xlab='cuantiles teóricos', ylab='cuantiles de muestra')
>points(q1, pch=19)
>points(q2, col="red", pch=19)
>qqline(urb, lty='dashed')
>qqline(rur, col="red", lty="dashed")
>legend('topleft', legend=c('urb', 'rur'), bty='n',
      col=c('black', 'red'), pch=19)
```



```
var.test(urb, rur, null.value=1, alternative="two.sided",
        conf.level=0.95)
```

F test to compare two variances

data: urb and rur

F = 0.24735, num df = 9, denom df = 9, p-value = 0.04936

alternative hypothesis: true ratio of variances is not equal to 1



```
95 percent confidence interval:
 0.06143758 0.99581888
sample estimates:
ratio of variances
 0.2473473
```

Como el p-valor es 0.04936 es menor que el nivel de significancia  $\alpha = 0.05$ , se puede concluir que las varianzas no son iguales.

## 06. Prueba de hipótesis para la diferencia de medias $\mu_1 - \mu_2$ con varianzas iguales

Para esta prueba se utiliza la función `t.test()`, cuya estructura es:

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Donde:

x: vector numérico con la información de la muestra 1,

y: vector numérico con la información de la muestra 2,

alternative: tipo de hipótesis alterna. Los valores disponibles son "two.sided" cuando la alterna es  $\neq$ , "less" para el caso  $<$  y "greater" para  $>$   $\mu$ :

valor de referencia de la prueba.

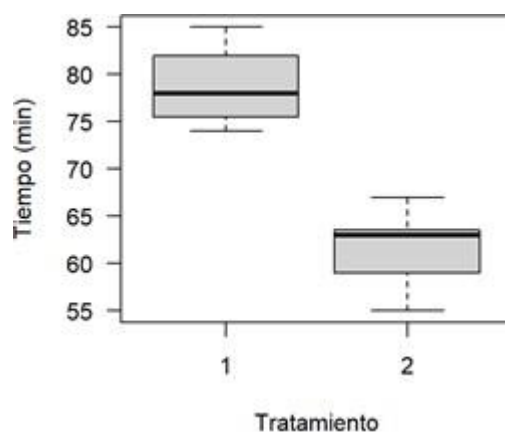
var.equal=TRUE: indica que las varianzas son desconocidas pero iguales.

conf.level: nivel de confianza para reportar el intervalo de confianza asociado (opcional).

Volvamos al ejemplo de la cocción de los frijoles. ¿Existen diferencias entre los tiempos de cocción de los frijoles con T1 y T2? Usar un nivel de significancia del 5%.



```
>datos <- data.frame(tiempo=c(T1, T2), trat=rep(1:2, each=7))
>boxplot(tiempo ~ trat, data=datos, las=1,
         xlab='Tratamiento', ylab='Tiempo (min)')
```



Se observa que las cajas de los boxplot no se traslapan, esto es un indicio de que las medias poblacionales,  $\mu_1$  y  $\mu_2$ , son diferentes, se observa también que el boxplot para el tratamiento T1 está por encima del T2.

Ahora planteamos las hipótesis

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

```
>t.test(x=T1, y=T2, alternative="two.sided", mu=0,
        paired=FALSE, var.equal=TRUE, conf.level=0.97)
Two Sample t-test
```

data: T1 and T2

t = 7.8209, df = 12, p-value = 4.737e-06

alternative hypothesis: true difference in means is not equal to 0

97 percent confidence interval:

11.94503 22.91212



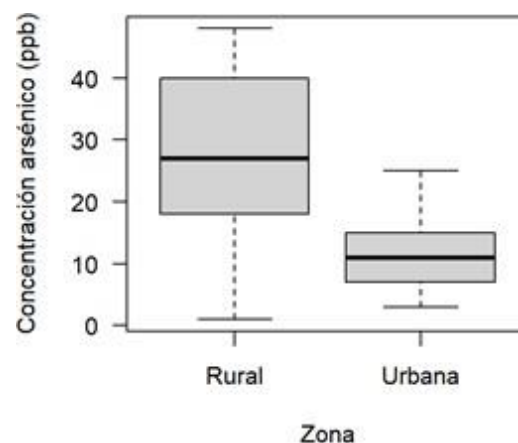
```
sample estimates:
mean of x mean of y
78.85714 61.42857
```

Concluimos que hay diferencias significativas entre los tiempos promedios de cocción con T1 y T2, resultado que ya se sospechaba al observar en el boxplot.

## 07. Prueba de hipótesis para la diferencia de medias $\mu_1 - \mu_2$ con varianzas diferentes

Retomemos el ejemplo de las mediciones de concentración de arsénico en muestras de agua.

```
>datos <- data.frame(Concentracion=c(urb, rur),
                     Zona=rep(c('Urbana', 'Rural'), each=10))
>boxplot(Concentracion ~ Zona, data=datos, las=1,
        xlab='Zona', ylab='Concentración arsénico (ppb)')
```





Nuestras hipótesis son:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

```
>t.test(x=urb, y=rur, alternative="two.sided", mu=0,
        paired=FALSE, var.equal=FALSE, conf.level=0.95)

Welch Two Sample t-test

data:  urb and rur
t = -2.7669, df = 13.196, p-value = 0.01583
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.694067  -3.305933
sample estimates:
mean of x mean of y
    12.5      27.5
```

Podemos concluir que hay diferencias significativas entre las concentraciones de arsénico del agua entre las dos zonas. La zona que presenta mayor concentración media de arsénico en el agua es la rural.

## Bibliografía

- BIOESTADÍSTICA Roberto Pastor-Barriuso Científico Titular Centro Nacional de Epidemiología, Instituto de Salud Carlos III, Madrid, 2012
- Test o Prueba de hipótesis .Prof Maria B. Pintarelli Recuperado de [https://www.mate.unlp.edu.ar/practicas/117\\_13\\_16072013221648.pdf](https://www.mate.unlp.edu.ar/practicas/117_13_16072013221648.pdf)



Autor: Myrian Aguilar. Esta obra está bajo una [Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional](https://creativecommons.org/licenses/by-nc-sa/4.0/). Mundos E.