

# *Inferencia e introducción a los modelos estadísticos con R*

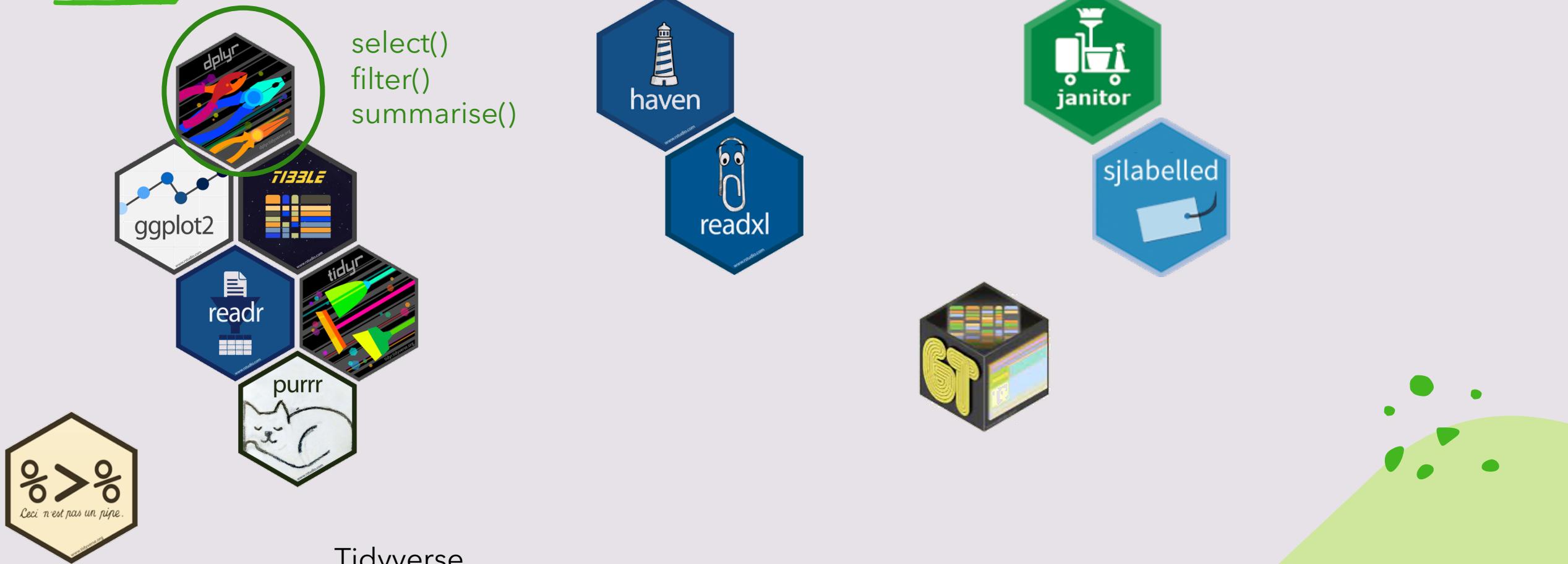
---

Día 2 y 3

Curso intersemestral de invierno

Dra. Ana Ruth Escoto Castillo

# *Después de la sesión 1: Repasso de los paquetes*





# *introducción a la inferencia*

---

Un par de apuntes antes de empezar

Curso intersemestral de invierno



# *Pasos de la inferencia*

---



# Algunos conceptos básicos



Población



Muestra



Individuo

**Variable:** es una característica que cambia en el tiempo o entre los individuos u objetos de estudio

# *¿Dios Juega a los dados?*

---

Pocas cosas en el mundo son verdaderamente aleatorias en el sentido de que ninguna cantidad de información nos permitirá predecir el resultado. Pero de acuerdo a la rama de la física llamada **mecánica cuántica**, el azar es la norma dentro de los átomos individuales. Aunque Albert Einstein ayudó a la teoría cuántica empezar, él siempre insistió en que la naturaleza debe tener alguna realidad fija, no sólo probabilidades. "Nunca creeré que Dios juega a los dados con el mundo", dijo el gran científico. Un siglo después del trabajo primera de Einstein sobre la teoría cuántica, parece que él estaba equivocado.



# *El azar y la probabilidad*

---

Llamamos a un fenómeno **aleatorio** si los resultados individuales son inciertos, pero hay sin embargo una distribución regular de los resultados en un gran número de repeticiones.

La probabilidad de cualquier resultado de un fenómeno aleatorio es la proporción de veces que el resultado se produciría en una muy larga serie de repeticiones



# *¿Qué es probabilidad?*

---

Usamos gráficas y medidas numéricas para describir conjuntos de datos que eran por lo general muestras.

Medimos "con qué frecuencia"

$$\text{Frecuencia relativa} = f/n$$

Cuando  $n$  se hace grande  $n \rightarrow \infty$

Muestra



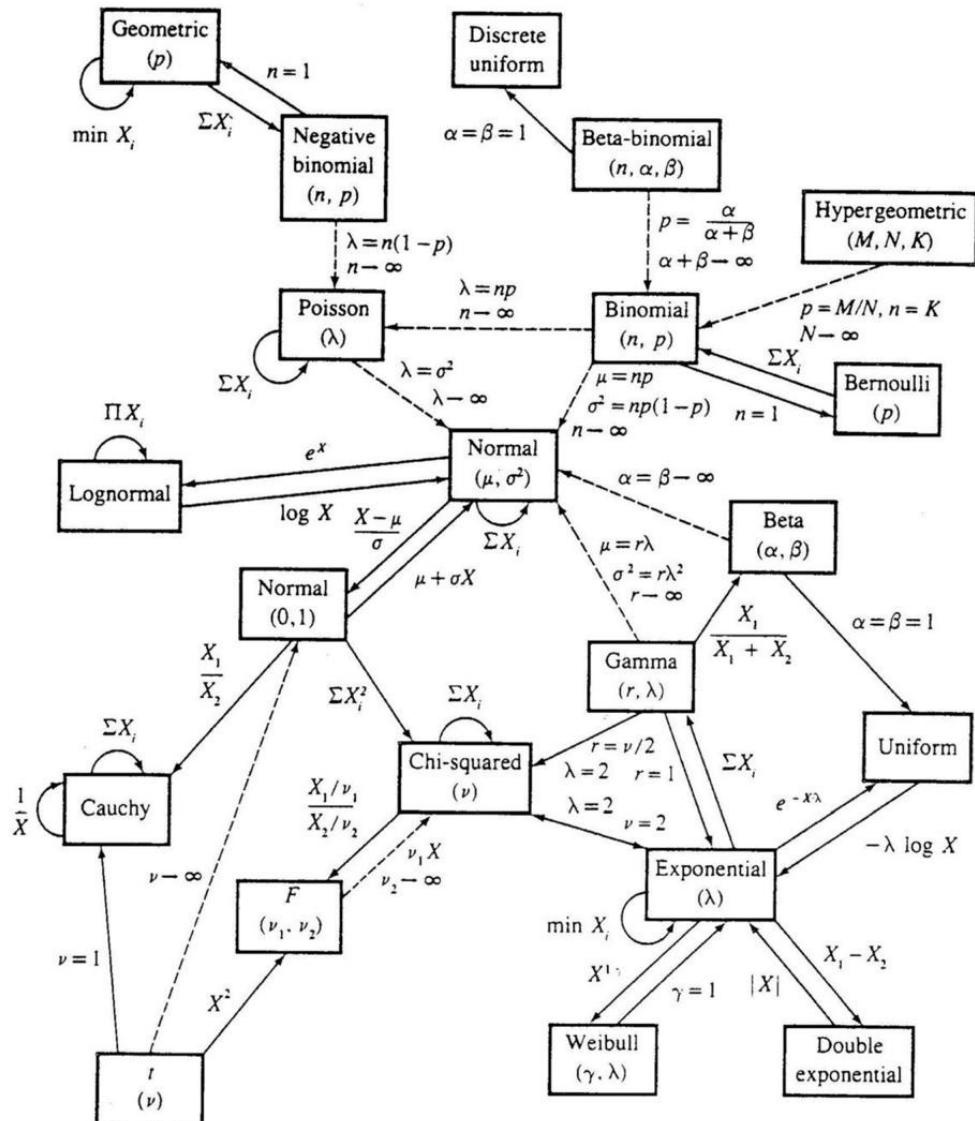
Población

Frecuencia relativa



Probabilidad

# *Modelos de probabilidad*



**Relationships among common distributions.** Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

# *Entonces...*

---

Las poblaciones son descritas por sus distribuciones de probabilidad y los parámetros.

Por ejemplo:

- Para las poblaciones cuantitativas, la ubicación y la forma son descritos por  $\mu$  y  $\sigma$ .
- Para poblaciones binomiales, la ubicación y la forma son determinados por  $p$ .

Si no se conocen los valores de los parámetros, **hacemos inferencia** sobre ellos utilizando información de la muestra.



# *POBLACIÓN, MUESTRA, DISEÑO DE MUESTREO*

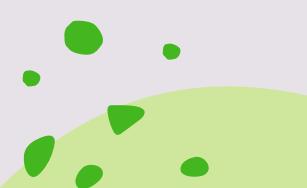
---

La población en un estudio estadístico es todo el grupo de personas sobre las que queremos información.

Una **muestra** es una parte de la población de la que en realidad nos recogemos información.

Se utiliza una muestra para sacar conclusiones sobre toda la población.

Un **diseño de muestreo** describe exactamente cómo elegir una muestra de la población.



# *Inferencia y por qué usamos muestreo*

---

- ❑ El proceso de elaboración de conclusiones sobre una población sobre la base de datos de la muestra se llama **inferencia**, porque inferimos información sobre la población de lo que sabemos acerca de la muestra.
- ❑ Usamos el muestreo aleatorio para
- ❑ Eliminar el sesgo de selección
  - ✓ Muestreo por conveniencia o muestras de respuesta voluntarias son engañosos debido a que estos métodos de elección de una muestra están sesgados.
- ❑ Para aplicar las leyes de la probabilidad que permiten la inferencia confiable acerca de la población.
  - ❑ Ley de los grandes números
  - ❑ Teorema del Límite Central



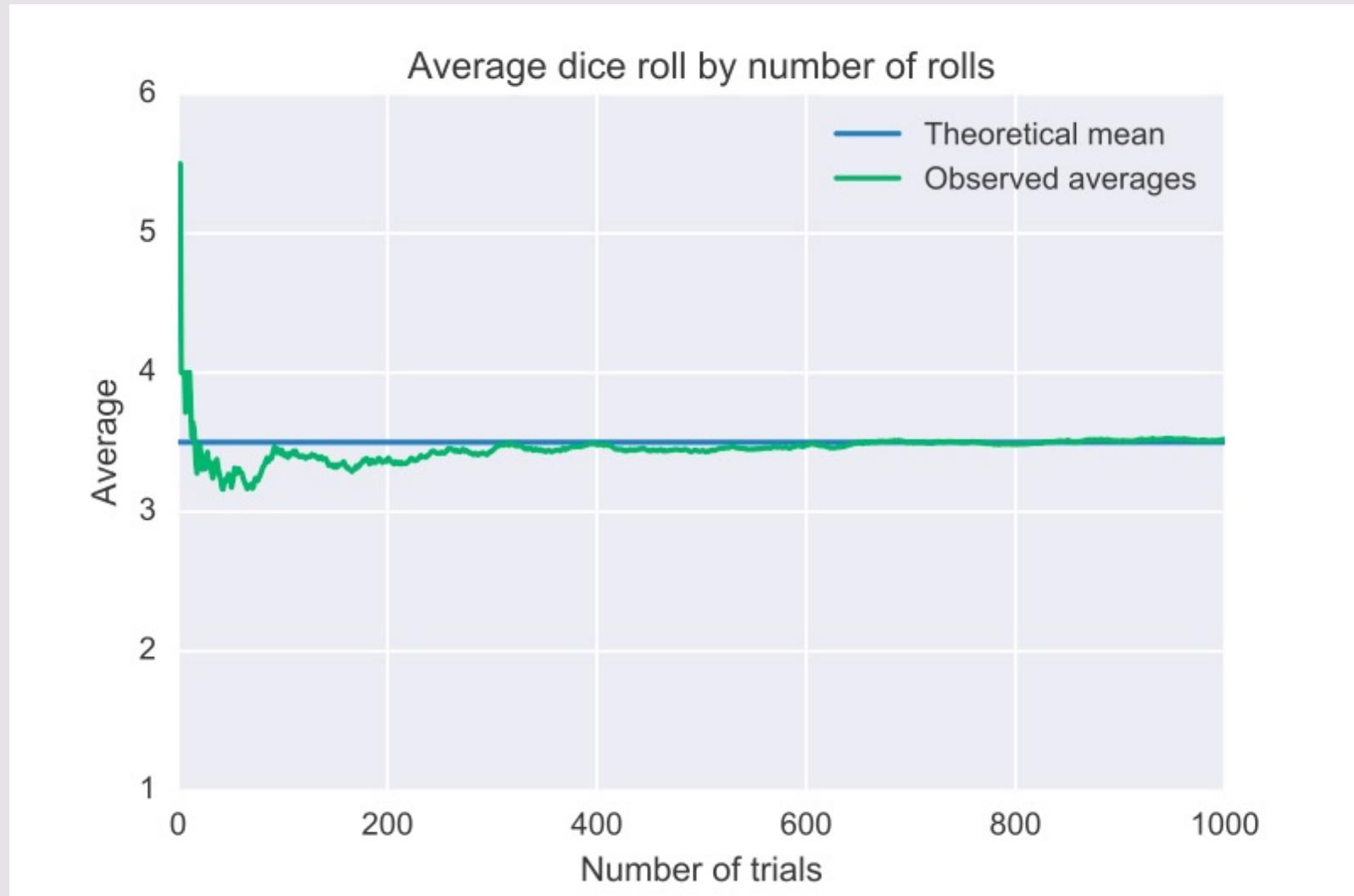
# *Convención*

---

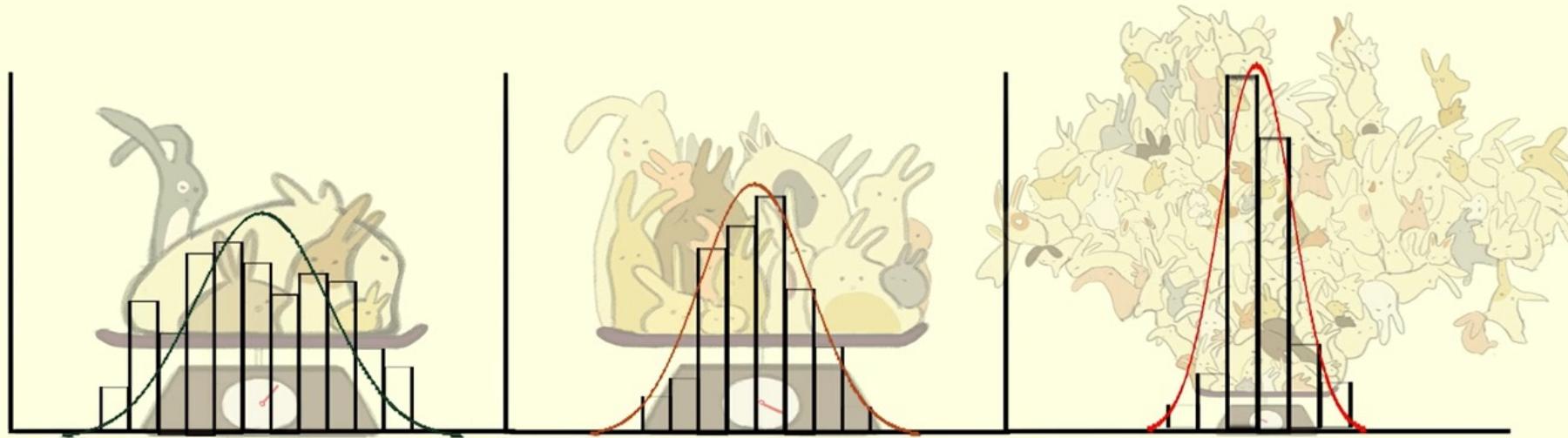
$\theta$  es un parámetro

$\hat{\theta}$  es un estadístico o estimador

# *Ley de los grandes números*



# Central Limit Theorem

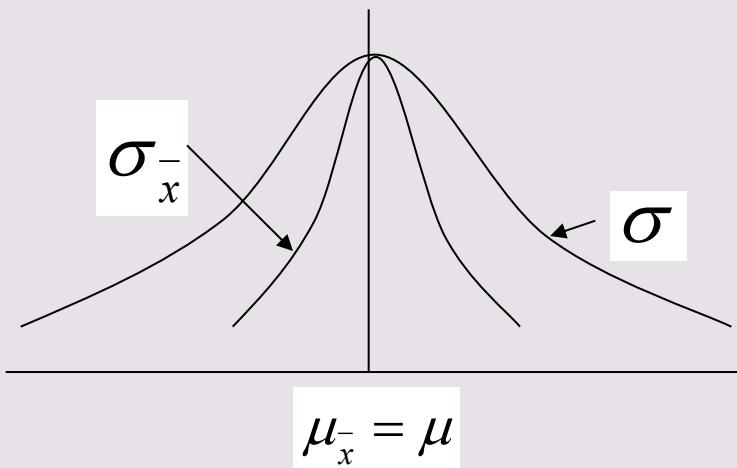


The averages of samples have approximately normal distributions

Sample size → Bigger  
Distribution of Averages → More normal and narrower

## **TEOREMA DEL LIMITE CENTRAL**

"Si se toman sucesivas muestras ( $k$ ) de tamaño  $n$  de una población que puede o no ser normal, la distribución de probabilidad de los promedio (sumas) de esas muestras, conforme  $n$  se vuelve grande, se aproxima a una distribución normal con:



$$\begin{aligned}\mu_{\bar{x}} &= \bar{x} = \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ Z &= \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\end{aligned}$$

# ***TIPOS DE INFERENCIA***

---

**Estimación:**

Estimar o predecir el valor del parámetro

"¿Cuál es (son) los valores más probables de  $\mu$  o  $p$ ?"

**Prueba de hipótesis:**

Decidir sobre el valor de un parámetro basado en una idea preconcebida.

"¿Acaso la muestra proviene de una población con  $\mu = 5$  ó  $p = 0.2$ ?"

**Modelos estadísticos - que también son pruebas estadísticas**

# *Tipos de inferencia*

---

Sea que se está estimando parámetros o probando hipótesis, los métodos estadísticos son importantes porque proporcionan:

- Los métodos para hacer la inferencia
- Una medida numérica de la bondad o la fiabilidad de la inferencia

# *Propiedades de los estimadores puntuales*

---

Ellas son:

- 1) Insesgabilidad**
- 2) Insesgabilidad de mínima varianza (eficiencia)**
- 3) Consistencia**
- 4) Distribución asintóticamente normal.**

# *Propiedades de los estimadores puntuales*

---

Dado que un estimador se calcula a partir valores de la muestra, éste varía de una muestra a otra en función de su distribución muestral.

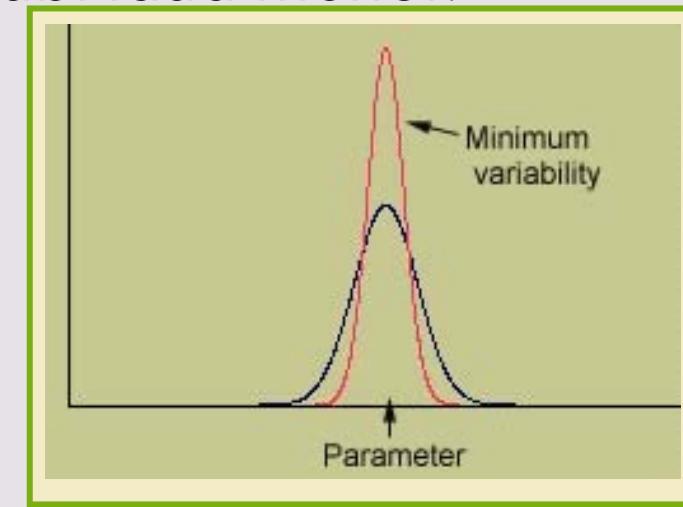
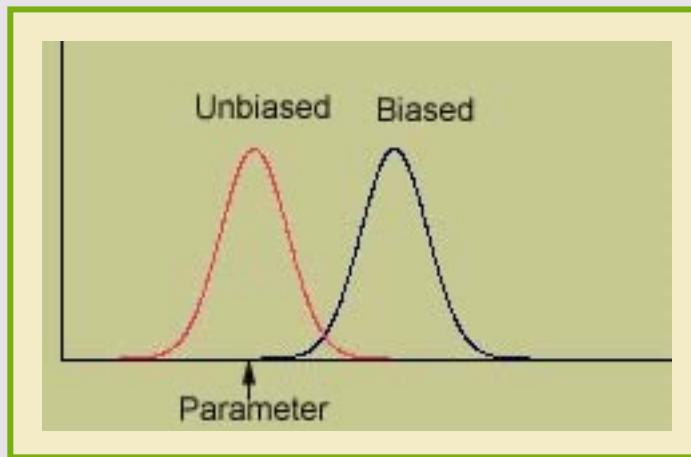
Un estimador es insesgado si la media de su distribución de muestreo es igual a el parámetro de interés.

- No sobreestimar o subestimar sistemáticamente el parámetro objetivo.

# *Propiedades de los estimadores puntuales*

---

De todos los estimadores insesgados, preferimos el estimador cuya distribución muestral tiene la dispersión o variabilidad menor.



# *Propiedades de los estimadores puntuales*

---

Ellas son:

- 1) Insesgabilidad**
- 2) Insesgabilidad de mínima varianza (eficiencia)**
- 3) Consistencia**
- 4) Distribución asintóticamente normal.**



**Sólo para muestras grandes**



# *Consistencia*

---

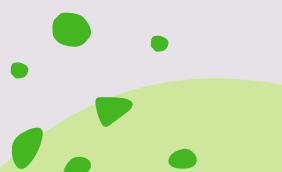
La condición de consistencia (conocida como consistencia simple) establece que para muestras grandes  $\hat{\theta}$  n tiende a aproximarse a  $\theta$ .

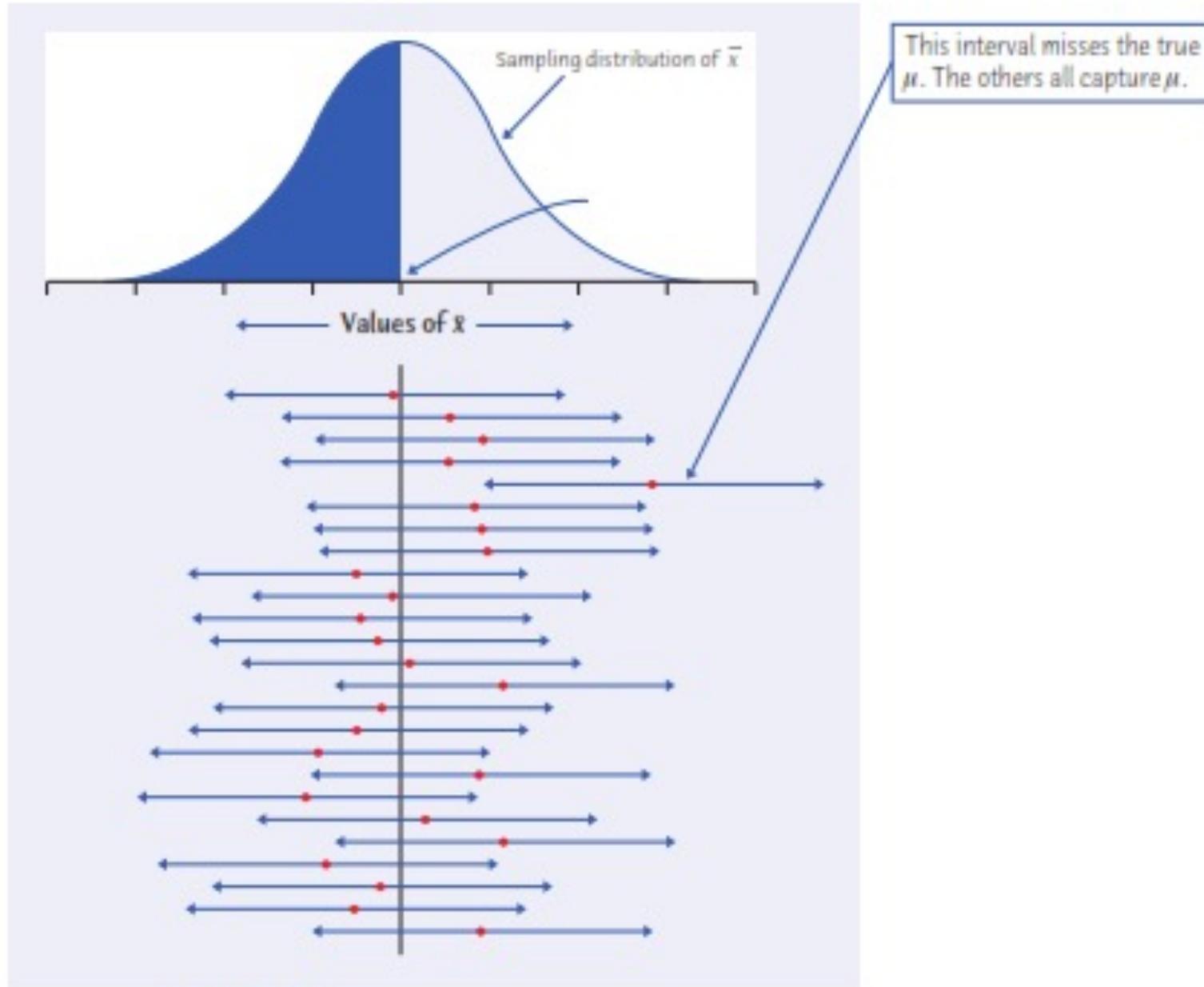
Existe no obstante otra condición de consistencia conocida como consistencia en error cuadrático que implica que tanto el sesgo como la varianza del estimador  $\hat{\theta}$  tienden a cero a medida que aumenta el tamaño de la muestra.

## *Distribución asintóticamente normal*

---

Un estimador es asintóticamente normal si además de ser insesgado y eficiente, cumple con la propiedad de tener **distribución normal** cuando el tamaño de la muestra se incrementa





**FIGURE 14.2**

Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that contains the population mean  $\mu$ .

# *Algunos elementos básicos de la inferencia*

---

Cálculos de intervalos o pruebas de hipótesis para una sola muestra cuantitativa

- Muestra grande
- Muestra chiquita

Cálculos de intervalos o pruebas de hipótesis para una sola muestra binomial (o proporción)

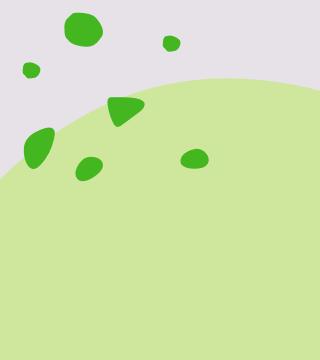
- Muestra grande

Cálculos de intervalos o pruebas de hipótesis para diferencia de dos muestras cuantitativas

- Muestra grande
- Muestras chicas

Cálculos para diferencia de dos muestras binomiales (o proporciones)

- Muestra grande



# *Gran regla para la estimación por intervalo*

---

Estimador  
puntual

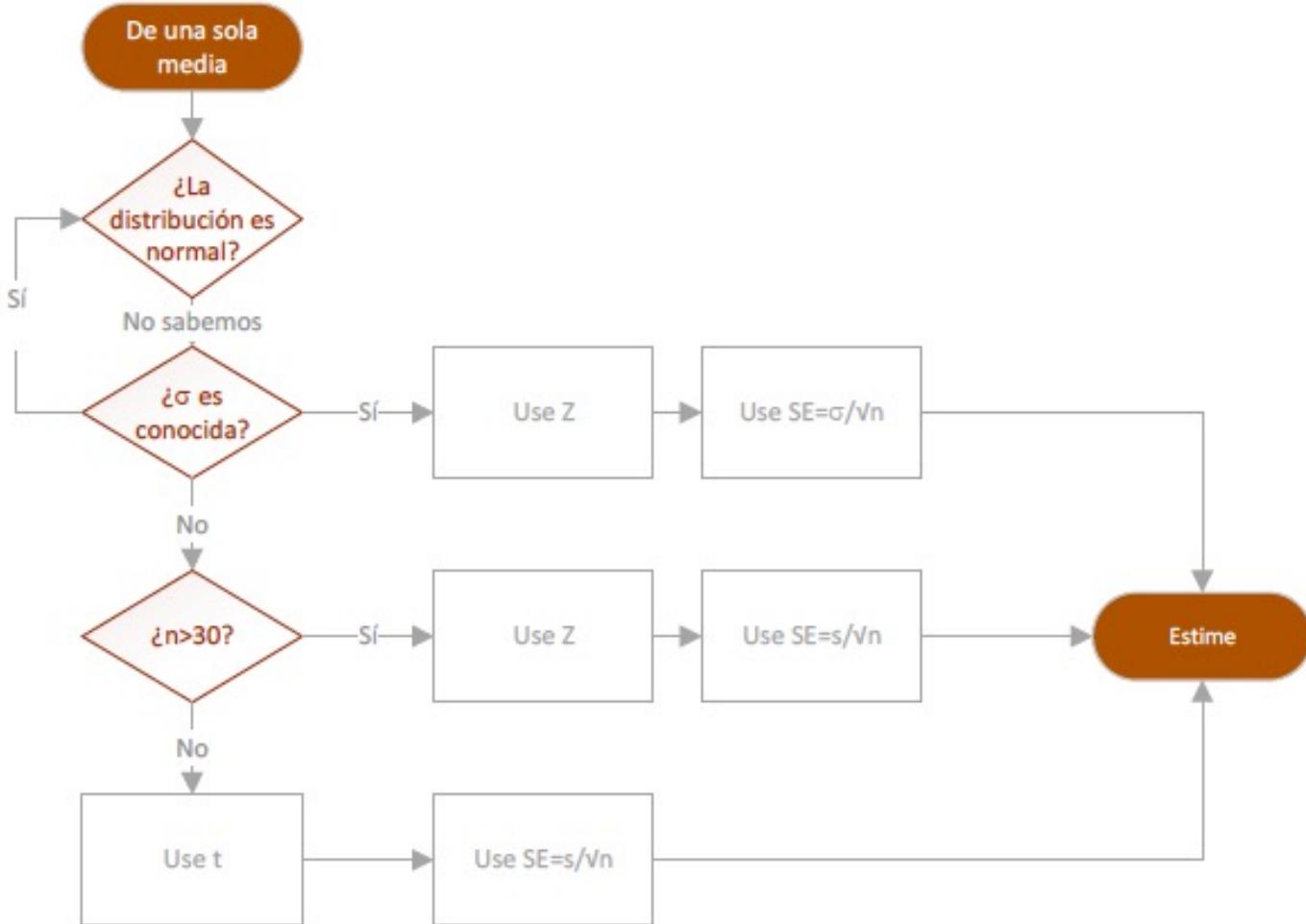


Un valor asociado  
con la  
probabilidad



Error estándar  
muestral

- Varianza
- Tamaño de la muestra



# *Pruebas de hipótesis*

---

It is a good morning exercise for a research scientist to discard a pet hypothesis every day before breakfast. It keeps him young.

–Konrad Lorenz (1903-1989)

# *Prueba de hipótesis*

---

Como en un juicio tribunal. Al tratar a una persona por un delito, el jurado tiene que decidir entre una de dos posibilidades:

- La persona es culpable.
- La persona es inocente.

Para empezar, se asume que la persona inocente. [En teoría]

El fiscal presenta pruebas, tratando de convencer al jurado de rechazar la hipótesis original de inocencia, y la conclusión de que la persona es culpable.



# *Introducción*

---

Involucra una suposición elaborada sobre uno o más parámetros de una o más poblaciones.

Usando la información muestral se verificará la suposición sobre los parámetros estudiados.

La hipótesis que se contrasta se llama hipótesis nula ( $H_0$ ).

Decisión	Conclusión
Se rechaza $H_0$	Se puede afirmar que $H_1$ es verdadera
No se rechaza $H_0$	No se puede afirmar que $H_1$ es verdadera



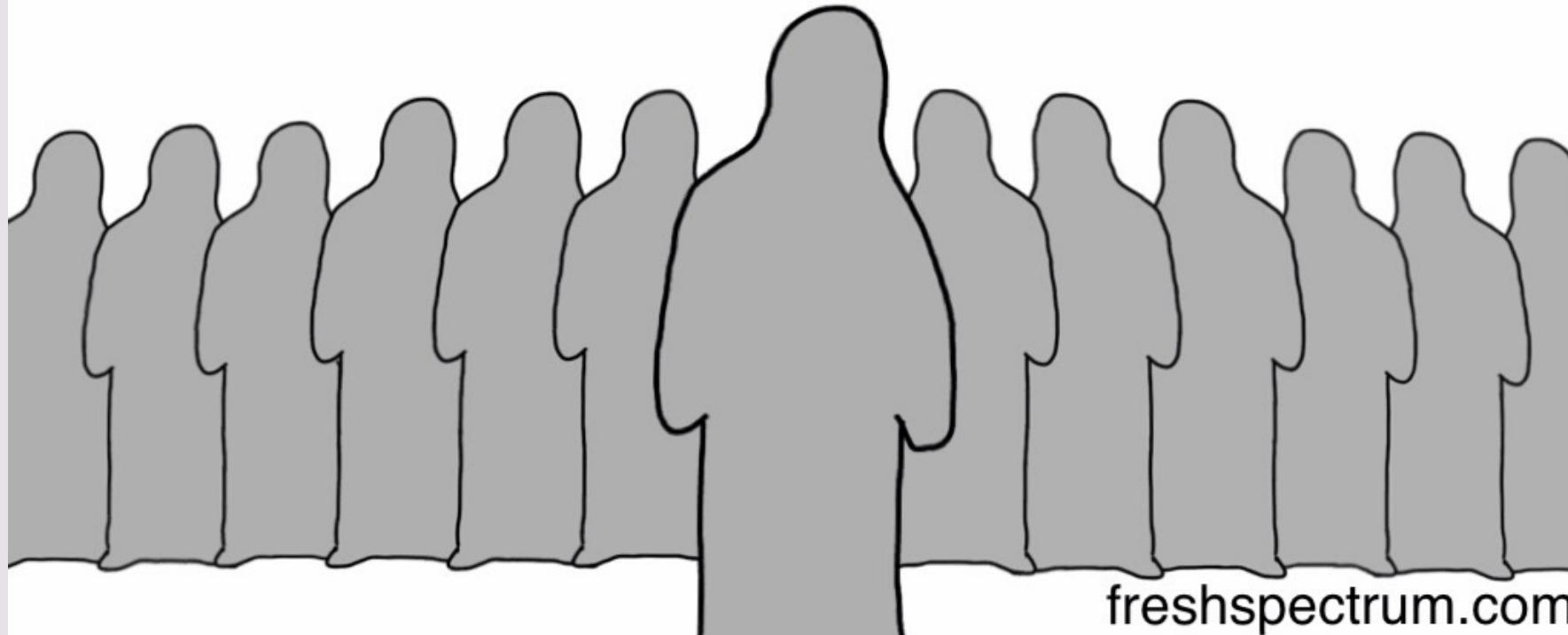
I am what is

The default, the status quo

I am already accepted, can only be rejected

The burden of proof is on the alternative

# I am the null hypothesis



# *Partes de una prueba estadística (o un test estadístico)*

---

1. La hipótesis nula,  $H_0$ :

Asume como cierta hasta que podamos demostrar lo contrario.

2. La hipótesis alternativa,  $H_a$ :

Será aceptado como cierto si podemos refutar  $H_0$

Tribunal de justicia:

$H_0$ : inocente

$H_a$ : culpable

Farmacéuticas:

$H_0$ :  $\mu$  no excede las cantidades mínimas

$H_a$ :  $\mu$  excede las cantidades mínimas

# *Partes de una prueba estadística (o un test estadístico)*

---

## 3. El estadístico de prueba:

Un sol estadístico calculado a partir de la muestra que nos permita rechazar o no rechazar  $H_0$ , y

## 4. 1 Valor-p:

Una probabilidad, calculada a partir de la estadística de prueba que mide si la estadística de prueba es probable o improbable, suponiendo que  $H_0$  es verdadera.

## 4.2 La región de rechazo:

Una regla que nos dice para qué valores de la estadística de prueba, o para los que p-valores, la hipótesis nula debe ser rechazada.



# *Partes de una prueba estadística (o un test estadístico)*

---

## 5. Conclusión:

"Rechazar  $H_0$ " o "No rechazar  $H_0$ ", junto con una declaración sobre la fiabilidad de su conclusión.

¿Cómo decidir cuándo rechazar  $H_0$ ?

Depende del nivel de significancia,  $\alpha$ , el riesgo máximo tolerable que desea tener de cometer un error, si decide rechazar  $H_0$ .

Por lo general, el nivel de significancia es  $\alpha = 0.01$  o  $\alpha = 0.05$ .



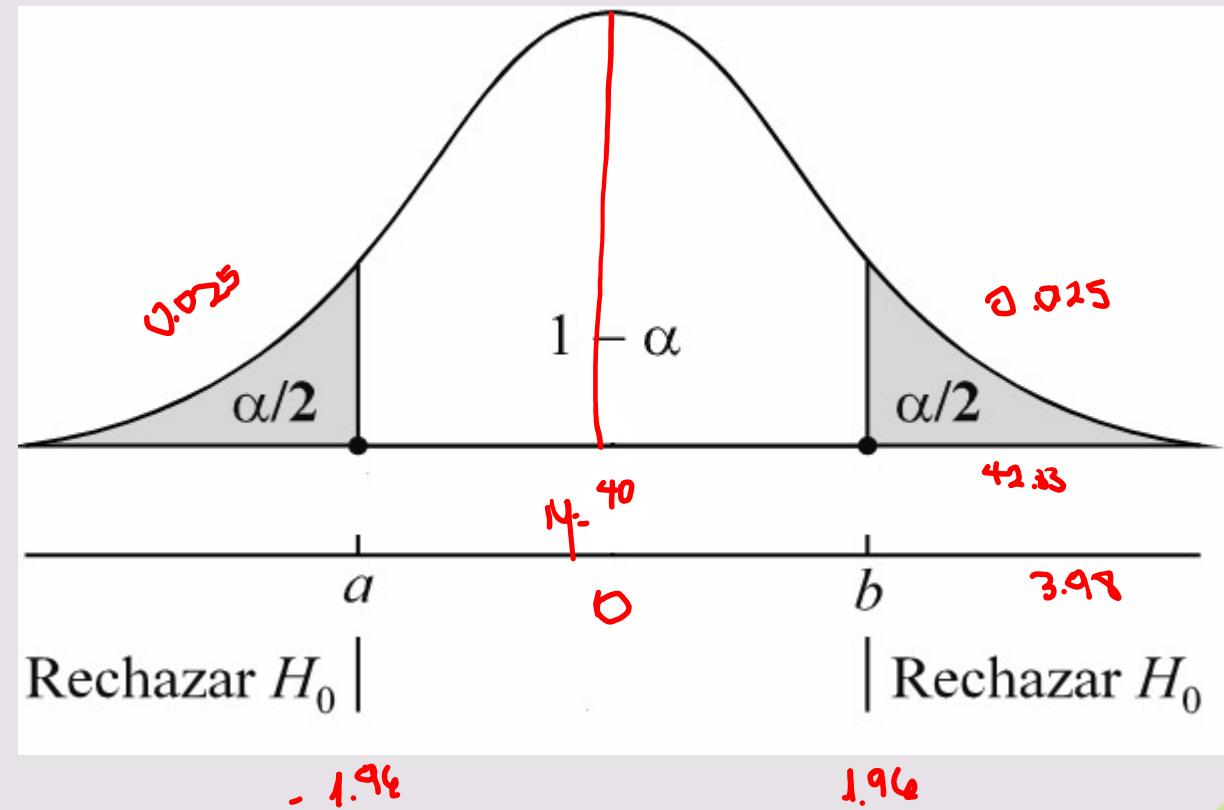
# Tipos de prueba de hipótesis

- Prueba bilateral o de dos colas:

$$\boxed{H_0 : \theta = \theta_0}$$

$$\boxed{H_1 : \theta \neq \theta_0}$$

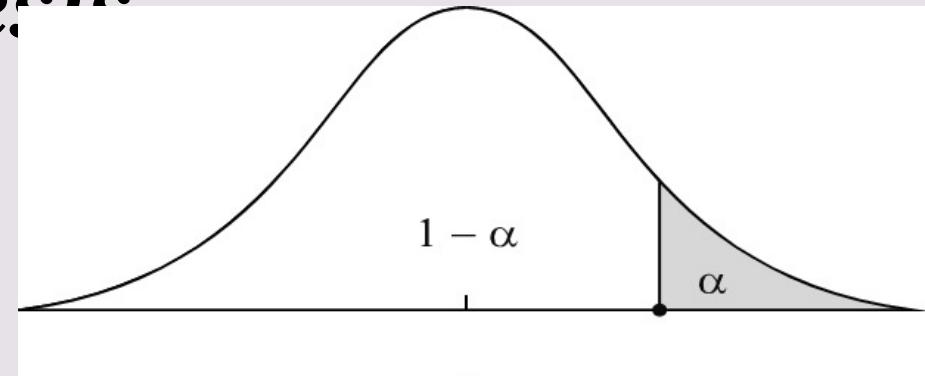
$$\alpha = 0.05$$
  
$$1 - \alpha = \text{Confianza}$$



# Tipos de prueba de hipótesis

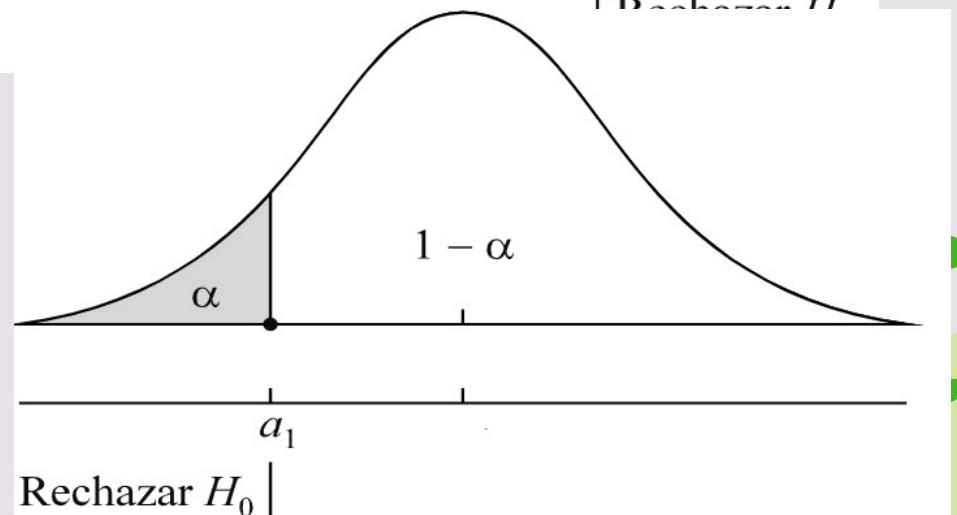
■ Prueba unilateral derecha:

$$H_0 : \theta \leq \theta_0$$
$$H_1 : \theta > \theta_0$$



- Prueba unilateral izquierda:

$$H_0 : \theta \geq \theta_0$$
$$H_1 : \theta < \theta_0$$



# *Estadístico de Prueba*

---

$\bar{x}$

Supongamos en primer lugar que  $H_0$  es verdadera. La media muestral  $\bar{x}$  es nuestra mejor estimación de  $m$ , y la usamos en una forma estandarizada como la estadística de prueba:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \approx \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Dado que  $\bar{x}$  tiene una distribución normal aproximada, con media  $\mu_0$  y una desviación estándar  $\sigma / \sqrt{n}$

# *Otras pruebas de hipótesis para muestras grandes*

---

Hay otras estadísticas que utilizamos para estimar parámetros poblacionales.

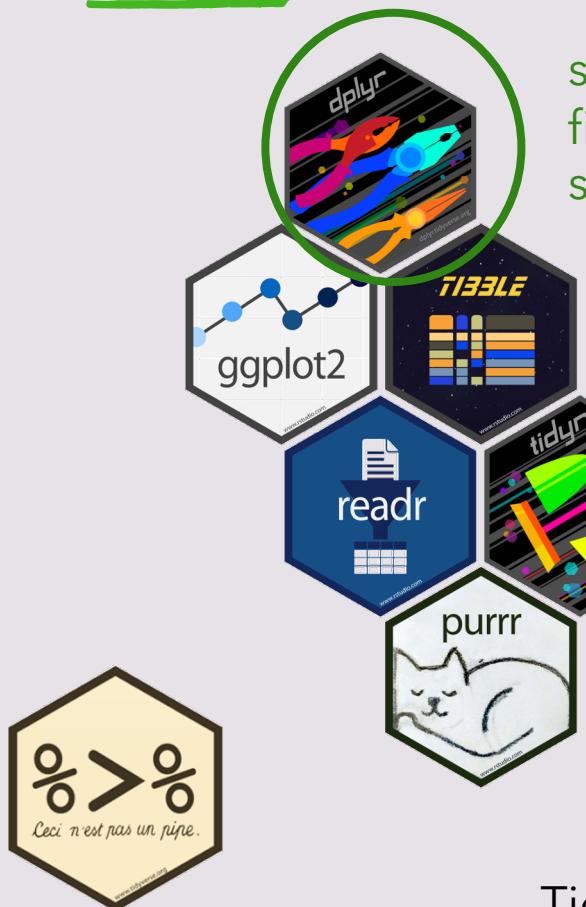
Estas estadísticas tenían distribuciones aproximadamente normales cuando el tamaño (s) de la muestra era grande.

Estas mismas estadísticas se pueden utilizar para poner a prueba hipótesis sobre esos parámetros, el uso de la estadística de prueba en general:

$$z = \frac{\text{estadístico calculado} - \text{valor hipotético}}{\text{error estándar del estadístico}}$$



# *Después de la sesión 2: Repaso de los paquetes*



select()  
filter()  
summarise()



Tidyverse

# *Diferencia entre dos medias*

---

Comparamos los dos promedios al hacer inferencias sobre  $\mu_1 - \mu_2$ , la diferencia en las dos medias poblacionales.

Si las dos medias de población son las mismas, entonces  $\mu_1 - \mu_2 = 0$ .

La mejor estimación de  $\mu_1 - \mu_2$  es la diferencia de las dos medias de la muestra,

$$\bar{x}_1 - \bar{x}_2$$



# *Diferencia entre dos medias*

A veces estamos interesados en comparar las medias de dos poblaciones.

- El promedio de crecimiento de las plantas alimenta mediante dos nutrientes diferentes.
- Las puntuaciones medias de los estudiantes enseñados con dos métodos de enseñanza diferentes.

Para hacer esta comparación necesitamos,

Una muestra aleatoria de tamaño  $n_1$  tomada de población 1 con media  $\mu_1$  y varianza  $\sigma_1^2$ .

Una muestra aleatoria de tamaño  $n_2$  tomada de población 2 con media  $\mu_2$  y varianza  $\sigma_2^2$ .

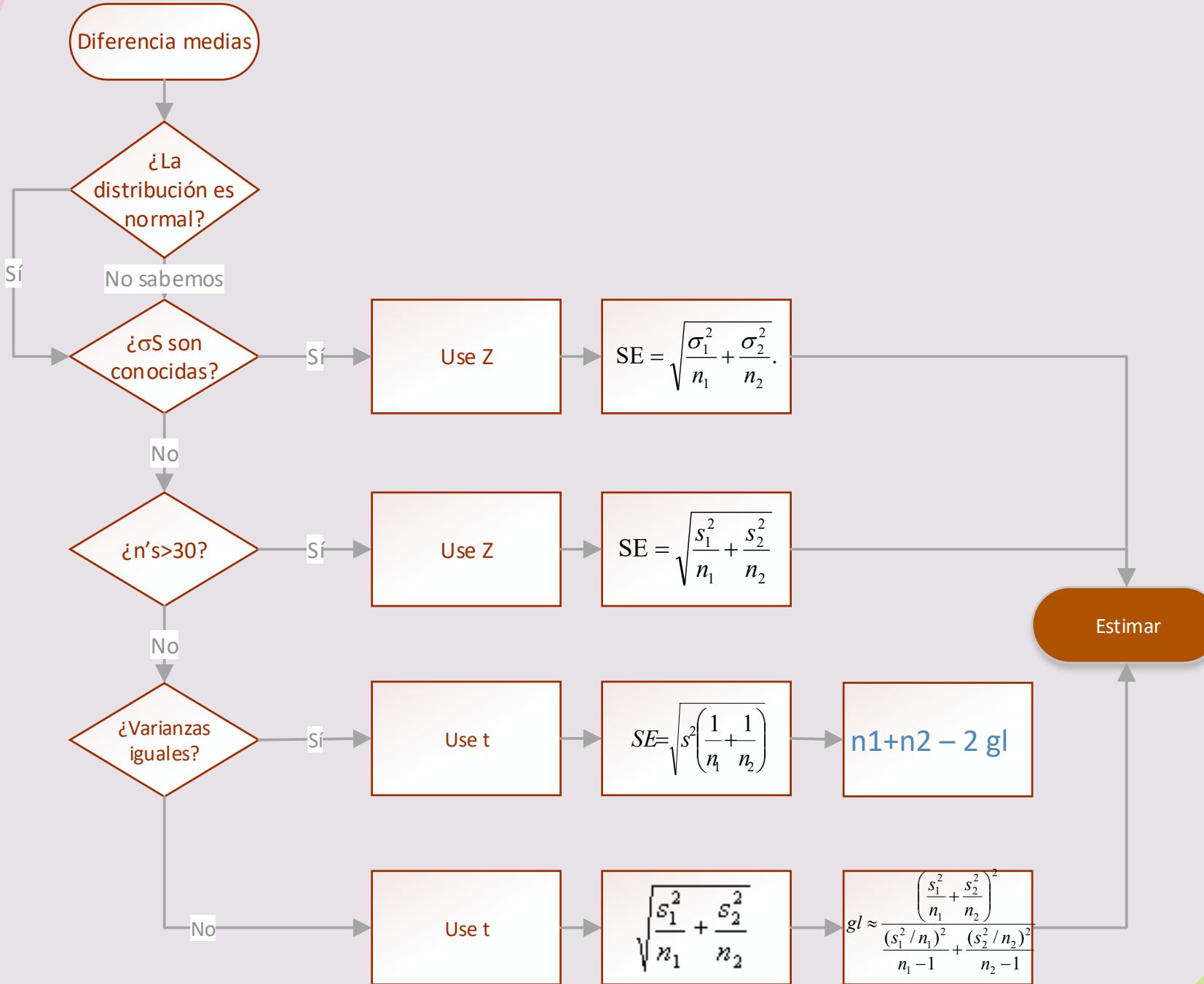
# *Estimando $\mu_1 - \mu_2$*

---

Para muestras grandes, las estimaciones puntuales y su margen de error, así como los intervalos de confianza se basan en la (z) la distribución normal estándar.

Intervalo de confianza para  $\mu_1 - \mu_2$  :

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



Estimador	Parámetro	Distribución
$\bar{x}$	$\mu$	Z o T-student
$\bar{x}_1 - \bar{x}_2$	$\mu_1 - \mu_2$	Z o T-student
r	$\rho$	Z o T-student
$\hat{\beta}_0, \hat{\beta}_i$	$\beta_0, \beta_1$	Z o T-student
$\hat{p}$	p	Z
$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	Z
$s^2$	$\sigma^2$	$\chi^2$
$\frac{s_1^2}{s_1^2}$	$\frac{\sigma_1^2}{s_1^2}$	F

$$y = b_0 + b_1 x$$

Elementos a considerar:

Muestreo aleatorio  
Tamaño de la muestra  
-- Varianzas



$\Phi_{11}$      $f_2$      $t$

# Prueba chi cuadrado

Cuando tenemos dos variables cualitativas o nominales podemos hacerla prueba chi-cuadrado o prueba de independencia. Ésta tiene unalógica un poco diferente a las pruebas que hemos hecho hasta el momento porque proviene de comparar la distribución de los datos, dado que no hay independencia entre las variables y los datos que se tienen.

chi-cuadrada cuyos grados de libertad están dados por  $(r - 1) * (c - 1)$ .

La hipótesis nula postula una distribución de probabilidad totalmente especificada, como el modelo matemático de la población que ha generado la muestra, por lo que, si la rechazamos, habrá evidencia estadística de la dependencia de las dos variables.

Ho: Las variables son independientes  
Ha: las variables no son independientes

$$\chi^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

# *Modelos estadísticos*

---

De forma habitual, cuando quienes investigan se plantean un diseño experimental y comienza con la recogida de datos es porque persigue el estudio de o verificación de un objetivo planteado sobre la población bajo estudio. Estos objetivos se suelen establecer en base a teorías o hipótesis que se desean verificar sobre le funcionamiento de la población bajo ciertas condiciones experimentales. Por ejemplo:

- Teorías que establezcan la posible relación entre dos características de la población.
- Teorías que plateen la idea de comportamientos distintas para una característica de la población en función de una variable que clasifica a los sujetos bajo estudio en diferentes grupos.

Es entonces cuando la modelización estadística interviene y el analista busca el mejor modelo que ajusta los datos disponibles y proporciona predicciones fiables.

El objetivo de la modelización estadística es el planteamiento de una expresión matemática que representa el comportamiento general de la población bajo estudio, teniendo en cuenta el diseño experimental establecido y el objetivo u objetivos que se desean verificar



# **Modelos estadísticos**

En función del tipo de variable respuesta, las predictoras, de la relación que se pueden establecer entre ellas a través de  $f$ , y del establecimiento de las estructuras aleatorias  $F$  para los errores tendremos diferentes tipos de modelos. A lo largo de esta materia veremos las diferentes posibilidades de modelización. A lo largo de las unidades siguientes iremos estudiando las características de los diferentes modelos, pero estos se pueden agrupar en dos grandes apartados:

- **Modelos Lineales (LM)**, que engloban los modelos de regresión, los modelos ANOVA y los modelos ANCOVA.
- **Modelos Lineales Generalizados (GLM)**, que engloba los modelos de respuesta binomial (modelos de regresión logística), modelos de respuesta poisson, modelos para tablas de contingencia (modelos log-lineales), y modelos de supervivencia.

Introduciremos además los modelos de suavizado y una breve introducción a los modelos de efectos aleatorios, que pueden ser utilizados en conjunción con los LM y los GLM.

# *Proceso*

El proceso de modelización y análisis estadístico de un banco de datos se puede estructurar según las siguientes pautas de actuación:

1. Contextualización del problema. Definición de objetivos y variables.
2. Diseño del experimento y recogida de información.
3. Registro y procesado previo de la información disponible.
4. Inspección gráfica e identificación de tendencias.
5. Consideración de hipótesis distribucionales y relacionales. Propuesta de modelización.
6. Ajuste del modelo. Comparación y selección del mejor modelo.
7. Diagnóstico y validación del modelo ajustado.
8. Valoración de la capacidad predictiva del modelo y predicción.
9. Interpretación y conclusiones.

# *El Análisis de Varianza (ANOVA)*

---

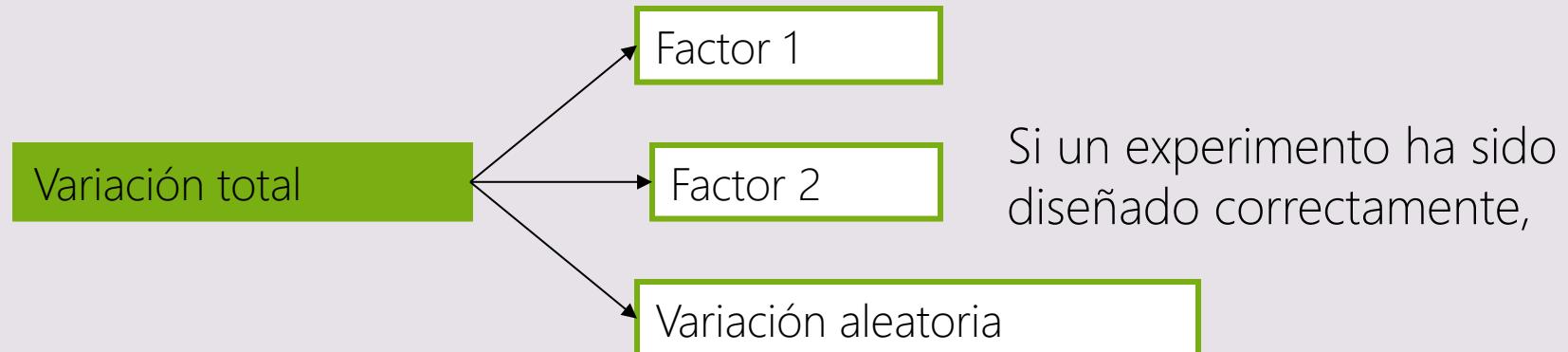
Todas las mediciones muestran una **variabilidad**.

La variación total de las medidas de respuesta de un experimento se pueden dividir en partes que pueden atribuirse a diversos **factores**.

Estas porciones se utilizan para juzgar el **efecto** de los diversos factores sobre la respuesta experimental.



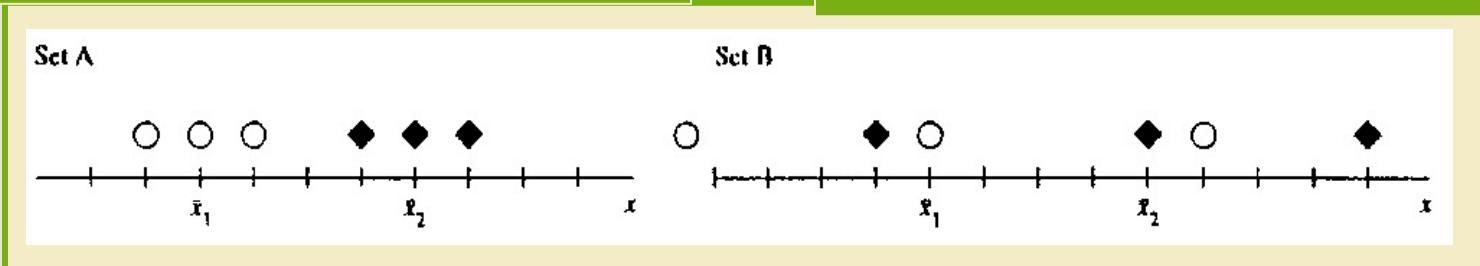
# Análisis de varianza



- Comparamos la variación debida a un factor a la variación aleatoria típica en el experimento

La variación entre las medias de la muestra es mayor que la variación típica dentro de las muestras.

La variación entre las medias de la muestra es aproximadamente la misma que la variación típica dentro de las muestras.



# Supuestos

---

1. Las observaciones dentro de cada población se distribuyen normalmente con una varianza común  $\sigma^2$ .
2. Los supuestos con respecto a los procedimientos de muestreo se especifican para cada diseño.

Los análisis de los procedimientos de la varianza son bastante robustos cuando los tamaños de muestra son iguales y cuando los datos son bastante en forma de montículo.



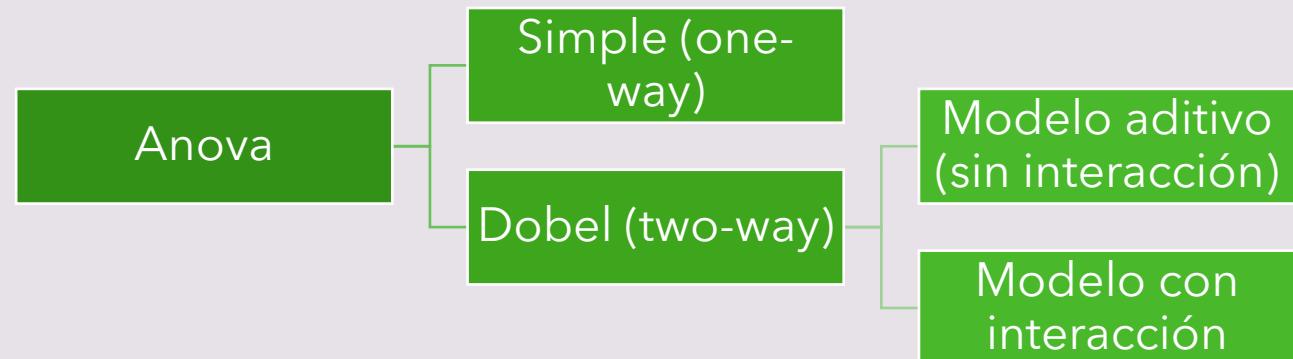
# Tres diseños

---

Diseño completamente al azar: una extensión de los dos muestra independiente t-test.

Diseño de bloques al azar: una extensión de la prueba de diferencias por pares.

a × b experimento factorial : estudiamos dos factores experimentales y su efecto en la respuesta.



# *Anova simple –one way*

## **El objetivo**

determinar el efecto que sobre alguna variable dependiente Y tienen distintos niveles de algún factor X (variable independiente y discreta).

## **El factor**

puede ser la temperatura, la empresa que ha producido el bien, el día de la semana, etc.

## **ANOVA implica**

obtener muestras aleatorias e independientes del valor de Y asociado a cada uno de los distintos niveles del factor X<sub>1</sub>, X<sub>2</sub>,..., X<sub>n</sub>.

Para determinar si los diferentes niveles del factor tienen un efecto **significativo** sobre el valor de la variable dependiente.

# *Anova simple –one way*

---

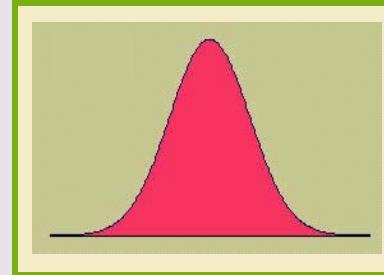
Se compara las medias de Y asociadas a los distintos niveles del factor (X<sub>1</sub>, X<sub>2</sub>,..., X<sub>n</sub>), compararemos una medida de la variación entre diferentes niveles con una medida de la variación dentro de cada nivel

Si el Suma de Cuadrados del Factor es significativamente mayor que el Suma de cuadrados de los errores concluiremos que las medias asociadas a diferentes niveles del factor son distintas.

Esto significa que el factor influye significativamente sobre la variable dependiente Y.

Si, por el contrario, el MS-factor no es significativamente mayor que el MS-error, no rechazaremos la hipótesis nula de que todas las medias, asociadas a diferentes niveles del factor, coinciden.



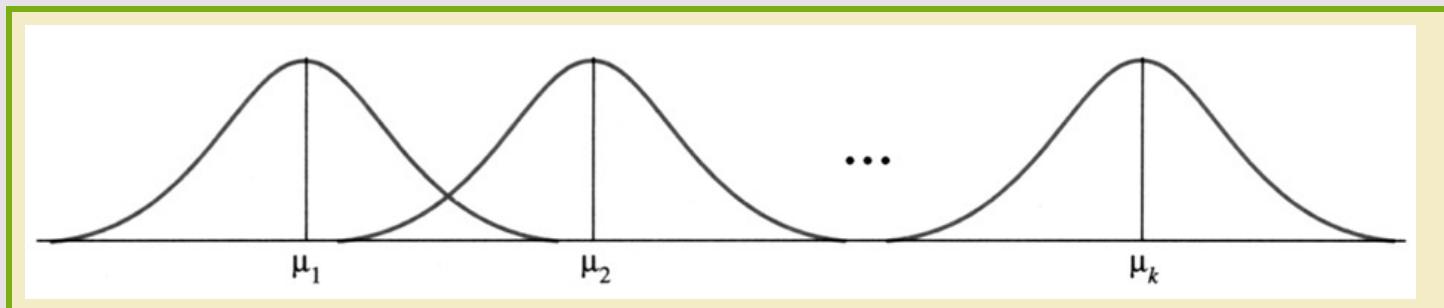


## Anova simple –one way

*One way*, implica una sola clasificación, en la que el factor k se fija en diferentes niveles.

Los niveles de k corresponden a k diferentes poblaciones normales, que son los “tratamientos”.

¿Las medias de k son diferentes, o al menos una es diferente a las demás?



# *Anova simple –one way*

---

Muestras aleatorias de tamaño  $n_1, n_2, \dots, n_k$  son extraídas de  $k$  poblaciones con medios  $\mu_1, \mu_2, \dots, \mu_k$  y con  $\sigma^2$  varianza común.

Sea  $x_{ij}$  la medición  $j$ -ésimo en la muestra de orden  $i$ .

La variación total en el experimento se mide por la suma total de cuadrados:

$$\text{Total SS} = \sum(x_{ij} - \bar{x})^2$$

*Suma de cuadrados*

# *Análisis de varianza o anova*

---

El total de la SS se divide en dos partes:

- **SST** (suma de los cuadrados de los tratamientos): mide la variación entre los medios k muestras.
- **SSE** (suma de los cuadrados del error): mide la variación dentro de las k muestras.

de una manera que:

$$\text{Total SS} = \text{SST} + \text{SSE}$$

Between + within  
Entre + dentro

# *Grados de libertad y media de los cuadrados*

---

Estas sumas de cuadrados se comportan como el numerador de una varianza de la muestra. Cuando dividida por los grados de libertad apropiados, cada uno proporciona un cuadrado medio, una estimación de la variación en el experimento.

Los grados de libertad son aditivos, al igual que las sumas de cuadrados.

$$\text{Total } gl = \text{Trt } gl + \text{Error } gl$$



# *La Tabla anova*

$$n_1+n_2+\dots+n_k - 1 = n - 1$$

Total df =

$$k - 1$$

Mean Squares

$$MST = SST/(k-1)$$

Tratamiento df =

$$n - 1 - (k - 1) = n - k$$

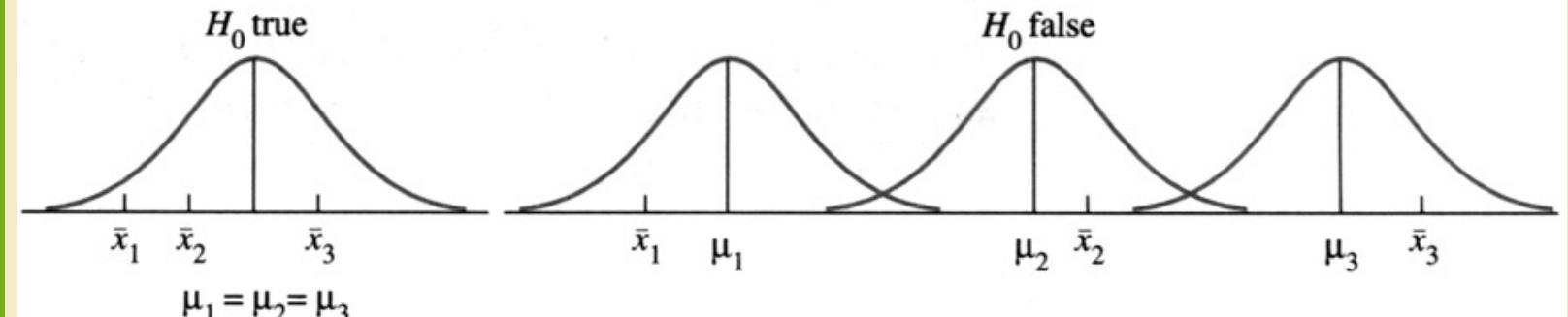
$$MSE = SSE/(n-k)$$

Error df =

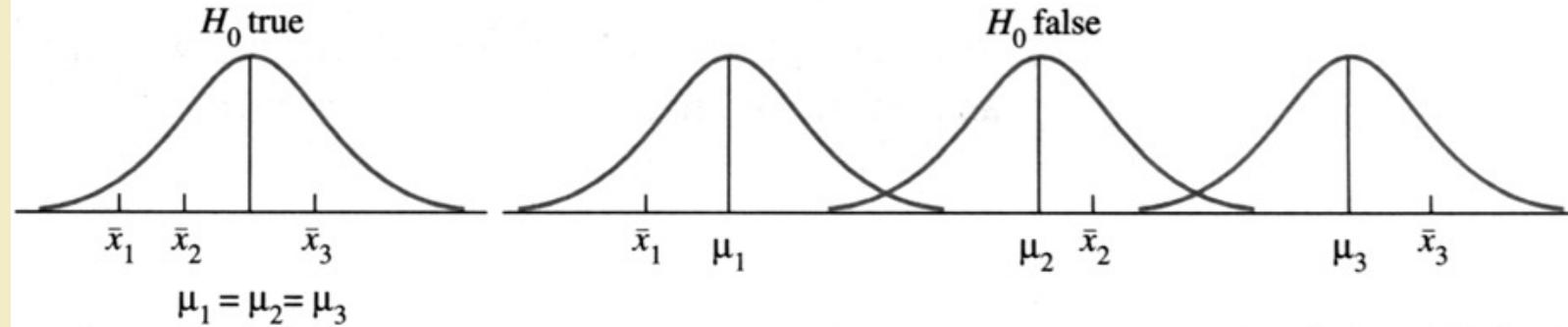
Fuente	df	SS	MS	F
Tratamientos	k - 1	SST	SST/(k-1)	MST/MSE
Error	n - k	SSE	SSE/(n-k)	
Total	n - 1	Total SS		

$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$  versus

$H_a : \text{al menos una es diferente}$



Recuerde que  $\sigma^2$  es la varianza común para todas las  $k$  poblaciones. La  $\text{MSE} = \text{SSE} / (n - k)$  es una estimación combinada de  $\sigma^2$ , a través de una media ponderada de todas las variaciones  $k$  muestras, si  $H_0$  es verdadera.



Si  $H_0$  es cierto, entonces la variación de las medias de la muestra, medida por  $MST = [SST / (k - 1)]$ , también proporciona una estimación imparcial de  $\sigma^2$ .

Sin embargo, si  $H_0$  es falso y las medias poblacionales son diferentes, entonces  $MST$ - que mide la variación entre las muestras- es inusualmente grande. El estadístico de prueba  $F = MST / MSE$  tiende a ser más grandes que lo habitual.