Introducción a la Bioestadística

Marta R. Hidalgo
Unidad de Bioinformática y Bioestadística

Octubre 2019

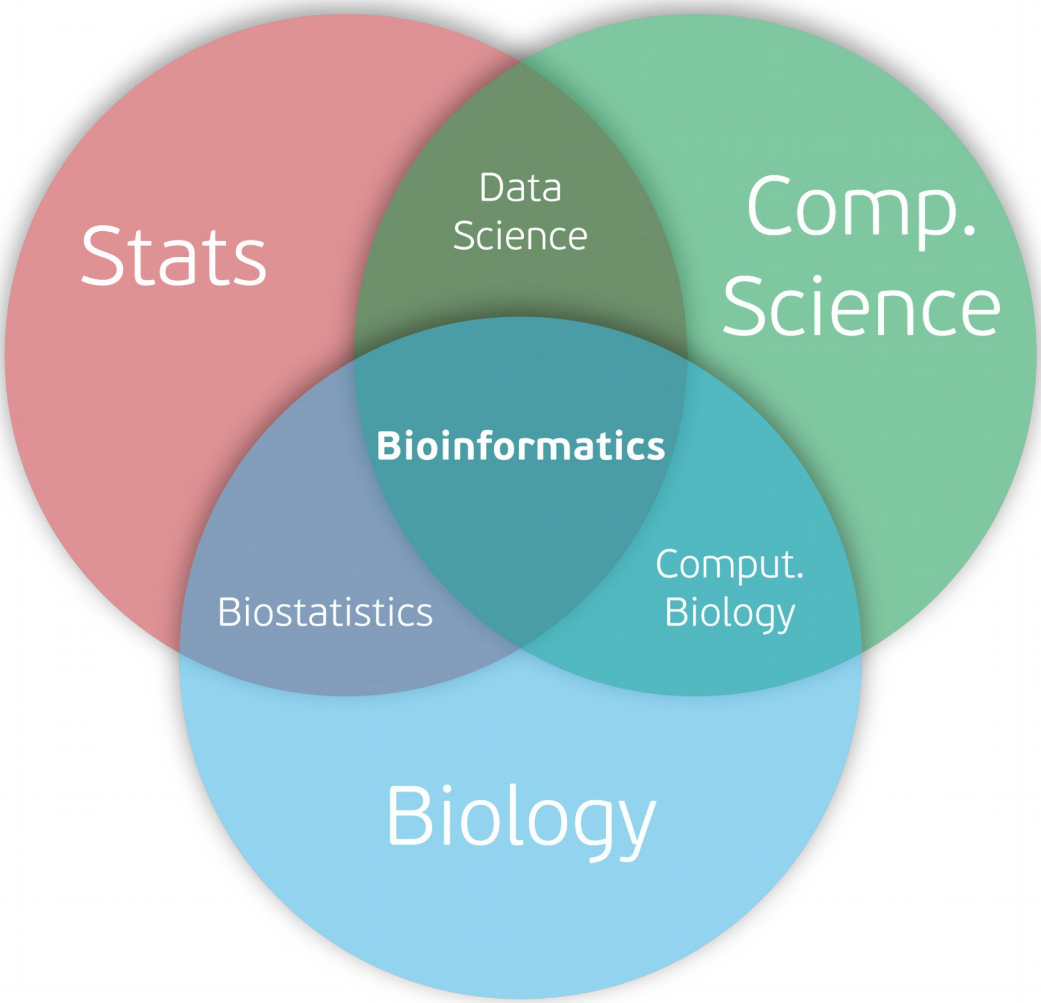
Unidad de Bioinformática y Bioestadística

PRINCIPE FELIPE CENTRO DE INVESTIGACION

CSIC Consejo Superior de Investigaciones Científicas

WODA
WEB-BASED OMICS DATA ANALYSIS

Bioinformática



Estadística

- Estadística
  - Ciencia cuyo objetivo es recoger, agrupar, presentar, analizar e interpretar **datos** (con variabilidad) mediante el uso de herramientas matemáticas e informáticas para extraer de ellos la máxima **información**, así como determinar el grado de **fiabilidad** de las conclusiones obtenidas.

Estadística

- Estadística
  - Ciencia cuyo objetivo es recoger, agrupar, presentar, analizar e interpretar **datos** (con variabilidad) mediante el uso de herramientas matemáticas e informáticas para extraer de ellos la máxima **información**, así como determinar el grado de **fiabilidad** de las conclusiones obtenidas.

Población

Estadística

- Estadística
  - Ciencia cuyo objetivo es recoger, agrupar, presentar, analizar e interpretar **datos** (con variabilidad) mediante el uso de herramientas matemáticas e informáticas para extraer de ellos la máxima **información**, así como determinar el grado de **fiabilidad** de las conclusiones obtenidas.

Estadística

- Estadística
  - Ciencia cuyo objetivo es recoger, agrupar, presentar, analizar e interpretar **datos** (con variabilidad) mediante el uso de herramientas matemáticas e informáticas para extraer de ellos la máxima **información**, así como determinar el grado de **fiabilidad** de las conclusiones obtenidas.

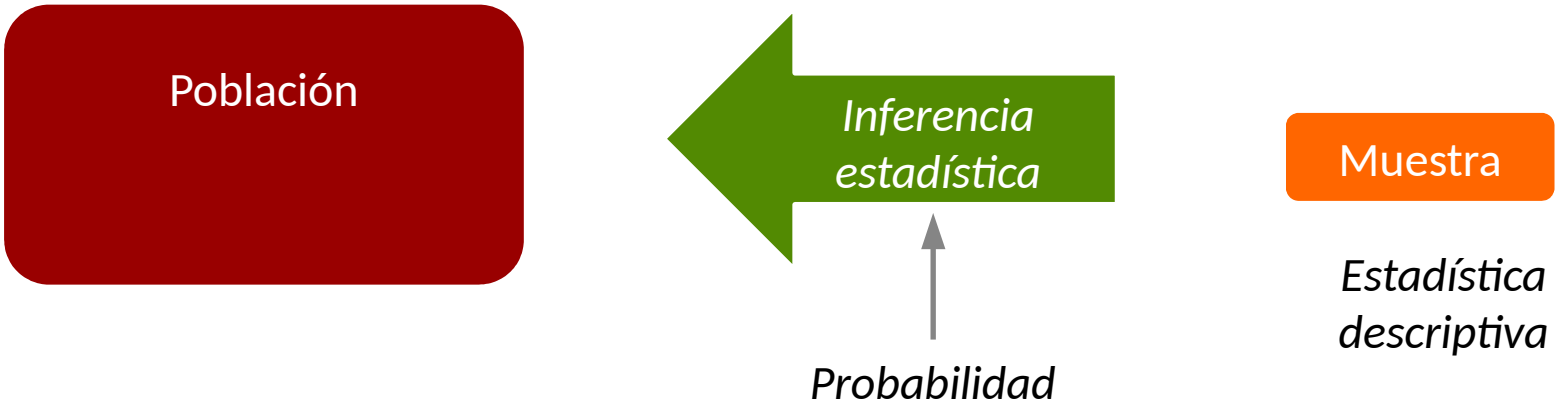
Población

Muestra

Estadística descriptiva

Estadística

- Estadística
  - Ciencia cuyo objetivo es recoger, agrupar, presentar, analizar e interpretar **datos** (con variabilidad) mediante el uso de herramientas matemáticas e informáticas para extraer de ellos la máxima **información**, así como determinar el grado de **fiabilidad** de las conclusiones obtenidas.



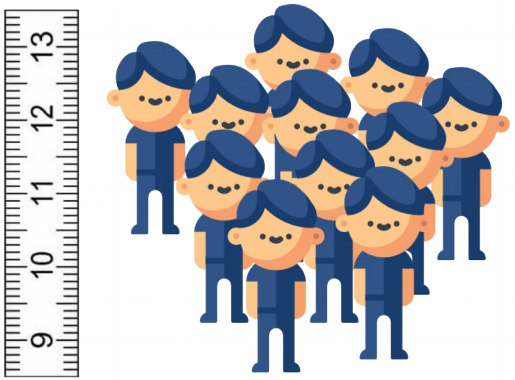
Estadística descriptiva

Estadística descriptiva

- La estadística descriptiva se encarga de resumir y presentar la información contenida en los datos
- Herramientas de la estadística descriptiva
  - Parámetros descriptivos
    - Localización
    - Dispersión
  - Tablas de frecuencias
  - Gráficos

Variables

- Una **variable estadística** es una característica cuya variación es susceptible de adoptar diferentes valores.



	Height
Human 1	1'70
Human 2	1'53
Human 3	2'01
Human 4	1'82
Human 5	1'65
Human 6	1'73
Human 7	1'91
Human 8	1.81

Variables

- Una **variable estadística** es una característica cuya variación es susceptible de adoptar diferentes valores.
- Numéricas
  - Discretas (procedentes de "contar") 0, 1, 2,...
    - *Número de hijos, número de pacientes, número de intervenciones ...*
  - Continuas (procedentes de "medir") nº reales
    - *Peso, altura, temperatura, edad, nivel de colesterol, ...*
- Categóricas
  - Nominal
    - *Sexo, tratamiento, tipo de dieta,...*
  - Ordinal
    - *Nivel de estudios, estadio de una enfermedad,...*

Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico

Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico

	Height
Human 1	1'70
Human 2	1'53
Human 3	2'01
Human 4	1'82
Human 5	1'65
Human 6	1'73
Human 7	1'91
Human 8	1'81
Human 9	1'62

Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico

	Height
Human 1	1'70
Human 2	1'53
Human 3	2'01
Human 4	1'82
Human 5	1'65
Human 6	1'73
Human 7	1'91
Human 8	1'81
Human 9	+ 1'62
	15'78 | 9
	1'753

Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico

	Height
Human 1	1'70
Human 2	1'53
Human 3	2'01
Human 4	1'82
Human 5	1'65
Human 6	1'73
Human 7	1'91
Human 8	1'81
Human 9	1'62

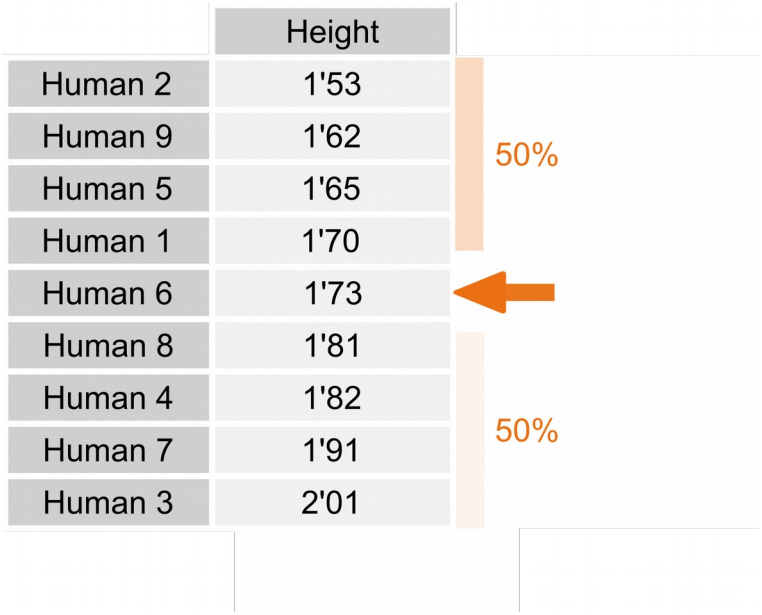
Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico

	Height
Human 2	1'53
Human 9	1'62
Human 5	1'65
Human 1	1'70
Human 6	1'73
Human 8	1'81
Human 4	1'82
Human 7	1'91
Human 3	2'01

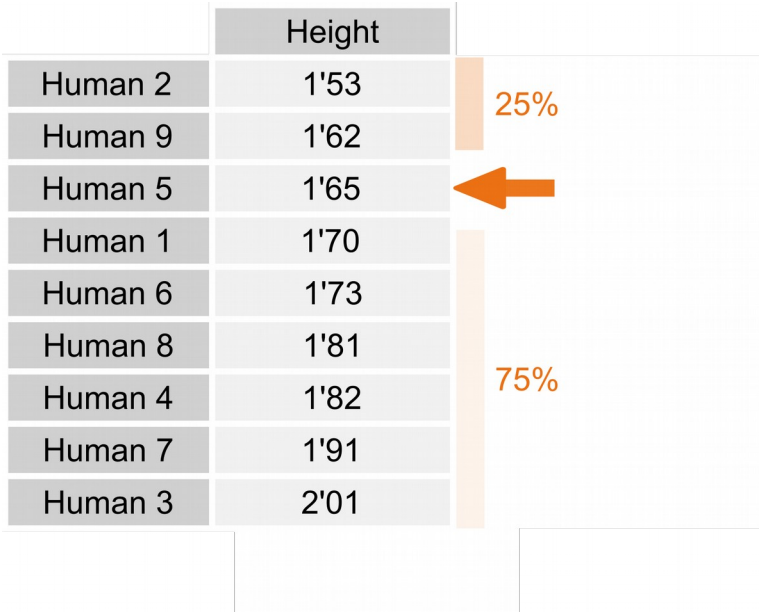
Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico



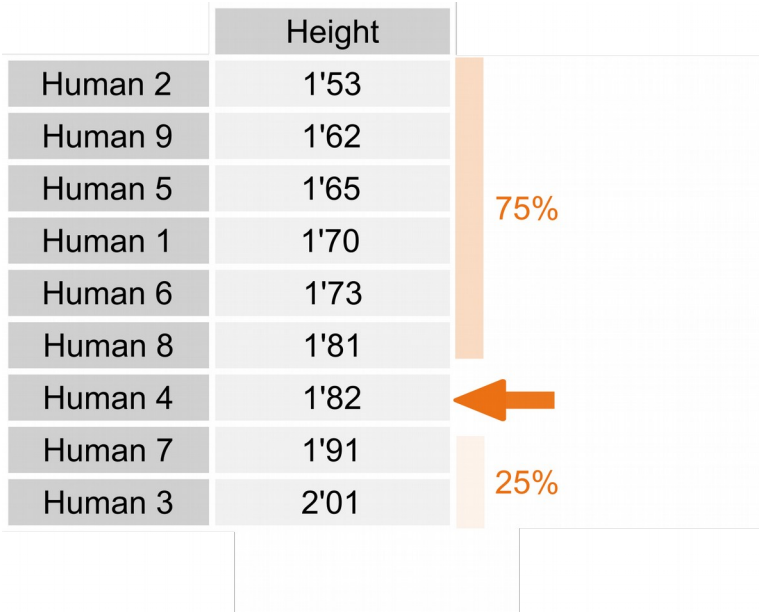
Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico



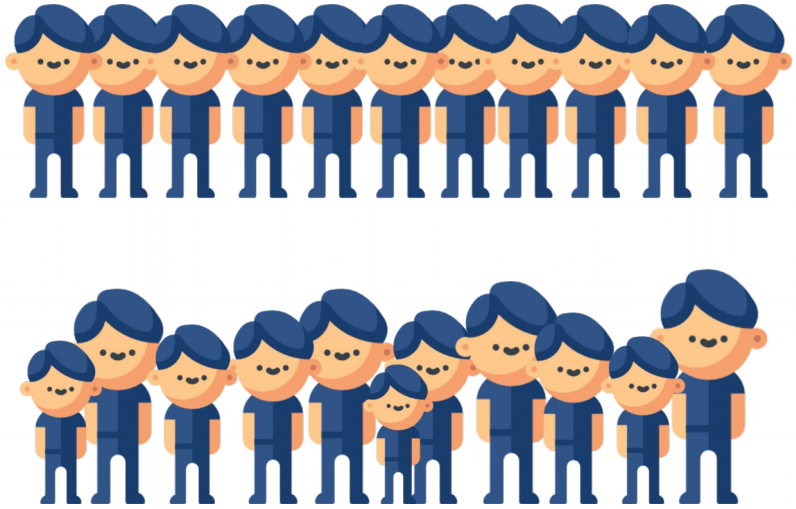
Parámetros descriptivos

Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico



Parámetros descriptivos

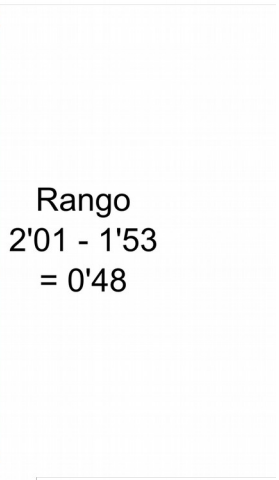
Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico



Parámetros descriptivos

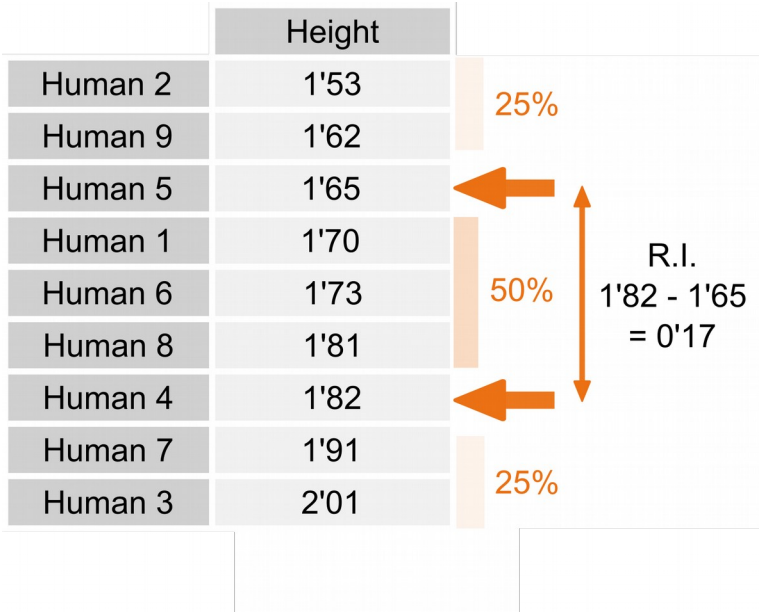
Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico

	Height
Human 2	1'53
Human 9	1'62
Human 5	1'65
Human 1	1'70
Human 6	1'73
Human 8	1'81
Human 4	1'82
Human 7	1'91
Human 3	2'01



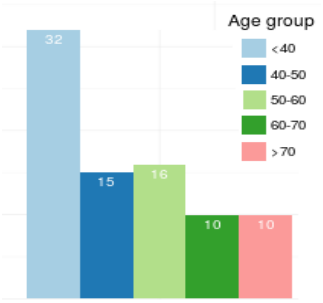
Parámetros descriptivos

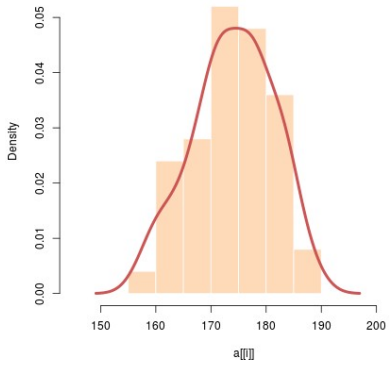
Tipo	Parámetro
Localización	Media
	Mediana
	Percentiles
Dispersión	Varianza
	Desviación típica
	Rango
	Rango intercuartílico

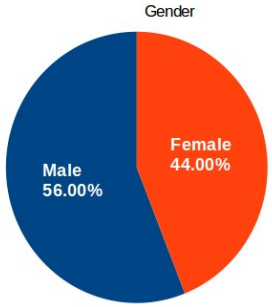


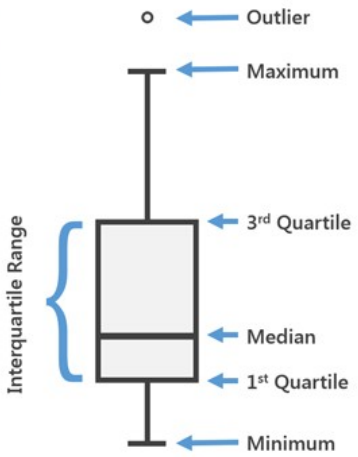
Gráficos

Gráfico	Tipo de datos
Diagrama de barras	Categóricos / Discretos (pocos)
Histograma	Continuos / Discretos (muchos)
Función de densidad	Continuos
Sectores	Categóricos
Boxplot simple	Continuos / Discretos (muchos)
Boxplot múltiple	Como boxplot simple, combinado con var. categórica



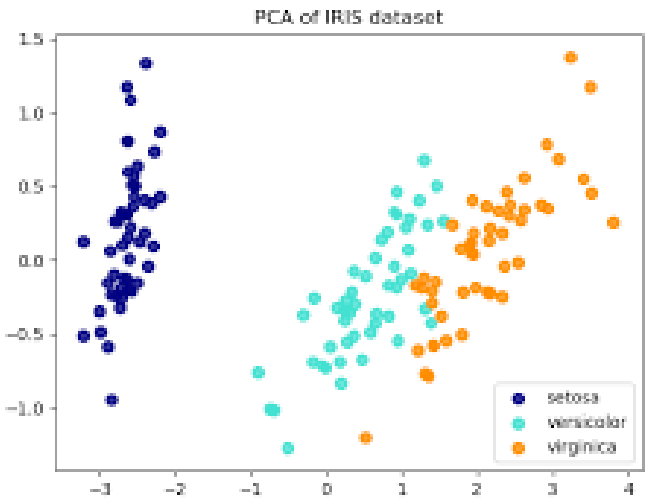






Análisis de Componentes Principales

- Técnica útil cuando se han medido muchas variables y algunas de ellas pueden estar relacionadas entre sí.
- Método de reducción de la dimensión, ya que construye unas “pocas” nuevas variables (llamadas Componentes Principales) que explican la mayor parte de la variabilidad de los datos originales.
- Las componentes principales (PCs) son combinaciones lineales de las variables originales.
- Las muestras parecidas se agruparán



Inferencia estadística

Inferencia estadística

- Rama de la estadística que trata de sacar **conclusiones de la población estudiada a partir de la** información proporcionada por una **muestra** representativa de la misma.

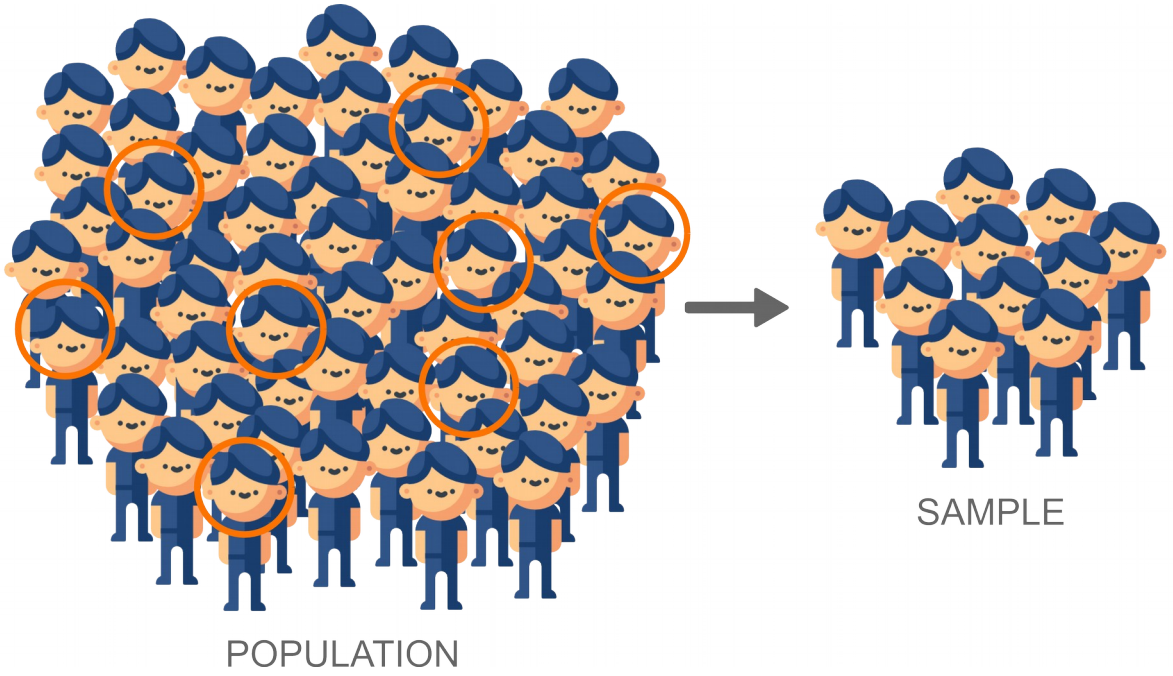
POPULATION

Inferencia estadística

- Rama de la estadística que trata de sacar **conclusiones de la población estudiada a partir de la** información proporcionada por una **muestra** representativa de la misma.

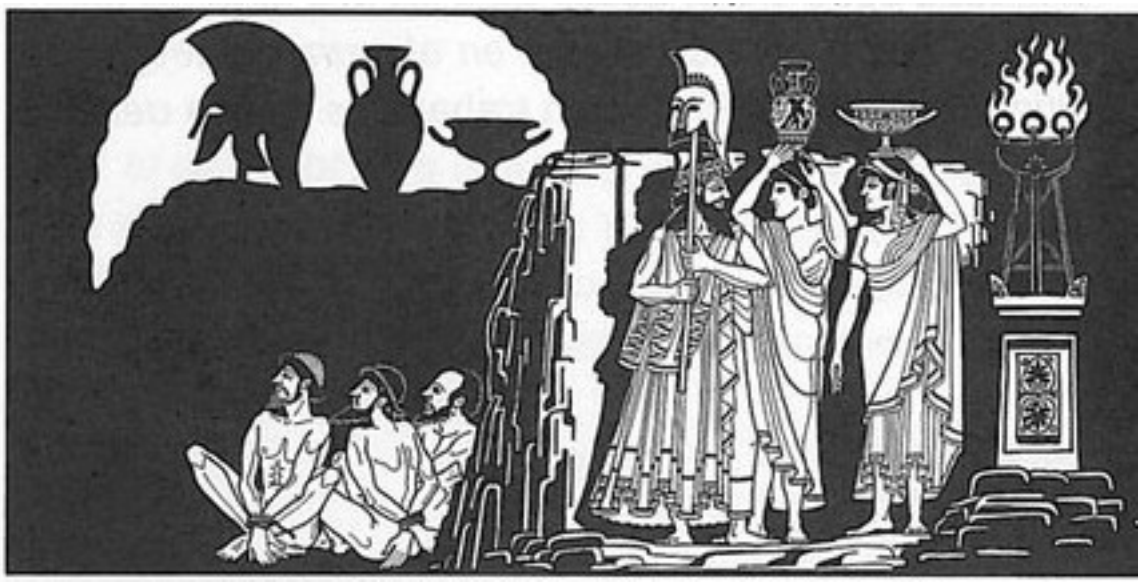
Inferencia estadística

- Rama de la estadística que trata de sacar **conclusiones de la población estudiada a partir de la** información proporcionada por una **muestra** representativa de la misma.



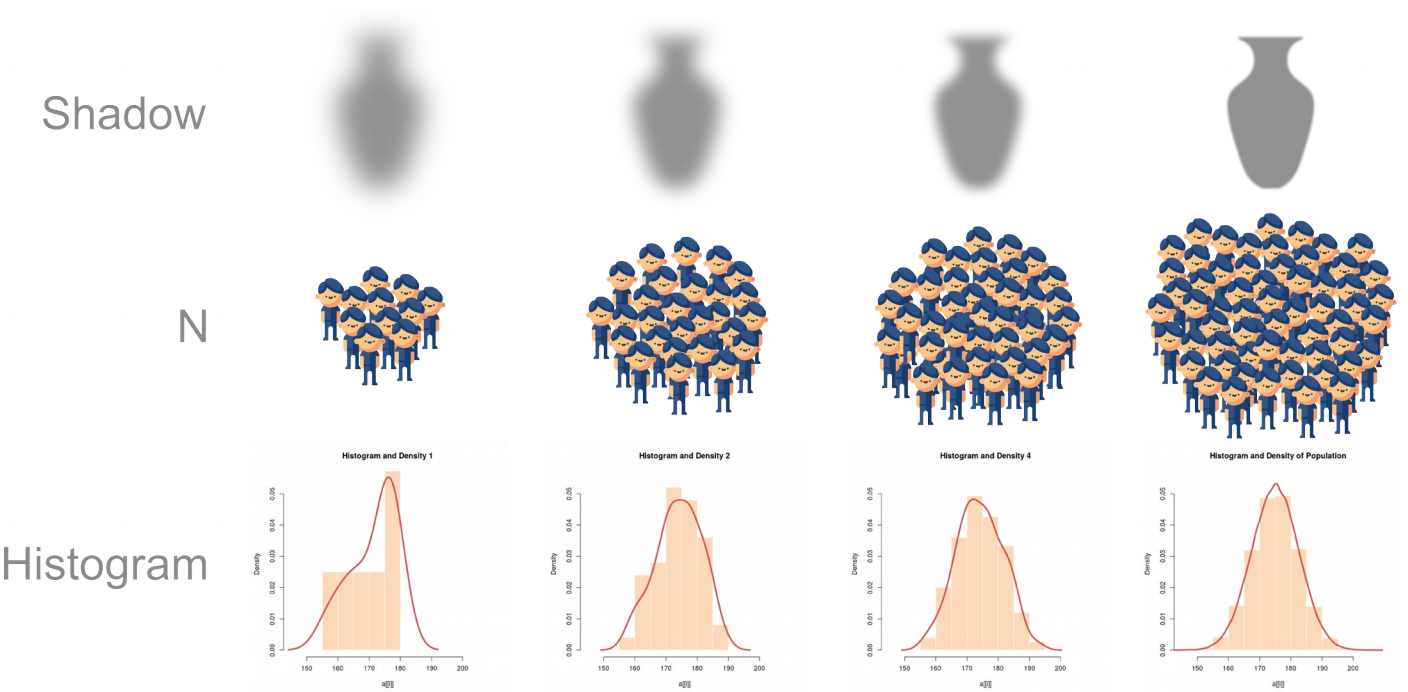
Inferencia estadística

- Rama de la estadística que trata de sacar **conclusiones de la población estudiada a partir de la** información proporcionada por una **muestra** representativa de la misma.



Inferencia estadística

- Rama de la estadística que trata de sacar **conclusiones de la población estudiada a partir de la** información proporcionada por una **muestra** representativa de la misma.



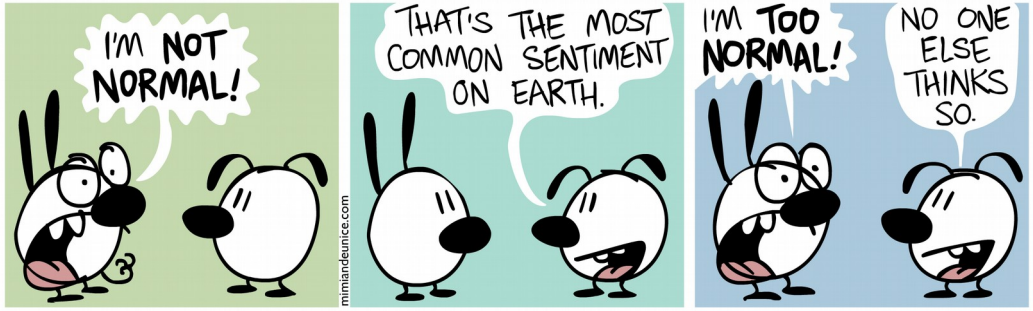
Inferencia estadística

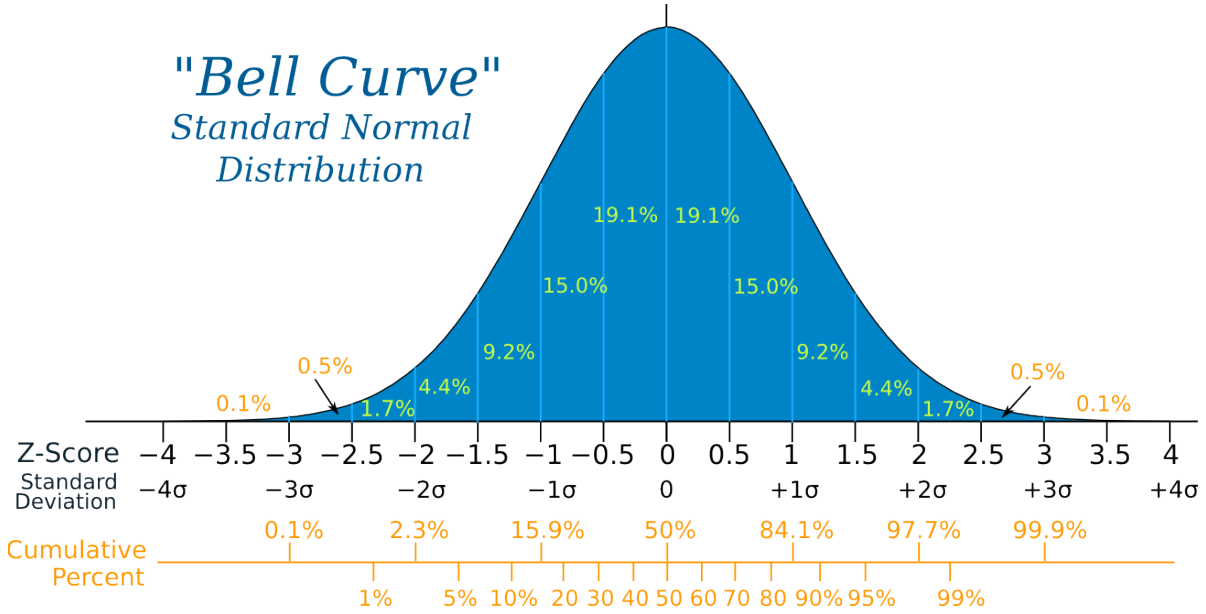
- **Herramientas de la inferencia estadística**:
  - Estimación puntual de un parámetro
    - Para obtener una primera aproximación de su valor
  - Estimación por intervalos
    - Un intervalo de confianza es un intervalo con una probabilidad alta de contener al verdadero valor del parámetro, que es desconocido
  - Contrastes de hipótesis

Métodos de inferencia estadística

- Paramétricos
  - Asumen que los datos siguen una cierta distribución de probabilidad
    - Distribución normal
    - Otras distribuciones
- No paramétricos
  - No asumen ninguna distribución para los datos
  - Suelen tener menos potencia estadística

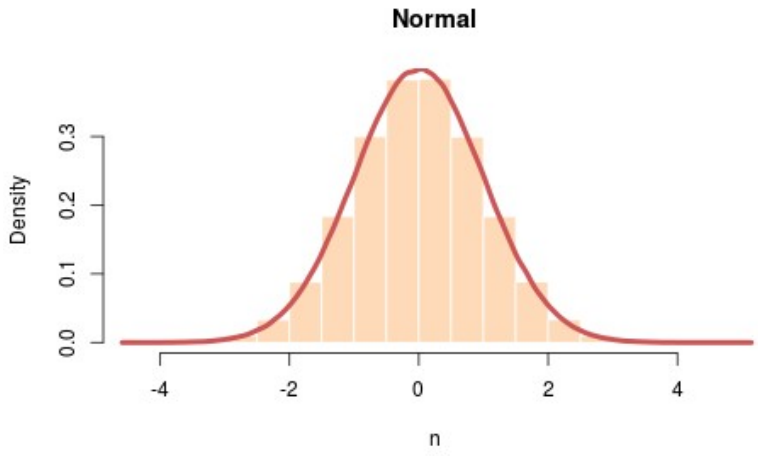
Distribución normal N(μ, σ)





Distribución normal N(μ, σ)

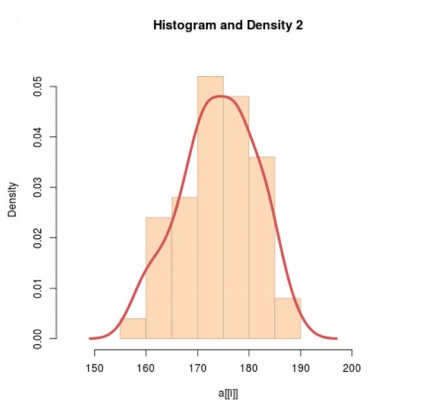
- ¿Cómo comprobar si nuestros datos siguen una distribución normal?
  - Histograma o gráfico de densidad
  - Gráfico probabilístico normal → Los puntos han de ajustarse a una línea recta
  - Test de normalidad (Test de Shapiro)

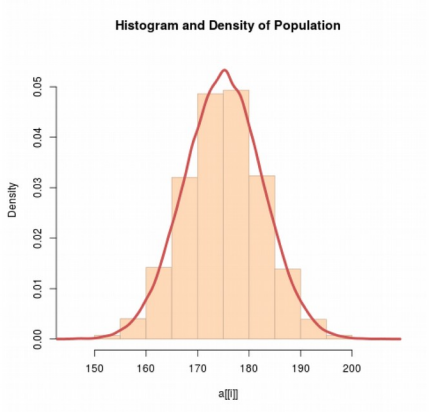


Intervalos de confianza

SAMPLE

POPULATION

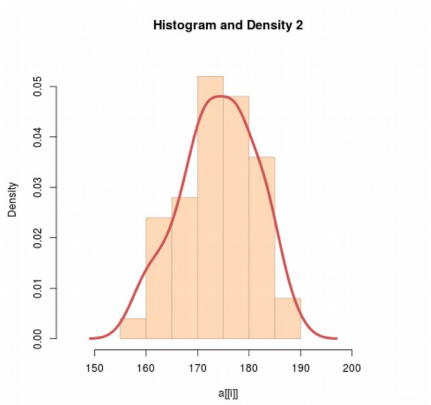


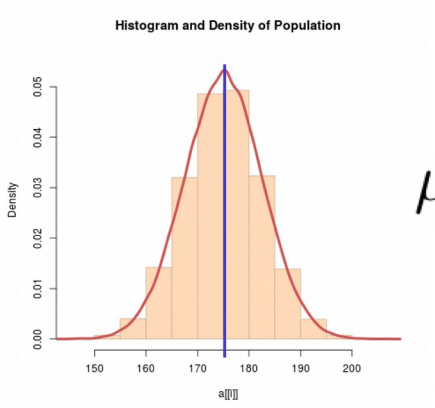


Intervalos de confianza

SAMPLE

POPULATION



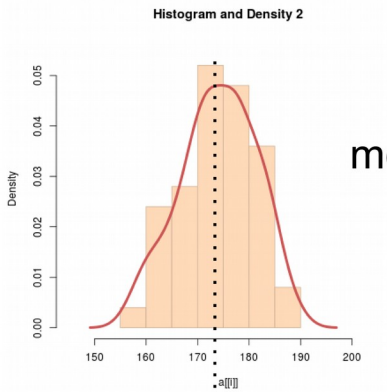


$\mu = 175$

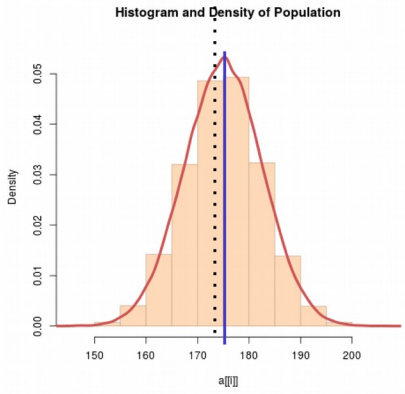
Intervalos de confianza

SAMPLE

POPULATION



mean = 173

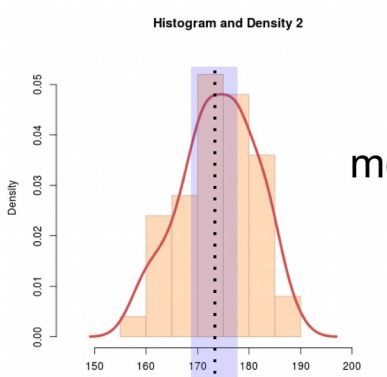


μ = 175

Intervalos de confianza

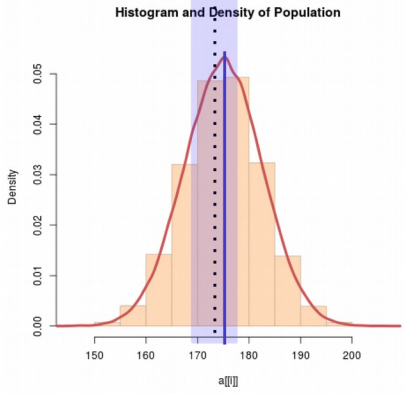
SAMPLE

POPULATION



mean = 173

C.I. = [168, 178]



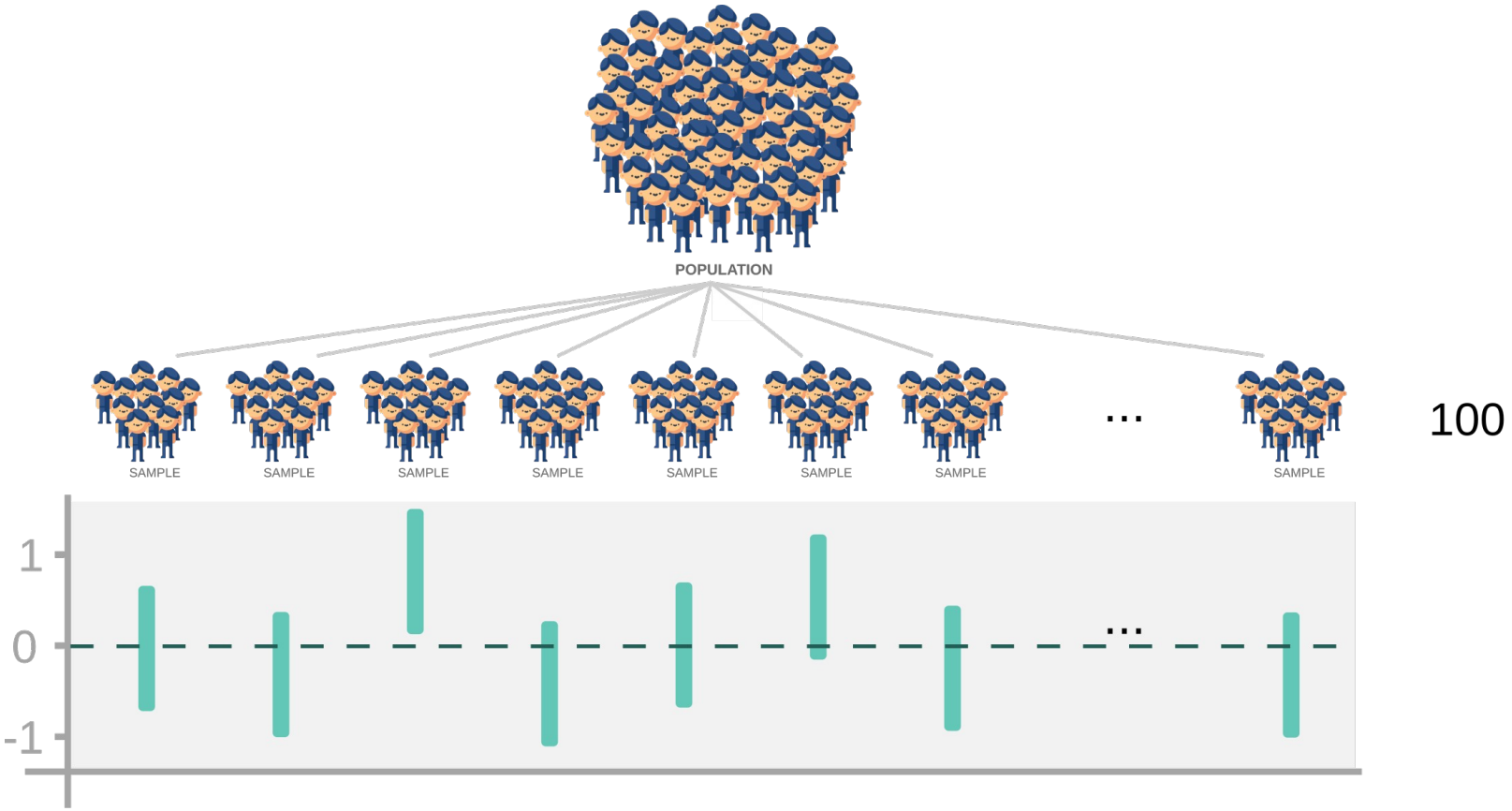
$\mu = 175$

Intervalos de confianza

- Intervalos de confianza
  - Distribución normal del parámetro
  - Nivel de confianza (95%, 90%,...)

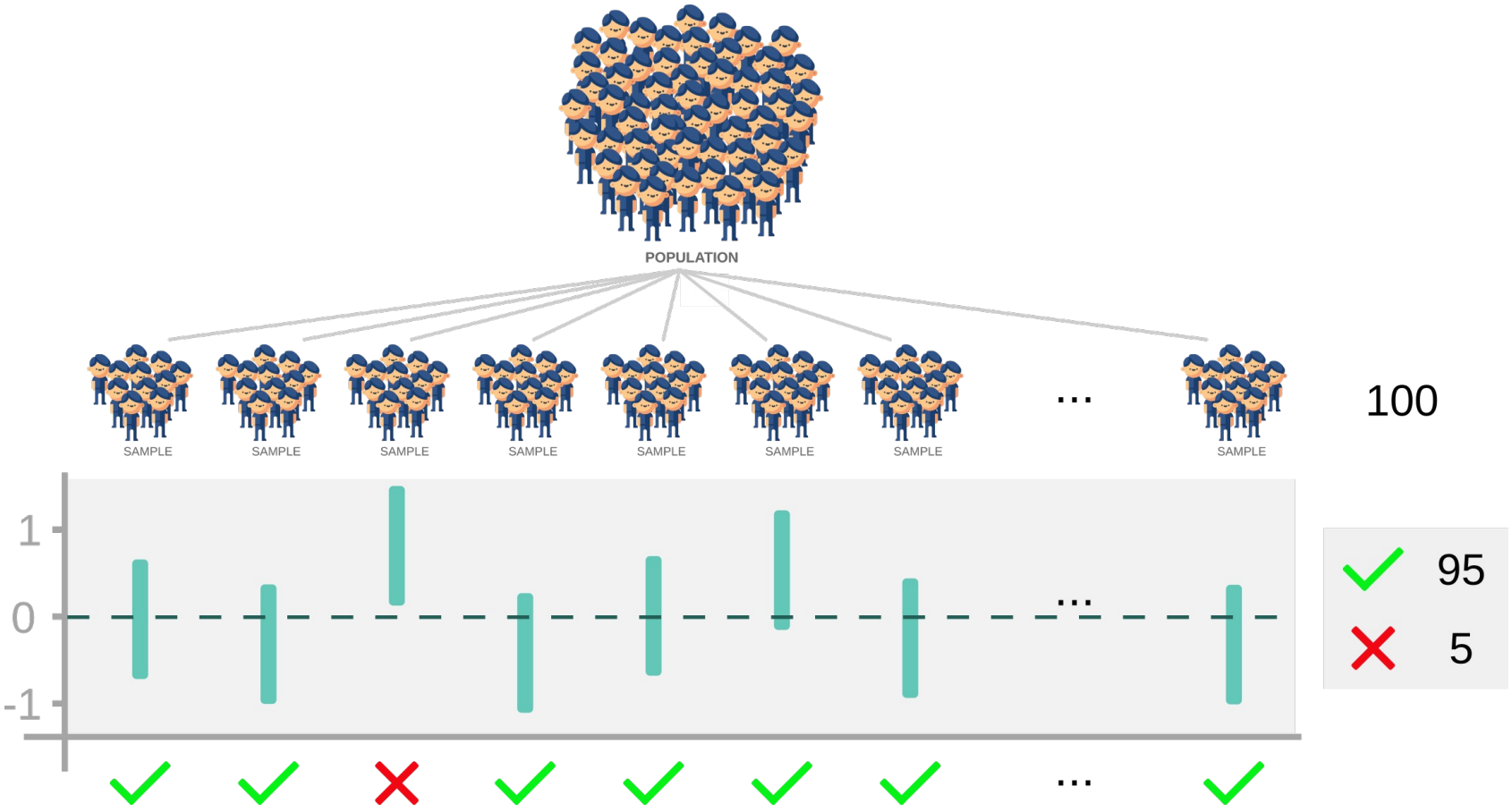
- El tamaño del intervalo dependerá del tamaño muestral, la varianza de los datos y del nivel de confianza elegido.
  - Mayor tamaño muestral → Menor tamaño del intervalo
  - Mayor varianza → Mayor tamaño del intervalo
  - Mayor confianza → Mayor tamaño del intervalo

Intervalos de confianza



100

Intervalos de confianza



Inferencia estadística

Objetivo	Diseño experimental	Parámetro a estudiar	Normalidad	No normalidad	No paramétrico
Comparar poblaciones	2 poblaciones	Media	t-test	Mann-Whitney test, Wilcoxon Signed Rank test	
		Varianza	F-test		
		Proporción	Z-test		
	> 2 poblaciones	Media	ANOVA	Kruskal-Wallis test, Friedman test	
Predecir/explicar una variable respuesta			Regresión lineal	Regresión lineal generalizada	Regresión no paramétrica (pe. Kaplan-Meier)
Relación entre dos o más variables	Categóricas		Fisher's Exact test, Chi2-test		
	Numéricas	Correlación lineal	Pearson	Spearman, Kendall	
		Otro tipo de relaciones	Regresión lineal	Regresión lineal generalizada	Regresión no paramétrica
	Categórica y numérica		ANOVA Regresión lineal	Regresión lineal generalizada	Regresión no paramétrica

Modelización estadística

- En general, un modelo es una representación a pequeña escala de la realidad.
- “Esencialmente, todos los modelos son incorrectos, pero algunos son útiles” (Box).
- “La formulación del problema es más esencial que su propia solución, que puede ser simplemente una habilidad matemática o experimental” (Einstein).
- Principio de la navaja de Occam: Un modelo estadístico debe ser lo más simple posible.

Contrastes de hipótesis

- **Hipótesis nula H_0** vs **Hipótesis alternativa H_1**
  - Hipótesis sobre la población (desconocida)
  - H_0 recoge aquello que nos creeremos mientras no haya fuertes evidencias que nos demuestren lo contrario
- Decisión a partir de los datos de la muestra
  - **Error de tipo I: P(** Rechazar H_0 cuando es cierta) → α
  - **Error de tipo II:** P(Aceptar H_0 cuando es falsa) → β
- **Estadístico de contraste**: Mide la discrepancia entre los datos muestrales y la hipótesis nula H_0
- p-valor: Probabilidad asociada a la muestra de cometer error de tipo I
  - **p-valor < α → Rechazar H_0**

Comparación de dos poblaciones normales

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

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

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

- La característica que queremos comparar entre las dos poblaciones es una variable continua con distribución normal

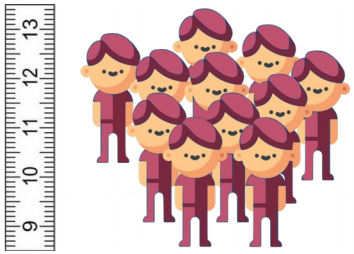
Comparación de dos poblaciones normales

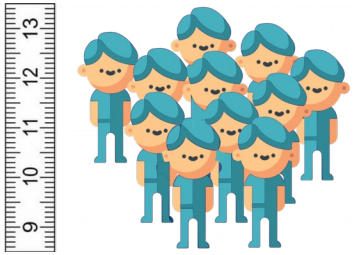
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

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

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

- La característica que queremos comparar entre las dos poblaciones es una variable continua con distribución normal





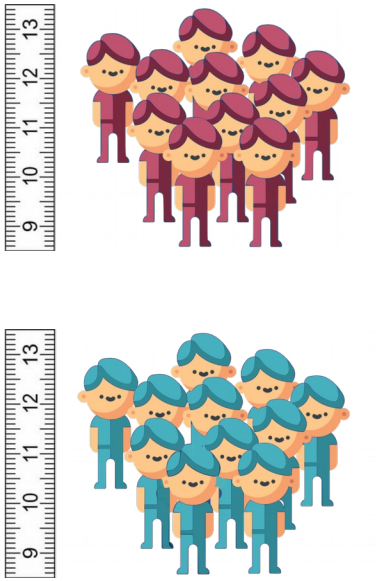
Comparación de dos poblaciones normales

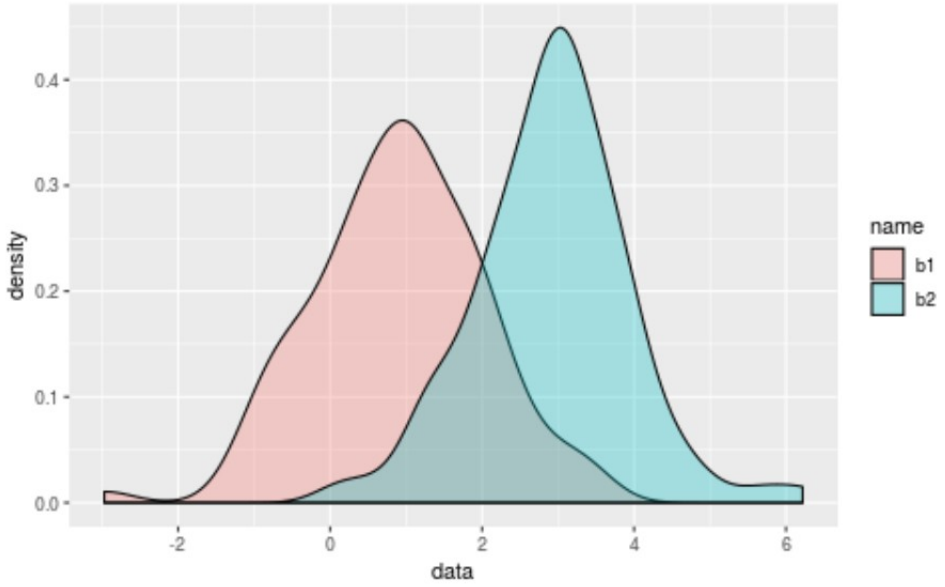
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

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

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

- La característica que queremos comparar entre las dos poblaciones es una variable continua con distribución normal





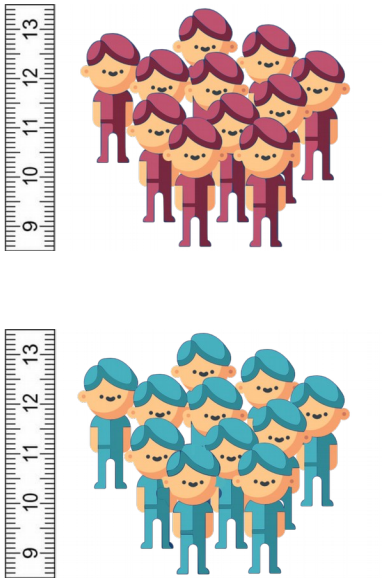
Comparación de dos poblaciones normales

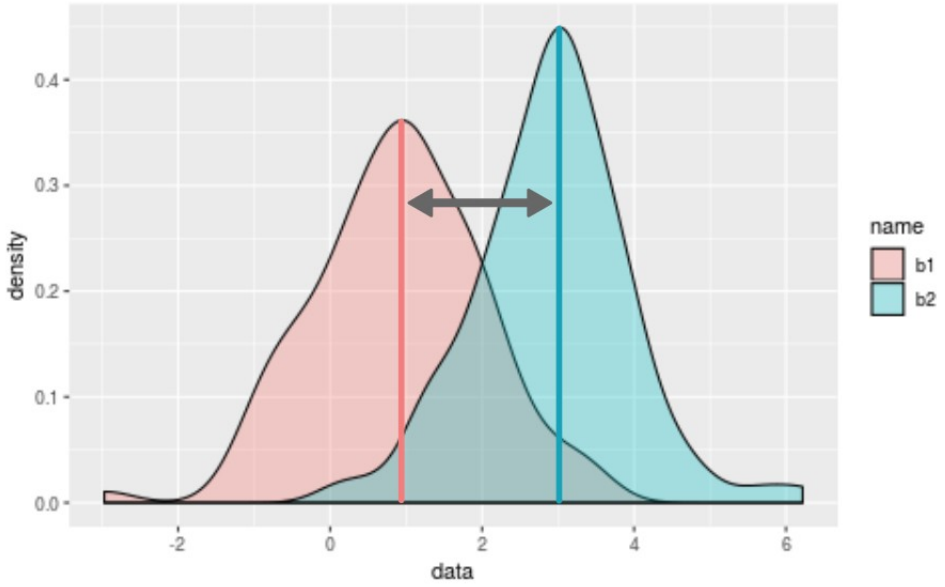
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

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

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

- La característica que queremos comparar entre las dos poblaciones es una variable continua con distribución normal





Comparación de dos poblaciones normales

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

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

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

- La característica que queremos comparar entre las dos poblaciones es una variable continua con distribución normal.
- T-test
  - ¿Unilateral o bilateral?
  - ¿Tenemos datos apareados o muestras independientes?
    - Ejemplo de datos apareados: Nivel de colesterol de un grupo de pacientes antes y después de un tratamiento
  - ¿Son iguales las varianzas de las poblaciones comparadas?
    - Test

Comparación de dos poblaciones no normales

H_0: Mediana$_1$ = Mediana$_2$

No podemos asumir normalidad

- La característica que queremos comparar entre las dos poblaciones es una variable cuantitativa (discreta o continua).
- Test Mann-Whitney
  - ¿Unilateral o bilateral?
  - ¿Tenemos datos apareados o muestras independientes? (Test de Wilcoxon para una muestra)
    - Ejemplo de datos apareados: Nivel de colesterol de un grupo de pacientes antes y después de un tratamiento
- Los tests no paramétricos suelen ser menos potentes que los paramétricos. Por tanto, si podemos asumir normalidad en nuestros datos, es más recomendable utilizar un test paramétrico.

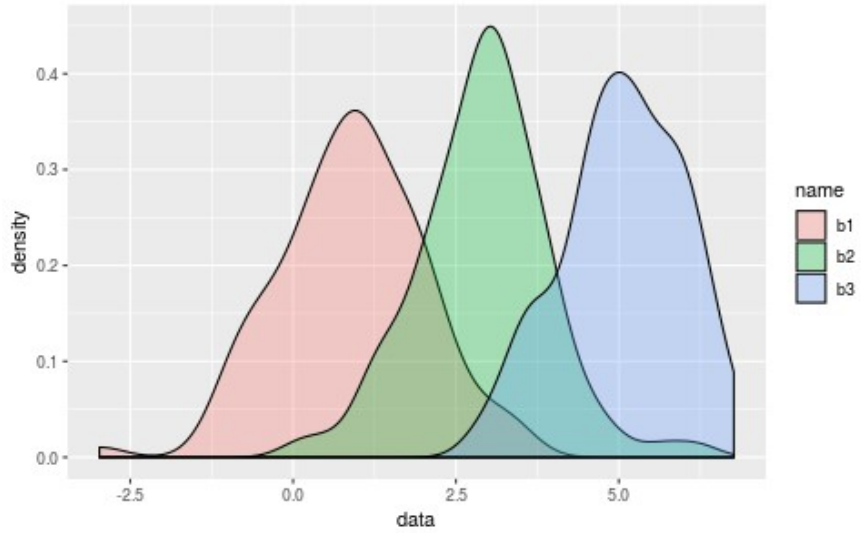
Comparación K poblaciones normales

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: Algún μ_i es distinto

ANOVA de 1 factor (de efectos fijos)

- Variable respuesta: Característica que queremos comparar entre los distintos grupos. Debe ser una variable aleatoria continua distribuida normalmente.

- Factor: Variable explicativa que indica los grupos o poblaciones que vamos a comparar. Es una variable categórica.



Relación entre dos variables categóricas

- Tests de independencia
  - Test Exacto de Fisher
  - Test Chi-2

- H_0: Las variables son independientes
  H_A: Las variables NO son independientes

- Tabla de contingencia

	Infected	Not infected	
Inoculated	3	276	279
Not inoculated	66	473	539
	69	749	818

Cholera Inoculation Study, 1894-96

Relación entre dos variables categóricas

- Tests de independencia
  - Test Exacto de Fisher
  - Test Chi-2

- H_0: Las variables son independientes
  H_A: Las variables NO son independientes

- Tabla de contingencia

	Infected	Not infected	
Inoculated	3 1%	276 99%	279
Not inoculated	66 12%	473 88%	539
	69	749	818

Cholera Inoculation Study, 1894-96

Relación entre dos variables numéricas

- Hay variables que tienen una relación entre ellas

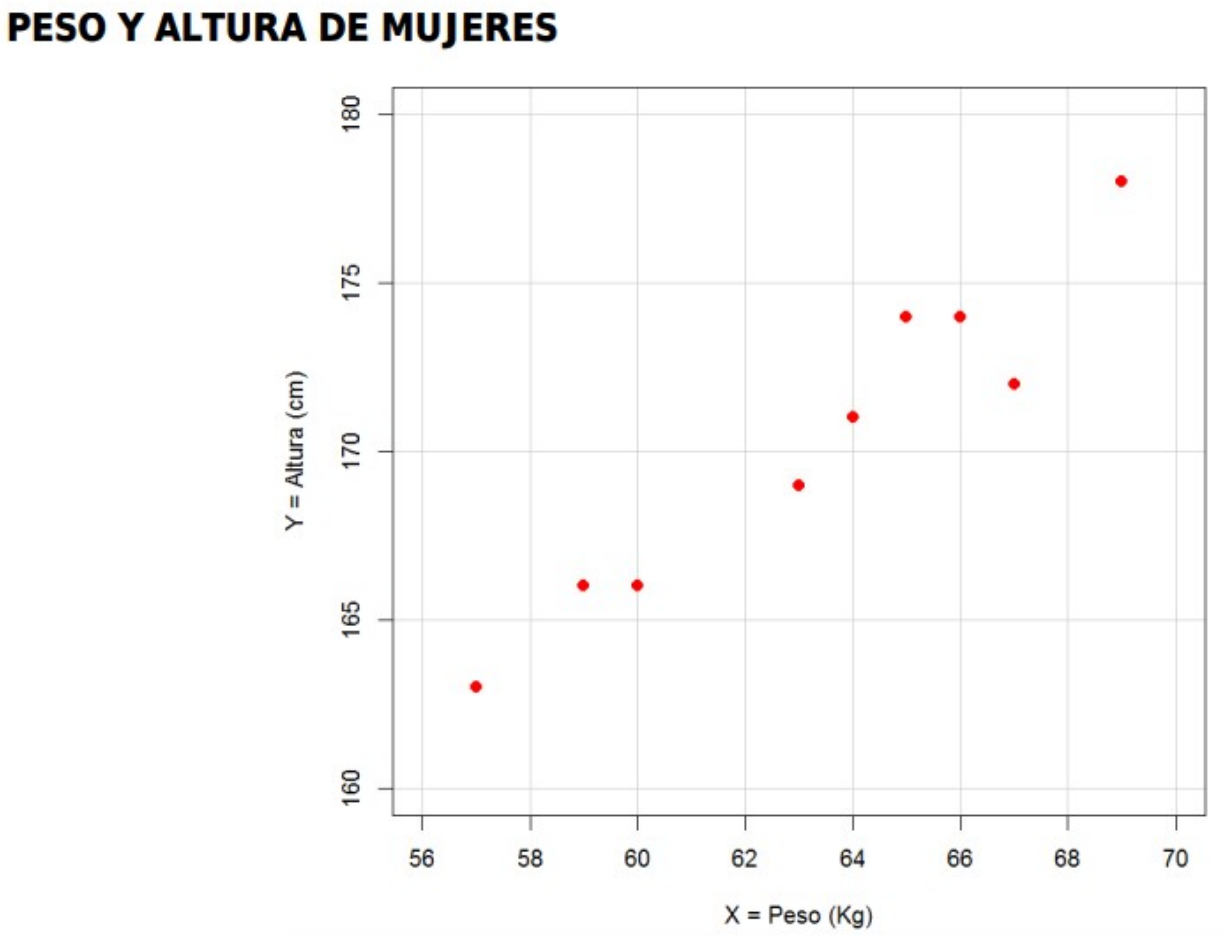
EJEMPLO CASO 2: PESO Y ALTURA DE MUJERES

La tabla siguiente muestra los pesos y las alturas de 9 mujeres obtenidas en una cierta farmacia de Valencia

	Peso (Kg)	Altura (cm)
	60	166
	69	178
	66	174
	64	171
	57	163
	67	172
	59	166
	65	174
	63	169
Media	63.33	170.33
Desviación típica	3.97	4.77

Diagrama de dispersión

- Mostramos las dos variables, cada una en un eje



Correlación

- Coeficientes de correlación lineal
  - Miden el grado de relación LINEAL entre dos variables
  - Toman valores entre -1 y 1
    - ~ 1 → Relación lineal positiva (ascendente)
    - ~ -1 → Relación lineal negativa (descendente)
    - ~ 0 → No existe relación LINEAL

- Métodos para calcular la correlación:
  - **Pearson**: Para variables "aproximadamente" normales
  - **Spearman / Kendall**: Se calcula mediante rangos por lo que aceptan cualquier tipo de variable y no están tan influidos por valores anómalos

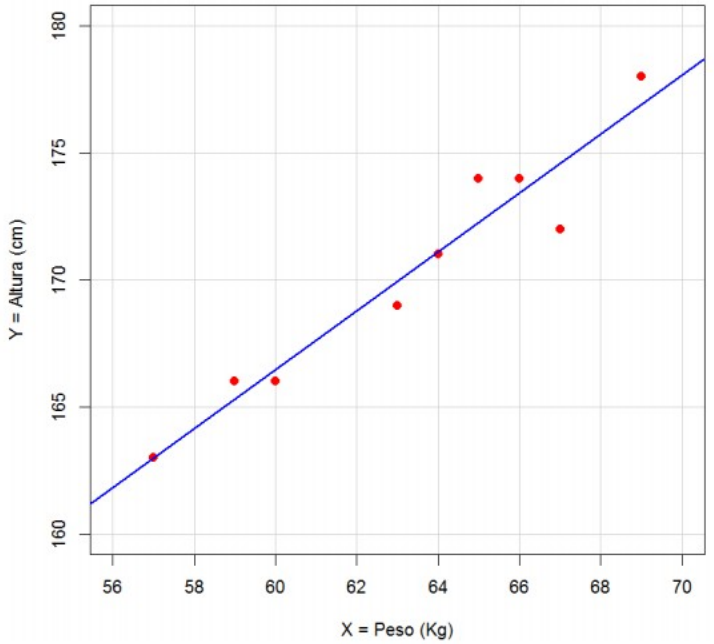
Modelos de regresión lineal

- Y → Variable respuesta (o dependiente)
  - Variables aleatoria con distribución normal
- X → Variable explicativa (o independiente)
  - Variables aleatorias o no

- Definimos recta de regresión:

$$Y = b_0 + b_1 \cdot X$$

Estimaremos los valores de los coeficientes de regresión b_i a partir de los datos de nuestra muestra.



Modelos de regresión lineal

$$Y = b_0 + b_1 \cdot X$$

- Hipótesis global del modelo:
  - $H_0 : b_0 = b_1 = 0$
  - Rechazar esta hipótesis equivale a aceptar que alguna de la variables explicativas del modelo tiene un efecto significativo sobre la variable respuesta. ¿Cuáles? Esto lo estudiantes con los contrastes de hipótesis siguientes.
- Hipótesis sobre cada uno de los coeficientes:
  - $H_0 : b_i = 0$
  - Rechazar esta hipótesis equivale a afirmar que la variable x_i tiene un efecto significativo sobre la variable respuesta y.

Referencias y links útiles

- Curso on-line sobre estadística aplicada
  - https://onlinecourses.science.psu.edu/stat500/

- Curso de Introducción al entorno R (David Conesa, UV)
  - https://www.uv.es/conesa/CursoR/cursoR.html

- Experimental Design and Data Analysis for Biologists. Gerry P. Quinn & Michael J. Keough. Cambridge University Press.