Introducción a la Bioestadística y R

Marta R. Hidalgo
Unidad de Bioinformática y Bioestadística

Junio 2019

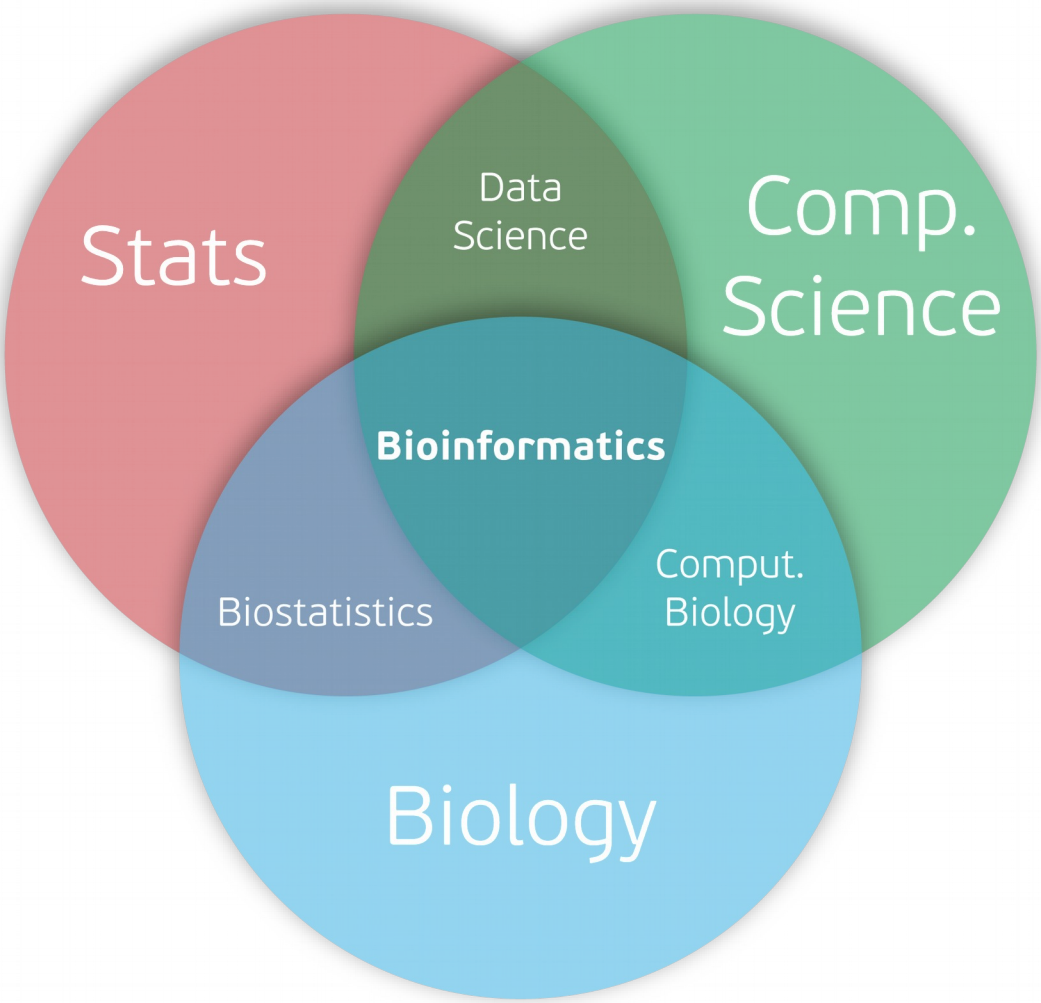
Unidad de Bioinformática y Bioestadística

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.

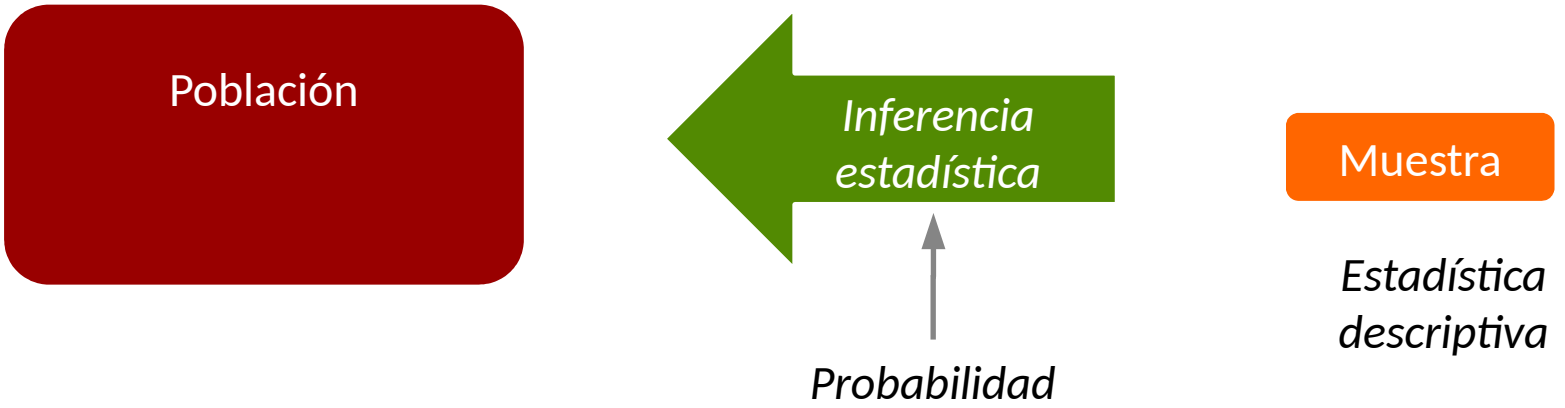
Población

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.



Programa Bioestadística

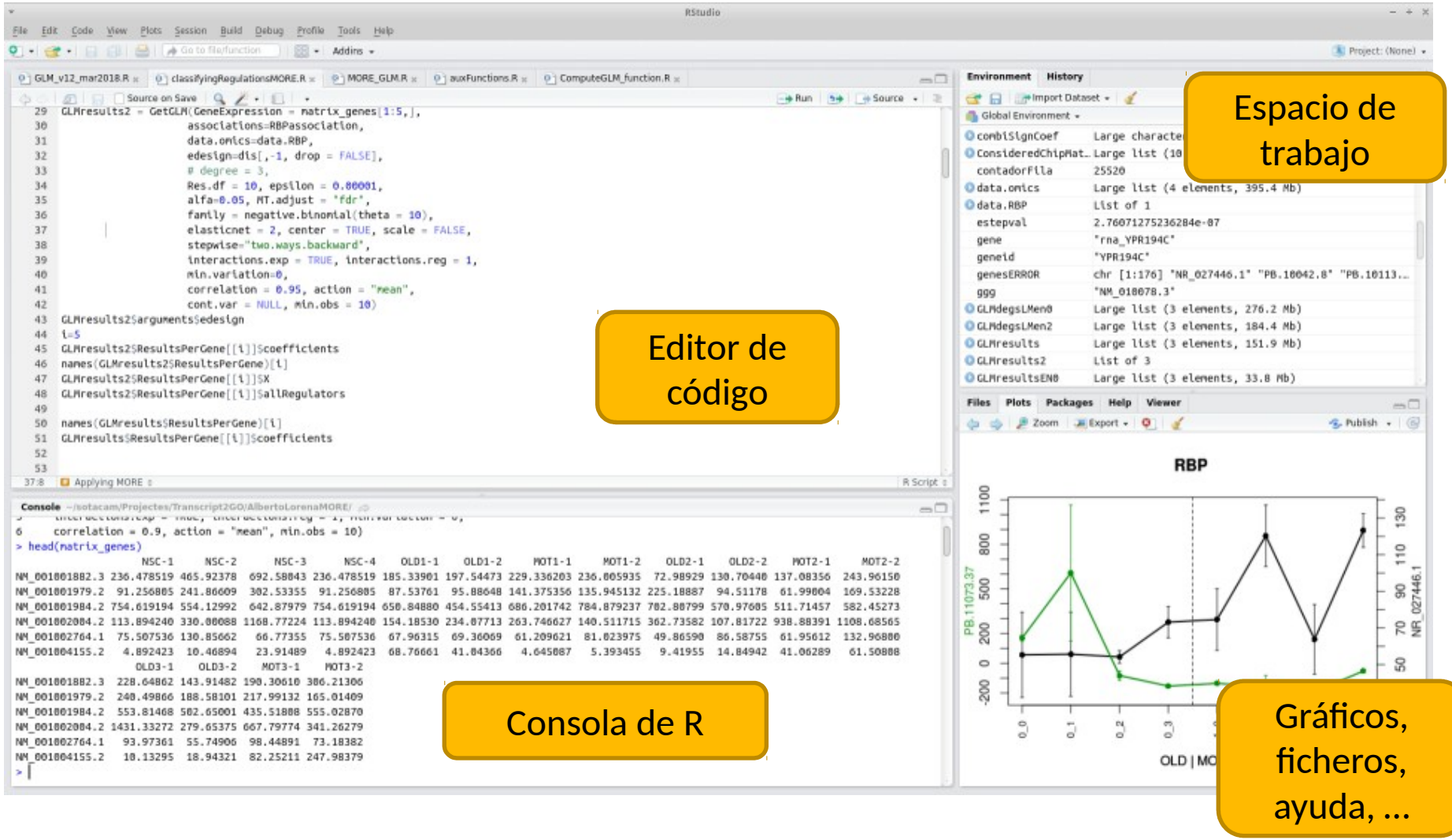
- Introducción a R y RStudio
- Estadística descriptiva
  - Tipos de datos
  - Parámetros descriptivos
  - Gráficos
  - Análisis de Componentes Principales
- Inferencia estadística
  - Introducción a la inferencia estadística
  - Resumen de métodos y su uso
  - Comparación de poblaciones
    - t-test
    - ANOVA
  - Predicción / explicación de una variable respuesta
    - Regresión lineal
  - Relación entre dos variables
    - Tests de independencia
    - Correlación

R y RStudio

Lenguaje de programación R

- R es un lenguaje de programación que permite implementar técnicas estadísticas, y además realizar análisis estadísticos y gráficos.
- Repositorios públicos
  - CRAN (https://cran.r-project.org/)
  - Bioconductor (https://www.bioconductor.org/)
  - GitHub
  - …
- Ventajas
  - Libre y gratuito
  - Ayuda, soporte
  - Flexibilidad
  - *Scripting*
  - …

RStudio


Espacio de trabajo
Editor de código
Consola de R
Gráficos, ficheros, ayuda, ...
RBP

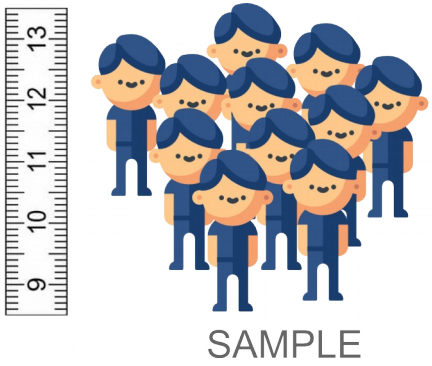
Estadística descriptiva

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,...*

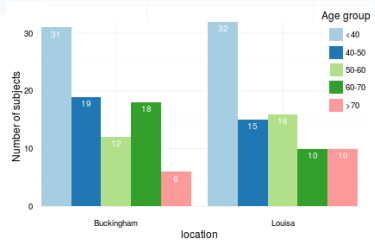
Parámetros descriptivos

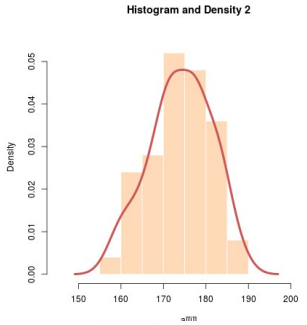
DATOS NUMÉRICOS

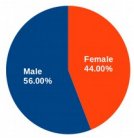
Tipo	Parámetro	Población	Muestra	Comando R
Localización	Media	μ	$\bar{x}$	`mean()`
	Mediana	Me		`median()`
	Percentiles	P_i		`quantile()`
Dispersión	Varianza	σ^2	s^2	`var()`
	Desviación típica	σ	s	`sd()`
	Rango	Max – Min		`summary()`
	Rango intercuartílico	$P_{75} - P_{25}$		`IQR()`

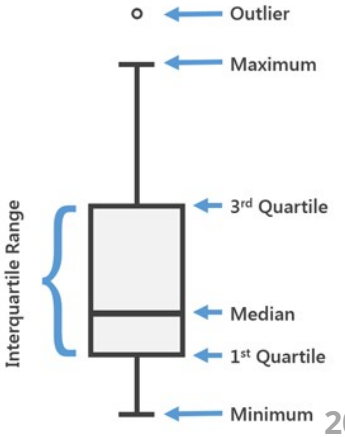
Gráficos

Gráfico	Tipo de datos	Comando R
Diagrama de barras	Categóricos Discretos (pocos)	`barplot()`
Histograma	Continuos Discretos (muchos)	`hist()`
Función de densidad	Continuos	`plot(density())`
Sectores	Categóricos	`pie()`
Boxplot simple	Continuos Discretos (muchos)	`boxplot()` `plot()`
Boxplot múltiple	Simple combinado con categórica	`boxplot()`



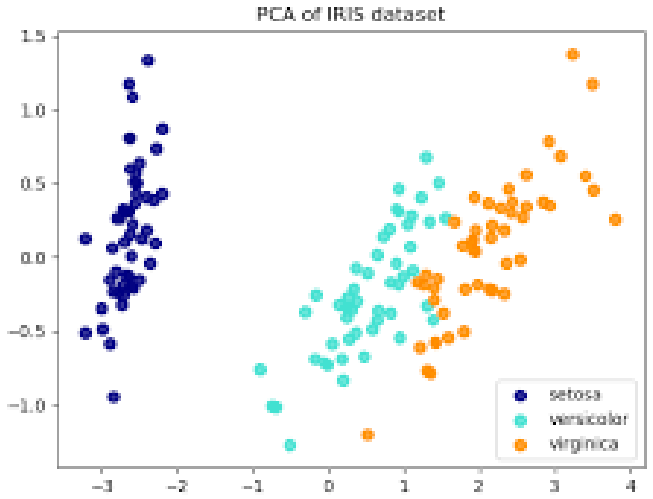






Análisis de Componentes Principales

- 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.



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.

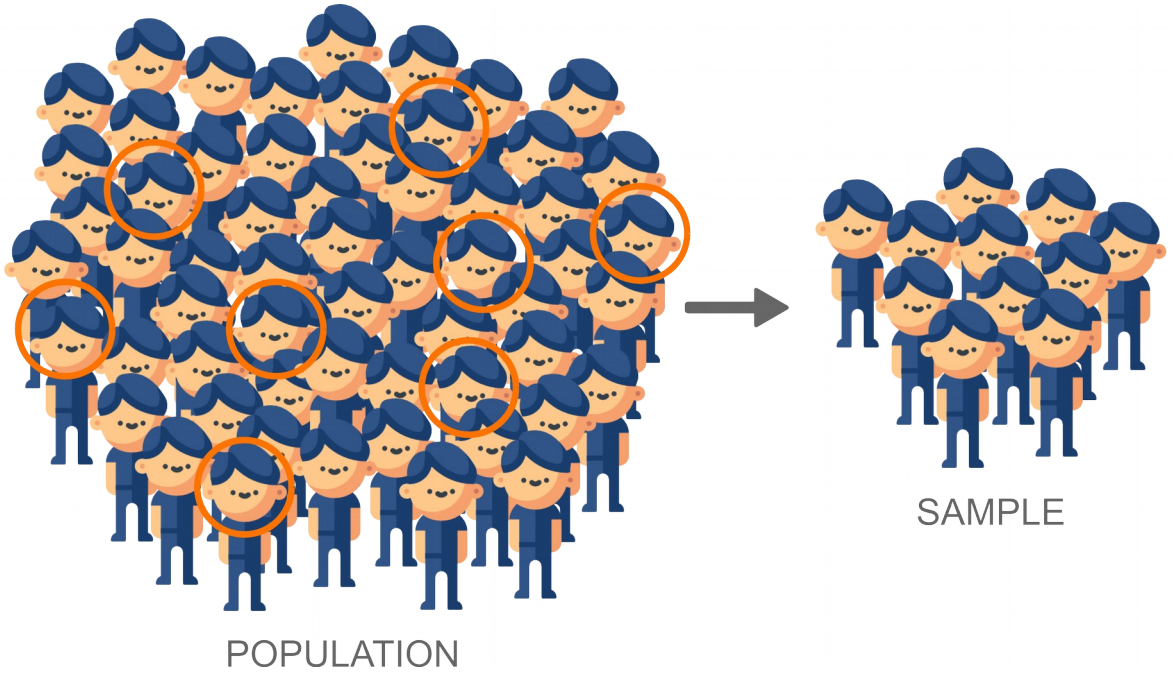
POPULATION

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.

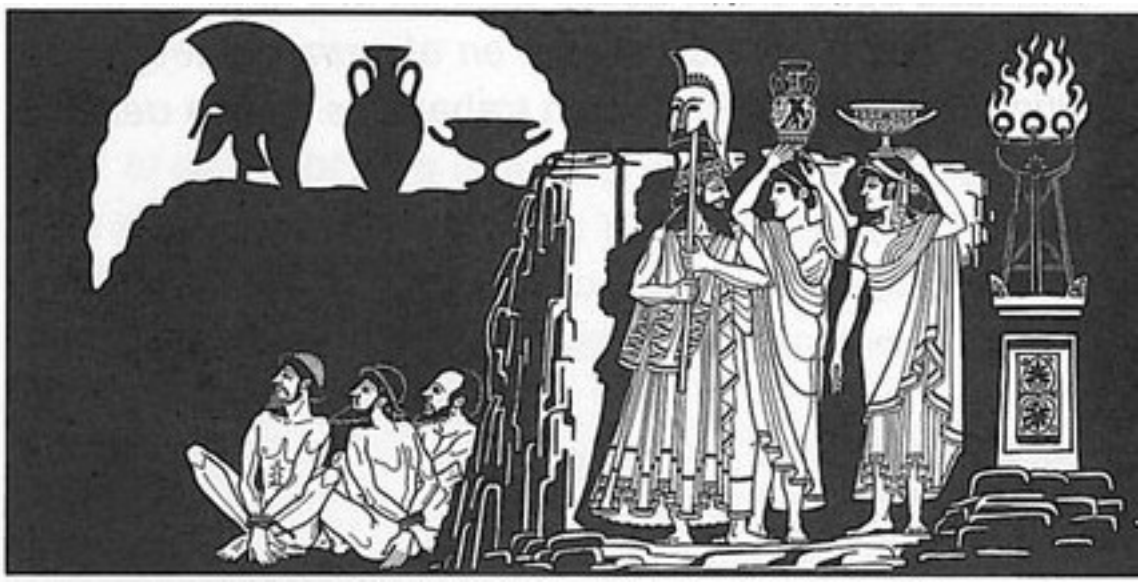
Inferencia estadística

- 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.



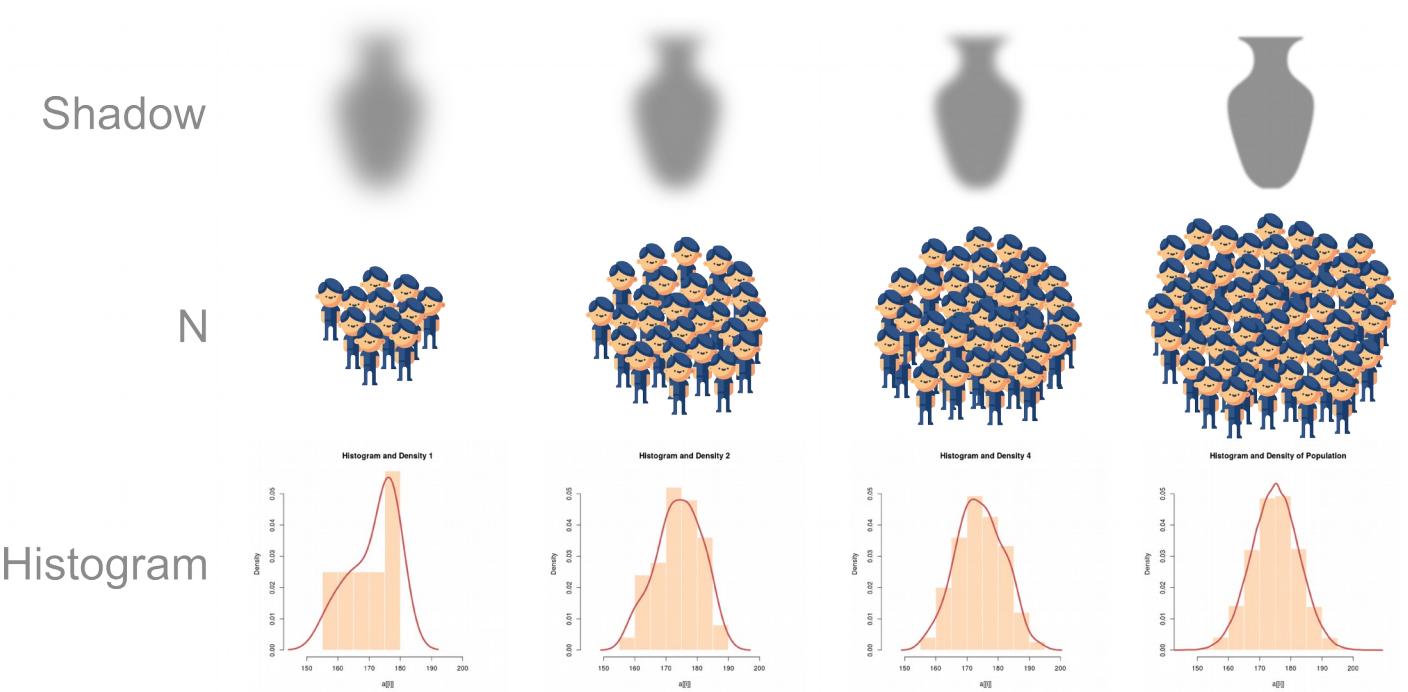
Inferencia estadística

- 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.



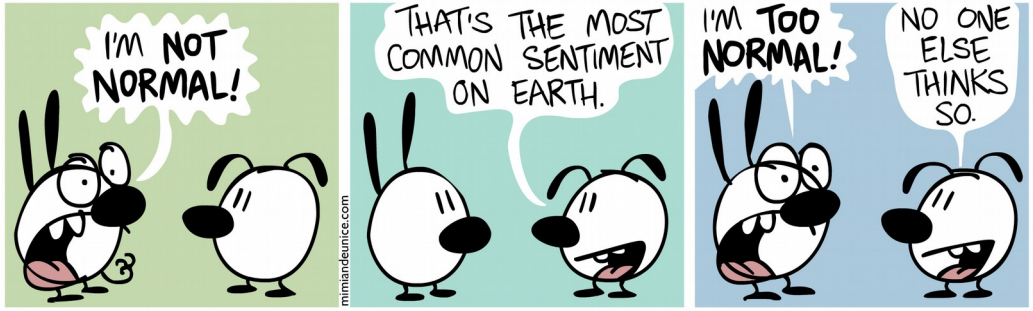
Inferencia estadística

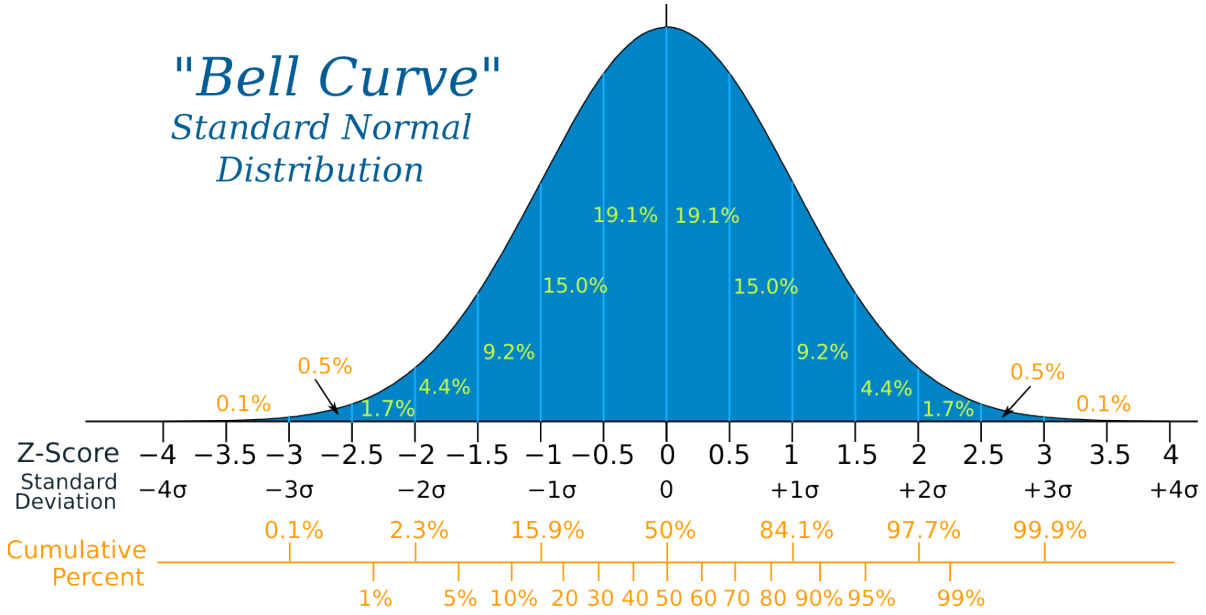
- **Herramientas de la inferencia estadística**:
  - Estimación puntual de un parámetro
    - Para obtener una primera aproximación de su valor
  - Estimación por intervalos
    - Un intervalo de confianza es un intervalo con una probabilidad alta de contener al verdadero valor del parámetro, que es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

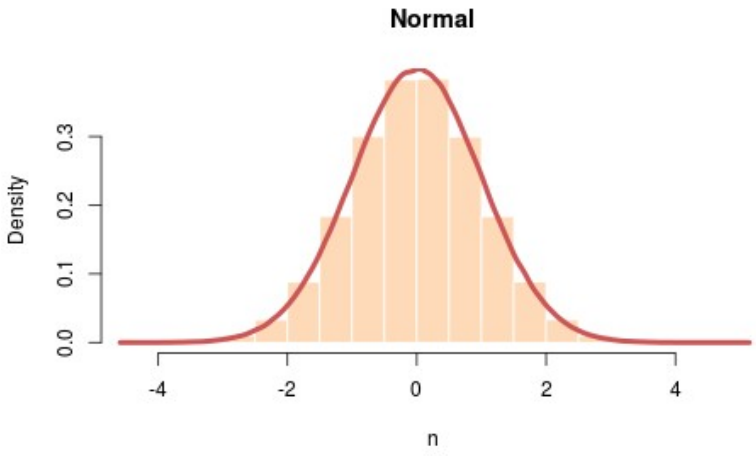
Distribución normal N(μ, σ)





Distribución normal N(μ, σ)

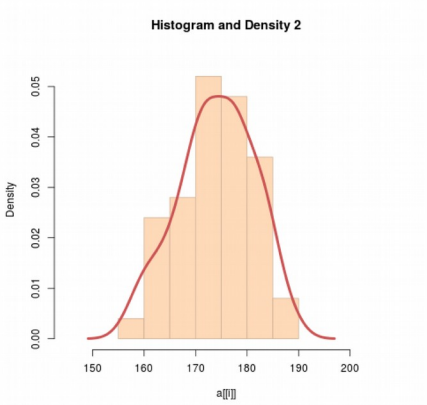
- ¿Cómo comprobar si nuestros datos siguen una distribución normal?
  - Histograma o gráfico de densidad
  - Gráfico probabilístico normal (*qqnorm()*) → Los puntos han de ajustarse a una línea recta
  - Test de normalidad → H_0: normalidad (*shapiro.test()*)

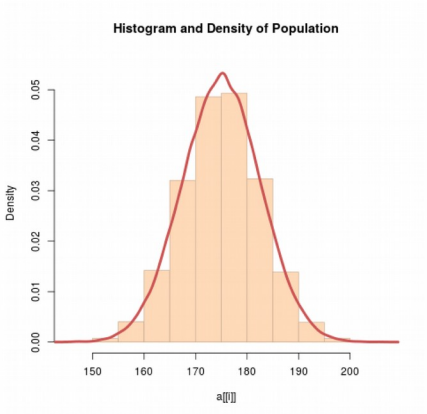


Intervalos de confianza

SAMPLE

POPULATION

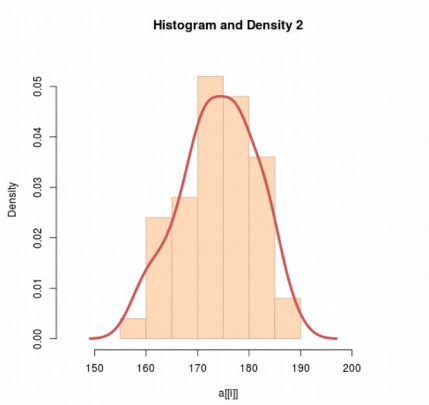


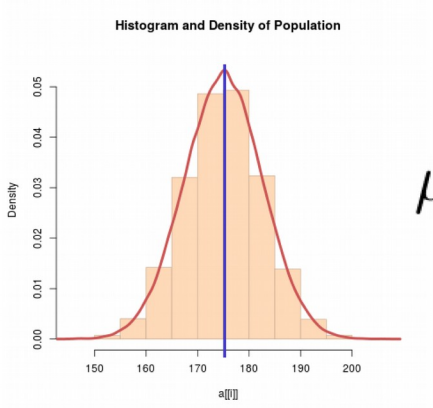


Intervalos de confianza

SAMPLE

POPULATION



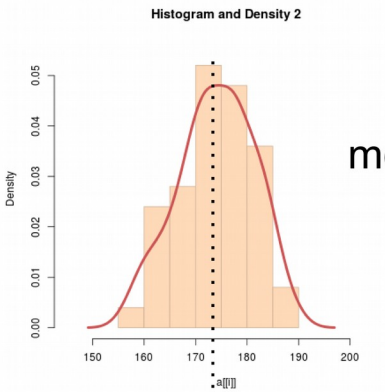


$\mu = 175$

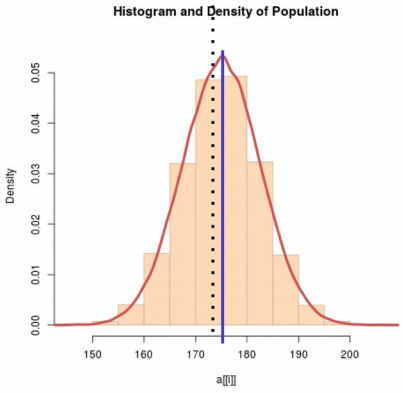
Intervalos de confianza

SAMPLE

POPULATION

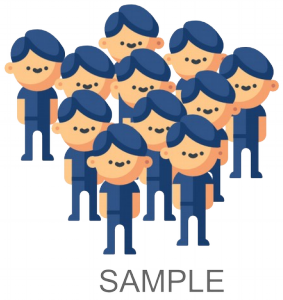


mean = 173



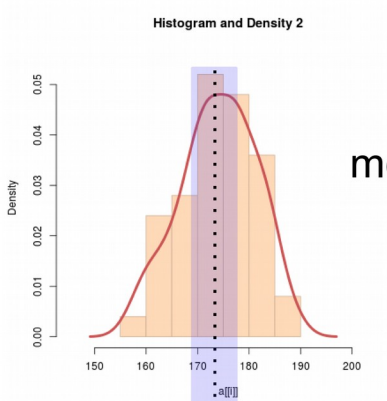
μ = 175

Intervalos de confianza



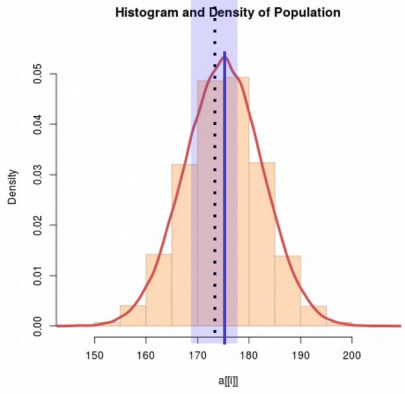
SAMPLE

POPULATION



mean = 173

C.I. = [168, 178]



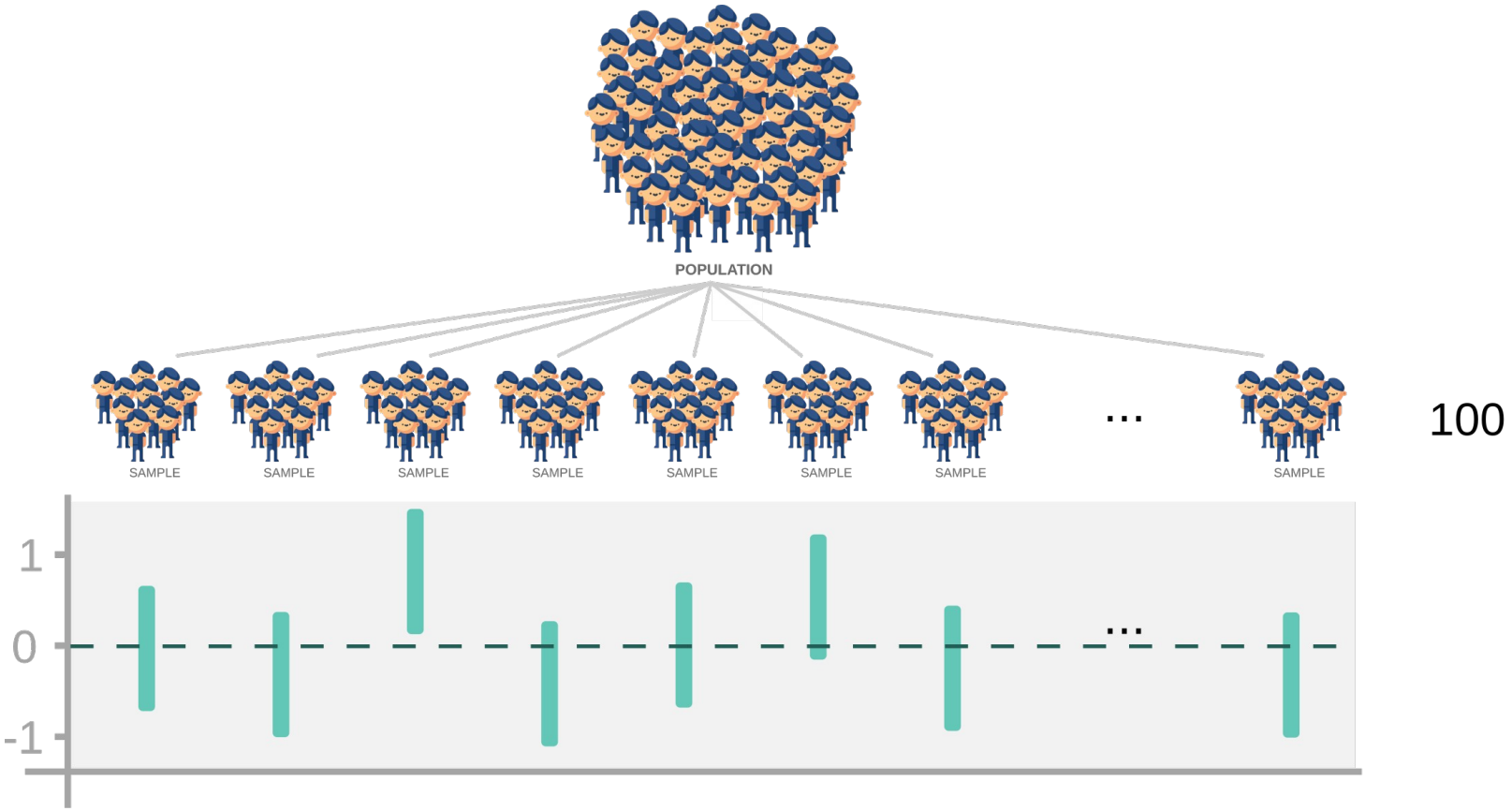
$\mu = 175$

Intervalos de confianza

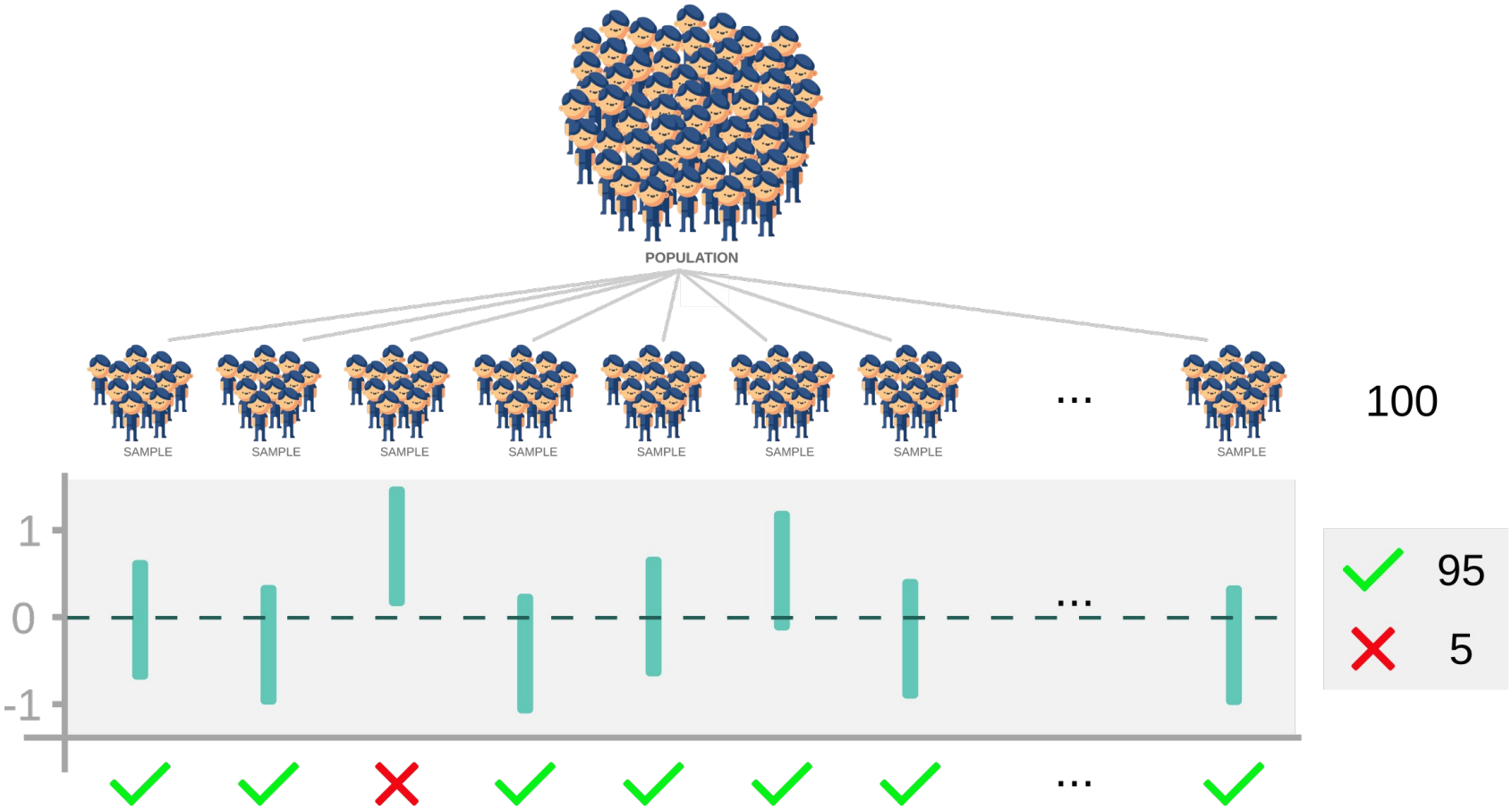
- Cómo crear intervalos de confianza: **t.test()**
  - Distribución normal del parámetro
  - **conf.level**: Nivel de confianza (95%, 90%,...)

- 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



Intervalos de confianza



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**

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.

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	

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

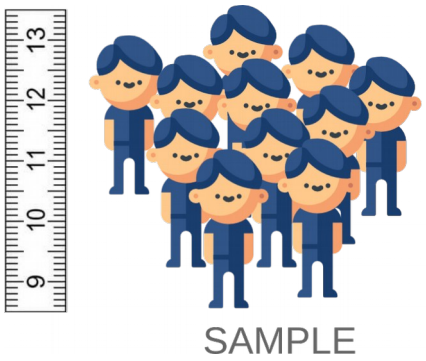
Comparación de dos poblaciones normales

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

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

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

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



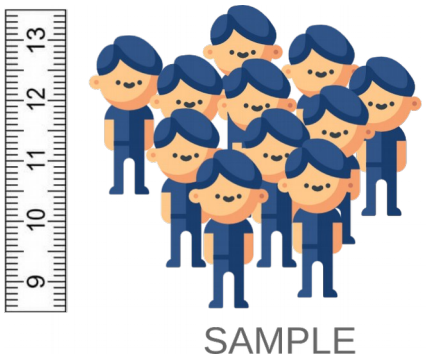
Comparación de dos poblaciones normales

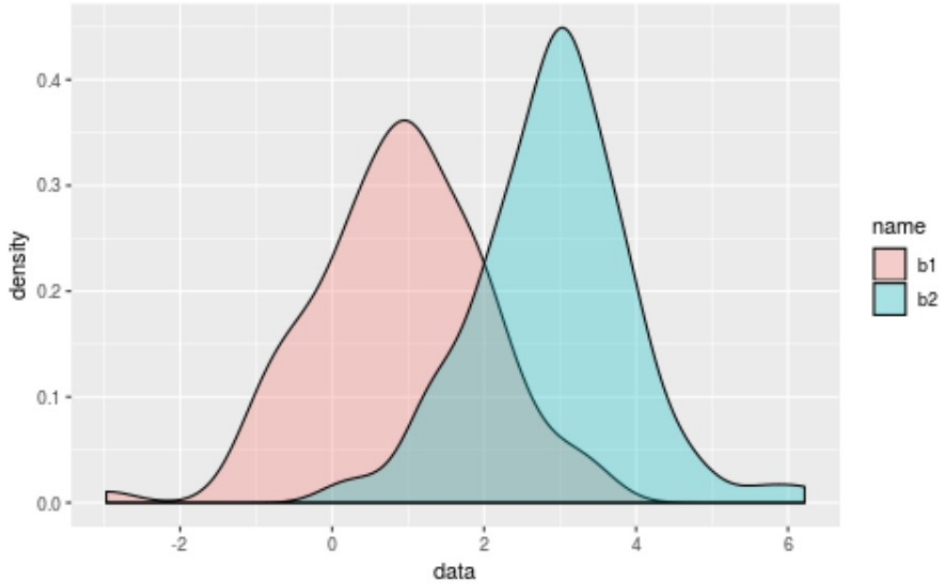
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

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

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

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





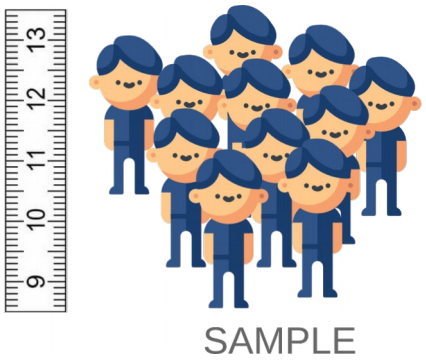
Comparación de dos poblaciones normales

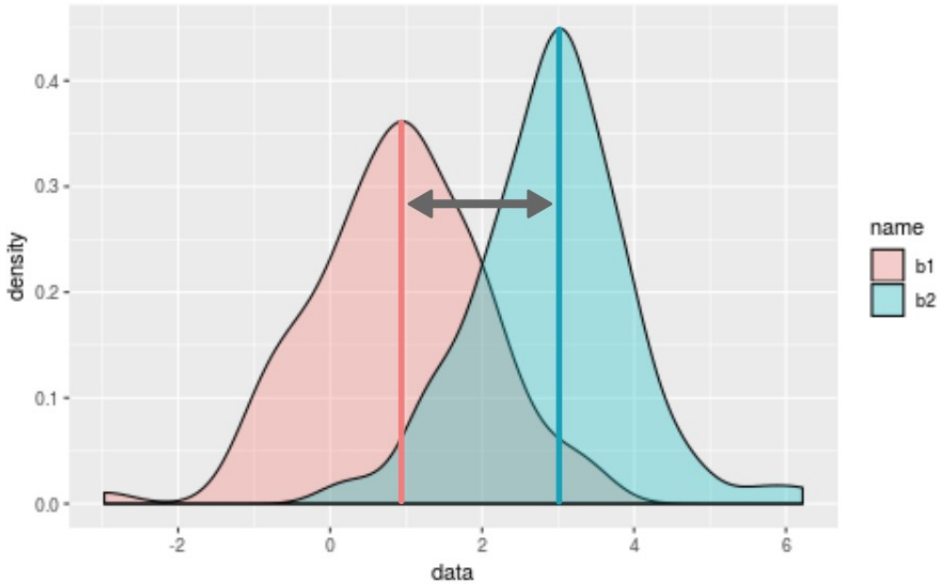
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

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

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

- 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()
  - **alternative** = c("two.sided", "greater", "less")
  - **paired**: ¿Tenemos datos apareados o muestras independientes?
    - Ejemplo de datos apareados: Nivel de colesterol de un grupo de pacientes antes y después de un tratamiento
  - **var.equal**: ¿Son iguales las varianzas de las poblaciones comparadas?
    - var.test() $H_0: \sigma_1^2 = \sigma_2^2$
- El estadístico de contraste depende de la diferencia de medias muestrales, de la variabilidad (que a su vez depende de los parámetros anteriores) y del tamaño muestral.

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 → wilcox.test()
  - alternative = c("two.sided", "greater", "less")
  - Paired: ¿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.

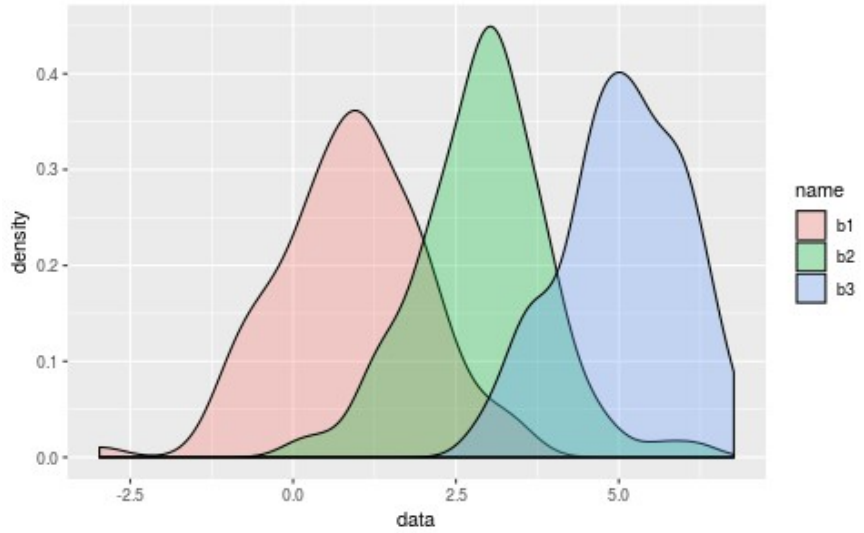
Comparación K poblaciones normales

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.



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.
- Análisis de la varianza (ANOVA)
  1. aov()
  2. summary()
- Validación del modelo → plot()
  - Normalidad: Si no se cumple, test de Kruskal Wallis → kruskal.test()
  - Homocedasticidad
  - Independencia
- Comparaciones a posteriori → TukeyHSD()

Inferencia estadística

Objetivo	Diseño experimental	Parámetro a estudiar	Normalidad	No normalidad	No paramétrico
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

Relación entre dos variables categóricas

- Tests de independencia
  - Fisher's Exact test → fisher.test()
  - Test Chi-2 → chisq.test()
- H_0: Las variables 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
  - Fisher's Exact test → fisher.test()
  - Test Chi-2 → chisq.test()
- H_0: Las variables 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

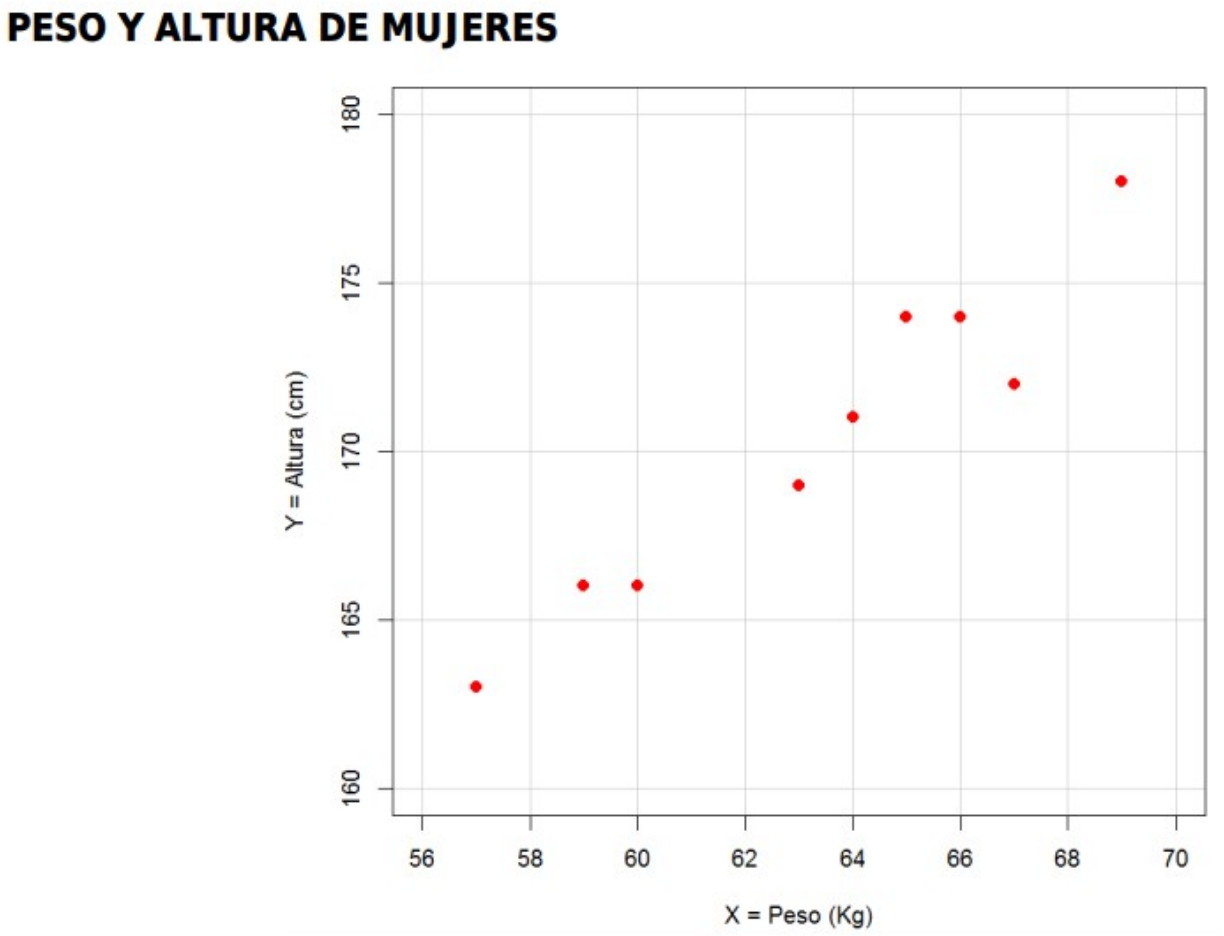
EJEMPLO CASO 2: PESO Y ALTURA DE MUJERES

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: **cor()**
  - **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

Inferencia estadística

Objetivo	Diseño experimental	Parámetro a estudiar	Normalidad	No normalidad	No paramétrico
Predecir/explicar una variable respuesta			Regresión lineal	Regresión lineal generalizada	Regresión no paramétrica (pe. Kaplan-Meier)

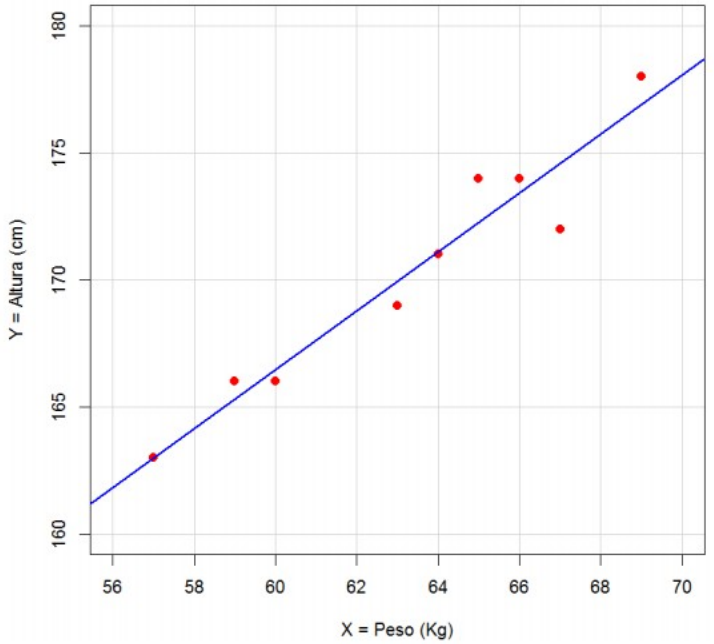
Modelos de regresión lineal

- 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.

Modelos de regresión lineal

- Funciones de R
  - Modelo ← lm(Y ~ X)
  - summary(Modelo)

```
> model <- lm(altura ~ peso)
> summary(model)

Call:
lm(formula = altura ~ peso)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58201 -0.47090  0.00529  0.68783  1.73545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.9471     7.6505  12.672 4.41e-06 ***
peso          1.1587     0.1206   9.609 2.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.354 on 7 degrees of freedom
Multiple R-squared:  0.9295,    Adjusted R-squared:  0.9195
F-statistic: 92.33 on 1 and 7 DF,  p-value: 2.781e-05
```

Modelos de regresión lineal

- Funciones de R
  - Modelo ← lm(Y ~ X)
  - summary(Modelo)

```
> model <- lm(altura ~ peso)
> summary(model)

Call:
lm(formula = altura ~ peso)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58201 -0.47090  0.00529  0.68783  1.73545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.9471     7.6505  12.672 4.41e-06 ***
peso          1.1587     0.1206   9.609 2.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.354 on 7 degrees of freedom
Multiple R-squared:  0.9295,    Adjusted R-squared:  0.9195
F-statistic: 92.33 on 1 and 7 DF,  p-value: 2.781e-05
```

Modelos de regresión lineal

- Bondad de ajuste del modelo
  - R^2: Porcentaje de variabilidad de la variable respuesta *y* que queda explicado por el modelo
  - R^2 ajustado: Sirve para comparar modelos anidados con distinto número de variables
- Validación del modelo
  - Se utilizan los residuos para validar el modelo
    - residuo = valor observado – valor predicho
  - Hipótesis del modelo
    - Normalidad
    - Homocedasticidad (igualdad de varianzas)
    - Independencia de las observaciones

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.

Ejercicios

Ejercicio

Usando los datos iris,

- Calcula el intervalo de confianza al 95% de la media de la variable Sepal.length.
- Calcula los intervalos de confianza al 90% y 99%. Cuál es mayor?

Comparación de dos poblaciones

- Ejercicio 1
  - ¿Es distinta la longitud media del sépalo entre las especies “setosa” y “virginica”?
- Ejercicio 2
  - ¿Es distinta la anchura media de los sépalos entre las mismas especies?
- Ejercicio 3
  - ¿Es distinta la anchura media del pétalo entre las especies “setosa” y “virginica”?

Comparación K poblaciones normales

- **Ejercicio:** Datos iris
  - ¿Es significativamente diferente la anchura media del sépalo para las 3 especies de Iris?
  - ¿Cuál es la especie con mayor longitud media de sépalo?
  - ¿Entre qué especies hay diferencias? ¿Entre qué pareja de especies hay mayor diferencia? ¿Y menor?

Relación entre dos variables categóricas

Ejercicio

- En un estudio reciente acerca del daltonismo, un grupo de investigadores examinó a un gran número de escolares noruegos obteniéndose los resultados de la siguiente tabla. Según estos datos, dirías que el daltonismo es independiente del sexo?

	Niños	Niñas	TOTAL
Daltónicos	725	40	765
No Daltónicos	8324	9032	17356
TOTAL	9049	9072	18121

Correlación

- **Ejercicio 1**
  - Usa la función cor() para calcular la correlación entre la altura y el peso de las mujeres del estudio explicado anteriormente.

- **Ejercicio 2:** Datos iris
  - Crees que hay relación entre las variables Petal.Length y Petal.Width? Dibuja el diagrama de dispersión y calcula el coeficiente de correlación.
  - Y entre las variables Sepal.Length y Sepal.Width? Dibuja el diagrama de dispersión y calcula la correlación.

Modelos de regresión lineal

- Ejercicio 1
  - Calcula la recta de regresión que describe la variable Petal.Length en función de la variable Petal.Width, en los datos iris.

- Ejercicio 2
  - ¿Tendría sentido calcular la recta de regresión para estimar el valor de la variable Sepal.Length en función de Sepal.Width? ¿Qué ocurre si lo hacemos?