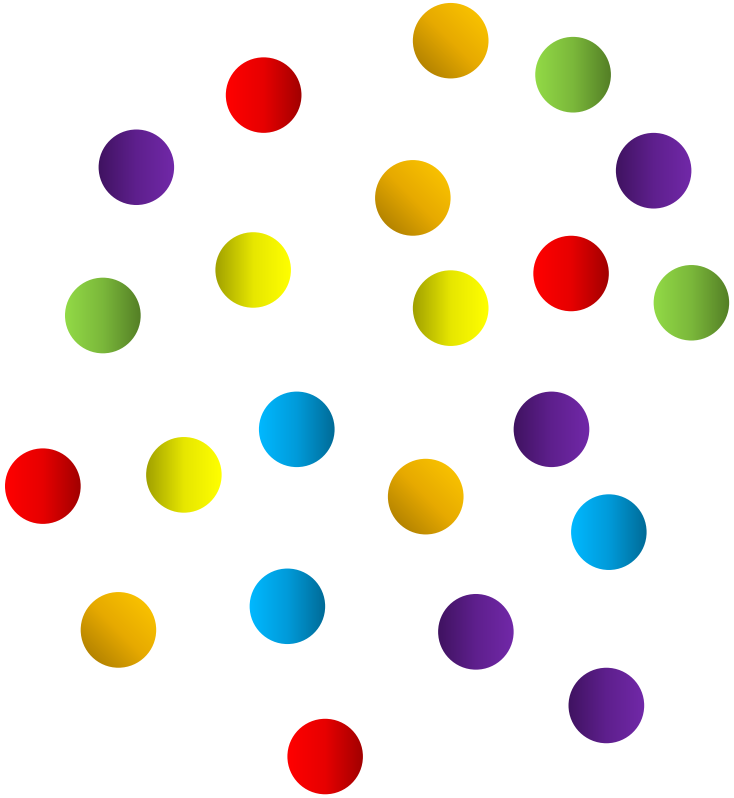
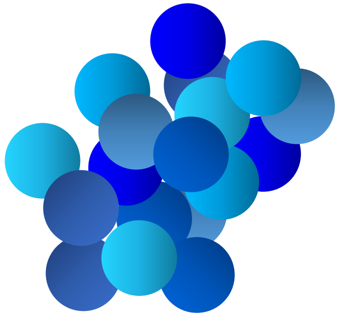
Pruebas de Hipótesis

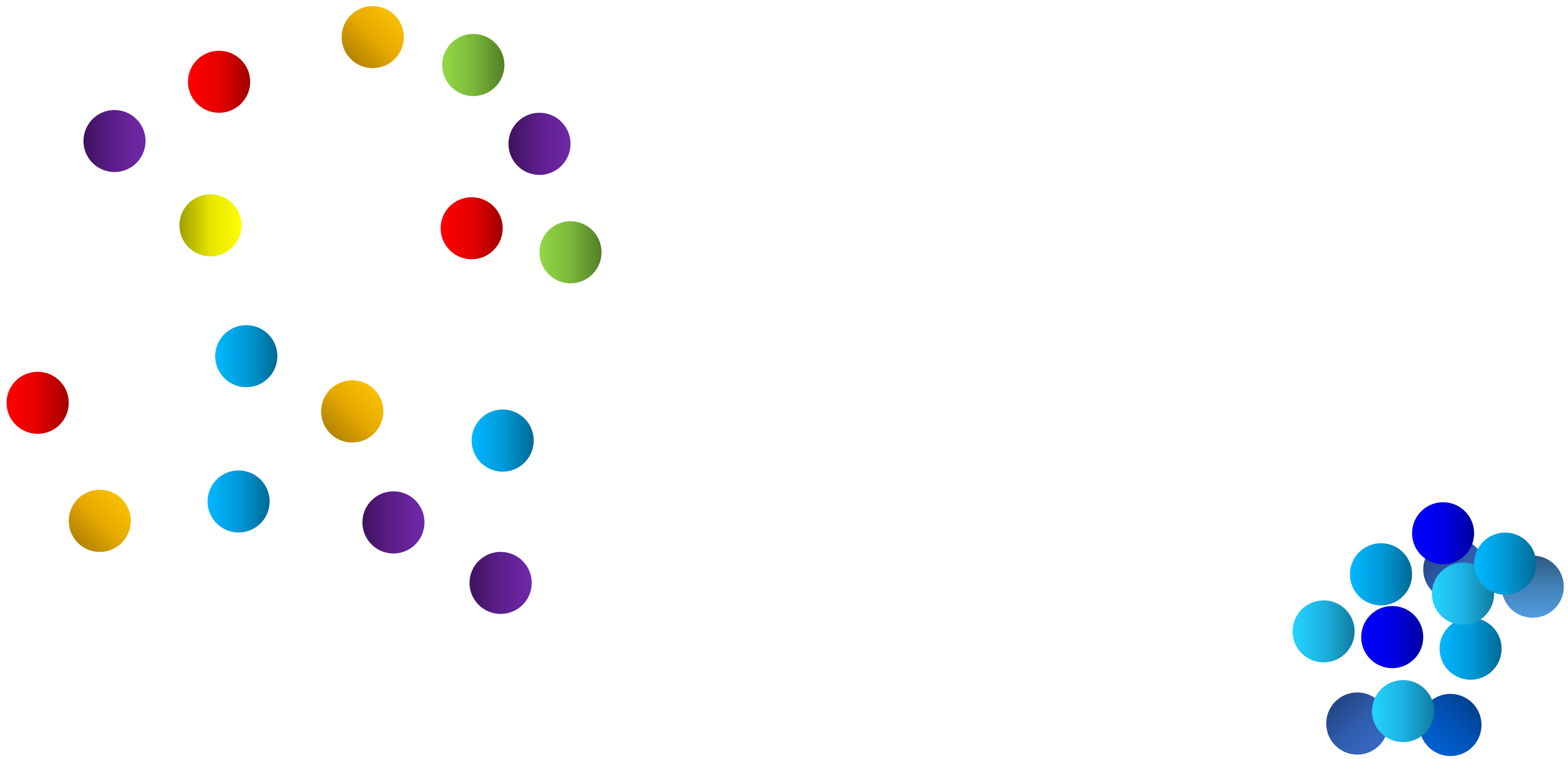
Diplomado Ciencia de Datos

Población





Muestra



Conceptos Importantes

Modelo: Representación simplificada de la realidad que contiene los aspectos mas importantes de la misma.

Modelo estadístico: Modelo que incorpora un elemento de aleatoriedad

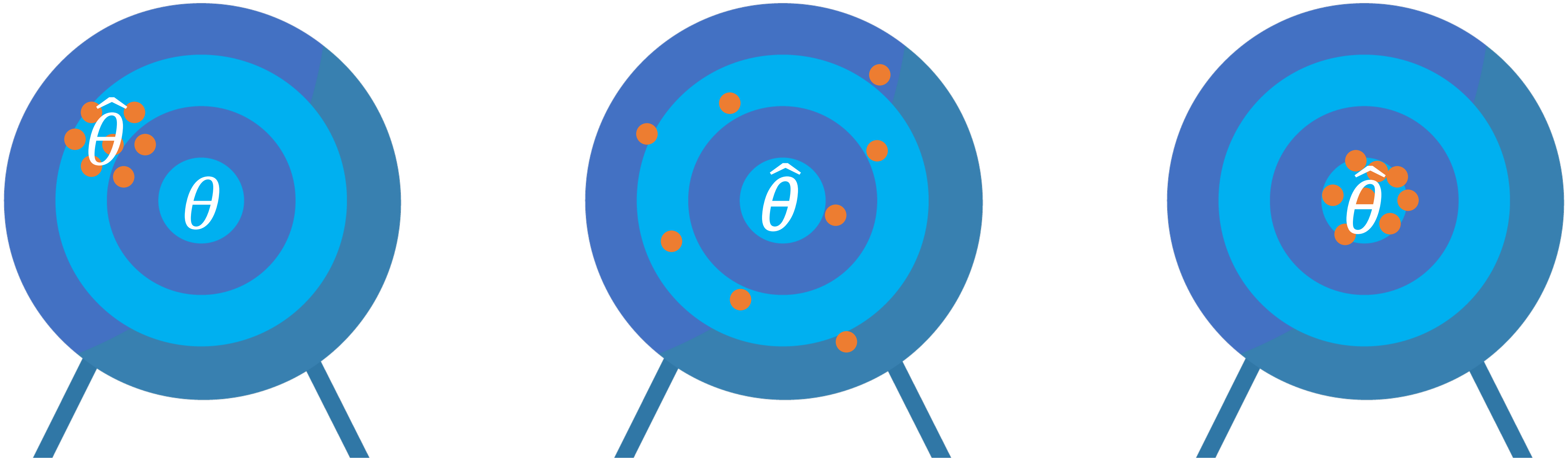
Parámetros: Cantidades fijas y usualmente desconocidas que indexa el modelo y representan características de la población

Estimador

Es una regla a menudo expresada como una fórmula, que indica el procedimiento que debe ser realizado con base en las mediciones contenidas en una muestra para encontrar el valor de una estimación

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

Sesgo y Error Cuadrático Medio



Sesgo y Error Cuadrático Medio

El sesgo de un estimador puntual $\hat{\theta}$ está dado por $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Un estimador puntual $\hat{\theta}$ es insesgado si $E(\hat{\theta}) = \theta$

El error cuadrático medio de un estimador puntual $\hat{\theta}$ es:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = V(\hat{\theta}) + \left[B(\hat{\theta})\right]^2$$

Prueba de hipótesis

Una prueba de hipótesis es un procedimiento a través del cual es posible especificar el hecho de aceptar o rechazar una afirmación. En general estas afirmaciones se realizan sobre la población y la elección entre aceptar o rechazar la afirmación es hecha con base en la muestra de datos.

El método científico

Observar un fenómeno y con definir una pregunta de interés

Plantear una hipótesis acerca de una posible respuesta a la pregunta planteada

Experimentar y recolectar datos

Analizar los datos y con base en ellos ver si la hipótesis se contradice o no

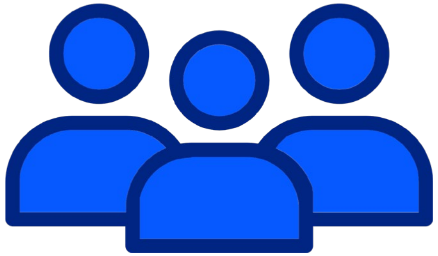
Establecer conclusiones y de ser posible dar respuesta a la pregunta de investigación

Pruebas de hipótesis y el método científico

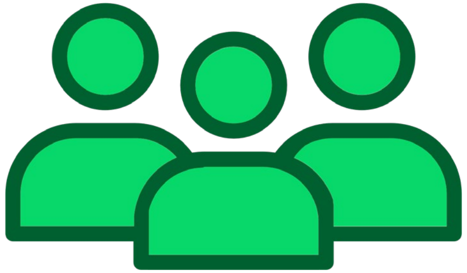
- Plantear una hipótesis o una pregunta de interés frente a un fenómeno
- Tomar una muestra representativa que refleje el comportamiento de dicho fenómeno
- Si lo que se observa en los datos de la muestra contradice la hipótesis que fue planteada se rechaza la hipótesis
- Si lo que se observa en los datos no contradice la hipótesis que fue planteada, no se rechaza la hipótesis

Ejemplo

Se quiere saber si un tratamiento nuevo es más efectivo que el que se aplica tradicionalmente para tratar una enfermedad.



Tratamiento A %



Tratamiento B %

De la muestra y nuestra decisión

¿Cómo utilizar las mediciones muestrales para tomar la decisión?

¿Cómo decidimos si la muestra no concuerda con la hipótesis planteada?

¿Cuándo rechazamos la hipótesis?, ¿cuándo debemos aceptarla?

¿Cuál es la probabilidad de que tomemos una mala decisión?

Estadístico de prueba

El estadístico de prueba es, al igual que los estimadores, una función de las mediciones muestrales en las que la decisión de la prueba estará basada. (Cumple la labor de ser una cantidad de referencia)

Región de rechazo

Elementos de una prueba de hipótesis

1. Hipótesis nula Ho
2. Hipótesis alternativa Ha
3. Estadístico de prueba
4. Región de rechazo

Error tipo 1 y tipo 2

Se comete un error tipo I si Ho es rechazada cuando Ho es verdadera.

$$P(\text{Error tipo I}) = \alpha$$

Se comete un error tipo II si Ho es aceptada cuando Ho es falsa.

$$P(\text{Error tipo II}) = \beta$$

Error tipo 1 y tipo 2 – Ejemplo 1

Se comete un error tipo I si Ho es rechazada cuando Ho es verdadera.

Se comete un error tipo II si Ho es aceptada cuando Ho es falsa.

Región de rechazo

Caso general

Hipótesis	$\begin{cases} Ho: \theta = \theta_0 \\ H_a: \theta \neq \theta_0 \end{cases}$	$\begin{cases} Ho: \theta = \theta_0 \\ H_a: \theta > \theta_0 \end{cases}$	$\begin{cases} Ho: \theta = \theta_0 \\ H_a: \theta < \theta_0 \end{cases}$
Región de rechazo	$\{\lvert Z \rvert > z_{\frac{\alpha}{2}}\}$	$\{Z > z_{1-\alpha}\}$	$\{Z < z_{\alpha}\}$

Estadístico de Prueba: $\boldsymbol{Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}}$

Ejemplo 1 – Media

Hipótesis	$\begin{cases} Ho: \mu = \mu_0 \\ H_a: \mu \neq \mu_0 \end{cases}$	$\begin{cases} Ho: \mu = \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$	$\begin{cases} Ho: \mu = \mu_0 \\ H_a: \mu < \mu_0 \end{cases}$
Región de rechazo	$\{\lvert Z\rvert > z_{\frac{\alpha}{2}}\}$	$\{Z > z_{1-\alpha}\}$	$\{Z < z_{\alpha}\}$

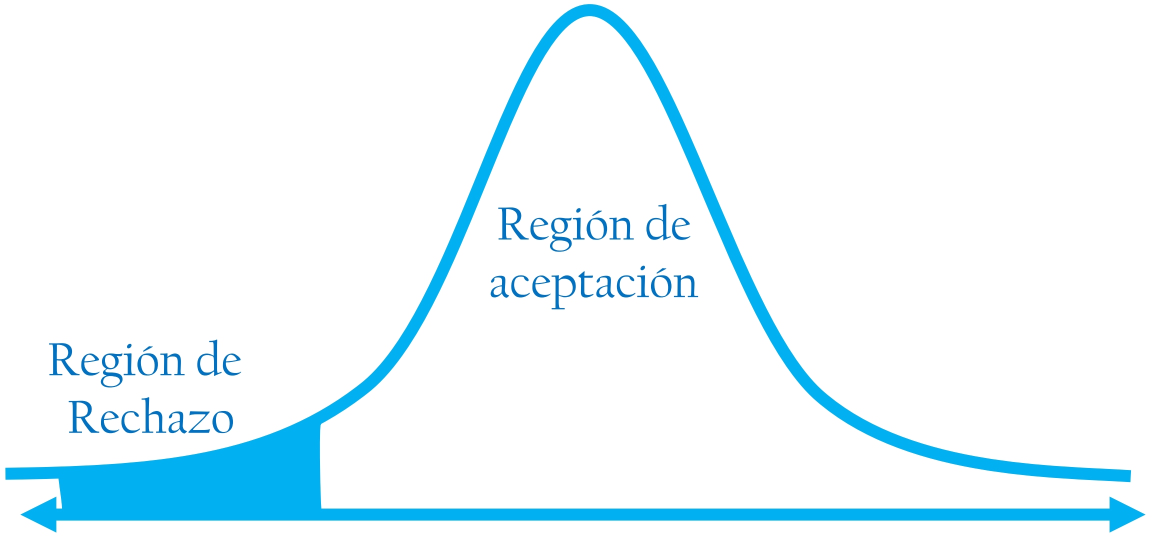
Estadístico de Prueba: $\boldsymbol{Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}$

Ejemplo 1 – Media

Se sabe que la presión arterial de las personas tiene distribución normal con una media de 120 mmHg y una desviación de 10 mmHg.

Se mide la presión arterial de 26 deportistas, encontrando una presión promedio de 104 mmHg.

¿Es posible afirmar que los deportistas tienen una media de presión arterial inferior a la del resto de la población?



$z_\alpha = -1{,}6448$

$$\begin{cases} Ho: \mu = 120 \\ H_a: \mu < 120 \end{cases}$$

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{104 - 120}{\frac{10}{\sqrt{26}}} = -8{,}158433$$

Region de Rechazo:

$$\{|Z| > z_\alpha\}$$

Decisión: Rechazar la hipótesis nula.

Interpretación: Existe evidencia estadísticamente significativa para concluir que los deportistas tienen una presión arterial inferior al resto de la población.

P-valor

El p-valor es la probabilidad de obtener un resultado más extremo o igual que el del estadístico de contraste bajo la hipótesis nula

SI el p-valor es menor que el nivel de significancia, se rechaza la hipótesis nula

Casos comunes

Parámetro – Caso	Hipótesis Nula	Estadístico de Prueba	Distribución	¿Cuándo se usa?
Media Poblacional	$\mu = \mu_0$	$\sqrt{n}\frac{(\bar{X} - \mu_0)}{\sigma}$	Z	- La varianza es conocida - El tamaño de muestra es mayor que 30 o los datos se distribuyen normalmente
Media Poblacional	$\mu = \mu_0$	$\sqrt{n}\frac{(\bar{X} - \mu_0)}{s}$	$t - student$ n-1 grados de libertad	- La varianza es desconocida - El tamaño de muestra es menor que 30
Proporción Poblacional	$p = p_0$	$\sqrt{n}\frac{(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}}$	Z	$n\hat{p}\ y\ n(1 - \hat{p}) \geq 10$

Parámetro – Caso	Hipótesis Nula	Estadístico de Prueba	Distribución	¿Cuándo se usa?
Diferencia de dos medias	$\mu_1 - \mu_2 = d$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	Z	- Las varianzas son conocidas - Los tamaños de muestra son mayores que 30
Diferencia de dos medias	$\mu_1 - \mu_2 = d$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	$t - student$ Con $\dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ grados de libertad	- Las varianzas son desconocidas y diferentes - Los tamaños de muestra son menores que 30
Diferencia de dos medias	$\mu_1 - \mu_2 = d$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}}}$	$t - student$ Con n-2 grados de libertad	- Las varianzas son desconocidas e iguales - Los tamaños de muestra son menores que 30
Diferencia de medias pareadas	$\mu_d = d$	$\sqrt{n}\dfrac{(\bar{D} - d)}{s_d}$	$t - student$ Con n-1 grados de libertad	- Muestras pareadas

Parámetro – Caso	Hipótesis Nula	Estadístico de Prueba	Distribución	¿Cuándo se usa?
Diferencia de dos proporciones	$p_1 - p_2 = d$	$\dfrac{(\hat{p}_1 - \hat{p}_2) - d}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}$	Z	$n\hat{p}\ y\ n(1-\hat{p}) \geq 10$
Varianza poblacional	$\sigma = \sigma_0$	$\dfrac{(n-1)s^2}{\sigma_0^2}$	$Chi\ cuadrado\ (\chi^2)$ n-1 grados de libertad	- Las observaciones se distribuyen de forma normal
Diferencia de varianzas	$\sigma_1^2 = \sigma_2^2$	$\dfrac{s_1^2}{s_2^2}$	$f - fisher$ n-1 grados de libertad en el numerador y n-1 grados de libertad en el denominador	- Las dos poblaciones son normales e independientes

Otras Pruebas de Hipótesis

Parámetro – Caso	Hipótesis Nula	Comando R	¿Cuándo se usa?
Coeficiente de correlación de Pearson	$\rho = 0$	`cor.test(x, y, method = “pearson")`	- Las dos variables son continuas
Coeficiente de correlación de Spearman	$\rho = 0$	`cor.test(x, y, method = "spearman")`	- Las dos variables son continuas
Independencia de variables categóricas	$\rho = 0$	`chisq.test(table(x,y))`	- Las dos variables son cuantitativas

Teorema del límite central

Sea $\{X_i\}$, i=1,...,N, una secuencia de variables aleatorias iid con $E(X) = \mu$ y V$(X) = \sigma^2$. Entonces:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \rightarrow N(0,1)$$

Método de Monte-carlo

Simular una variable aleatoria exponencial

Calcular su promedio

Repetir muchas veces el experimento

Obtener estadísticas de la distribución del promedio

Regresión lineal simple

Modelos

Una teoría o hipótesis a menudo predice una relación entre dos variables. ¿Cómo evalúan los científicos si los datos apoyan o refutan una relación ?

Video

Regresión

La forma más común de análisis de regresión es la regresión lineal en la que se calcula la "mejor línea recta" para un conjunto de datos x, y utilizados para explicar la relación entre ellos.

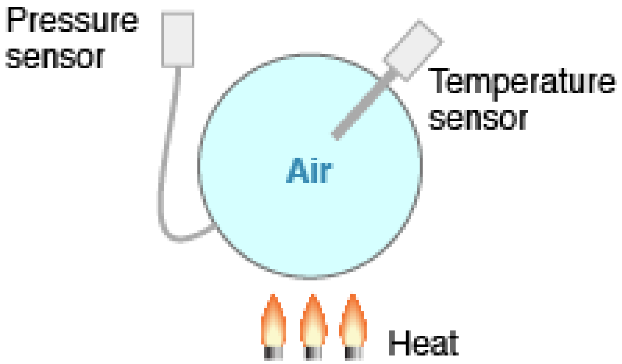
Regresión

Ley de gas ideal: La ley del gas ideal predice que la presión de un gas aumenta linealmente a medida que cambia la temperatura.

Datos

Linear regression

Experiment
Measure pressure
of fixed volume of air
as temperature changes



Data

Temperature (ºC)	Pressure (Pa)
20	111
22	111
25	106
33	112
44	117
47	122
59	123
70	128

Definiciones

Modelo: Representación simplificada de la realidad que contiene los aspectos mas importantes de la misma

Modelo estadístico: Modelo que incorpora un elemento de aleatoriedad

Parámetros: Cantidades fijas y usualmente desconocidas que indexa el modelo y representan características de la población

Análisis de regresión

Un modelo de regresión es un modelo estadístico en que alguna característica distribucional de la variable de interés es afectada por otras variables.

Media (Gauss)
Mediana (Laplace)

Y : Variable de interés
Variable respuesta
Variable dependiente
Regresando

Hecho importante
X afecta el comportamiento de Y

X : Variable independiente
Regresor
Variable exploratoria
Covariables

Si la media de Y es afectada por X tenemos entonces que:

$$E(Y) = \mu = f(x)$$

El caso más importante ocurre cuando la función de x es lineal

$$\mu = \beta_0 + \beta_1 x$$

Caso más importante

Si la media de Y es afectada por X tenemos entonces que:

$$E(Y) = \mu = f(x)$$

El caso más importante ocurre cuando la función de x es lineal

$$\boxed{\mu = \beta_0 + \beta_1 x}$$

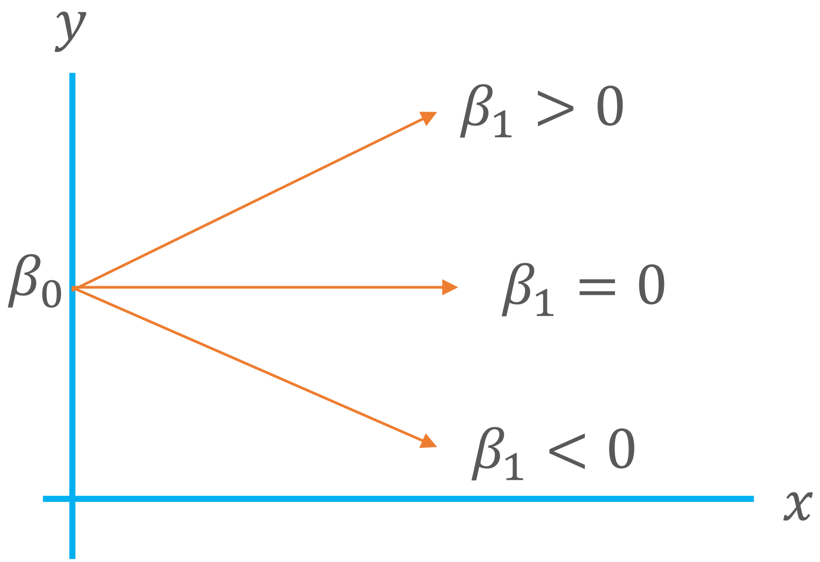
Modelo indexado por parámetros desconocidos y fijos

Modelo de regresión lineal simple

$$\mu = \beta_0 + \beta_1 x$$

$Y \sim alguna\ distribución(\beta_0 + \beta_1 x, \sigma^2)$

$\beta_1, \beta_2, \sigma^2$: Parámetros



Objetivo:
A partir de los datos hacer inferencia
sobre los parámetros

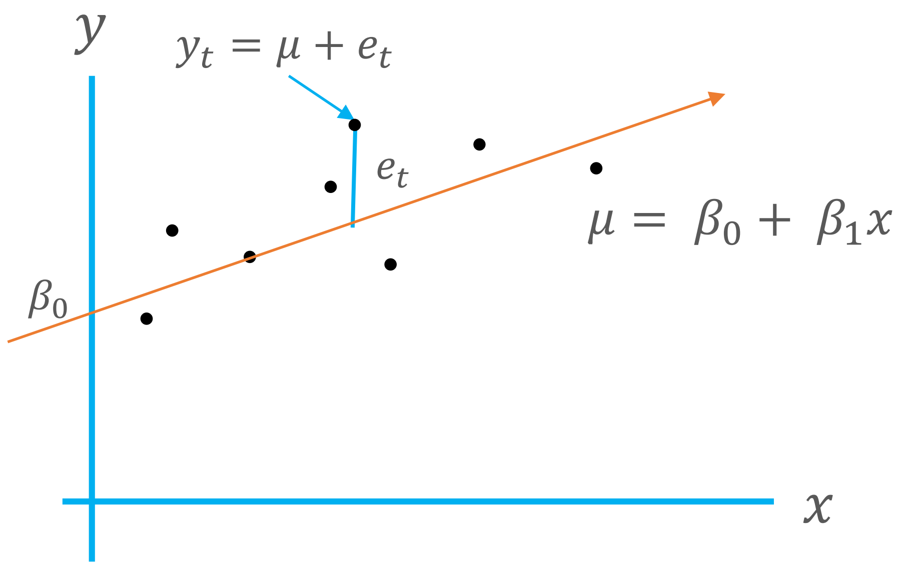
Modelo de regresión lineal simple

$$\mu_t = \beta_0 + \beta_1 x_t \quad \longleftrightarrow \quad y_t = \beta_0 + \beta_1 x_t + e_t$$

t = 1, 2,..., N

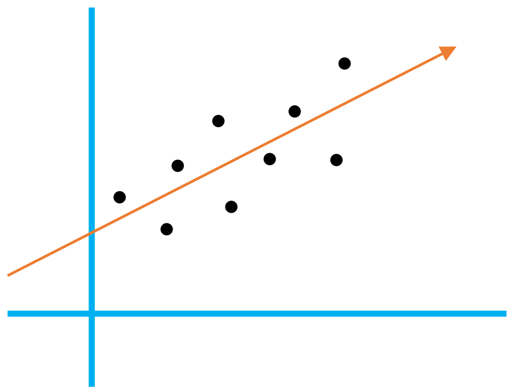
N: Tamaño de la muestra

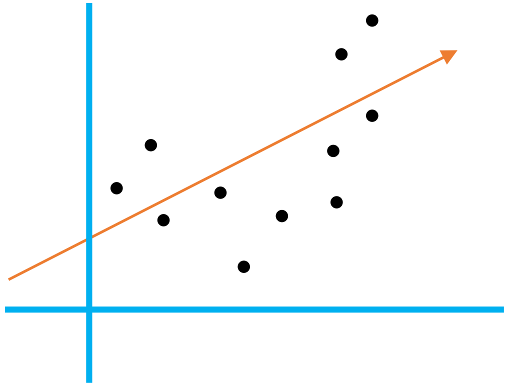
Error: No observable

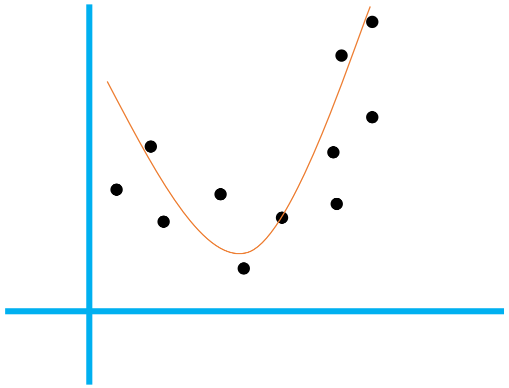


Supuestos del modelo de regresión lineal

[1] El modelo propuesto ofrece una buena descripción de la realidad





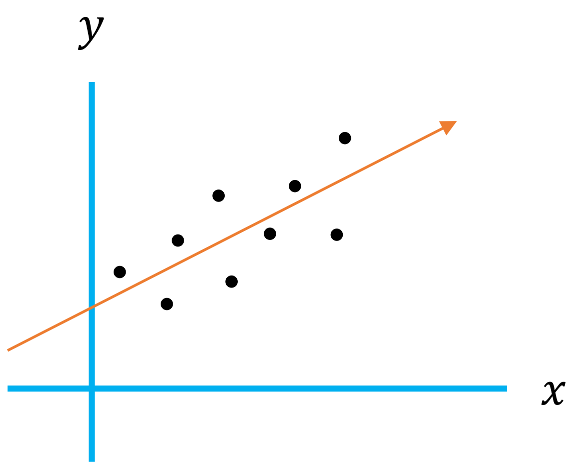


Implicación

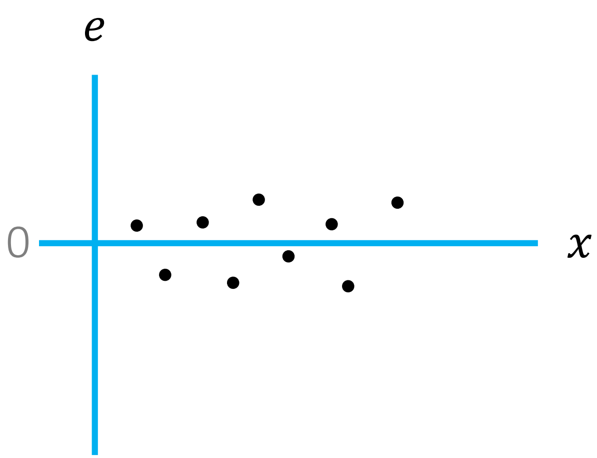
El modelo es adecuado para describir
el fenómeno de interés

Supuestos del modelo de regresión lineal

[2] La media de los errores es cero



$$y_t - \mu = e_t$$

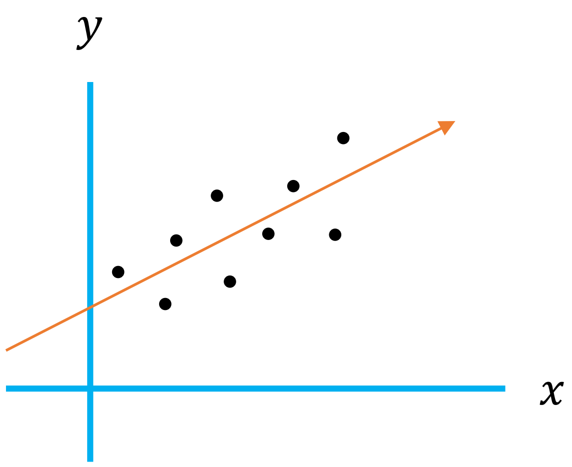


Implicación:
El valor esperado de la variable respuesta es efectivamente la media
*Para x fija

$$\mathrm{E}(\mathrm{y_t}) = \mu_t = \beta_0 + \beta_1 x_t$$

Supuestos del modelo de regresión lineal

[3] La varianza de los errores es constante

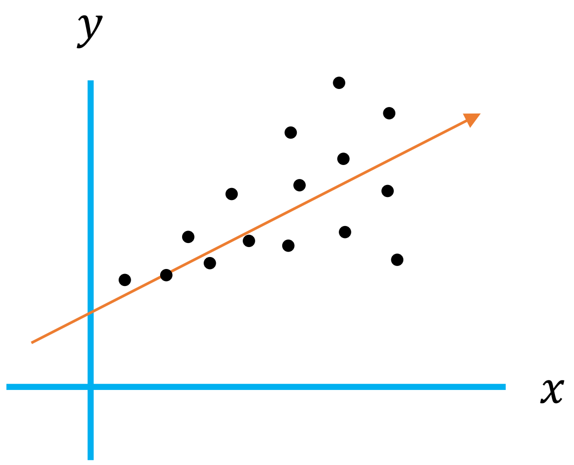


Homocedástico

$var(e_t) = \sigma^2$

Implicación

La varianza de la variable respuesta es también fija

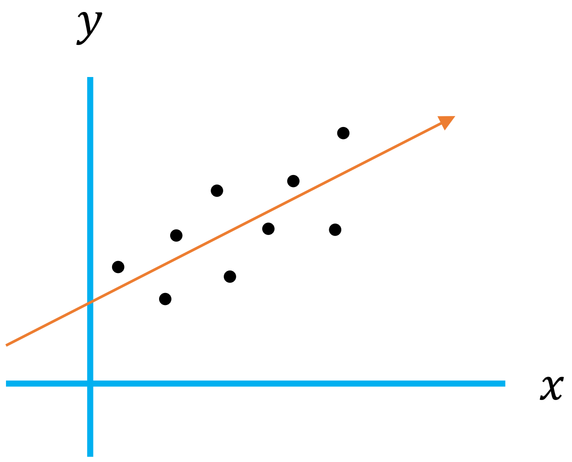


Heterocedástico

$var(e_t) = \sigma_t^2$

Supuestos del modelo de regresión lineal

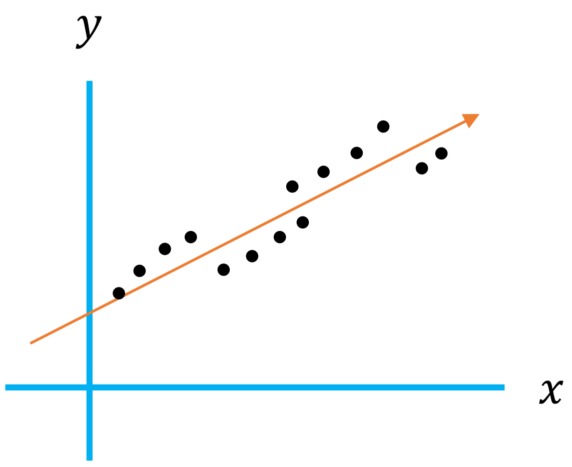
[4] No hay covarianza entre errores diferentes



$cov(e_i, e_j) = 0; \; i \neq j$

Implicación

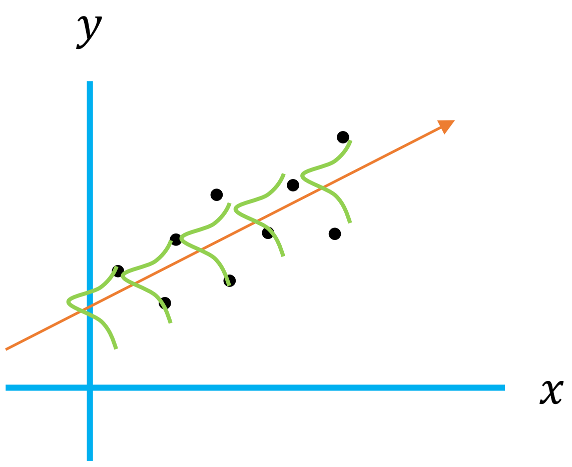
El comportamiento de una de las observaciones no influye en el comportamiento de las otras



$cov(e_i, e_j) \neq 0; \; i \neq j$

Supuestos del modelo de regresión lineal

[5] (Dependiendo del método de estimación) Los errores tienen distribución normal



Implicación

Permite construir intervalos de confianza en caso de que la estimación sea intervalar

El modelo de regresión debe ser lineal en los parámetros, es decir en las cantidades desconocidas. Para el modelo lineal simple los parámetros desconocidos son β_0 y β_1

¿Qué es considerado lineal?

El modelo de regresión debe ser lineal en los parámetros, es decir en las cantidades desconocidas. Para el modelo lineal simple los parámetros desconocidos son β_0 y β_1

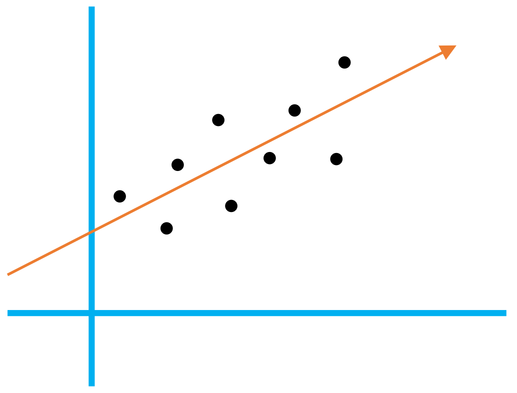
$$y_t = \beta_0 + \beta_1 x_t + e_t \qquad y_t = \beta_0 + \beta_1 x_t^2 + e_t \qquad y_t = \beta_0 + \beta_1 \ln(x_t) + e_t$$

$$\ln(y_t) = \beta_0 + \beta_1 x_t + e_t \qquad y_t = \beta_0 + \beta_1^2 x_t + e_t \qquad y_t = \beta_0 + \ln(\beta_1) x_t + e_t$$

¿Qué representan los parámetros?

β_0 Intercepto
Media de y cuando x = 0
* Preguntarse si tiene sentido su interpretación dependiendo del problema

β_1 Inclinación
Variación en la media de la variable respuesta cuando la regresora aumenta una unidad



$$consumo_t = \beta_0 + \beta_1 ingresos_t + e_t$$

β_0 Consumo promedio de una persona cuyos ingresos son cero

β_1 Aumento promedio en el consumo cuando los ingresos aumentan en una unidad ("Propensión marginal al consumo")

Estimación por mínimos cuadrados

[Objetivo] Encontrar estimadores de los parámetros del modelo de regresión lineal simple.

[¿Cómo?] Minimizando cosas “malas” o maximizando cosas “buenas”

Malo: Errores positivos y negativos se cancelan $\sum_{t=1}^{N} e_t$

Regresión L1 (Difícil de estimar por esta vía) $\sum_{t=1}^{N} |e_t|$

Regresión L2 (Estimación de mínimos cuadrados) $\sum_{t=1}^{N} e_t^2$

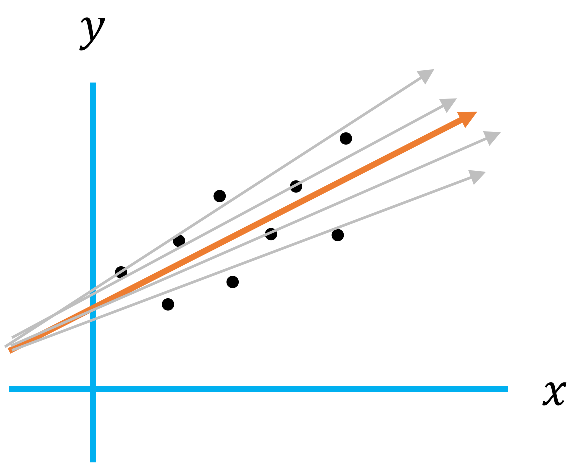
Errores
Minimizar los errores o una medida agregada de los errores

Verosimilitud
Máximizar la probabilidad de que la muestra en conjunto sea la observada

Estimación por mínimos cuadrados

En la estimación por mínimos cuadrados se buscan los argumentos que minimicen la suma de los cuadrados de los errores

$$\sum_{t=1}^{N} e_t^2 = \sum_{t=1}^{N} (y_t - (\beta_0 + \beta_1 x))^2$$



Obtención de los estimadores

1) Desarrollar el cuadrado
2) Derivar respecto a β_0 y β_1 e igualar a cero
3) Despejar los valores de β_0 y β_1 del sistema de ecuaciones obtenido en 2.
4) Usar $\hat{\beta}_0$ y $\hat{\beta}_1$ como estimadores para calcular la media a predecir

$\mu = \hat{\beta}_0 + \hat{\beta}_1 x$

El procedimiento completo puede ser encontrado en [Ref 1 – pag 540]

Estimadores de mínimos cuadrados

Intercepto	Pendiente
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$\hat{\beta}_1 = \dfrac{N\sum x_t y_t - \sum x_t \sum y_t}{N\sum x_t^2 - (\sum x_t)^2}$
$var(\hat{\beta}_0) = \sigma^2 \dfrac{\sum x_t^2}{N\sum (x_t - \bar{x})^2}$	$var(\hat{\beta}_1) = \sigma^2 \dfrac{1}{\sum (x_t - \bar{x})^2}$
$cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \dfrac{\bar{x}}{\sum (x_t - \bar{x})^2}$	

Cuanto mayor sea σ^2, menos precisos son los estimadores

A mayor tamaño de muestra, las varianzas de los estimadores tienden a cero (Consistentes)

Entre más variables los valores de x, más precisos son los estimadores

Práctica en R

```
rm(list=ls)
modelo <- lm(iris$Sepal.Length ~ iris$Petal.Width + iris$Petal.Length)
summary(modelo)
plot(modelo)
```

Descomposición del error

$$Y'Y = \hat{Y}\hat{Y} + \hat{e}'\hat{e}$$

Suma de cuadrados de los errores

(SST)

$$\boxed{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \boxed{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} + \boxed{\sum_{i=1}^{n}\hat{e}_i^2}$$

Suma de los errores **no** explicados por la regresión

SSE

Suma de cuadrados de los errores explicados por la regresión

(SSR)

$$SST = SSR + SSE$$

Coeficiente de determinación

$$\frac{SST}{SST} = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$$1 = \boxed{\frac{SSR}{SST}} + \frac{SSE}{SST}$$

Coeficiente de determinación R^2

$$R^2 = 1 - \frac{SSE}{SST}$$

% de variabilidad explicado por la regresión

Propiedades:

- ✓ Es **no** decreciente respecto al numero de variables regresoras
- ✓ Es una proporción (toma valores entre 0 y 1)
- ✓ Mide la variación proporcional en la variación total de y que es explicada por el modelo

Coeficiente de determinación ajustado

$$\bar{R}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

Su finalidad es medir el poder de discriminación de un modelo

Propiedades:

- ✓ Puede decrecer cuando aumenta la cantidad de variables regresoras
- ✓ No es una proporción (puede tomar valores negativos)
- ✓ No se interpreta, sólo se usa como criterio de decisión entre diferentes modelos