


Estadística y modelación de sistemas socioecológicos en R



Laboratorio
Nacional
de Ciencias
de la Sostenibilidad

Dra. Yosune Miquelajauregui Graf

Plan del día

1. Introducción a la estadística
 2. Estadística descriptiva
 3. Regresión lineal
 4. Ejercicio
- 

Introducción a la estadística

¿Para qué hacer estadística?

1. Estimar parámetros

¿Cuál es la densidad del ocelote en el Área Natural Protegida, Reserva de la Biósfera Janos, en Chihuahua?

2. Hacer pruebas de hipótesis

¿Las diferencias de crecimiento entre los niños que viven en zonas urbanas y los que viven en zonas rurales se debe al azar o es resultado de un efecto del tipo de alimentación?

3. Hacer inferencias

Cerca del 20% de la población mexicana votará por Morena en las próximas elecciones del 2018.

Introducción a la estadística

Población y muestra

La población estadística es el conjunto de elementos sobre el cual basamos nuestras conclusiones. Desconocido.

(hombres de 20 a 35 años en la UNAM)

Sin embargo no conocemos los parámetros que caracterizan a la población (e.x. la altura)

Estrategia 1 : Medir la altura de todos los hombres de la UNAM (poco práctico logísticamente)

Estrategia 2 : Utilizar una muestra de 50 hombres de 20 a 35 años seleccionados aleatoriamente.



Introducción a la estadística

Población y muestra

Se puede inferir sobre la población a partir de la muestra:

Si $\bar{x} = 1.7$ m; podríamos decir que la media de la muestra \bar{x} es un estimador de la media de la población y que 1.7 es un estimado de ese valor (el estimador produce un estimado).

Característica de la población : PARÁMETRO (e.x. la media poblacional μ)


Característica de la muestra : ESTIMADOR O ESTADÍSTICO (e.x. la media muestral \bar{x})

Estadística descriptiva

Medidas de tendencia central


Distintos estadísticos permiten caracterizar una muestra y de estimar los parámetros de la población:

Medidas de tendencia central (posición):

1. Media - mean ()
 2. Mediana - median ()
 3. Moda - mode ()
- 

Estadística descriptiva

Medidas de dispersión

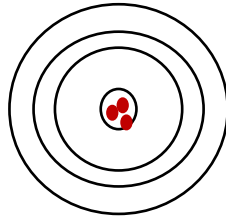
1. Varianza - `var ()`
 2. Desviación estándar - `sd ()`
 3. Rango - `range ()`
 4. Suma de cuadrados del error (SCE)
- 

Estadística descriptiva

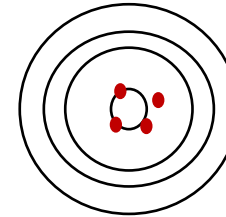
Medidas de precisión

1. Error tipo o error estándar del estimador (SE) – una medida de la imprecisión de los valores estimados. El SE mide la variabilidad de las diferentes estimaciones, si el muestreo se repite un gran número de veces.

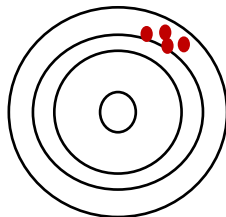
Preciso y exacto



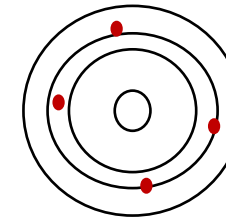
No preciso pero exacto



Preciso pero no exacto



Ni preciso ni exacto



Estadística descriptiva

Sesgo

1. Diferencia entre el valor esperado del estimador y el valor real del parámetro a estimar. Es deseable que un estimador sea insesgado.
2. El sesgo representa la tendencia de los estimadores de un parámetro a diferir sistemáticamente.

$$Sesgo = E(\hat{\theta}) - \theta$$

donde $\hat{\theta}$ es el estimador de un parámetro, $E(\hat{\theta})$ es el valor esperado del estimador del parámetro y θ es el valor del parámetro.

Estadística descriptiva

Variables aleatorias

1. Una variable cuyos valores observados son resultado de un proceso aleatorio (experimento aleatorio).
2. Las variables aleatorias pueden ser discretas o continuas:

Discretas: Toma únicamente valores enteros:

- a) Binarias (e.x. presencia/ausencia, muerto/vivo)
- b) Categóricas y ordinales (e.x. pequeño, mediano, grande)
- c) Número de individuos (e.x. 0, 1, 23, 54)

Continuas: Toma un número infinito de valores dentro del intervalo dado (e.x. distancia, temperatura, largo).

Estadística descriptiva

Distribuciones estadísticas

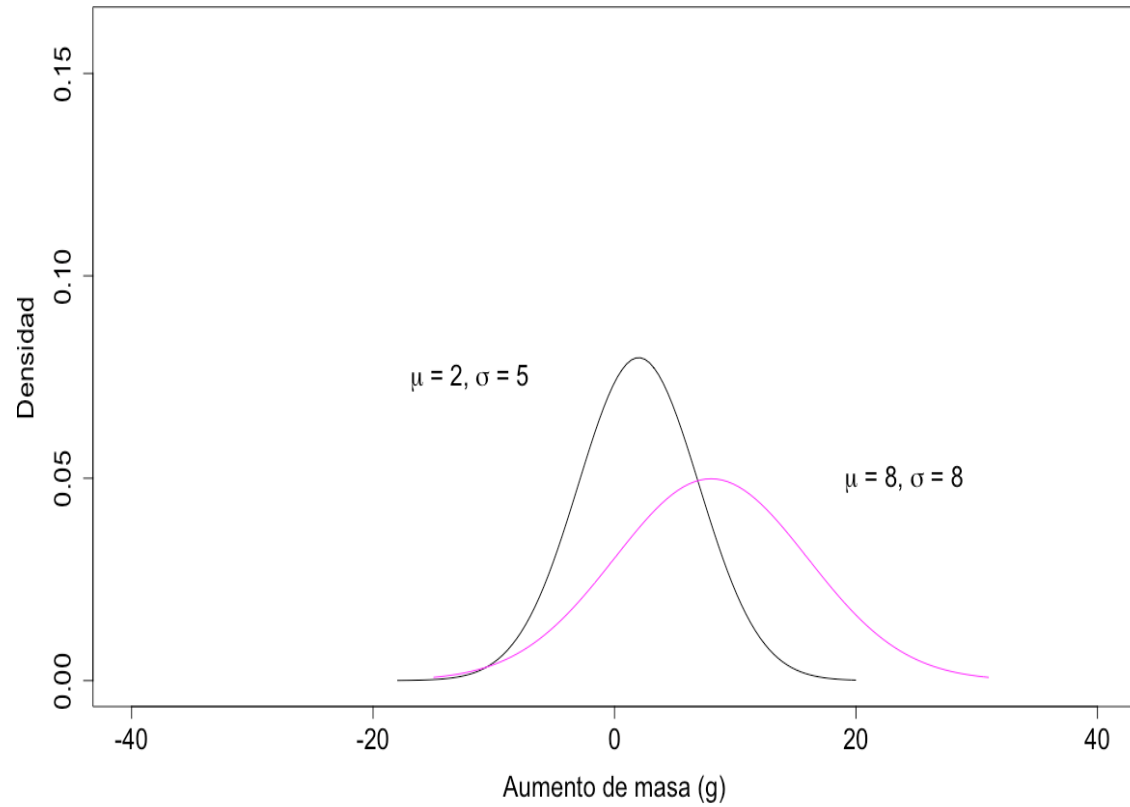
La mayor parte de los análisis estadísticos dependen de una distribución estadística (análisis paramétricos tales como prueba T, ANOVA, regresión lineal y múltiple).

Los análisis paramétricos involucran una serie de supuestos asociados a los parámetros de la distribución.

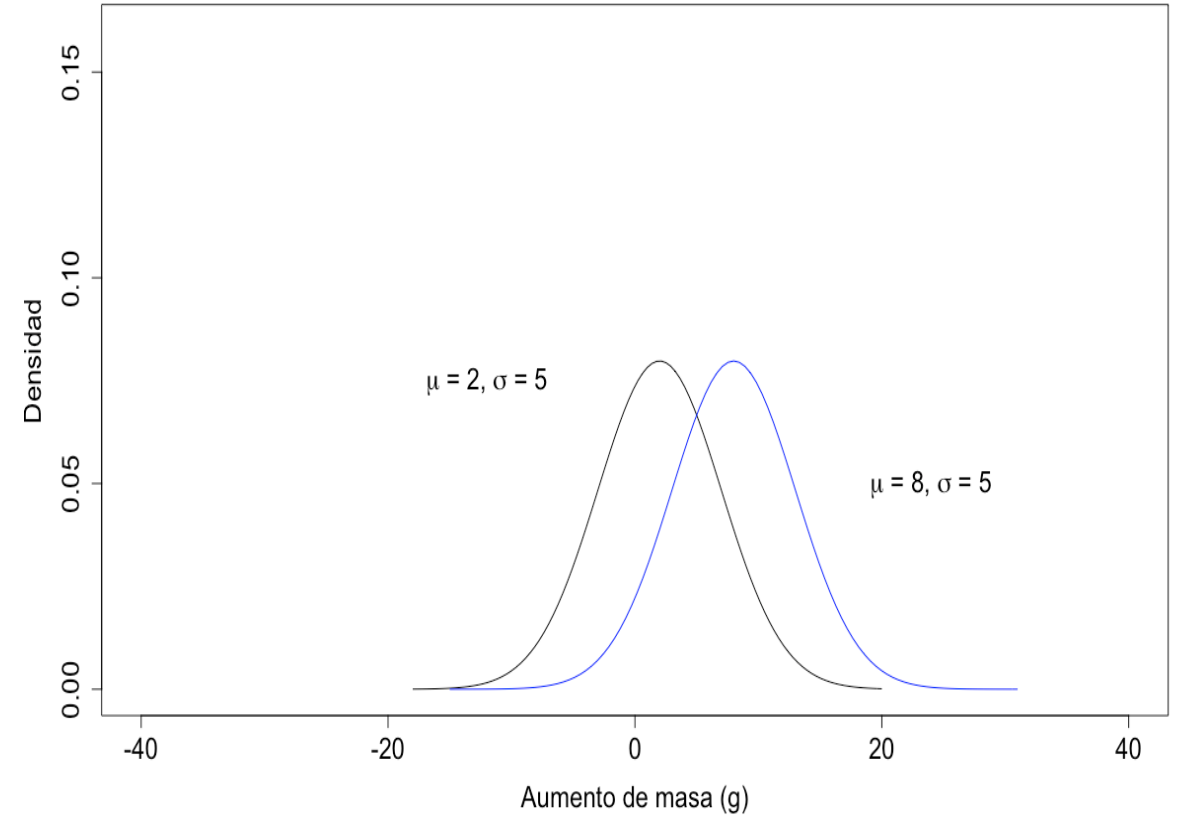
Por ejemplo, la prueba de T para dos grupos independientes supone que las muestras son aleatorias y que provienen de poblaciones normales cuyas varianzas son iguales.

Estadística descriptiva

Comparación entre dos grupos

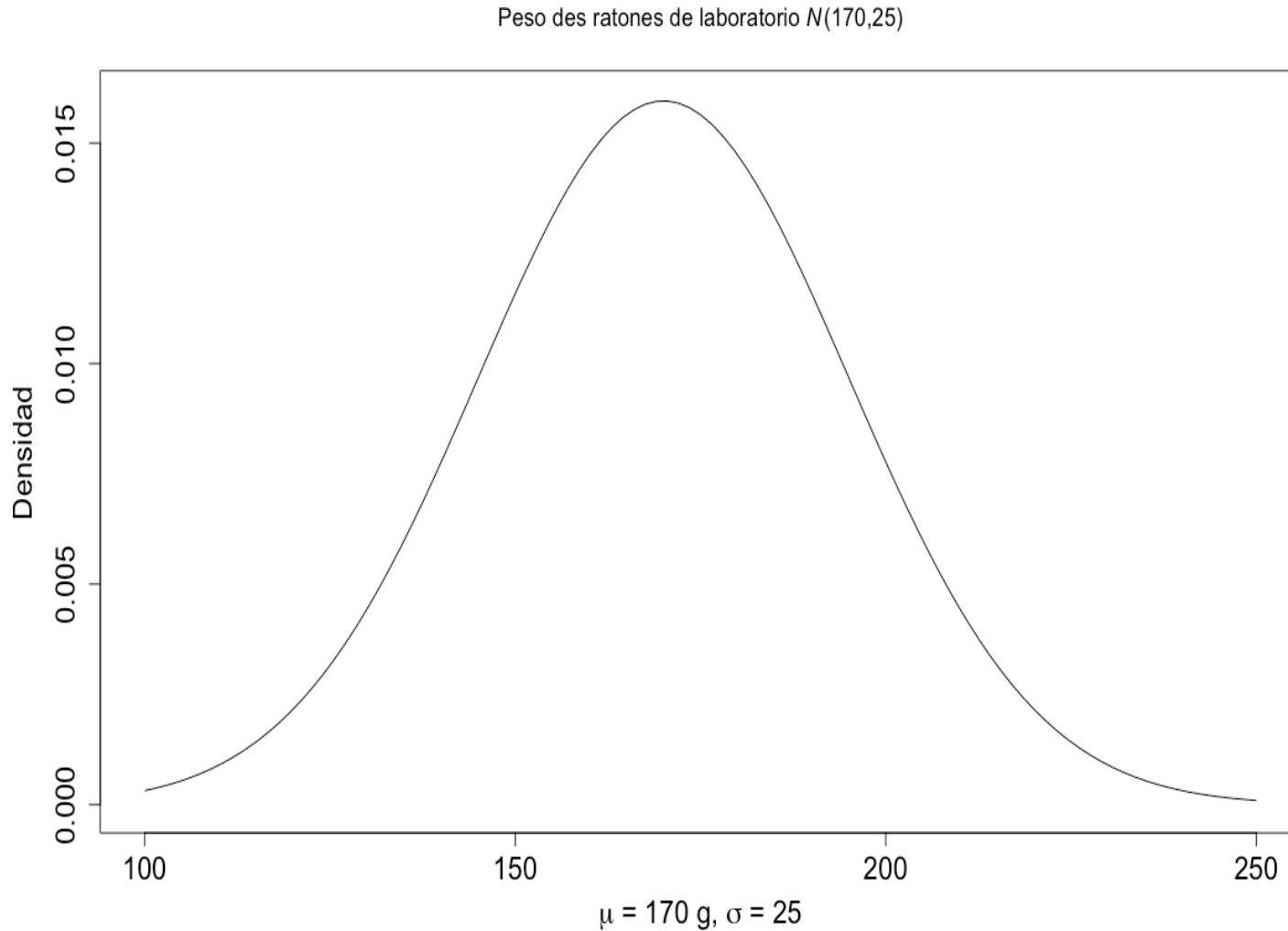


Comparación entre dos grupos



Estadística descriptiva

Distribución normal

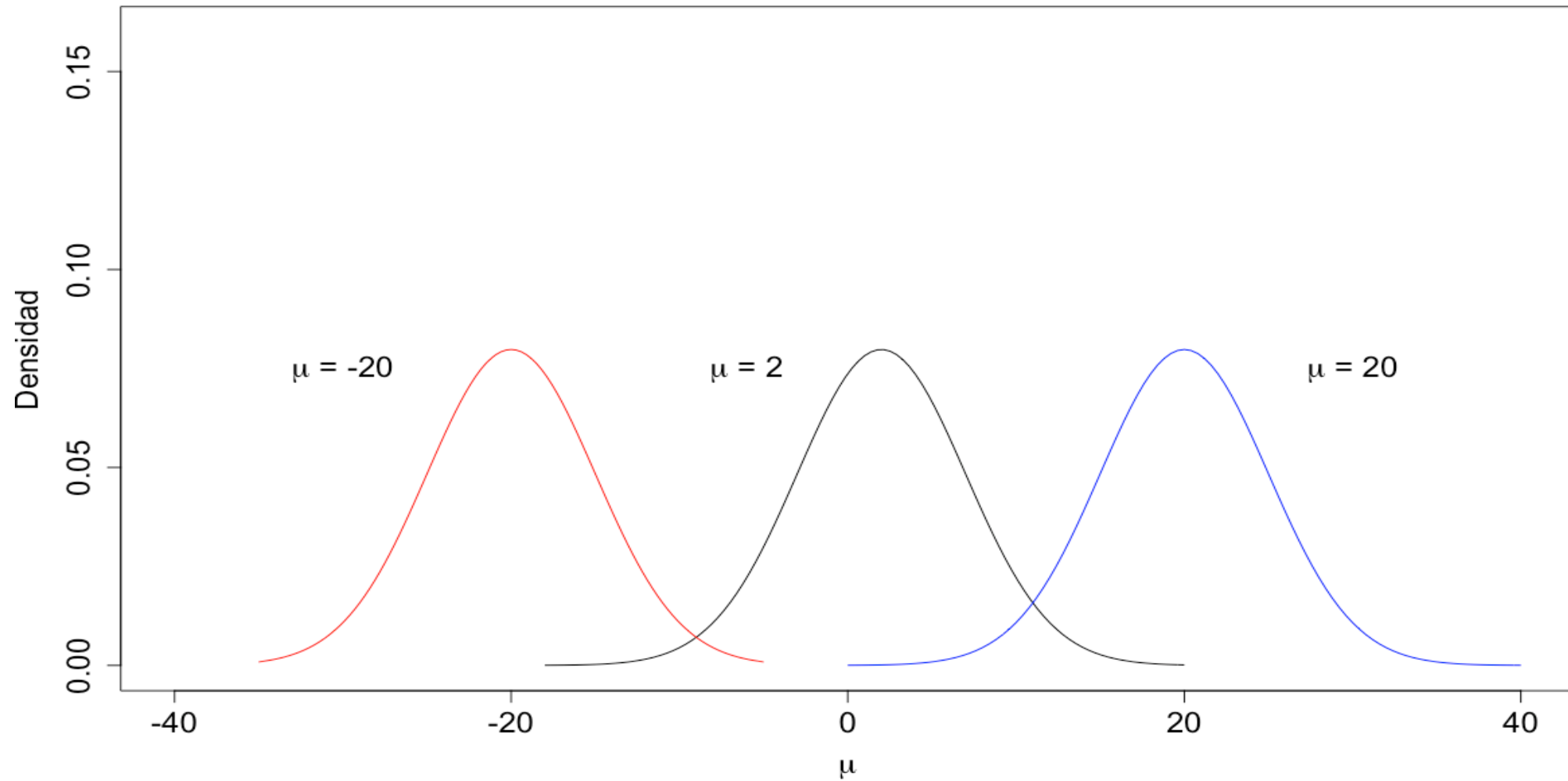


Características:

- Distribución continua
- La suma del área bajo la curva es 1
- Distribución simétrica
- 90% de las observaciones se encuentran a 1.64σ de μ
- 95% de las observaciones se encuentran a 1.96σ de μ
- 99% de las observaciones se encuentran a 2.58σ de μ

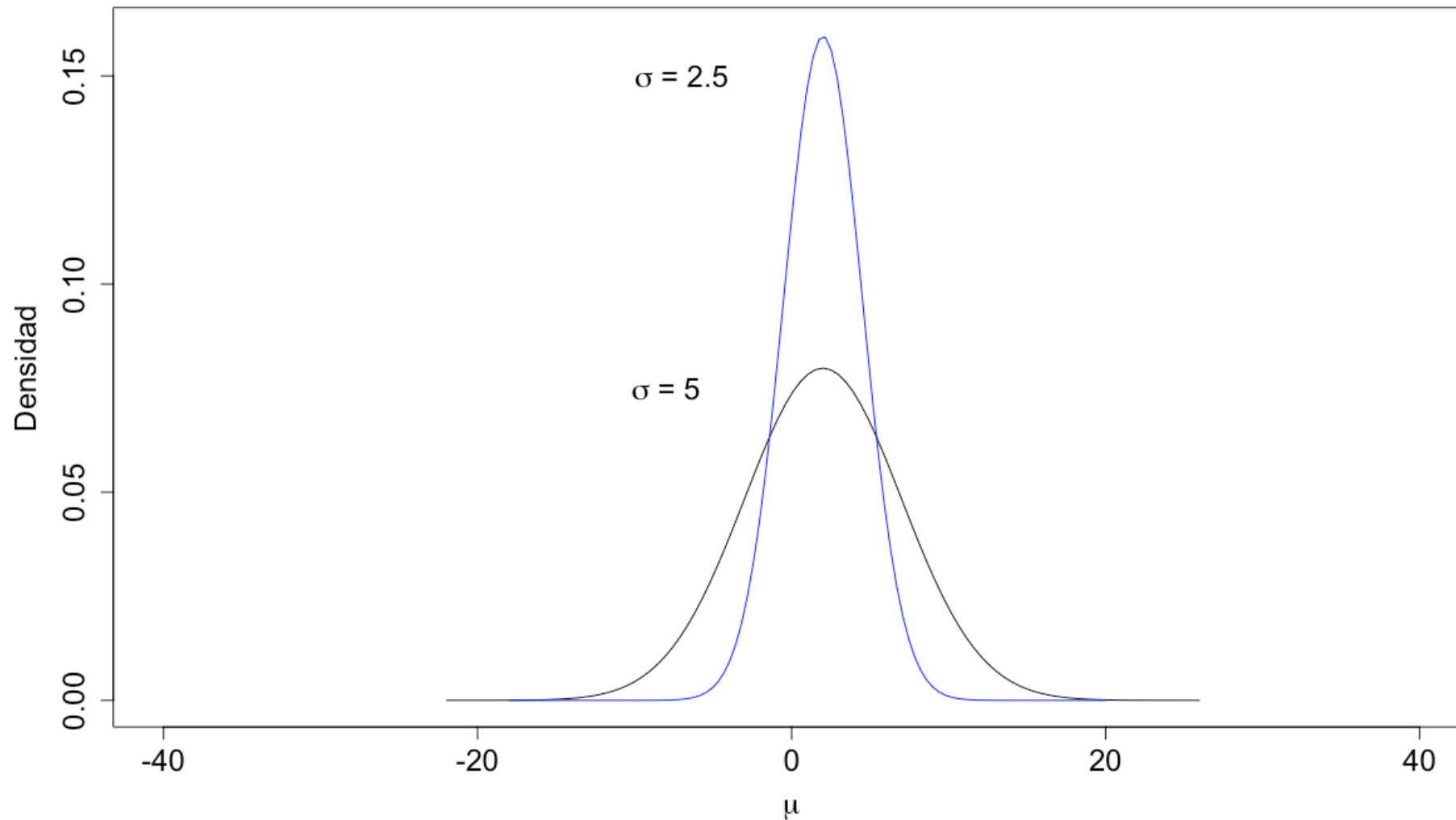
Estadística descriptiva

Distribución normal: la media determina la posición



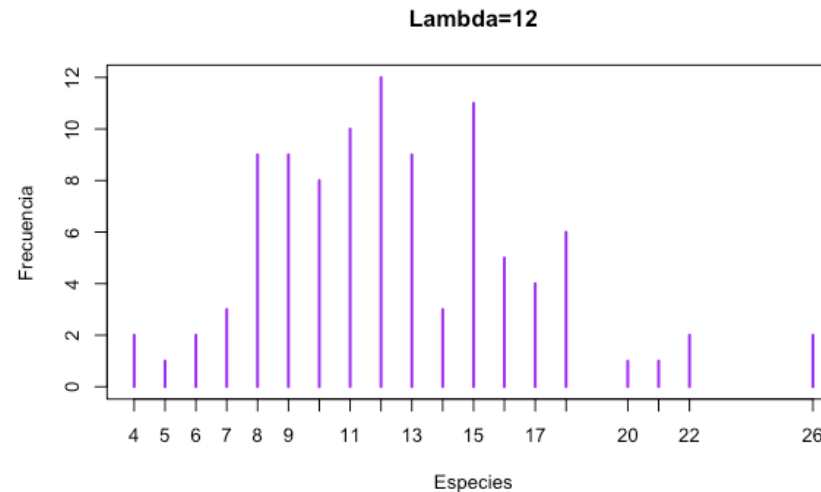
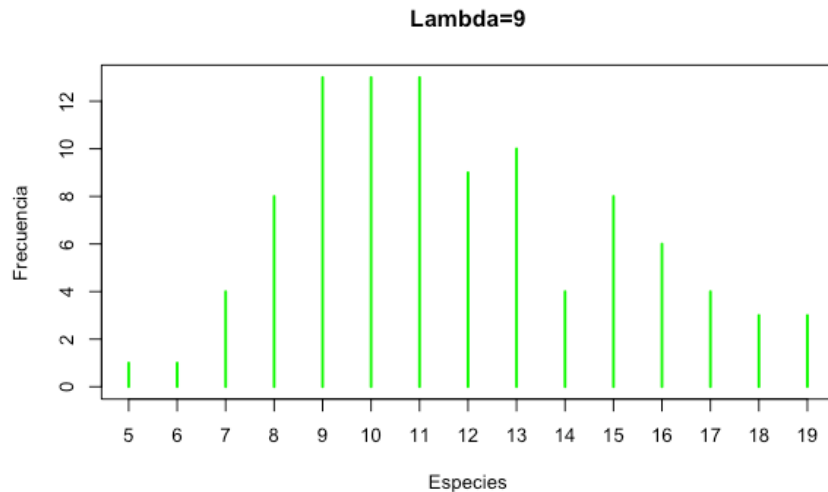
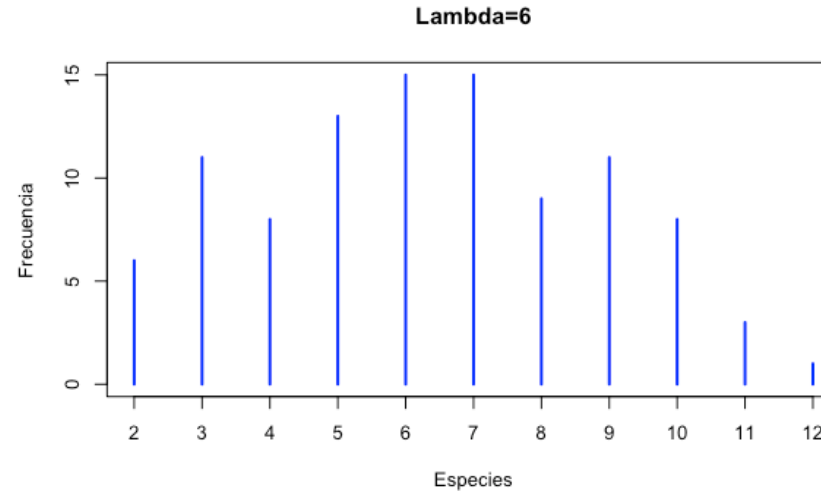
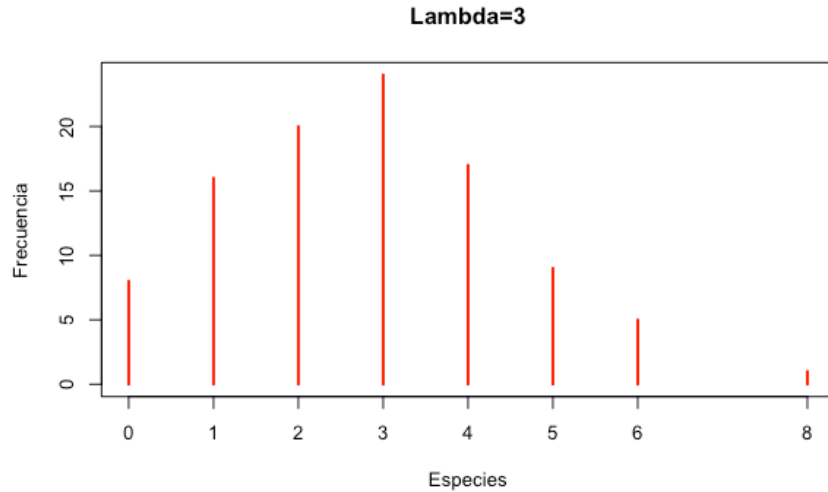
Estadística descriptiva

Distribución normal: la varianza determina la forma



Estadística descriptiva

Distribución Poisson

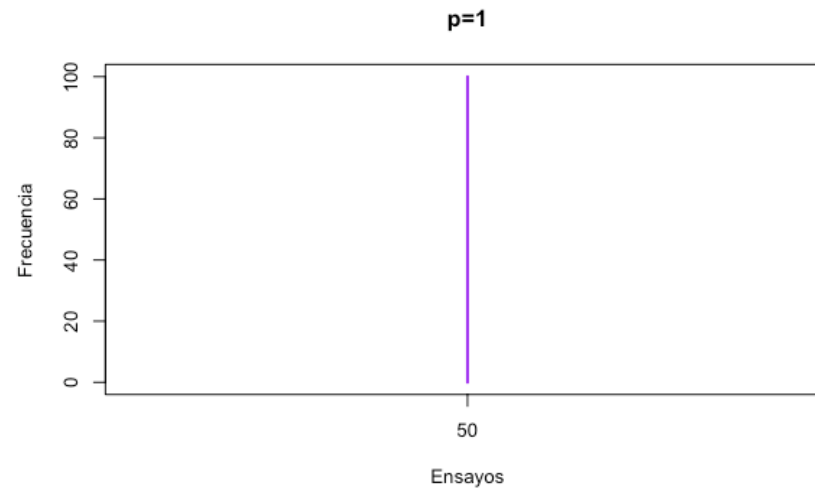
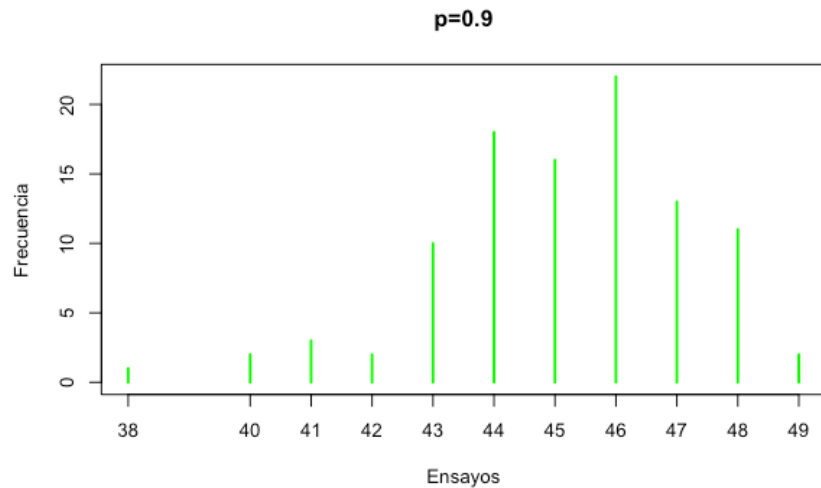
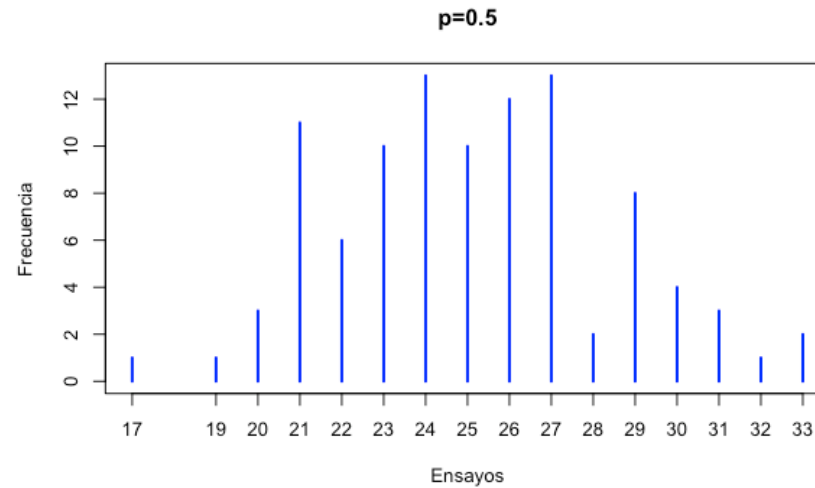
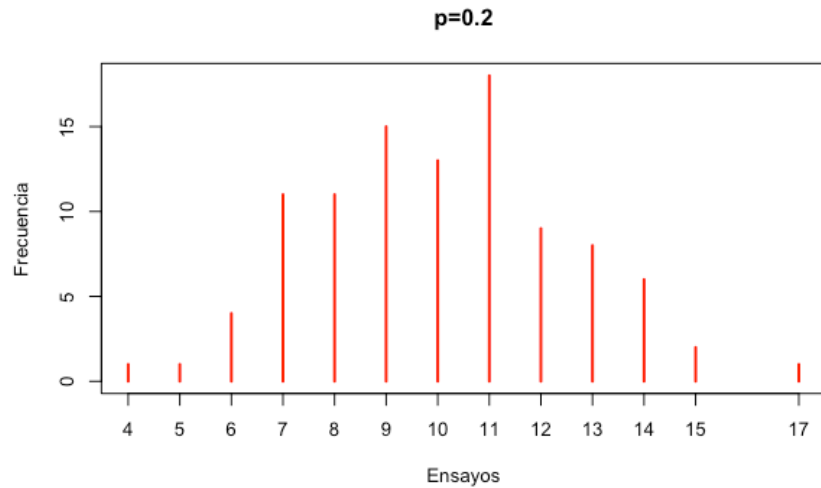


Características:

- Se aplica a fenómenos discretos de la naturaleza
- Definida por el parámetro λ , que representa el número de veces que se espera que ocurra el fenómeno.

Estadística descriptiva

Distribución Binomial



Características:

- Discreta que cuenta el número de éxitos x en una secuencia de n ensayos independientes entre sí, con una probabilidad fija p de ocurrencia del éxito entre los ensayos.

Correlación y regresión

Correlación : Indica la magnitud y dirección de una relación lineal entre dos variables estadísticas. No hace distinción entre la variables dependiente e independiente. Se utiliza frecuentemente el coeficiente de correlación de Pearson como índice de la relación.

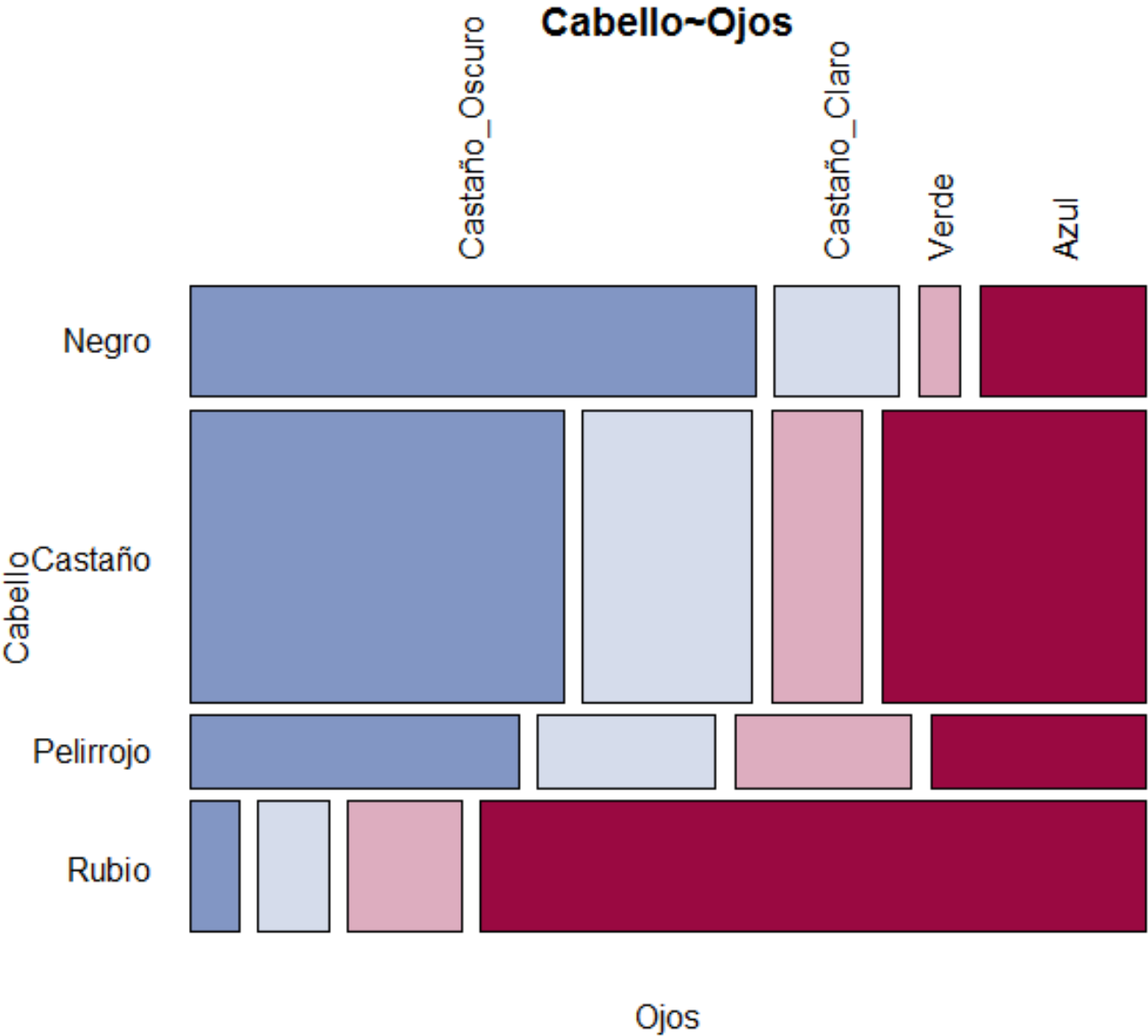
Regresión: Hace distinción entre la variable dependiente e independiente. Su objetivo es el de “predecir” los valores de la variable dependiente basado en los valores de la variable independiente

Correlación: visualización

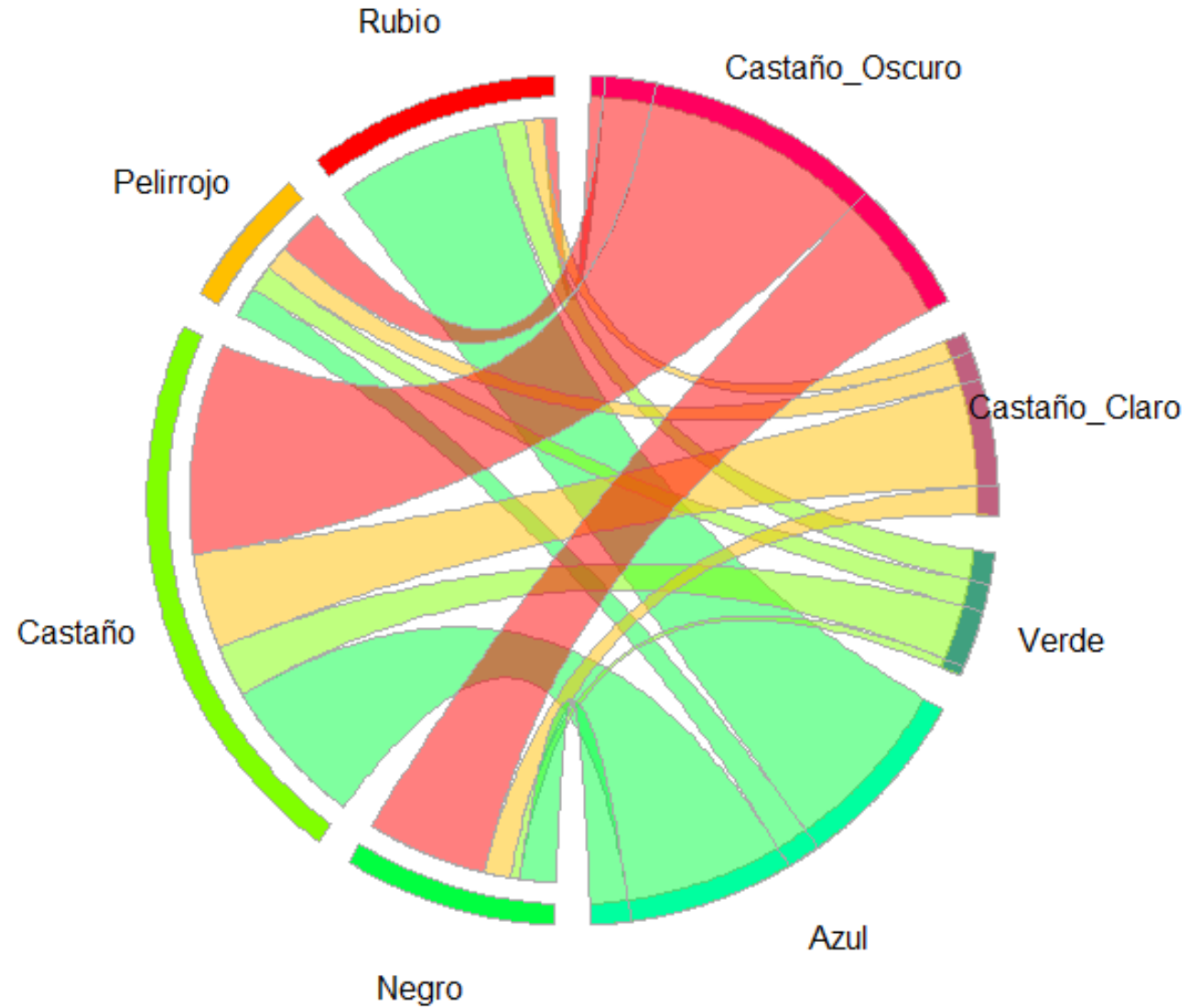
Porcentaje de personas reportadas en cada combinación
Cabello-Ojos

	Ojos	Castaño oscuro	Castaño claro	Verde	Azul
Cabello					
Negro		11.5	2.5	0.8	3.4
Castaño		20.1	9.1	4.9	14.2
Pelirrojo		4.4	2.4	2.4	2.9
Rubio		1.2	1.7	2.7	15.9

Correlación: visualización

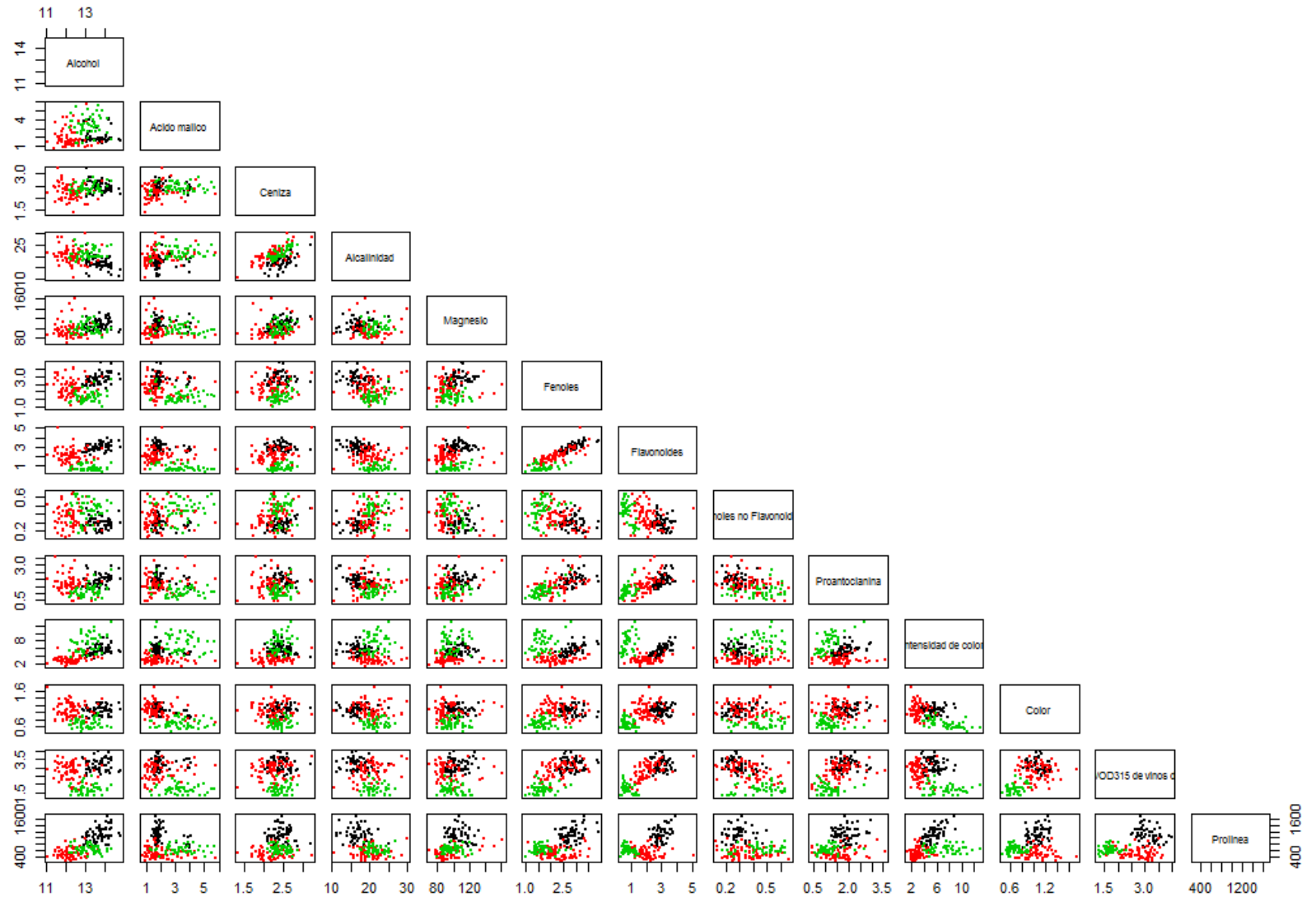


Correlación: visualización



Correlación: visualización

Es posible visualizar las correlaciones entre cada par de variables independientemente. Sin embargo, puede resultar un proceso poco intuitivo.



Correlación: ejemplo

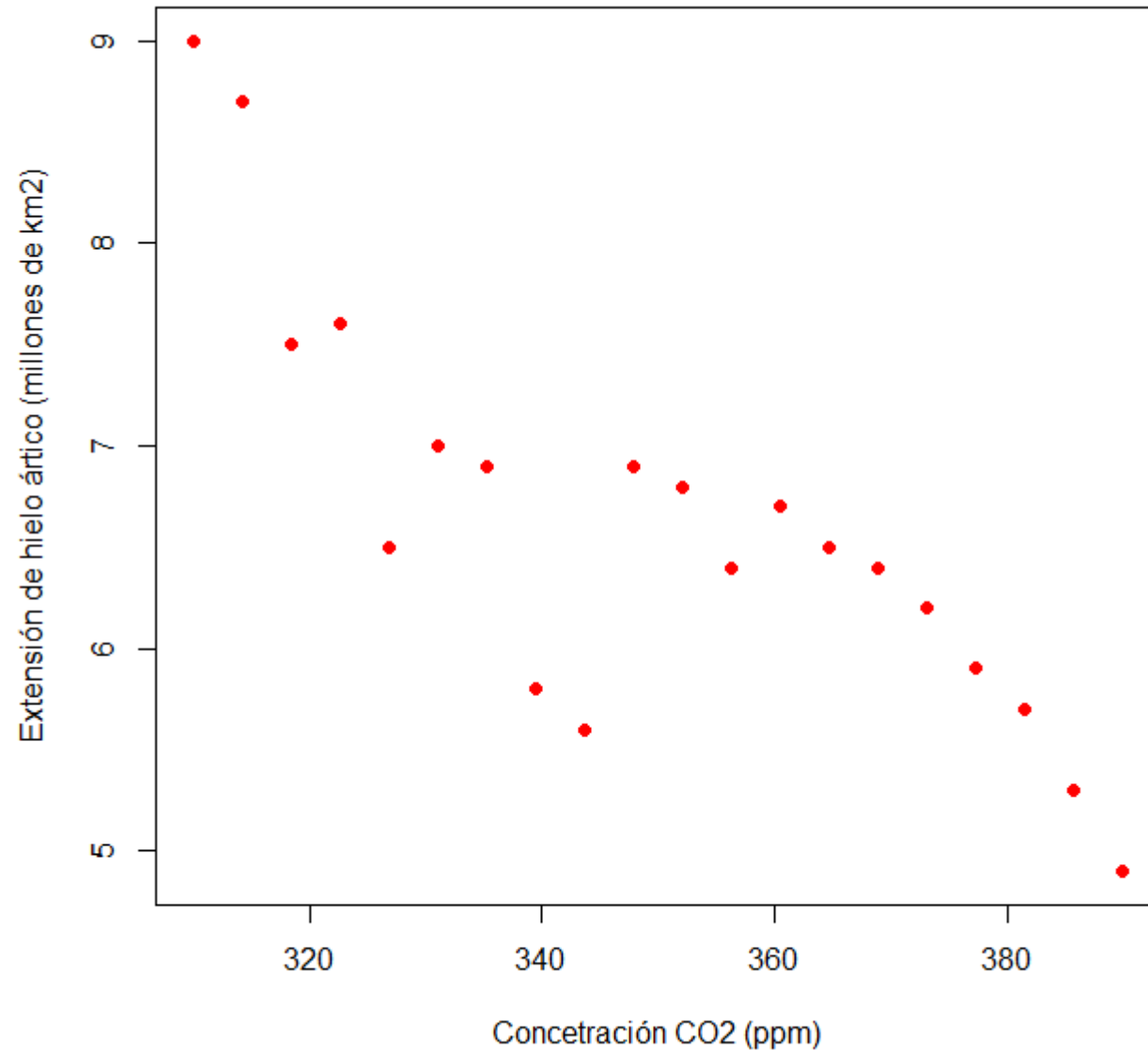
Coeficiente de correlación de Pearson

Índice que puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas.

$$r = \frac{(\sum xi * yi) - (n * \bar{x}\bar{y})}{(n - 1) * \sigma_x * \sigma_y}$$



Correlación: ejemplo



Correlación: ejemplo

CO2	Extensión hielo ártico	(xi*yi)
310	9	2790
314.21	8.7	2733.627
318.42	7.5	2388.15
322.63	7.6	2451.988
326.84	6.5	2124.46
331.05	7	2317.35
335.26	6.9	2313.294
339.47	5.8	1968.926
343.68	5.6	1924.608
347.89	6.9	2400.47343
352.11	6.8	2394.31604
356.32	6.4	2280.42112
360.53	6.7	2415.52621
364.74	6.5	2370.7892
368.95	6.4	2361.26336
373.16	6.2	2313.57898
377.37	5.9	2226.47356
381.58	5.7	2174.9946
385.79	5.3	2044.68435
390	4.9	1911
350.00	6.62	45905.924
24.91	1.03	
20.00		

$$r = \frac{(\sum xi * yi) - (n * \bar{x} \bar{y})}{(n - 1) * \sigma_x * \sigma_y}$$

n= 20

\bar{x} = 350.00

\bar{y} = 6.62

σ_x = 24.91

σ_y = 1.03

$$\sum xi * yi = 45905.924$$

$$r = 45905.924 - (20 * (350.0 * 6.62)) / 19 * 24.91 * 1.03$$

$$= 45905.94 - 46340.0 / 487.488 = \mathbf{-0.89}$$



Regresión lineal simple

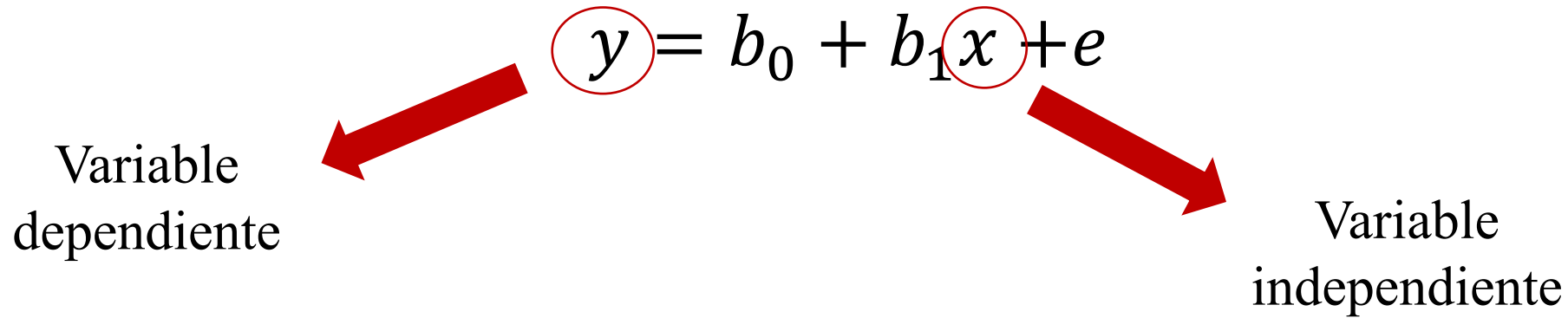
Dos o más variables pueden estar involucradas en el análisis de regresión y correlación. Si solamente están involucradas **dos variables**, se dice que la técnica es una **regresión simple**. Cuando están implicadas **tres o más variables**, se tratará de una **regresión múltiple**.

La técnica de regresión se refiere al procedimiento de obtener una ecuación con fines de estimación o predicción. La regresión lineal se refiere a una función de relación que puede expresarse gráficamente mediante una línea recta:

$$y = b_0 + b_1x + e$$

Regresión lineal simple

- Investigar si existe una **relación estadística** entre la variable dependiente y la variable independiente.
- Estudiar la **forma** de la relación lineal (positiva o negativa)
- Estudiar la **fuerza** de la relación a través del coeficiente de determinación (r^2)




The diagram shows the simple linear regression equation $y = b_0 + b_1x + e$. The variable y is circled in red, and a red arrow points from it to the text "Variable dependiente" on the left. The variable x is also circled in red, and a red arrow points from it to the text "Variable independiente" on the right.

$$\text{Variable dependiente } y = b_0 + b_1x + e \text{ Variable independiente}$$

Regresión lineal simple

- Investigar si existe una **relación estadística** entre la variable dependiente y la variable independiente.
- Estudiar la **forma** de la relación lineal (positiva o negativa)
- Estudiar la **fuerza** de la relación a través del coeficiente de determinación (R^2)


$$y = b_0 + b_1x + e$$


Parámetros a
estimar

b_0 : ordenada al origen
 b_1 : pendiente de la recta

Regresión lineal simple

- Investigar si existe una **relación estadística** entre la variable dependiente y la variable independiente.
- Estudiar la **forma** de la relación lineal (positiva o negativa)
- Estudiar la **fuerza** de la relación a través del coeficiente de determinación (R^2)


$$y = b_0 + b_1x + e$$


Error

Regresión lineal simple: supuestos

CONSIDERACIONES PREVIAS

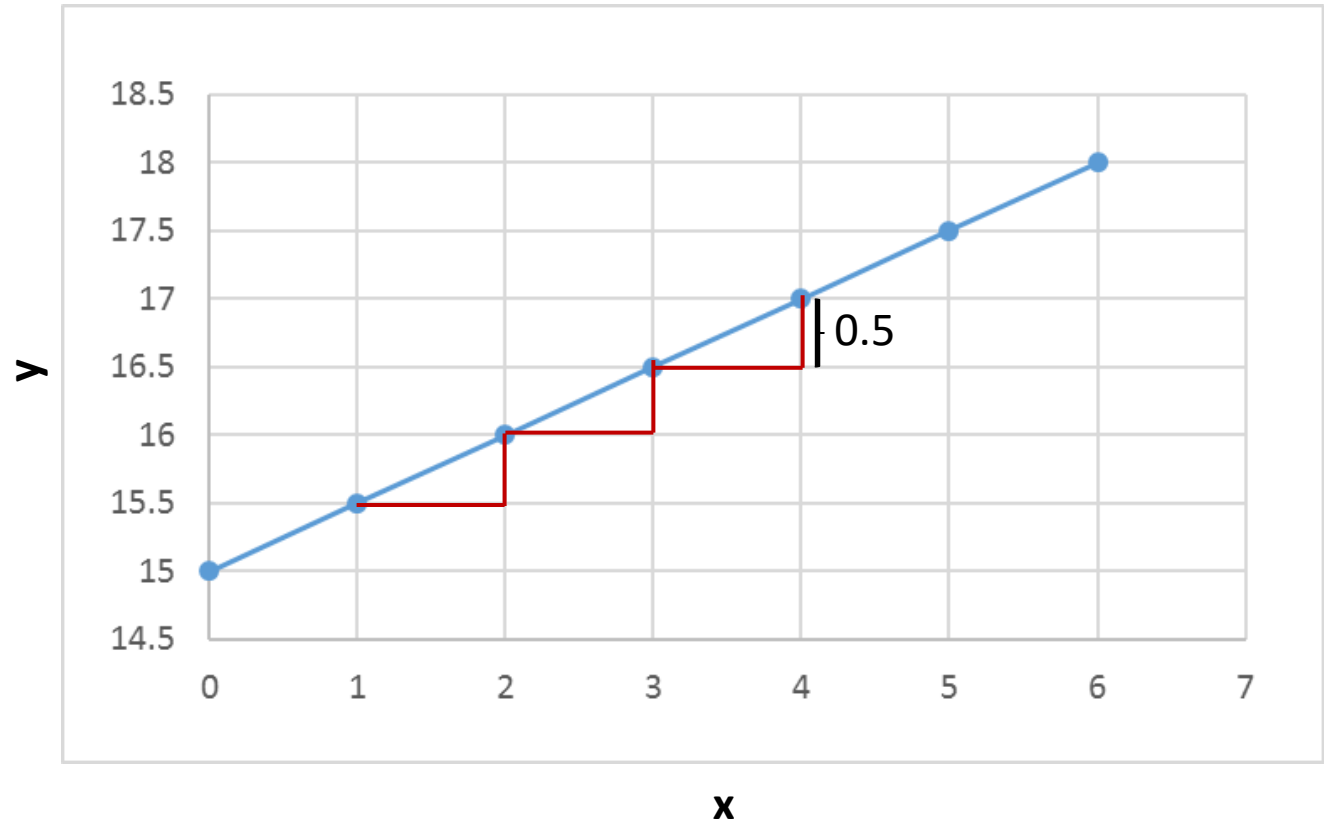
Verificar que los datos:

1. Sean independientes.
 2. La variable dependiente debe ser continua.
 3. La variable independiente puede ser categórica o continua.
- 

Regresión lineal simple

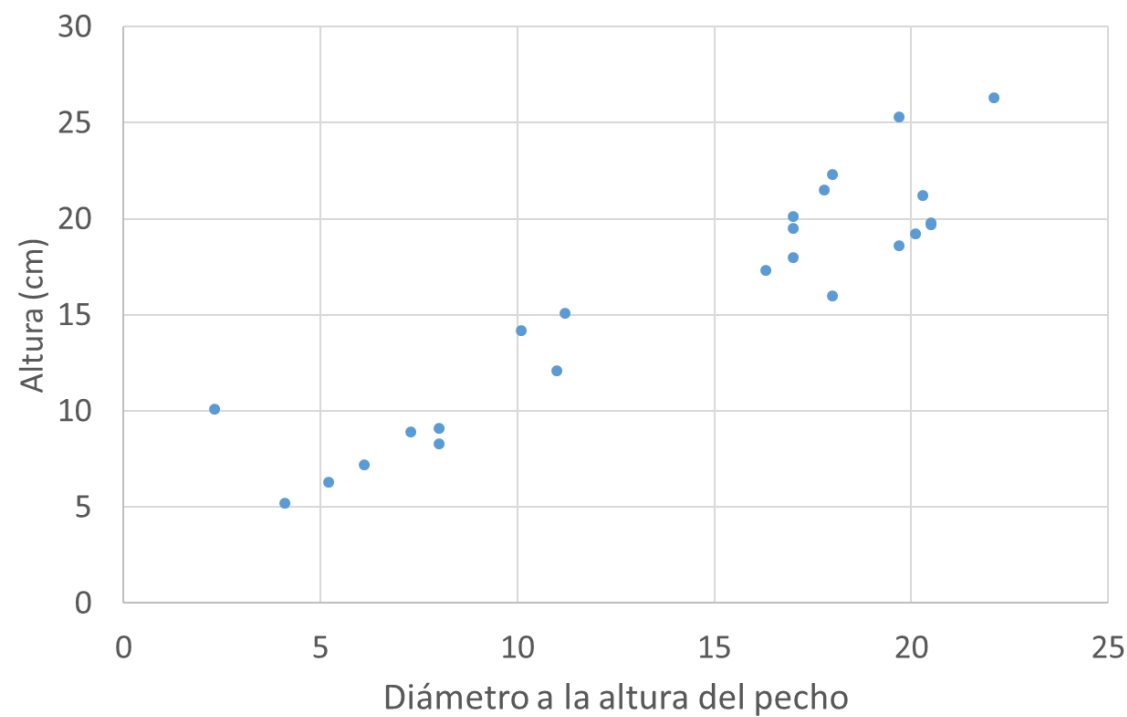
$$y = 15 + 0.5x$$

x	y
0	15
1	15.5
2	16
3	16.5
4	17
5	17.5
6	18



Regresión lineal simple: ejemplo

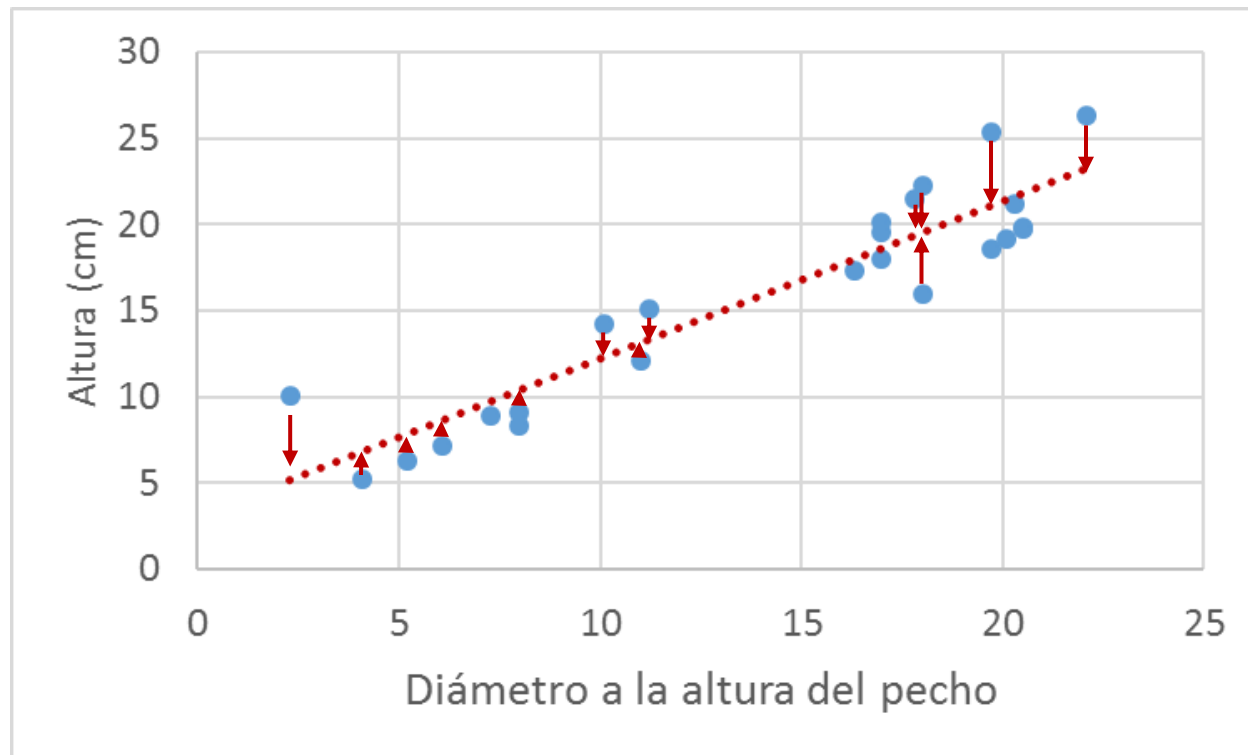
Un ingeniero forestal toma medidas del diámetro de los arboles y de su altura en una parcela de 1.0 ha. Ya que la altura es más difícil de medir que el diámetro, se busca entonces ajustar una línea que permita predecir la altura a partir del diámetro.



Observacion	Diámetro a la altura del pecho	Altura
1	10.1	14.2
2	11.2	15.1
3	19.7	25.3
4	20.3	21.2
5	17.8	21.5
6	17	18
7	11	12.1
8	4.1	5.2
9	5.2	6.3
10	8	9.1
11	2.3	10.1
12	20.1	19.2
13	18	16
14	22.1	26.3
15	16.3	17.3
16	20.5	19.8
17	17	20.1
18	18	22.3
19	17	19.5
20	19.7	18.6
21	20.5	19.7
22	6.1	7.2
23	7.3	8.9
24	8	8.3

Método de mínimos cuadrados

- Criterio para seleccionar la recta que “mejor” se ajuste a los datos
- El criterio de mínimos cuadrados implica que la recta elegida **minimiza** la suma de los cuadrados de las distancias verticales entre los puntos de la muestra y la recta



Método de mínimos cuadrados

- 1. Obtener n
- 2. Obtener la media de la variable dependiente (\bar{y}) y la independiente (\bar{x})
- 3. Obtener la suma de cuadrados de x (**SCX**) y la de y (**SCY**)

Observación	dap (xi)	altura (yi)	(xi- \bar{X})	(xi- \bar{X}) ²	(y- \bar{Y})	(y- \bar{Y}) ²
1	10.1	14.2	-3.95	15.64	-1.69	2.85
2	11.2	15.1	-2.85	8.15	-0.79	0.62
3	19.7	25.3	5.65	31.88	9.41	88.60
4	20.3	21.2	6.25	39.01	5.31	28.22
5	17.8	21.5	3.75	14.03	5.61	31.50
6	17	18	2.95	8.68	2.11	4.46
7	11	12.1	-3.05	9.33	-3.79	14.35
8	4.1	5.2	-9.95	99.09	-10.69	114.22
9	5.2	6.3	-8.85	78.40	-9.59	91.92
10	8	9.1	-6.05	36.65	-6.79	46.07
11	2.3	10.1	-11.75	138.16	-5.79	33.50
12	20.1	19.2	6.05	36.55	3.31	10.97
13	18	16	3.95	15.57	0.11	0.01
14	22.1	26.3	8.05	64.74	10.41	108.42
15	16.3	17.3	2.25	5.04	1.41	2.00
16	20.5	19.8	6.45	41.55	3.91	15.31
17	17	20.1	2.95	8.68	4.21	17.75
18	18	22.3	3.95	15.57	6.41	41.12
19	17	19.5	2.95	8.68	3.61	13.05
20	19.7	18.6	5.65	31.88	2.71	7.36
21	20.5	19.7	6.45	41.55	3.81	14.54
22	6.1	7.2	-7.95	63.27	-8.69	75.47
23	7.3	8.9	-6.75	45.62	-6.99	48.83
24	8	8.3	-6.05	36.65	-7.59	57.57
n= 24						
Media	14.05	15.89				
SCX				894.34		
SCY						868.69

Método de mínimos cuadrados

4. Obtener la suma del producto
(SPXY)

$$(x_i - \bar{x}) * (y_i - \bar{y})$$

Observación	dap (xi)	altura (yi)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	10.1	14.2	-3.95	15.64	-1.69	2.85	6.67
2	11.2	15.1	-2.85	8.15	-0.79	0.62	2.25
3	19.7	25.3	5.65	31.88	9.41	88.60	53.14
4	20.3	21.2	6.25	39.01	5.31	28.22	33.18
5	17.8	21.5	3.75	14.03	5.61	31.50	21.02
6	17	18	2.95	8.68	2.11	4.46	6.22
7	11	12.1	-3.05	9.33	-3.79	14.35	11.57
8	4.1	5.2	-9.95	99.09	-10.69	114.22	106.39
9	5.2	6.3	-8.85	78.40	-9.59	91.92	84.89
10	8	9.1	-6.05	36.65	-6.79	46.07	41.09
11	2.3	10.1	-11.75	138.16	-5.79	33.50	68.03
12	20.1	19.2	6.05	36.55	3.31	10.97	20.03
13	18	16	3.95	15.57	0.11	0.01	0.44
14	22.1	26.3	8.05	64.74	10.41	108.42	83.78
15	16.3	17.3	2.25	5.04	1.41	2.00	3.17
16	20.5	19.8	6.45	41.55	3.91	15.31	25.22
17	17	20.1	2.95	8.68	4.21	17.75	12.41
18	18	22.3	3.95	15.57	6.41	41.12	25.30
19	17	19.5	2.95	8.68	3.61	13.05	10.64
20	19.7	18.6	5.65	31.88	2.71	7.36	15.31
21	20.5	19.7	6.45	41.55	3.81	14.54	24.57
22	6.1	7.2	-7.95	63.27	-8.69	75.47	69.10
23	7.3	8.9	-6.75	45.62	-6.99	48.83	47.19
24	8	8.3	-6.05	36.65	-7.59	57.57	45.94
n= 24							
Media	14.05	15.89					
SCX				894.34			
SCY						868.69	
SPXY							817.57

Método de mínimos cuadrados

Para estimar los coeficientes b_0 y b_1 de la regresión $y = b_0 + b_1x + e$

b_0 = ordenada al origen (el valor de Y cuando X= 0)

b_1 = pendiente de la recta, su signo indica si la relación es positiva o negativa

$$b_1 = \frac{SPXY}{SCX} = 817.56/894.33 = \mathbf{0.914}$$

$$b_0 = \bar{y} - (b_1 * \bar{x}) = (15.88) - (0.914 * 14.05) = 15.88 - 12.84 = \mathbf{3.04}$$

$$\mathbf{y = 3.04 + 0.914X + e}$$

Método de mínimos cuadrados

Interpretación:

$$y = 3.04 + 0.914X + e$$

La ordenada al origen: 3.04 m es el valor de altura promedio de los árboles con diámetro = 0

La pendiente 0.914 nos dice que por cada metro que aumenta la altura del árbol, se espera un cambio de 0.914 cm en el diámetro.

Método de mínimos cuadrados

A partir de los coeficientes, podemos calcular los residuales (desviaciones) de la ecuación así como la suma de las desviaciones cuadradas de los datos (**SCE**).

Observaci	dap (xi)	Altura (yi)	\hat{y} (predicha)	$(y_i - \hat{y})$	$(y_i - \hat{y})^2$
1	10.10	14.20	12.27	1.93	3.72
2	11.20	15.10	13.28	1.82	3.32
3	19.70	25.30	21.05	4.25	18.10
4	20.30	21.20	21.59	-0.39	0.16
5	17.80	21.50	19.31	2.19	4.80
6	17.00	18.00	18.58	-0.58	0.33
7	11.00	12.10	13.09	-0.99	0.99
8	4.10	5.20	6.79	-1.59	2.52
9	5.20	6.30	7.79	-1.49	2.23
10	8.00	9.10	10.35	-1.25	1.57
11	2.30	10.10	5.14	4.96	24.58
12	20.10	19.20	21.41	-2.21	4.89
13	18.00	16.00	19.49	-3.49	12.19
14	22.10	26.30	23.24	3.06	9.37
15	16.30	17.30	17.94	-0.64	0.41
16	20.50	19.80	21.78	-1.98	3.91
17	17.00	20.10	18.58	1.52	2.32
18	18.00	22.30	19.49	2.81	7.88
19	17.00	19.50	18.58	0.92	0.85
20	19.70	18.60	21.05	-2.45	5.98
21	20.50	19.70	21.78	-2.08	4.31
22	6.10	7.20	8.62	-1.42	2.00
23	7.30	8.90	9.71	-0.81	0.66
24	8.00	8.30	10.35	-2.05	4.21
SCE					121.30

Análisis de varianza: ANOVA

La variabilidad total (SCY) se divide en dos partes: la cantidad que ha sido eliminada por la recta de regresión (SCR) y la cantidad que permanece a pesar de la recta de regresión (SCE)

$$\bar{x}=14.05$$

$$\bar{Y}= 15.88$$

$$SCX=894.33$$

$$SCY=868.68$$

$$SPXY=817.56$$

$$SCE=121.30$$

$$SCY=SCR+SCE$$

$$SCR= SCY-SCE$$

$$SCR= 868.68-121.30= 747.38$$

Mientras más variación se elimine mediante la recta de regresión, más cercana será la relación entre X y Y. Como resultado, la estimación de Y se volverá más precisa.

Coeficiente de determinación: R^2

Es una medida relativa de la asociación lineal entre X y Y. Se mide como la proporción de la variabilidad total de Y explicada por su relación con X.

$$0 \leq R^2 \leq 1$$

$$R^2 = SCR/SCY = \\ 747.38/868.68 = 0.86$$

Conclusión: 86% de la variabilidad total de Y puede atribuirse a una relación lineal con X.

Tabla de ANOVA

Tabla ANOVA				
Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	$SCR = \sum (\hat{y} - \bar{Y})^2$	1	$MCR_{reg} = SCR/1$	MCR_{reg} / MCE
Error	$SCE = \sum (y_i - \hat{y})^2$	n-2	$MCE = SCE/n-2$	
Total	$SCY = \sum (y_i - \bar{Y})^2$	n-1	$SCY/n-1$	


Los grados de libertad (GL) son la cantidad de información suministrada por los datos que se utiliza para estimar los parámetros desconocidos y calcular la variabilidad de las estimaciones. Los GL están determinados por el número de observaciones (n) y el número de parámetros.

Tabla de ANOVA

Tabla ANOVA				
Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	747.38	1	747.38	747.3/ 5.5= 135.8
Error	121.30	24-2=22	5.5	
Total	868.6	24-1=23	37.7	

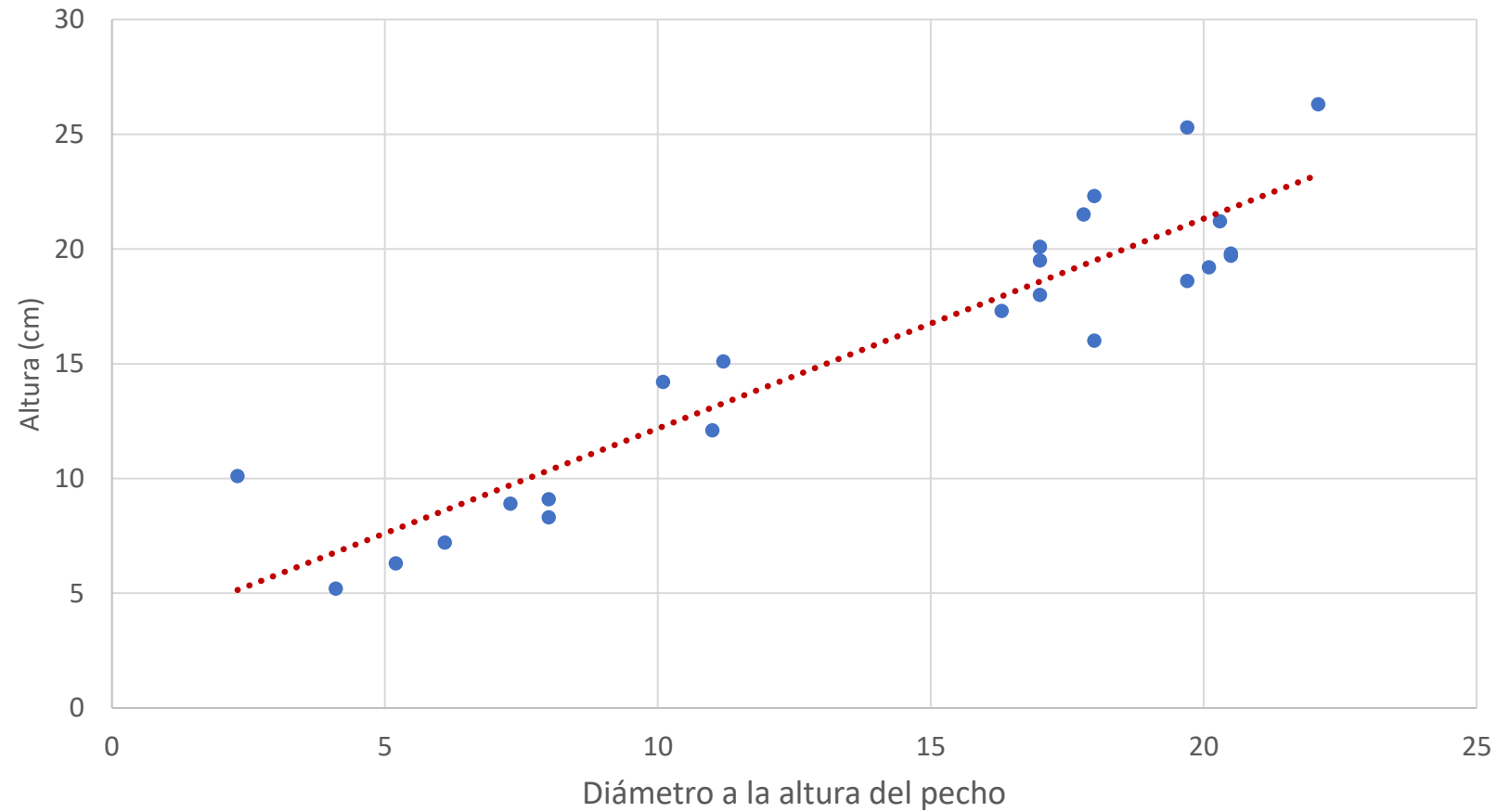
Supuestos de la regresion lineal

Antes de interpretar la tabla de ANOVA debemos verificar los siguientes supuestos:

1. La relación entre Y y X es lineal (**LINEALIDAD**)
 2. Independencia de las observaciones (**INDEPENDENCIA**)
 3. Igualdad de varianzas (**HOMOSCEDASTICIDAD**)
 4. Normalidad de los residuos o desviaciones tipificadas (**NORMALIDAD**)
- 

Supuestos de la regresion lineal

1. LINEALIDAD



Supuestos de la regresion lineal

2. INDEPENDENCIA

Los residuales son independientes entre sí, es decir, no hay correlación entre errores consecutivos. Se utiliza la prueba Durbin-Watson para evaluar el grado de independencia.

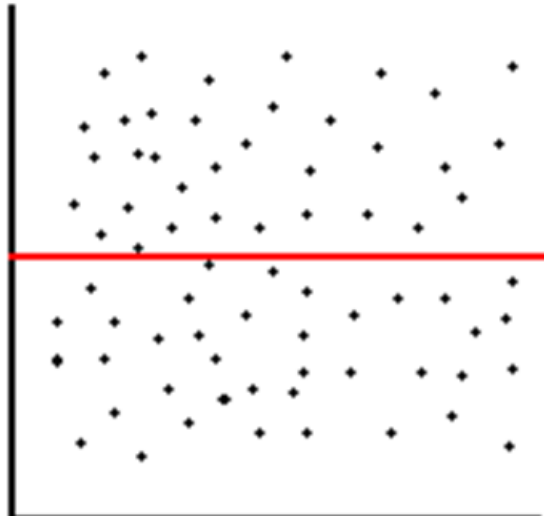
$1.5 \leq DW \leq 2.0$ Independencia

$DW \leq 1.5$ correlación negativa

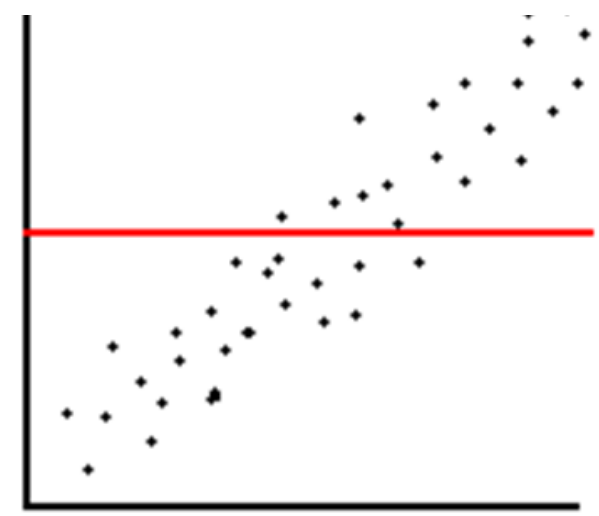
$DW \geq 2$ correlación positiva

DW Statistic = 1.84

No hay correlación



Hay correlación



Supuestos de la regresion lineal

3. HOMOSCEDASTICIDAD

Igualdad de varianzas de los residuos y los pronósticos.
Es decir, la variación de los residuos debe ser uniforme
en todo el rango de valores de los pronósticos.

Podemos también aplicar la prueba Breusch-Pagan

H_0 : No hay heteroscedasticidad

H_a : Hay heteroscedasticidad

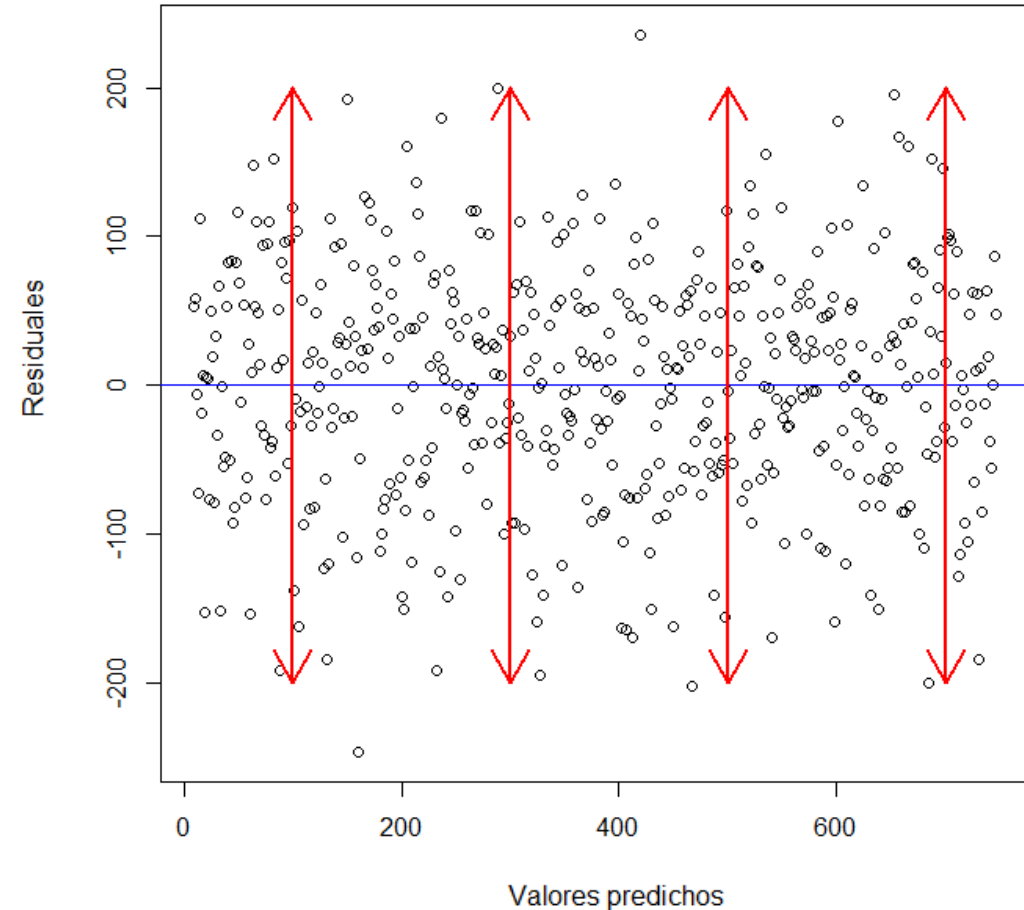
Non-constant Variance Score Test

Variance formula: \sim fitted.values

Chisquare = 0.003751778 Df = 1 p = 0.9511587

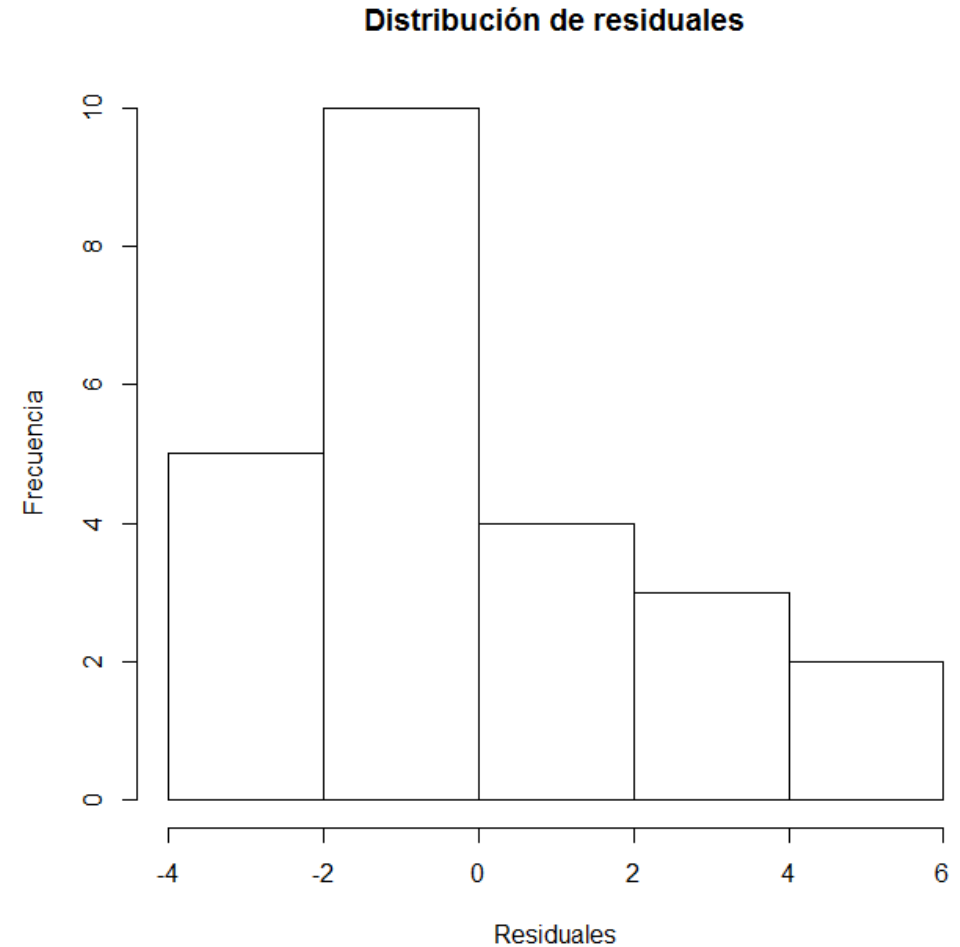
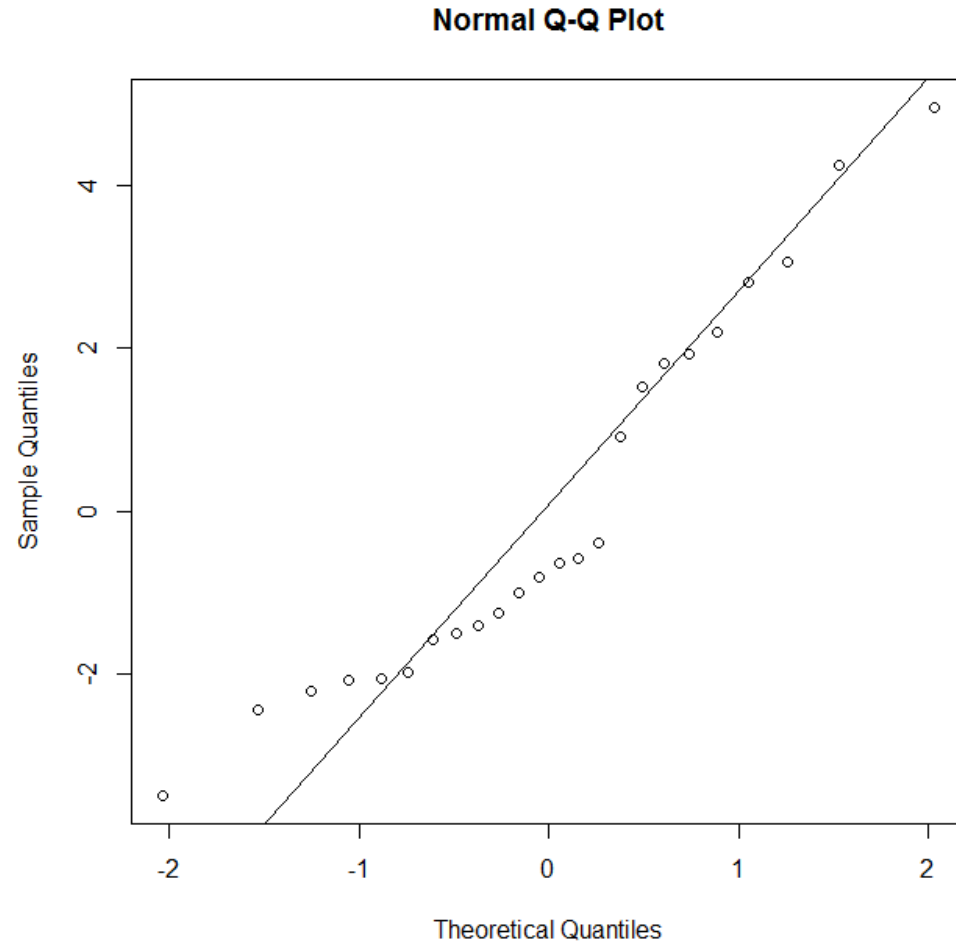
Si la $p < \alpha = 0.05$, entonces se rechaza H_0

Conclusión: No hay evidencia que sugiera heteroscedasticidad



Supuestos de la regresion lineal

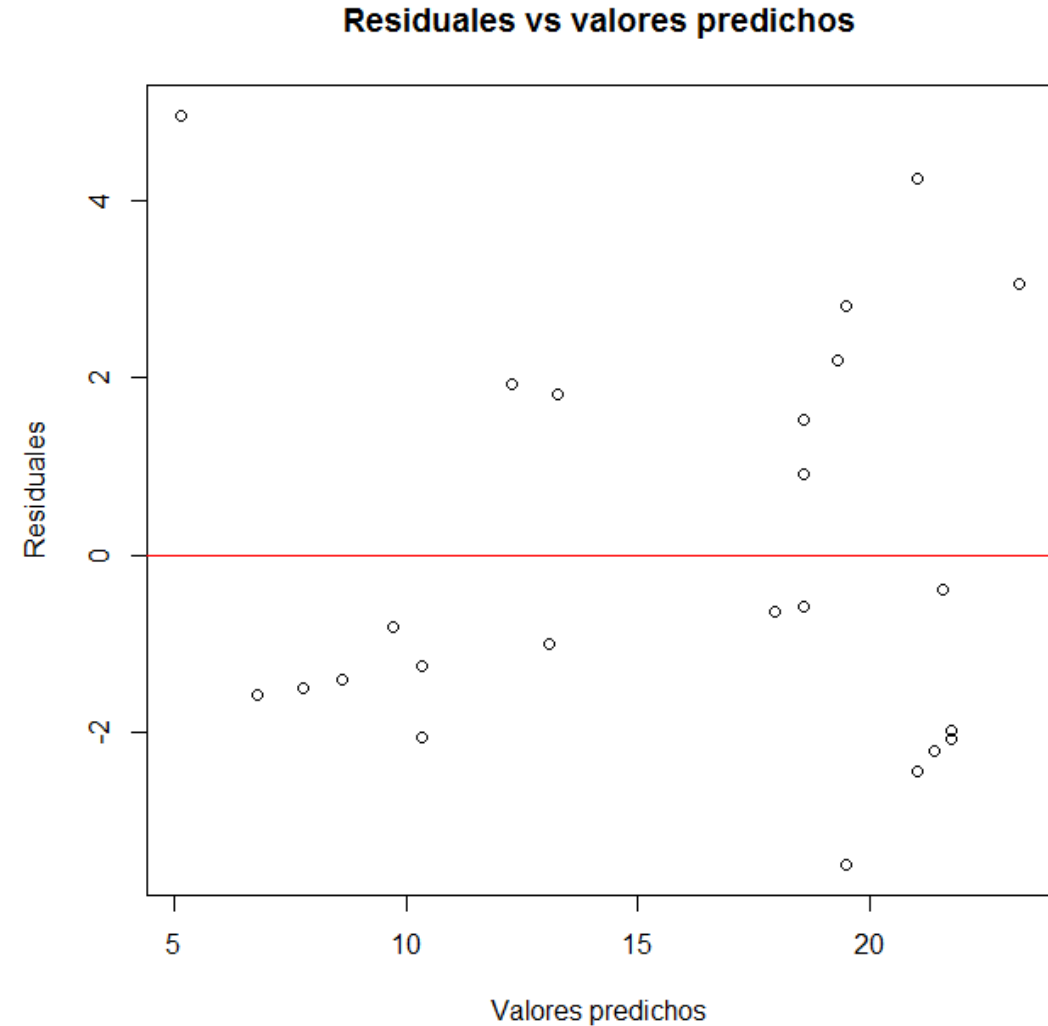
4. NORMALIDAD DE RESIDUALES



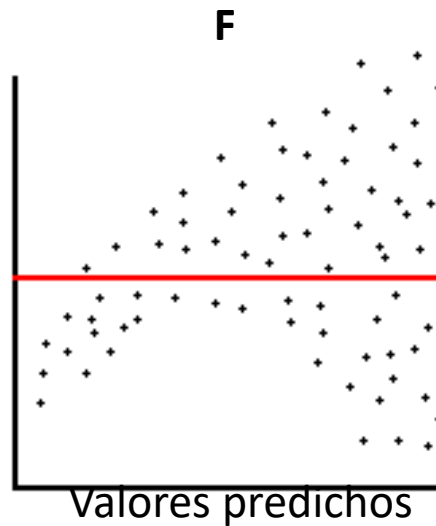
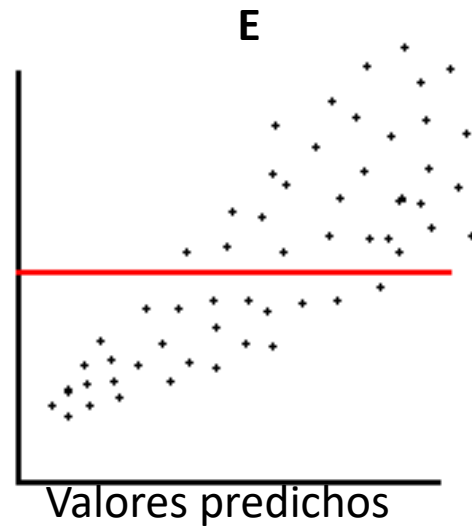
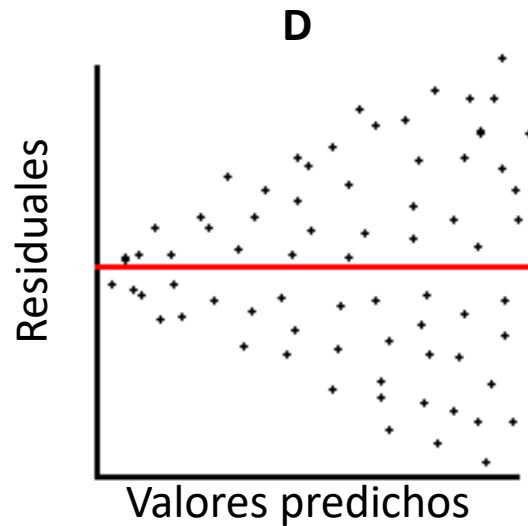
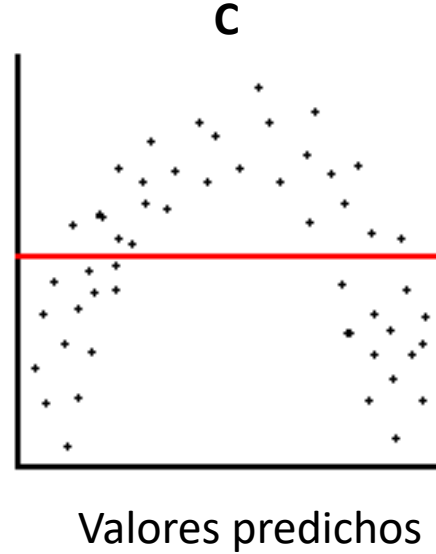
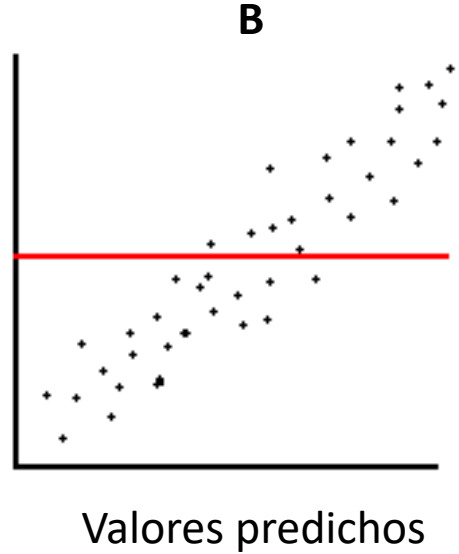
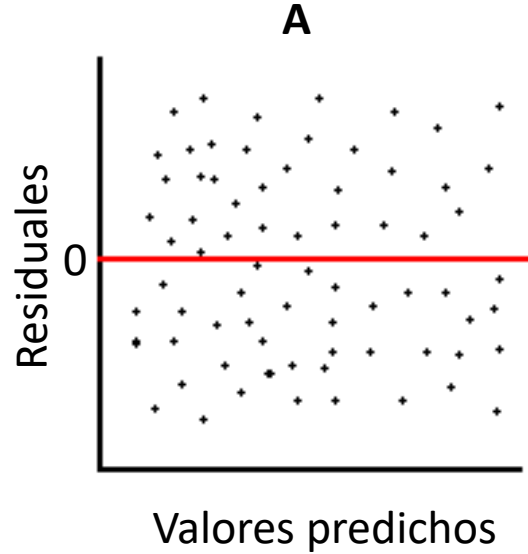
Supuestos de la regresion lineal

4. NORMALIDAD DE RESIDUALES

- Lo ideal es observar una distribución uniforme alrededor de la línea base cero. No patrones (e.j. trapezoide)



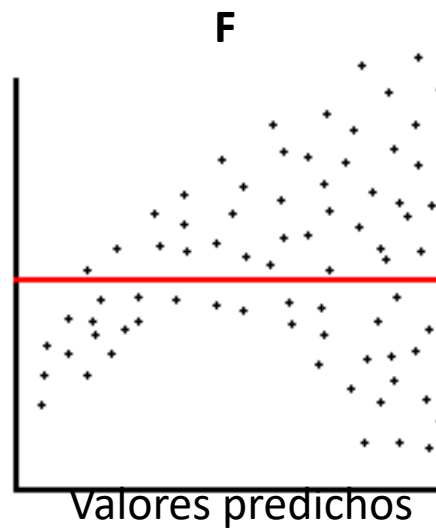
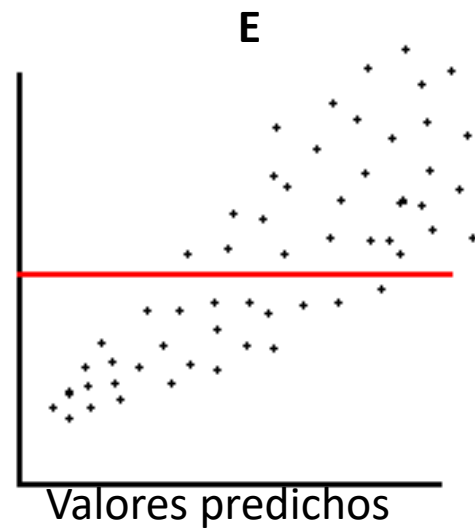
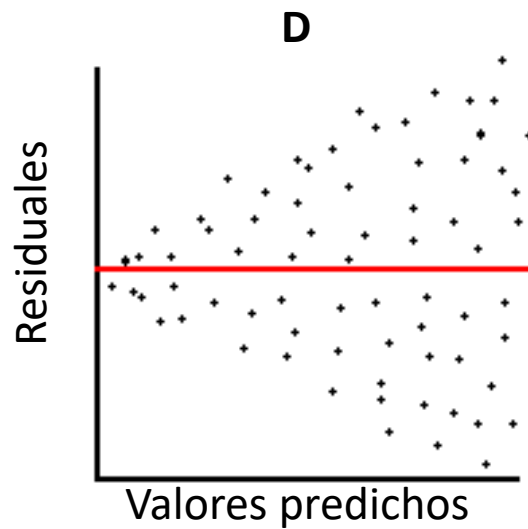
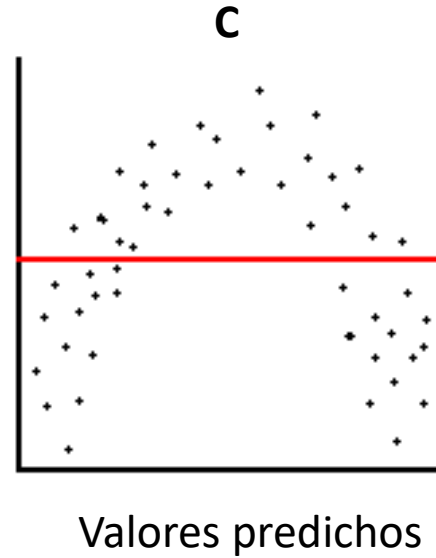
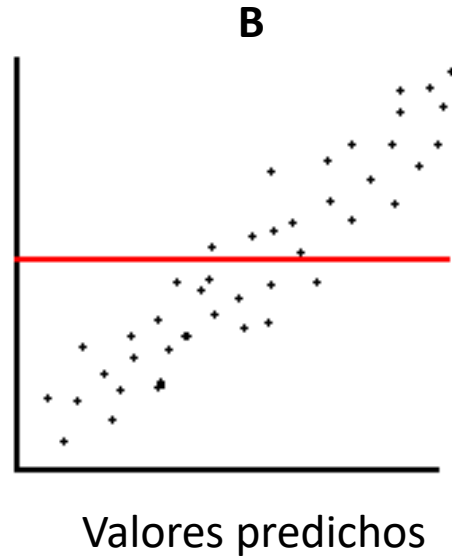
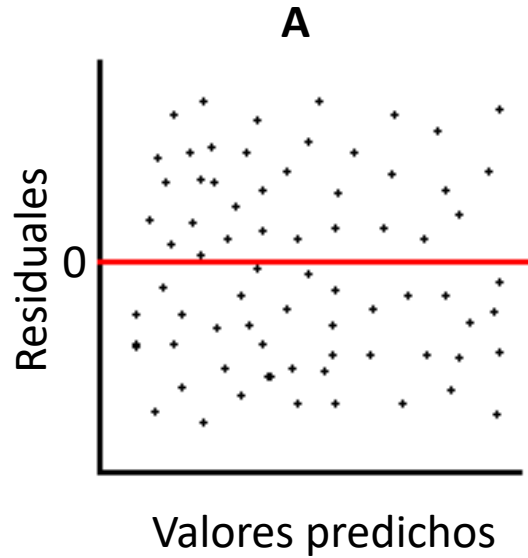
Residuales



La verificación de residuales representa un paso clave para determinar el comportamiento óptimo del modelo de regresión y el respeto de los supuestos de la regresión lineal.

¿Qué tipo de distribución de residuales indican un buen comportamiento del modelo de regresión?

Residuales



La gráfica A presenta una distribución uniforme de residuales alrededor del cero. No muestra sesgo ni autocorrelación.

Prueba de significancia de la regresión

Nos interesa saber si la variable independiente influye significativamente en el comportamiento de la variable dependiente.

Ho : la regresión no es significativa

Ha : la regresión es significativa

Fijado un nivel de significación α , se rechaza Ho si $F > F_{\alpha,1,n-2}$

Tabla ANOVA				
Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	747.38	1	747.38	747.3/ 5.5= 135.8
Error	121.30	24-2=22	5.5	
Total	868.6	24-1=23	37.7	

Prueba de significancia de la regresión

Distribución F de Fisher

$$F_{0.05, 1, 22} = 4.30 \text{ (valor crítico)}$$

Si $F > \text{valor crítico}$, entonces rechazamos H_0

Conclusión: $135.8 > 4.30$

Rechazamos H_0 y la regresión es significativa

n ₁	5 % (normal) y 1 % (negritas) puntos para la distribución de F																										n ₂
	n: grados de libertad (para el mayor cuantil medio)																										
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞			
11	9.64	3.98	3.28	2.88	2.68	2.58	2.51	2.45	2.40	2.35	2.30	2.25	2.21	2.16	2.10	2.05	2.01	1.97	1.93	1.89	1.85	1.81	1.77	1.73	1.70	11	
12	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.29	4.21	4.10	4.02	3.94	3.86	3.81	3.74	3.71	3.66	3.62	3.60	12		
13	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.05	3.97	3.86	3.78	3.70	3.62	3.57	3.50	3.47	3.41	3.38	3.36	13		
14	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.22	2.21	14		
15	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.86	3.78	3.66	3.59	3.51	3.43	3.38	3.31	3.27	3.22	3.19	3.17	15		
16	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13	16		
17	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.35	3.27	3.22	3.15	3.11	3.06	3.03	3.00	17		
18	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.42	2.38	2.33	2.29	2.25	2.20	2.18	2.14	2.12	2.10	2.08	2.07	18		
19	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.56	3.49	3.37	3.29	3.21	3.13	3.08	3.01	2.98	2.92	2.89	2.87	19		
20	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.37	2.33	2.28	2.24	2.19	2.15	2.12	2.09	2.07	2.04	2.02	2.01	20		
21	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.45	3.37	3.26	3.18	3.10	3.02	2.97	2.90	2.86	2.81	2.78	2.75	21		
22	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.10	2.08	2.04	2.02	1.99	1.97	1.96	22		
23	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.35	3.27	3.16	3.08	3.00	2.92	2.87	2.80	2.76	2.71	2.68	2.65	23		
24	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.06	2.04	2.00	1.98	1.95	1.93	1.92	24		
25	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.27	3.19	3.08	3.00	2.92	2.84	2.78	2.71	2.68	2.62	2.59	2.57	25		
26	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.26	2.21	2.16	2.11	2.07	2.03	2.00	1.96	1.94	1.91	1.89	1.88	26		
27	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.71	2.64	2.60	2.55	2.51	2.49	27		
28	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.22	2.18	2.12	2.08	2.04	1.99	1.97	1.93	1.91	1.88	1.86	1.84	28		
29	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.13	3.05	2.94	2.86	2.78	2.69	2.64	2.57	2.54	2.48	2.44	2.42	29		
30	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.16	2.10	2.05	2.01	1.96	1.94	1.90	1.88	1.84	1.83	1.81	30		
32	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.07	2.99	2.88	2.80	2.72	2.64	2.58	2.51	2.48	2.42	2.38	2.36	32		
33	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.17	2.13	2.07	2.03	1.98	1.94	1.91	1.87	1.85	1.82	1.80	1.78	33		
34	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.36	2.33	2.31	34		
35	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.15	2.11	2.05	2.01	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76	35		
36	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.54	2.48	2.41	2.37	2.32	2.28	2.26	36		
37	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.13	2.09	2.03	1.98	1.94	1.89	1.86	1.82	1.80	1.77	1.75	1.73	37		
38	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.37	2.33	2.27	2.24	2.21	38		
39	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.11	2.07	2.01	1.96	1.92	1.87	1.84	1.80	1.78	1.75	1.73	1.71	39		
40	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.33	2.29	2.23	2.19	2.17	40		
41	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.09	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.73	1.71	1.69	41		
42	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.86	2.78	2.66	2.58	2.50	2.42	2.36	2.29	2.25	2.19	2.16	2.13	42		
43	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.08	2.04	1.97	1.93	1.88	1.84	1.81	1.76	1.74	1.71	1.69	1.67	43		
44	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.82	2.75	2.63	2.55	2.47	2.38	2.33	2.26	2.22	2.16	2.12	2.10	44		
45	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.82	1.79	1.75	1.73	1.69	1.67	1.65	45		
46	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.79	2.72	2.60	2.52	2.44	2.35	2.30	2.23	2.19	2.13	2.09	2.06	46		
47	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.01	1.94	1.90	1.85	1.81	1.77	1.73	1.71	1.67	1.65	1.64	47		
48	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.77	2.69	2.57	2.49	2.41	2.33	2.27	2.20	2.16	2.10	2.06	2.03	48		
49	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.70	1.66	1.64	1.62	49		
50	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.74	2.66	2.55	2.47	2.39	2.30	2.25	2.17	2.13	2.07	2.03	2.01	50		
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.01	1.97	1.91	1.86	1.82	1.77	1.74	1.69	1.67	1.63	1.61	1.59	32		
52	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.86	2.80	2.70	2.62	2.50	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96	52		

Nivel de significancia

Cargamento de naranjas con
síntomas de enfermedad fúngica



Nivel de significancia

Cargamento de naranjas con
síntomas de enfermedad fúngica



Muestra aleatoria representativa



Nivel de significancia

Suponiendo una distribución binomial (n =tamaño muestra, p =probabilidad de éxito)

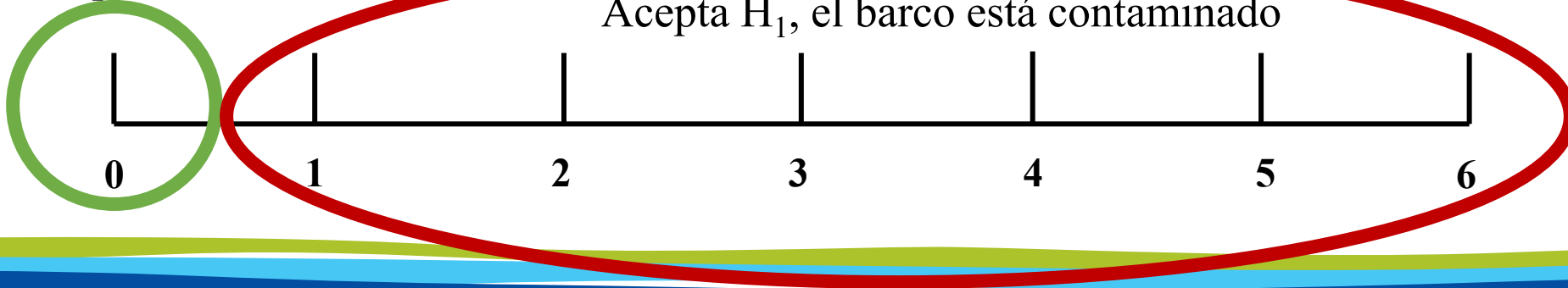
$n=6$

$H_0: p=0$ (no hay naranjas infectadas en el cargamento)

$H_1: p>0$ (si hay infección)



Acepta H_0



Nivel de significancia

- Debido a que la muestra es un subconjunto de los datos poblacionales, no se puede estar completamente seguros de si la hipótesis nula es cierta o no. Tan sólo podemos inferir, con base en la evidencia, si la hipótesis es probable o no.
- Se habla entonces de rechazar o no rechazar (es decir aceptar) la hipótesis nula en función del resultado de alguna prueba estadística.
- Cuando se plantean tales pruebas, siempre existe la probabilidad de concluir incorrectamente. Hay **dos tipos de errores**:

TIPO I: H_0 se rechaza cuando es verdadera (falsa positivo)

TIPO II: H_0 no se rechaza cuando es falsa (falso negativo)

Pruebas de hipótesis

2. Se selecciona un **nivel de significancia**

- Para decidir entre la hipótesis nula y la alternativa podemos crear zonas de aceptación y rechazo con probabilidades conocidas de error



© Can Stock Photo - csp41882095

Ejemplo: Queremos determinar si la persona es un zombie

H_0 : La persona no es un zombie, por lo tanto es normal

H_1 : La persona es un zombie

Pruebas de hipótesis

2. Se selecciona un nivel de significancia

- Para decidir entre la hipótesis nula y la alternativa podemos crear zonas de aceptación y rechazo con probabilidades conocidas de error



Si la persona es normal y la prueba nos dice que es normal entonces **acertamos!!!**

Pruebas de hipótesis

2. Se selecciona un nivel de significancia

- Para decidir entre la hipótesis nula y la alternativa podemos crear zonas de aceptación y rechazo con probabilidades conocidas de error



Si la persona es un zombie y la prueba nos dice que es zombie entonces **acertamos!!!**

Pruebas de hipótesis

2. Se selecciona un nivel de significancia

- Para decidir entre la hipótesis nula y la alternativa podemos crear zonas de aceptación y rechazo con probabilidades conocidas de error



ERROR TIPO I

Rechazamos H_0 cuando es verdadera

Si la persona es normal y la prueba nos dice que es zombie entonces cometemos **error tipo I**

Pruebas de hipótesis

2. Se selecciona un nivel de significancia

- Para decidir entre la hipótesis nula y la alternativa podemos crear zonas de aceptación y rechazo con probabilidades conocidas de error



ERROR TIPO II

Acepta H_0 cuando es falsa

Si la persona es un zombie y la prueba nos dice que es normal entonces cometemos **error tipo II**

Pruebas de hipótesis

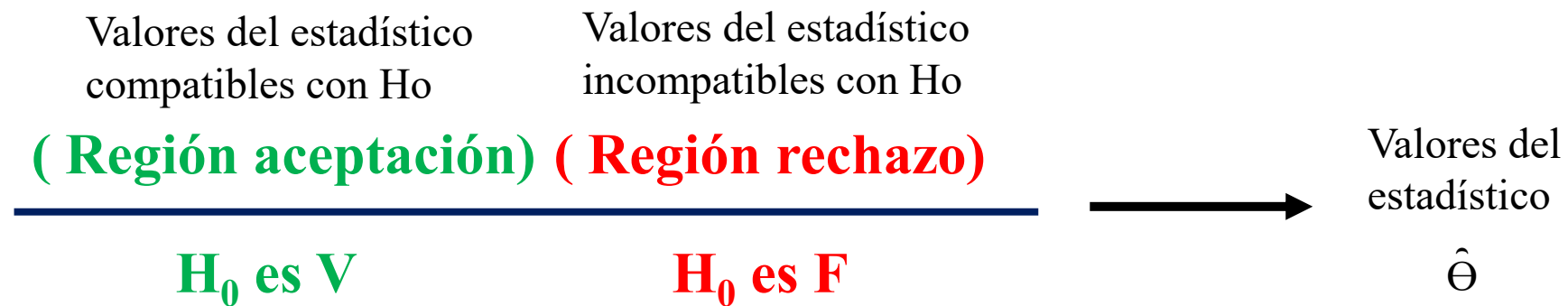
2. Se selecciona un nivel de significancia

- El nivel aceptable para el error tipo I se designa como **alpha** (α), mientras que para el error tipo II como **beta** (β).
- Nivel de significancia es entonces el nivel aceptable para el error tipo I. Corresponde a la probabilidad máxima de cometer el error tipo I. Típicamente, se utiliza un nivel de significancia de $\alpha = .05$ (aunque otros niveles pueden ser utilizados $\alpha = .01$).
- $\alpha = .05$ significa que estamos dispuestos a tolerar 5% del error tipo I (falso positivo). Es decir, estamos dispuestos a aceptar el hecho que en 1 de cada 20 muestras se rechaza la hipótesis nula cuando ésta sea verdadera, o en otras palabras, se da por **FALSA** una hipótesis **VERDADERA**.

Pruebas de hipótesis

2. Se selecciona un nivel de significancia

- El poder estadístico se define como $1-\beta$, es decir, la probabilidad de encontrar un efecto cuando dicho efecto existe. La probabilidad de rechazar correctamente la hipótesis nula.
- Región crítica: Es la región del espacio probabilístico que corresponde al rechazo de la hipótesis nula, es decir, el conjunto de valores del estadístico que mejor explica la hipótesis alternativa. **El nivel de significancia**, es entonces, la **probabilidad de que el estadístico se encuentre dentro de la región crítica** cuando se supone la hipótesis nula.



Significación estadística

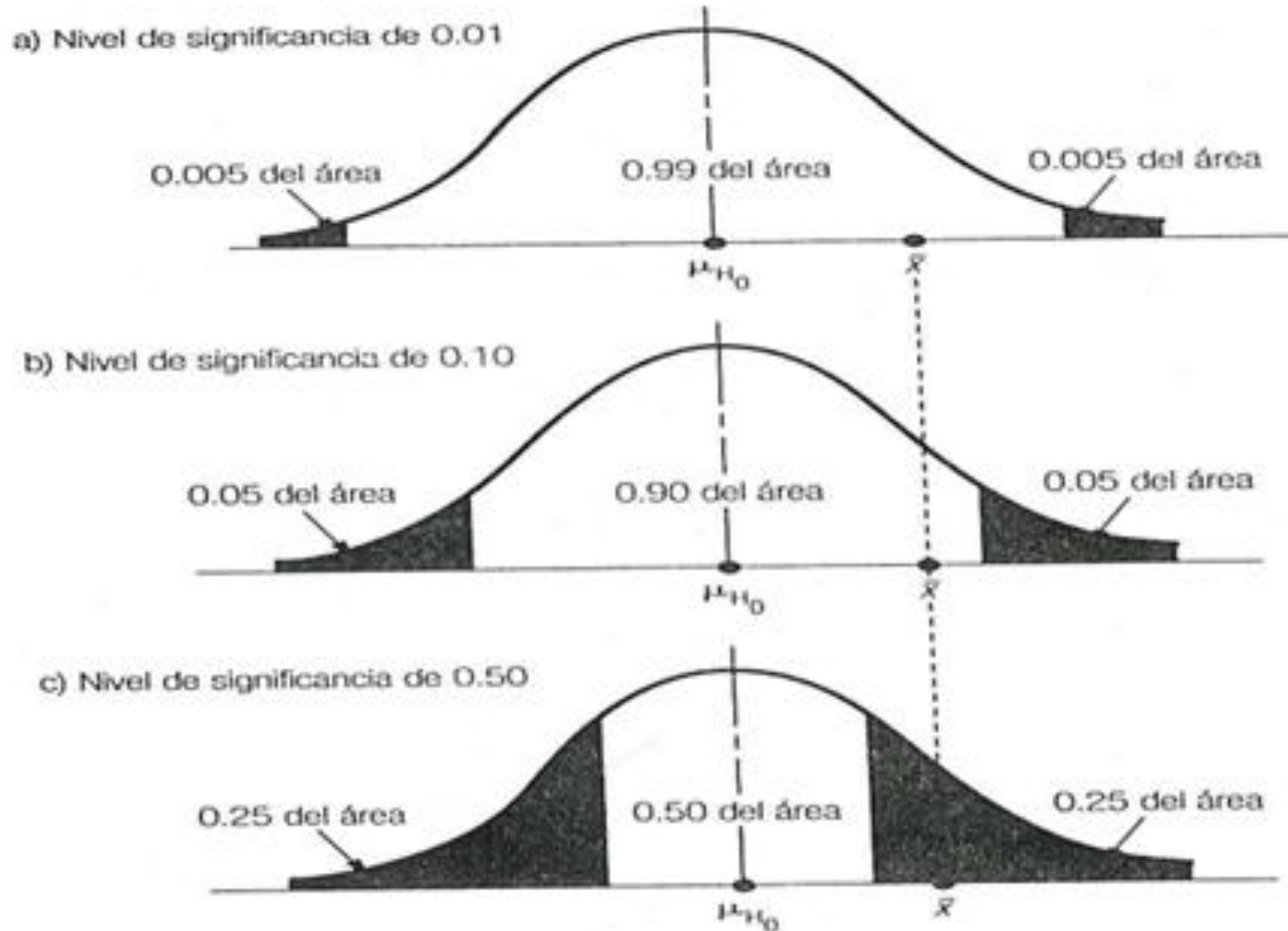
El nivel de significación estadística alfa (α) representa la probabilidad de que los resultados observados puedan ser debidos al azar, es decir, la probabilidad de cometer el error tipo 1.

Error tipo 1: Probabilidad de rechazar la hipótesis nula cuando es verdadera (“falso positivo”).

Un nivel de significación de 1% ($\alpha = 0.01$) indica que existe un 1% de rechazar la hipótesis nula cuando es cierta (99% de aceptarla cuando es cierta).

Es decir, si hiciéramos un experimento 100 veces, cometeríamos 1 vez el error tipo 1.

Significación estadística



Intervalos de confianza para los coeficientes

- Cuando trabajamos con muestras, siempre necesitamos una “banda de confianza” alrededor de los estimados de los coeficientes de regresión.
- Estas bandas se conocen como intervalos de confianza
- Los intervalos de confianza siempre tienen asociados un nivel de confianza $(1 - \alpha)$, típicamente 95%.
- Bajo los supuestos del modelo lineal (normalidad, homogeneidad de varianzas, linealidad e independencia) los coeficientes son estimadores insesgados de los “coeficientes verdaderos” .

Para β_0 : $b_0 \pm t_{n-2, \alpha/2} * SE(\beta_0)$

Para β_1 : $b_1 \pm t_{n-2, \alpha/2} * SE(\beta_1)$

donde $n-2, \alpha / 2$ es el percentil de la distribución t de Student con $n - 2$ grados de libertad que deja a su derecha un área $\alpha/2$. SE es el error estándar asociado a cada coeficiente.

Intervalos de confianza para los coeficientes

Para β_1 : $b_1 \pm t_{n-2, \alpha/2} * SE(\beta_1)$

$$SE(\beta_1) = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{SCE}{SCX}}$$

Observaci	dap (xi)	Altura (yi)	\hat{y} (predicha)	$(y_i - \hat{y})$	$(y_i - \hat{y})^2$
1	10.10	14.20	12.27	1.93	3.72
2	11.20	15.10	13.28	1.82	3.32
3	19.70	25.30	21.05	4.25	18.10
4	20.30	21.20	21.59	-0.39	0.16
5	17.80	21.50	19.31	2.19	4.80
6	17.00	18.00	18.58	-0.58	0.33
7	11.00	12.10	13.09	-0.99	0.99
8	4.10	5.20	6.79	-1.59	2.52
9	5.20	6.30	7.79	-1.49	2.23
10	8.00	9.10	10.35	-1.25	1.57
11	2.30	10.10	5.14	4.96	24.58
12	20.10	19.20	21.41	-2.21	4.89
13	18.00	16.00	19.49	-3.49	12.19
14	22.10	26.30	23.24	3.06	9.37
15	16.30	17.30	17.94	-0.64	0.41
16	20.50	19.80	21.78	-1.98	3.91
17	17.00	20.10	18.58	1.52	2.32
18	18.00	22.30	19.49	2.81	7.88
19	17.00	19.50	18.58	0.92	0.85
20	19.70	18.60	21.05	-2.45	5.98
21	20.50	19.70	21.78	-2.08	4.31
22	6.10	7.20	8.62	-1.42	2.00
23	7.30	8.90	9.71	-0.81	0.66
24	8.00	8.30	10.35	-2.05	4.21
SCE					121.30

Intervalos de confianza para los coeficientes

Para β_1 : $b_1 \pm t_{n-2, \alpha/2} * SE(\beta_1)$

$$SE = \sqrt{\frac{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-2}}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \sqrt{\frac{121.30}{22} SCX}$$

Tabla ANOVA				
Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	747.38	1	747.3	747.3/ 5.5= 135.8
Error	121.30	24-2=22	5.5	
Total	868.6	24-1=23	37.7	

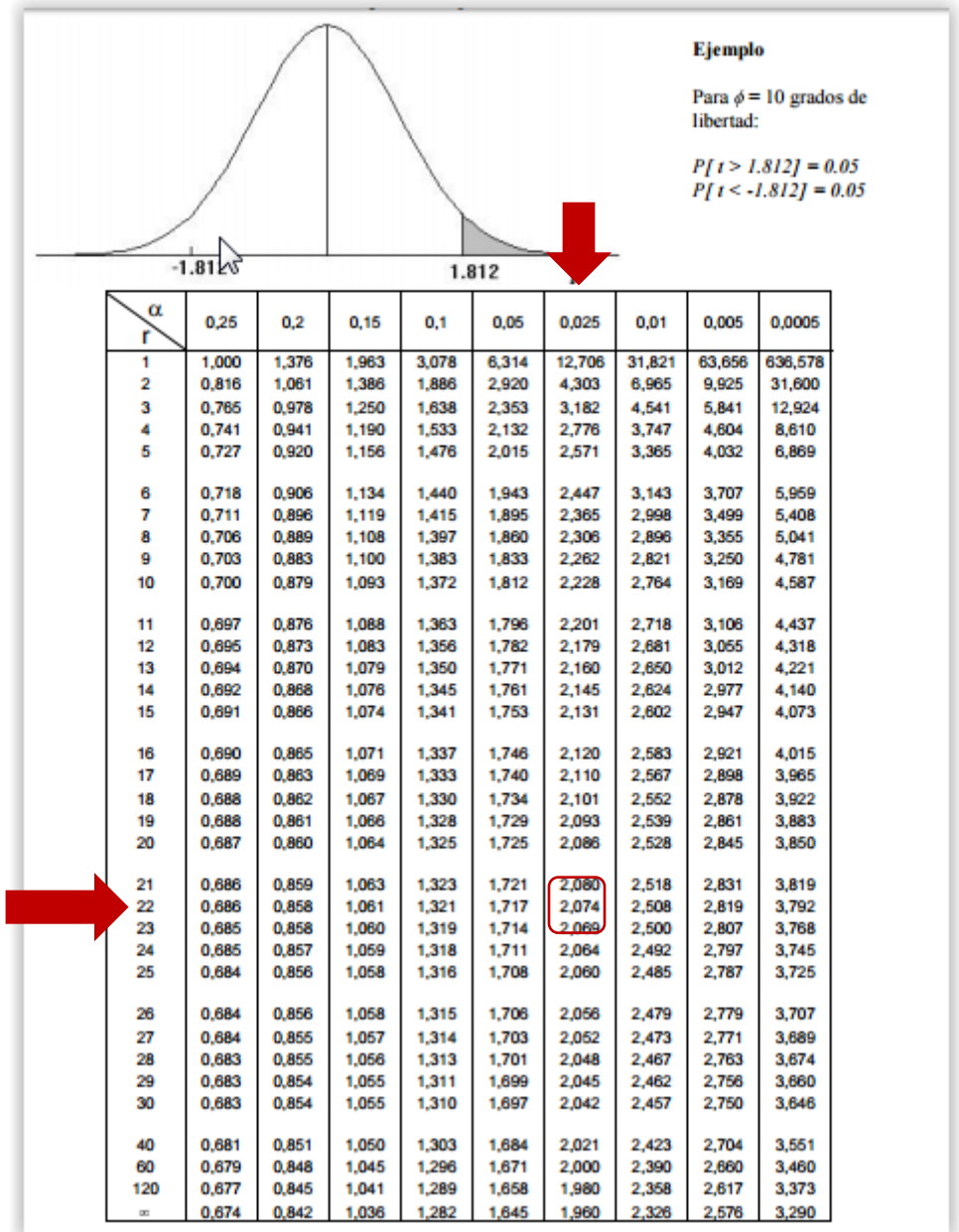
Intervalos de confianza para los coeficientes

Para β_1 : $b_1 \pm t_{22, 0.025} * SE(\beta_1)$

$$SE(\beta_1) = \sqrt{\frac{5.5}{894.33}} = \sqrt{0.0061} = 0.078$$

Para β_1 (95% IC):

$$b_1 \pm 2.074 * 0.078 = 0.91 \pm 0.16 = (0.75, 1.07)$$



Intervalos de confianza para los coeficientes

Para β_0 : $b_0 \pm t_{n-2, \alpha/2} * SE(\beta_0)$

$$SE(\beta_0) = \frac{\overset{\text{SCE}}{\sum_{i=1}^n (y_i - \hat{y})^2}}{n - 2} * \left[\frac{1}{n} + \frac{\bar{x}^2}{\underset{\text{SCX}}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

n = 24

SCE = 121.30

SCX = 894.33

$\bar{x} = 14.05$

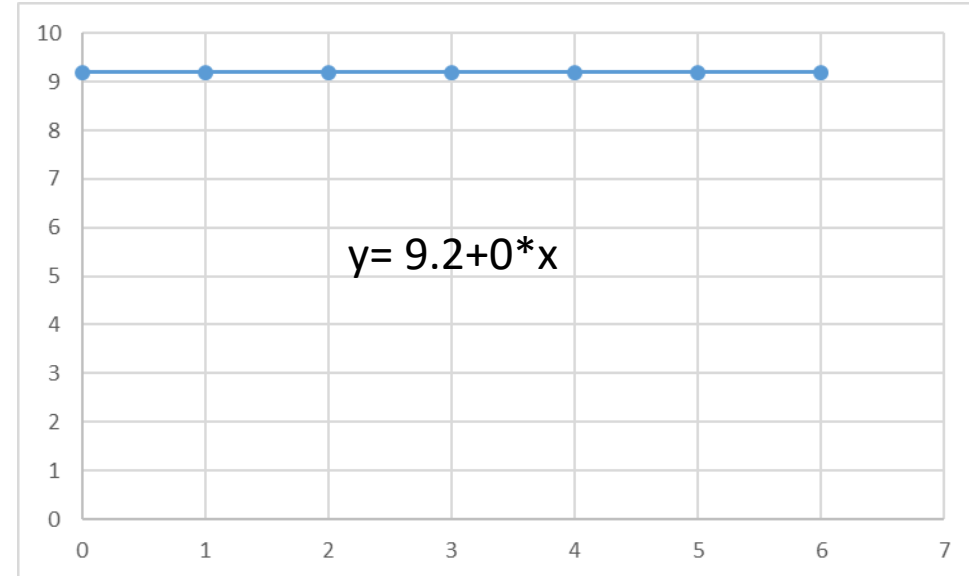
$$\begin{aligned} &= \frac{121.30}{22} * \left[\frac{1}{24} + \frac{(14.05)^2}{894.33} \right] = 5.5 \left[0.041 + \frac{197.40}{894.33} \right] = 5.5 [0.041 + 0.22] \\ &= 5.5 * 0.261 = 1.43 \end{aligned}$$

Para β_0 (95% IC):

$$b_0 \pm 2.074 * 1.43 = 3.04 \pm 2.96 = (0.08, 6.0)$$

Intervalos de confianza para los coeficientes

- El intervalo de confianza de la pendiente también nos permite decidir si la relación entre la variable dependiente y la independiente es estadísticamente significativa (al 95%).
- ¿Cómo? – verificando si el 0 está contenido en el intervalo de confianza
- Si el 0 **NO** está contenido en el intervalo de confianza entonces la relación es significativa
- ¿Qué significa que la pendiente $b_1 = 0$?

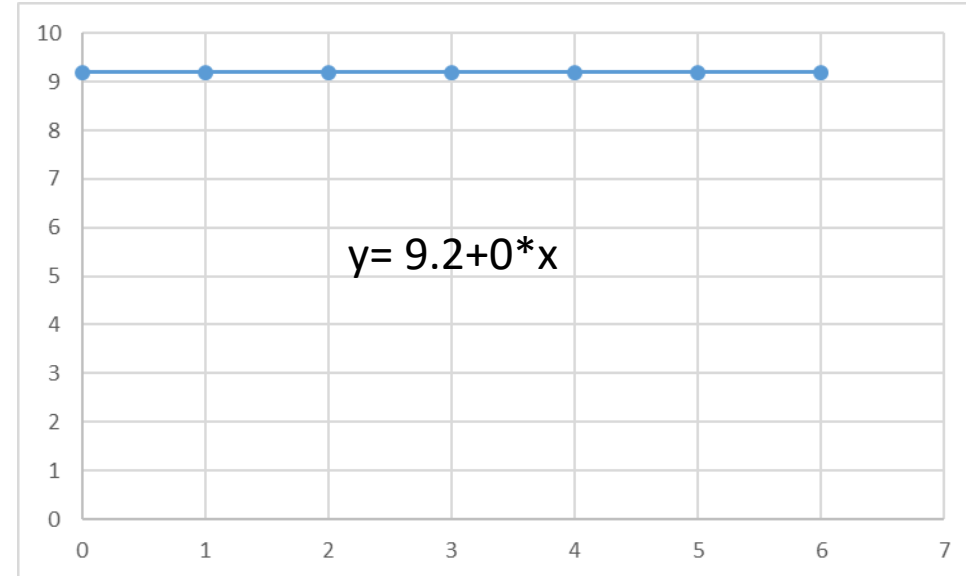


Intervalos de confianza para los coeficientes

- El intervalo de confianza de la pendiente también nos permite decidir si la relación entre la variable dependiente y la independiente es estadísticamente significativa (al 95%).
- ¿Cómo? – verificando si el 0 está contenido en el intervalo de confianza
- Si el 0 **NO** está contenido en el intervalo de confianza entonces la relación es significativa
- ¿Qué significa que la pendiente $b_1 = 0$?
- Significa que no hay relación entre ambas


$$b_1 (0.75, 1.07)$$

Conclusión: la relación entre x y y es significativa



Supuestos de la regresión lineal

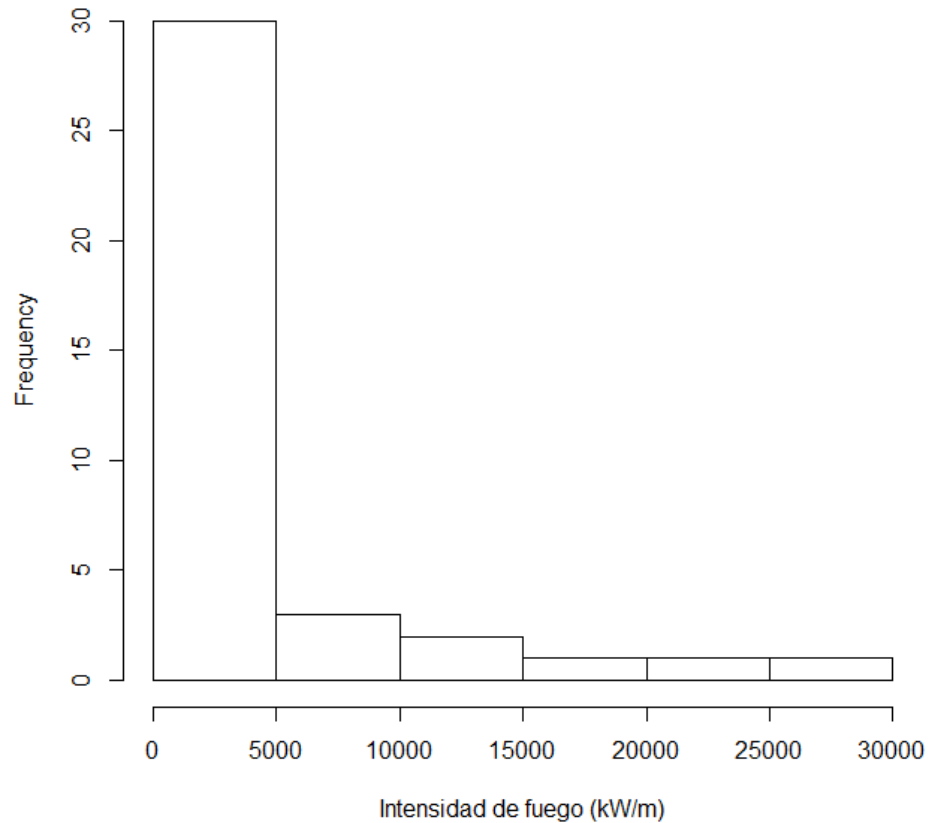
¿Que hacer cuando los supuestos no se respetan?

1. Transformar la variable dependiente e independiente (logaritmo, raíz cuadrada)
 2. Añadir nuevas variables
 3. Utilizar otro modelo que permita incluir la heterogeneidad de varianzas
 4. Realizar regresiones que utilicen otros tipos de distribuciones (Poisson, binomial)
 5. Utilizar métodos no paramétricos
- 

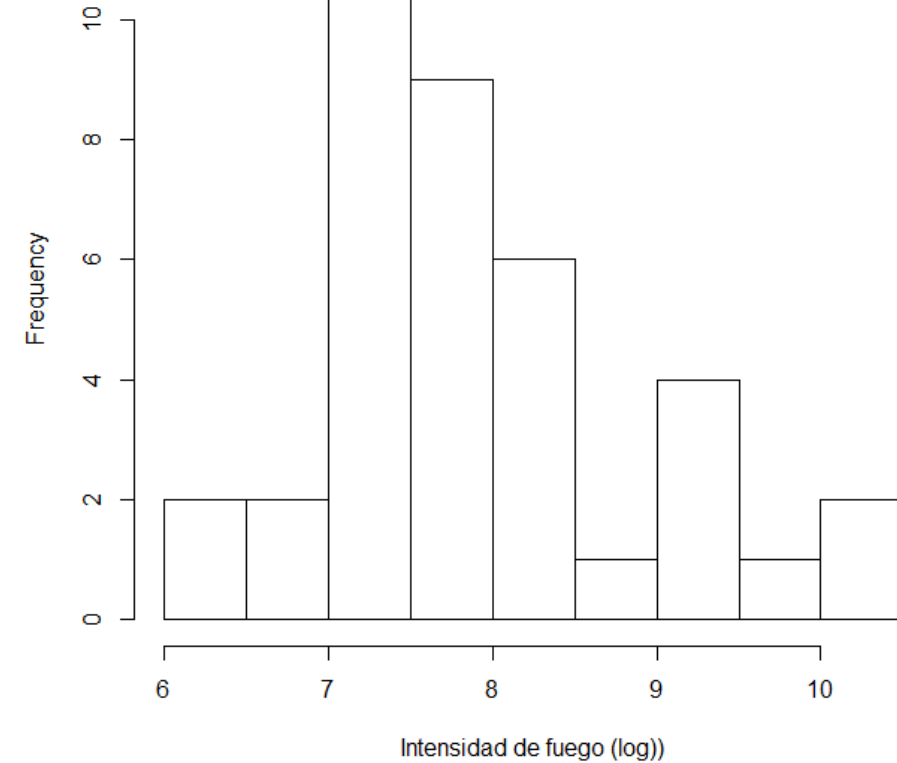
Supuestos de la regresión lineal

Transformación logarítmica

Histograma1



Histograma2



Ejercicio 6

1. Importar la base de datos `Extension_hielo.csv` que contiene información sobre la concentración de CO_2 atmosférico (ppm) y la extensión de hielo ártico (millones de km^2)
2. Construir una función que evalúe lo siguiente: A) SCX, SCY, SPXY, b) ordenada al origen, c) pendiente, d) SCE, e) coeficiente de determinación
3. Evaluar la base de datos `Extension_hielo.csv`