Modelos lineales

Tratan de explicar el comportamiento de una variable aleatoria mediante su relación lineal con los valores de otras que pueden influirla

Elementos del modelo básico:

Variable a explicar = constante común + suma de efectos de las variables o factores de influencia + Errores o Variaciones aleatorias

DISEÑO DE EXPERIMENTOS

Las variables explicativas (factores) son cualitativas

Tema 1: Análisis de la varianza unifactorial

Analiza y compara el comportamiento de una variable continua **Y** en distintos niveles (poblaciones o grupos o tratamientos) de **un factor** (variable explicativa)

Ejemplo: producción de un cultivo en parcelas iguales con distintos fertilizantes

Tema 2: Análisis de la varianza con varios factores

Analiza y compara el comportamiento de una variable continua **Y** en distintos niveles de **varios factores** (variables explicativas) y las posibles interacciones entre ellos.

Ejemplo: altura de una especie de árboles en distintos suelos y distintos climas.

REGRESIÓN

Las variables explicativas son cuantitativas

Tema 3: Regresión lineal simple

Analiza el comportamiento de una variable continua **Y** a través de los valores de otra variable continua **X** (variable explicativa)

Ejemplo: peso de un caimán en relación a su longitud medida por fotos desde el aire.

Tema 4: Regresión lineal múltiple

Analiza el comportamiento de una variable continua **Y** a través de los valores de otras variables continuas $\mathbf{X_1 \ldots X_k}$ (variables explicativas)

Ejemplo: crecimiento de un tipo de cultivo en función de las cantidades de distintas sustancias en el agua que lo riega.

Elementos básicos del procedimiento estadístico

- **Modelo:** planteamiento y definición de las variables que intervienen y sus propiedades teóricas
- **Muestra aleatoria:** número de observaciones que van a realizarse, procedimiento a seguir. Modelo teórico y sus propiedades.
- **Datos:** (muestra realizada) valores numéricos obtenidos al realizar efectivamente las observaciones previstas.
- **Aplicación de las técnicas estadísticas** adecuadas al diseño establecido

Antecedentes

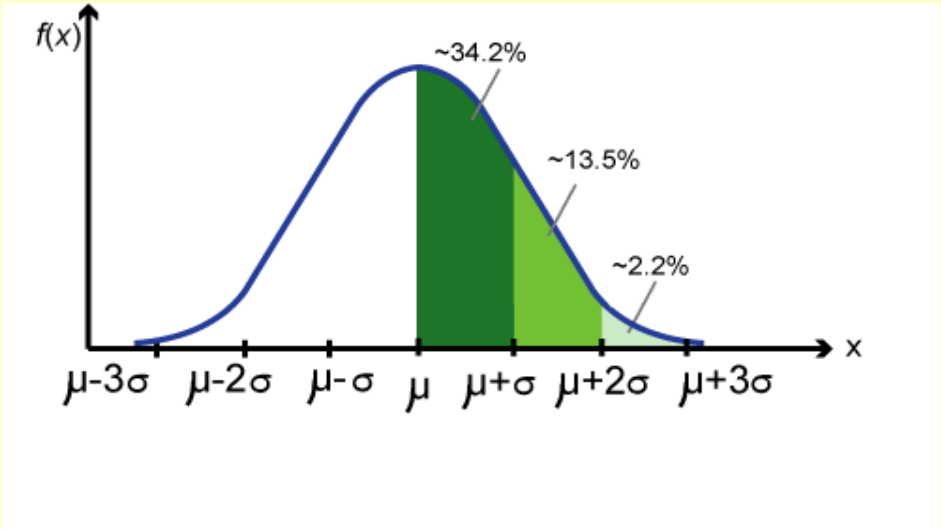
- **Modelos de probabilidad:**

  Variables aleatorias (dependientes de parámetros desconocidos). Independencia.
- **Estimación de los parámetros**
- **Intervalos de confianza**
- **Contraste de Hipótesis**

El modelo básico en este curso es la distribución Normal de parámetros μ y σ

La distribución Normal (μ,σ)

Diremos que X sigue una distribución N(μ,σ) si es una variable continua cuya función de densidad es:



μ = E(X) = valor medio de la variable X

σ^2 = Var(X) = varianza de la variable X

Normal tipificada N(0,1)

$Z = (X-\mu)/\sigma$ o $X = \sigma Z + \mu$

Estimación de los parámetros

Muestra aleatoria: $(X_1,\ldots,X_n)$

X_i = resultado que obtendremos
al realizar la i-ésima observación de X
(variables aleatorias i.i.d. Normales)

Datos (o muestra realizada): $(x_1,\ldots,x_n)$

x_i = resultado obtenido
al realizar la i-ésima observación de X (números)

Estimador de μ

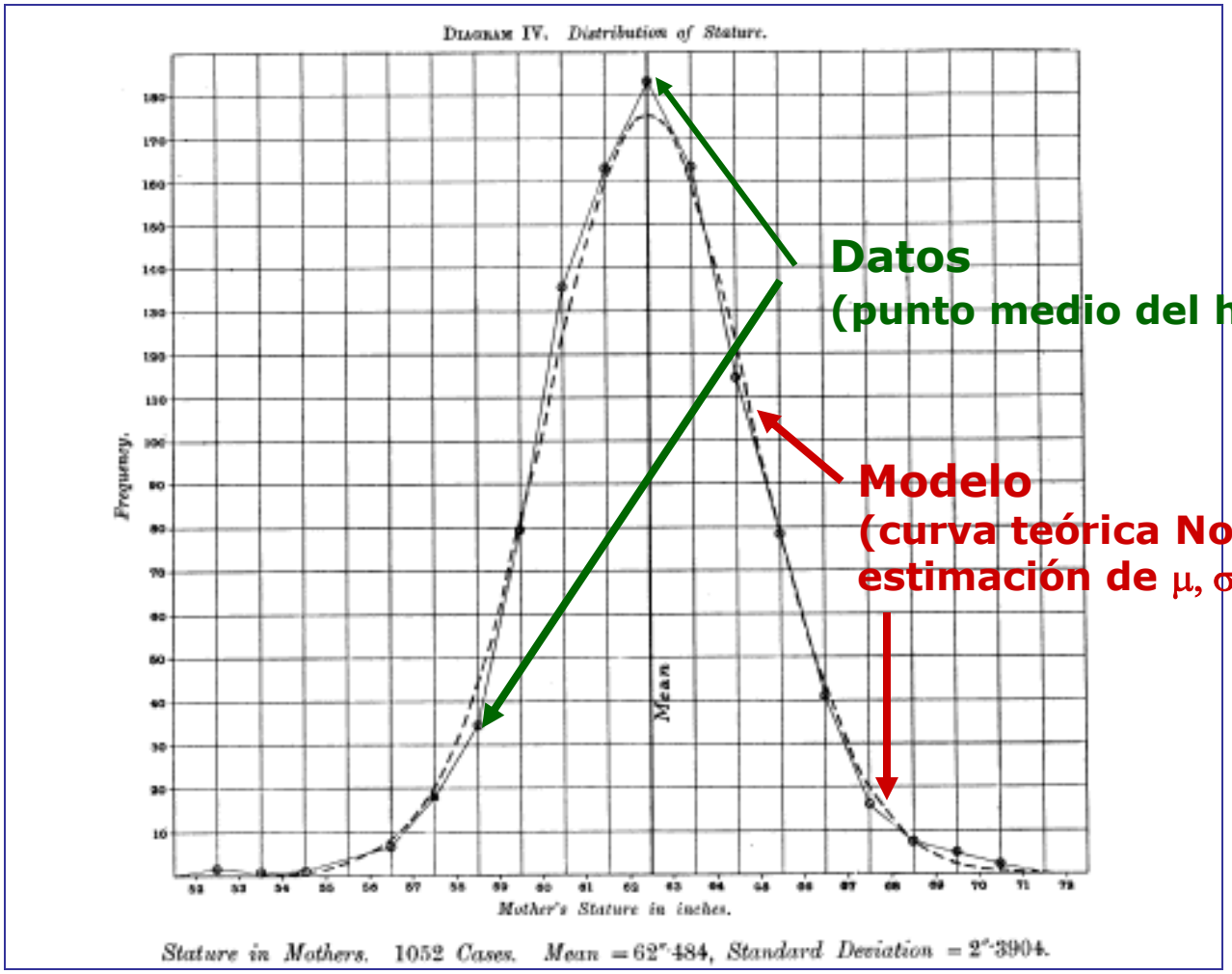
Media muestral $\mathbf{\bar{X}} = \Sigma X_i/n$

Estimadores de σ^2

Varianza muestral $\mathbf{V_x} = \Sigma(X_i - \bar{X})^2/n$

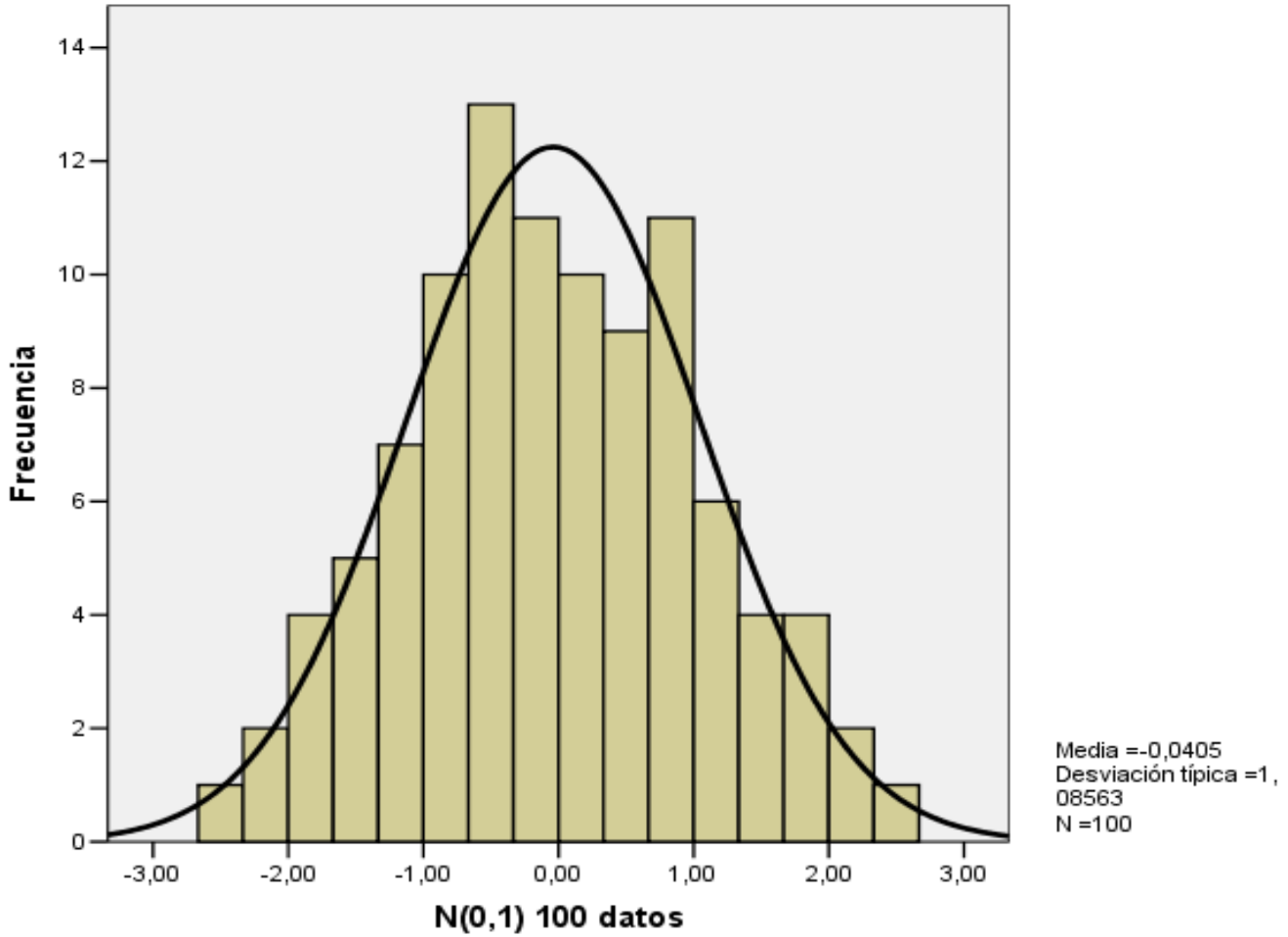
Cuasivarianza muestral $\mathbf{S^2_x} = \Sigma(X_i - \bar{X})^2/(n-1)$

Ajuste de una distribución Normal (K. Pearson, 1903)
X= estatura de una mujer elegida al azar



n= 1.052 mujeres, media estimada = 62’48 pulgadas, desviación típica estimada = 2’39 pulgadas

Histograma y curva Normal ajustada a 100 datos simulados con ordenador de una variable N(0,1)



El efecto del azar y el tamaño de la muestra: simulaciones

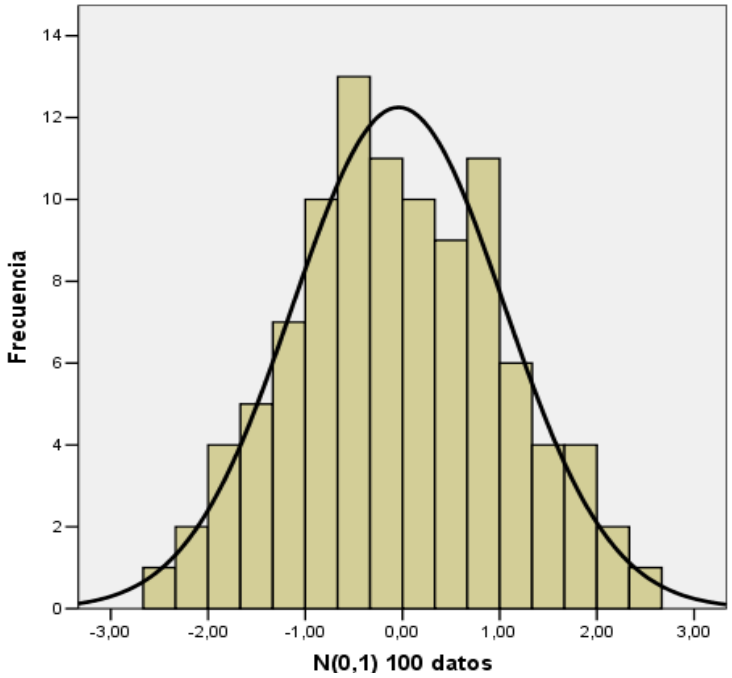
con 100 datos	N(0,1)	N(2,1)	N(4,1)
Media	-0,040	1,965	4,048
Mediana	-0,085	1,931	3,977
Desviación estándar	1,086	1,006	1,062
Varianza de la muestra	1,179	1,013	1,129
Coeficiente de asimetría	0,091	0,070	0,263
Mínimo	-2,578	-0,474	1,933
Máximo	2,376	4,374	6,324
Cuenta	100	100	100

con 10 datos	N (0,1)	N (2,1)	N (4,1)
Media	0,046	1,638	3,951
Mediana	0,005	1,850	3,885
Desviación estándar	1,292	0,862	1,169
Varianza de la muestra	1,670	0,744	1,367
Coeficiente de asimetría	-0,359	-1,215	0,180
Mínimo	-2,184	-0,234	2,369
Máximo	1,733	2,656	5,583
Cuenta	10	10	10

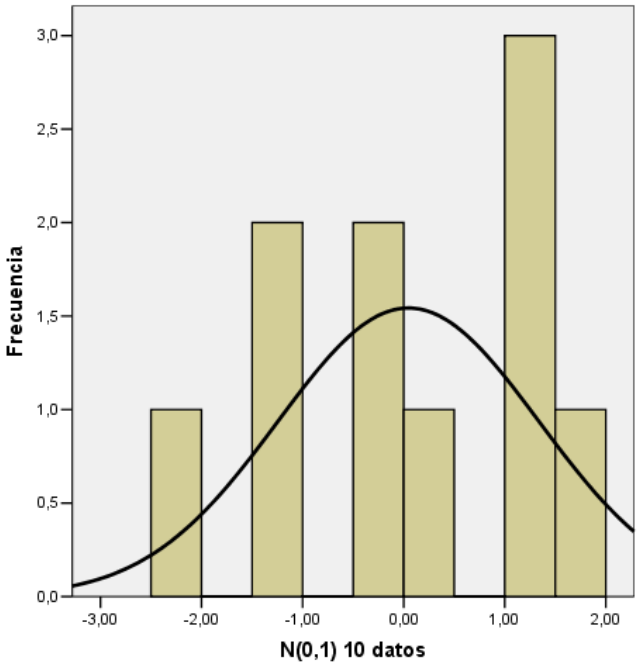
con 100 datos	N(0,10)	N(2,10)	N(4,10)
Media	-0,405	1,652	4,476
Mediana	-0,849	1,313	3,773
Desviación estándar	10,856	10,062	10,625
Varianza de la muestra	117,859	101,253	112,884
Coeficiente de asimetría	0,091	0,070	0,263
Mínimo	-25,776	-22,738	-16,674
Máximo	23,757	25,744	27,237
Cuenta	100	100	100

con 10 datos	N(0,10)	N(2,10)	N(4,10)
Media	0,465	-1,616	3,505
Mediana	0,050	0,502	2,854
Desviación estándar	12,924	8,623	11,693
Varianza de la muestra	167,030	74,357	136,722
Coeficiente de asimetría	-0,359	-1,215	0,180
Mínimo	-21,836	-20,335	-12,312
Máximo	17,331	8,562	19,825
Cuenta	10	10	10

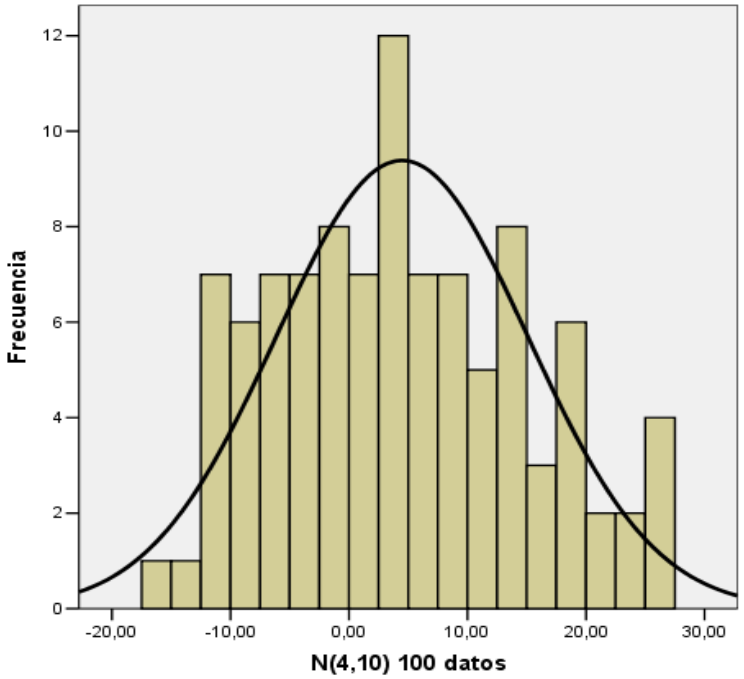
Con un generador de números aleatorios (Excel) hemos simulado datos de varias variables Normales con distintas medias (0, 2 y 4) y distintas desviaciones típicas (1 y 10)



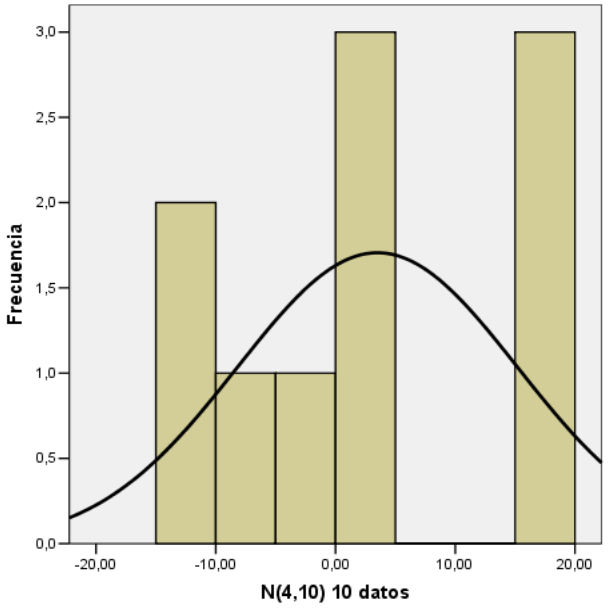
Media =-0,0405
Desviación típica =1,08563
N =100



Media =0,0465
Desviación típica =1,2924
N =10



Media =4,476
Desviación típica =10,62467
N =100



Media =3,5054
Desviación típica =11,69282
N =10

Gráfico P-P Normal de N(0,1) 100 datos

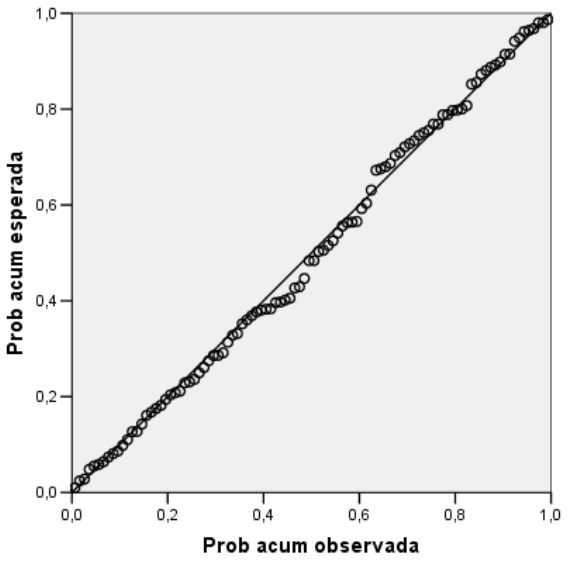


Gráfico P-P Normal de N(0,1) 10 datos

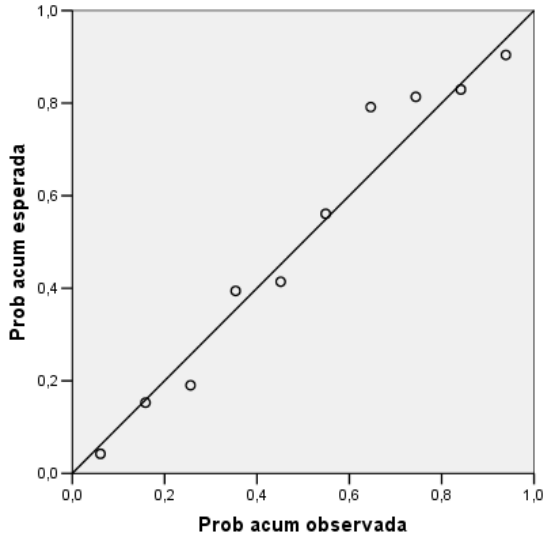


Gráfico P-P Normal de N(4,10) 100 datos

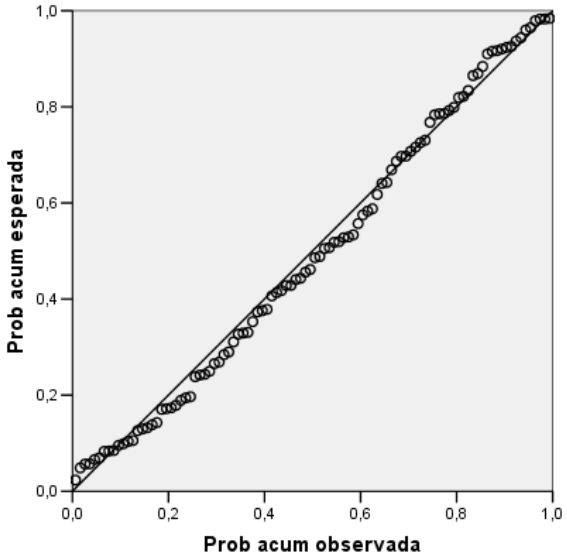
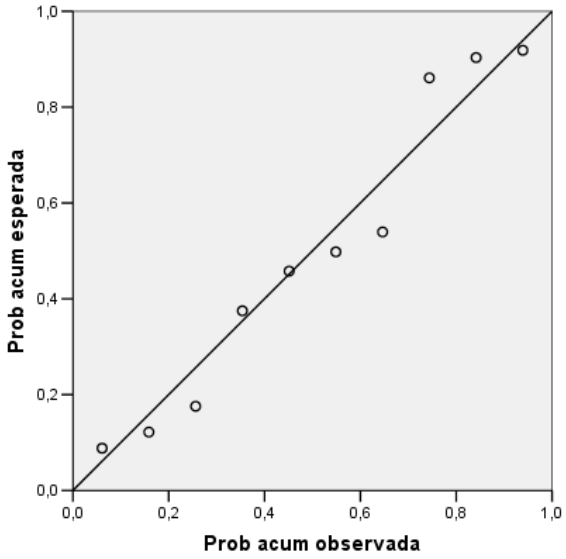


Gráfico P-P Normal de N(4,10) 10 datos



Intervalos de confianza

2 poblaciones Normales e independientes

$X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$ independientes

$(X_1, \ldots, X_{n_1})$ m.a.s. de X; se calcula $\bar{x}$ y s_1^2.

$(Y_1, \ldots, Y_{n_2})$ m.a.s. de Y; se calcula $\bar{y}$ y s_2^2.

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Intervalo de confianza $1 - \alpha$ para $\mu_1 - \mu_2$:

$$I = \left(\bar{x} - \bar{y} \pm t_{n_1+n_2-2;\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \qquad \sigma_1,\ \sigma_2 \text{ desconocidas, } \sigma_1 = \sigma_2$$

Intervalo de confianza $1 - \alpha$ para σ_1^2/σ_2^2: $I = \left(\dfrac{s_1^2/s_2^2}{F_{n_1-1;n_2-1;\alpha/2}} \ , \ (s_1^2/s_2^2)\, F_{n_2-1;n_1-1;\alpha/2} \right)$

En los temas 1 y 2 extenderemos estas ideas al caso de 2 o más poblaciones Normales e independientes

Contrastes de Hipótesis

2 poblaciones Normales e independientes

$X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$ independientes

$(X_1, \ldots, X_{n_1})$ m.a.s. de X; se calcula $\bar{x}$ y s_1^2.

$(Y_1, \ldots, Y_{n_2})$ m.a.s. de Y; se calcula $\bar{y}$ y s_2^2.

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Contraste *t* de igualdad de medias con σ_1, σ_2 desconocidas pero iguales

$$H_0 : \mu_1 = \mu_2 \ (\sigma_1 = \sigma_2) \qquad R = \left\{ |\bar{x} - \bar{y}| > t_{n_1+n_2-2;\alpha/2}\ s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

Para resolver el contraste anterior hay que contrastar previamente:

$$H_0 : \sigma_1 = \sigma_2 \qquad R = \left\{ s_1^2 / s_2^2 \notin \left[F_{n_1-1;n_2-1;1-\alpha/2}\ ,\ F_{n_1-1;n_2-1;\alpha/2} \right] \right\}$$

Obsérvese la relación entre estos contrastes y los intervalos anteriores

Tema 1
Análisis de la varianza unifactorial

- **Definición de la variable a explicar** (también llamada variable respuesta)
- **Definición de los I distintos niveles (poblaciones, cualidades, grupos, tratamientos...) de la variable explicativa (factor)**
- **Modelo:**

$$Y_i = \mu_i + U = \mu + \alpha_i + U_i \qquad i = 1, 2, \ldots, I$$

Donde:

Y_i representa la respuesta de la variable en el i-ésimo nivel del factor explicativo.

$\mu_i = E(Y_i)$ es el valor medio de $Y_i = \mu + \alpha_i$ $\qquad \Sigma\alpha_i = 0$

μ_i a veces se descompone como $\mu_i = \mu + \alpha_i$ $(i = 1, 2, \ldots, I)$ donde α_i representa el efecto que sobre la media global μ tiene del nivel i

U_i es la variación aleatoria de las Y_i (v.a. independientes y con la misma distribución $N(0,\sigma)$ para todo i)

Que U_i siga una distribución $N(0,\sigma)$ implica que:

Y_i sigue una distribución $N(\mu_i, \sigma)$

$$\sigma^2 = Var(U_i) = Var(Y_i) \text{ igual para todo i}$$

La elección de los I niveles de la variable explicativa puede hacerse de dos maneras:

1. ***Niveles fijos*: los distintos tratamientos o poblaciones son seleccionados por el experimentador.**
   Por ejemplo, si se trata de estudiar el efecto sobre una enfermedad de distintos medicamentos, los medicamentos son elegidos por el experimentador.

2. ***Niveles aleatorios*: los distintos tratamientos o poblaciones son seleccionados al azar entre todos los posibles.**
   Por ejemplo, si se trata de estudiar el efecto de un contaminante sobre distintas razas de perros, se pueden seleccionar al azar I razas entre todas las posibles.

En las propiedades estadísticas del Análisis de la Varianza unifactorial no hay diferencia entre la selección fija o aleatoria de los niveles.

Muestra aleatoria y datos

1 - **Muestra aleatoria**: Y_{ij} resultado que obtendremos en la j-ésima observación dentro del i-ésimo nivel del factor explicativo.

$i = 1,2,\ldots,I$ $\quad$ $j = 1,2,\ldots,n_i$

n_i es el tamaño de la muestra en el nivel i

Si todas las muestras tienen el mismo tamaño el diseño se llama **equilibrado**

Las observaciones se realizarán al azar e independientemente unas de otras.

2 – **Datos**: y_{ij} resultado obtenido en la j-ésima observación dentro del i-ésimo nivel del factor explicativo.

nº total de datos: $n = n_1 + \ldots + n_I$

Muestra aleatoria

Factor	1	2	...	I
	Y_{11}	Y_{21}	...	Y_{I1}
	Y_{12}	Y_{22}	...	Y_{I2}
	$\vdots$	$\vdots$		$\vdots$
	$\vdots$	Y_{2n_2}		$\vdots$
	$\vdots$			Y_{In_I}
	Y_{1n_1}			
	$\downarrow$	$\downarrow$		$\downarrow$
	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$		$\bar{Y}_{I.}$
	$\hat{S}_1^2$	$\hat{S}_2^2$		$\hat{S}_I^2$

Datos

Factor	1	2	3
	20	15	19
	18	17	11
	21	22	18
	22	24	22
	19		17
	25		
n_i	6	4	5
$\bar{y}_{i.}$	20.8	19.5	17.4
s_i^2	6.2	17.6	16.2

$$Y_{ij} \sim N(\mu_i; \sigma^2) \text{ independientes; } \; i = 1, ..., I; \; j = 1., ..., n_i; \; \sum_i n_i = n$$

Análisis estadístico:
Estimación de los parámetros desconocidos

Parámetros desconocidos del modelo (I+1) :

$\mu_1, \ldots, \mu_I \quad \sigma$

Estimaciones de los parámetros:

$$\hat{\mu}_i = \bar{y}_{i.} = \frac{1}{n_i} \sum_j y_{ij} \ , \ i = 1, \ldots, I$$

$$\hat{\sigma}^2 = S_R^2 = \frac{1}{n - I} \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

es $(n_i\text{-}1)S^2_i$

Análisis estadístico:
Estimación de los parámetros desconocidos

Intervalos de confianza

$$IC_{1-\alpha}(\mu_i) = \left(\bar{y}_{i.} \pm t_{n-I;\alpha/2} \; S_R \sqrt{\frac{1}{n_i}} \right)$$

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-I)S_R^2}{\chi^2_{n-I;\alpha/2}} \; ; \frac{(n-I)S_R^2}{\chi^2_{n-I;1-\alpha/2}} \right)$$

Análisis estadístico: requisitos previos

1. **Normalidad:** los datos obtenidos en cada nivel del factor se ajustan razonablemente a una distribución Normal (gráficos y contrastes)
   Y_i sigue una distribución $N(\mu_i, \sigma)$ para cada i
2. **Homocedasticidad:** la variabilidad de los datos en cada nivel del factor es similar (contraste de igualdad de varianzas)
   $\sigma^2 = Var(Y_i)$ igual para todo i
3. **Linealidad:** los residuos se distribuyen homogéneamente alrededor del cero (gráfico de residuos) **$E(U_i) = 0$ para todo i**
4. **Independencia:** las observaciones se realizan de forma independiente unas de otras (diseño de la obtención de datos)

SI HAY DESVIACIONES SIGNIFICATIVAS SOBRE ESTOS REQUISITOS LOS RESULTADOS POSTERIORES PUEDEN SER INCORRECTOS

Residuos tipificados

Los requisitos previos del modelo exigen que las U_i sigan una $N(0,\sigma)$ o lo que es lo mismo:

$U_i/\sigma = (Y_i - \mu_i)/\sigma$ seguirá una $N(0,1)$

Utilizando los datos de la muestra (y_{ij}: valores observados de Y_i) y tipificándolos con las estimaciones adecuadas de los parámetros obtendremos una lista de números que, si los requisitos previos son ciertos, se distribuirán aproximadamente según la curva Normal de media 0 y varianza 1

Obtención de los residuos tipificados con SPSS (ejemplo con los datos de la diapositiva 18)

1- Introducir los datos identificando el nivel del factor y nombrando las variables

2- Pulsar “analizar”, luego:
“modelo lineal general” y “univariante”

3- Identificar la variable dependiente (datos), identificar el factor (nivel)

4- Pulsar “guardar” y marcar “residuos tipificados” (aparece una nueva columna en el editor de datos (ZRE_1 residuo estandarizado)

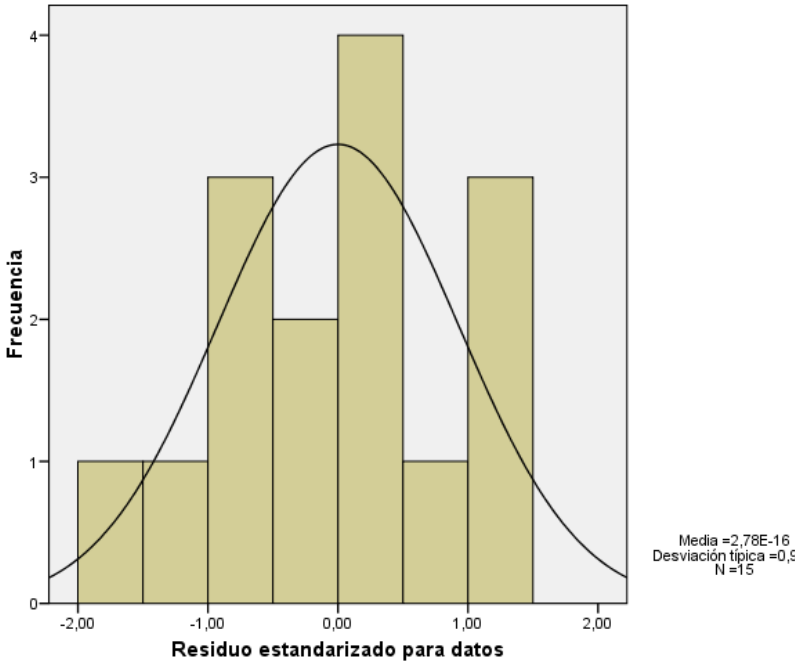
Datos	Nivel	Residuo tipificado
20,00	1	-,24
18,00	1	-,80
21,00	1	,05
22,00	1	,33
19,00	1	-,52
25,00	1	1,18
15,00	2	-1,28
17,00	2	-,71
22,00	2	,71
24,00	2	1,28
19,00	3	,45
11,00	3	-1,82
18,00	3	,17
22,00	3	1,31
17,00	3	-,11

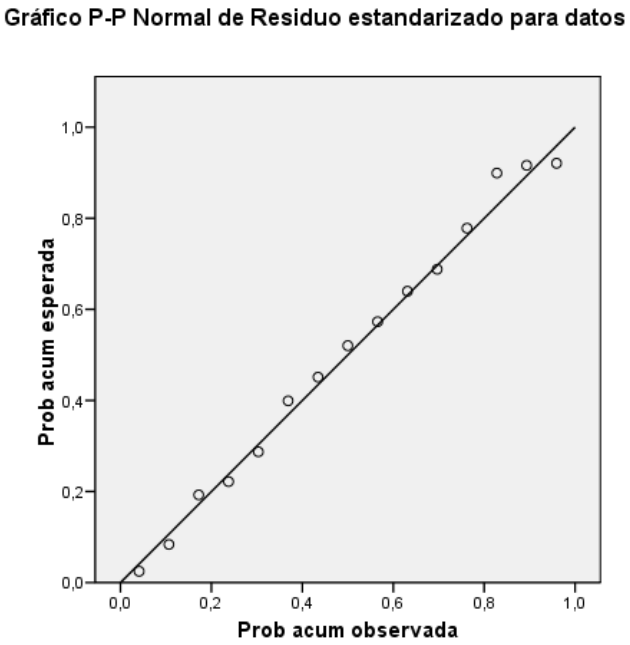
Nota: según la versión de SPSS puede haber algunas diferencias

Gráficos de los residuos (Normalidad)

Histograma:
Pulsar “gráficos”, “histograma”
Identificar como variable “residuo estandarizado”, marcar “mostrar curva Normal” y “aceptar”

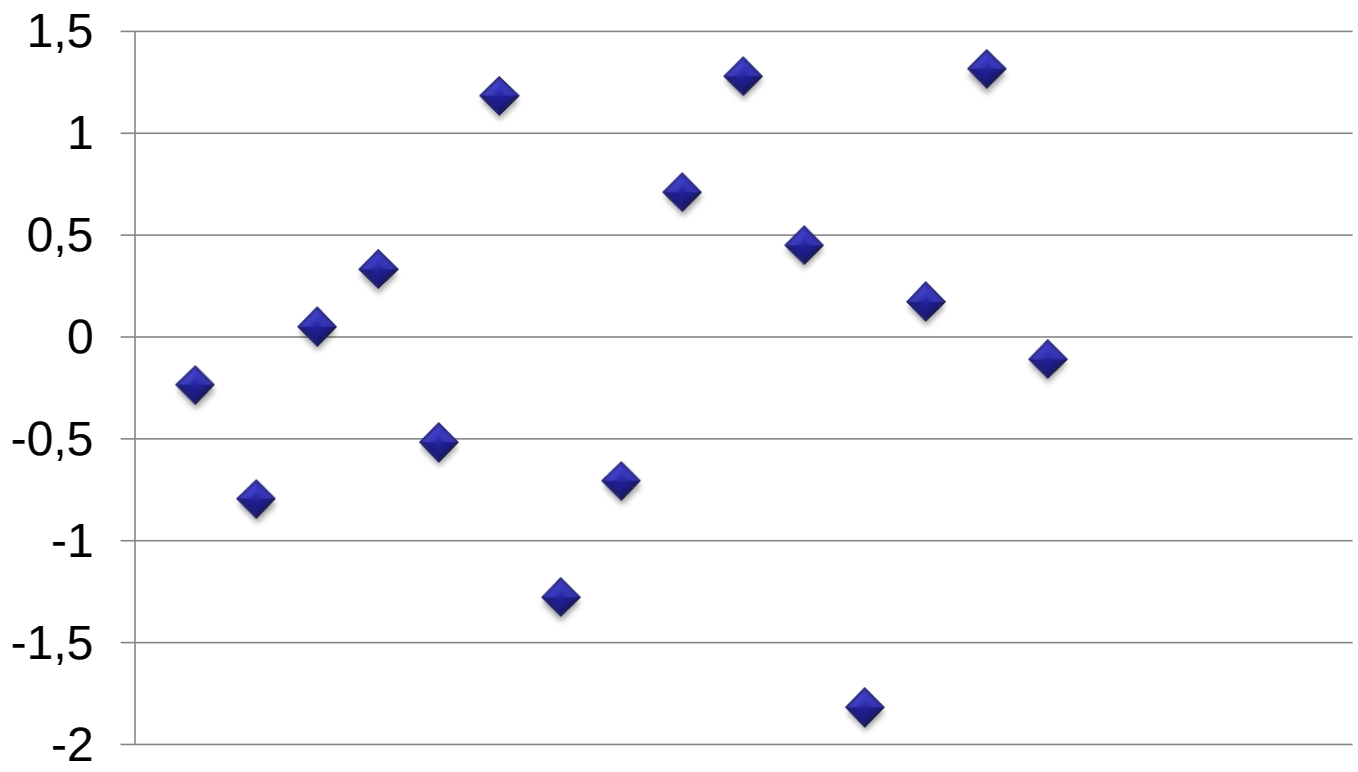
Gráfico P-P:
Pulsar “analizar”, “estadísticos descriptivos” y “gráficos P-P”
Identificar como variable “residuo estandarizado” y “aceptar”





Nota: según la versión de SPSS puede haber algunas diferencias

Gráficos de los residuos (diagrama de dispersión)



Análisis estadístico: ANOVA

(Análisis de la Varianza)

$$(Y_{ij} - \bar{Y}_{..}) = \underbrace{(Y_{ij} - \bar{Y}_{i.})}_{\text{Desviaciones intra-grupos}} + \underbrace{(\bar{Y}_{i.} - \bar{Y}_{..})}_{\text{Desviaciones entre-grupos}}$$

Elevando al cuadrado, sumando y simplificando

$$SCE = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SCR = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

$$SCT = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$$

Se cumple que: SCE + SCR = SCT

SCE Suma de cuadrados explicada (variabilidad debida a que hay distintos niveles del factor)

SCR Suma de cuadrados residual (variabilidad interna dentro de cada nivel)

SCT Suma de cuadrados total (variabilidad total de todos los datos)

Análisis estadístico: ANOVA

(Contraste de igualdad de medias)

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$ (equivalentemente H_0: $\alpha_1=\alpha_2=\ldots=\alpha_I=0$)

(todas las medias son iguales, el factor no influye)

$H_1 : \mu_i \neq \mu_j$ para algún par i,j

(las medias difieren en al menos dos de los niveles, el factor influye)

Tabla ANOVA

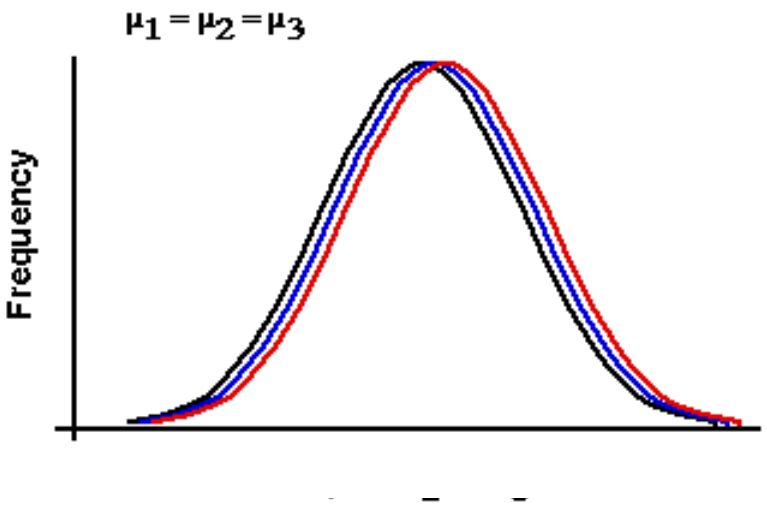
Suma de cuadrados	g.l.	Varianza	Estadístico
$SCE = \sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$I-1$	$\frac{SCE}{I-1}$	$F = \frac{SCE/(I-1)}{SCR/(n-I)}$
$SCR = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$	$n-I$	$S_R^2 = \frac{SCR}{n-I}$	
$SCT = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$	$n-1$		

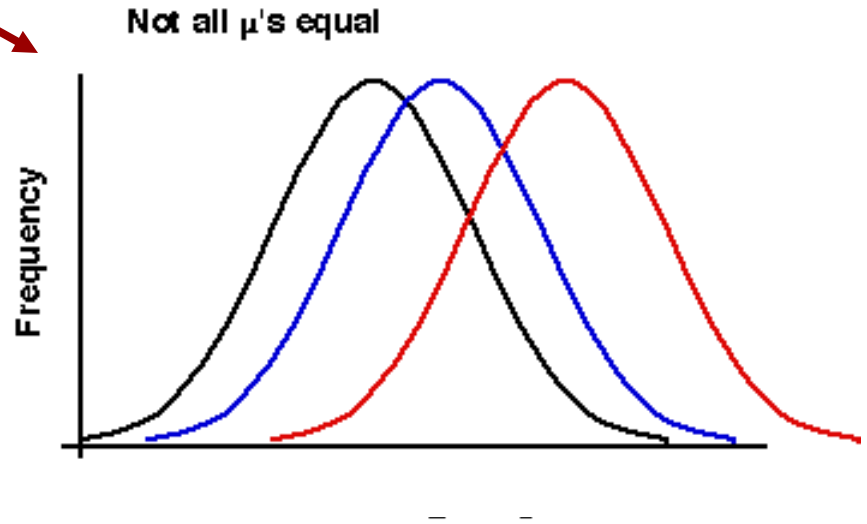
A nivel de significación α, rechazamos cuando

$$F > F_{I-1,n-I,\alpha}$$

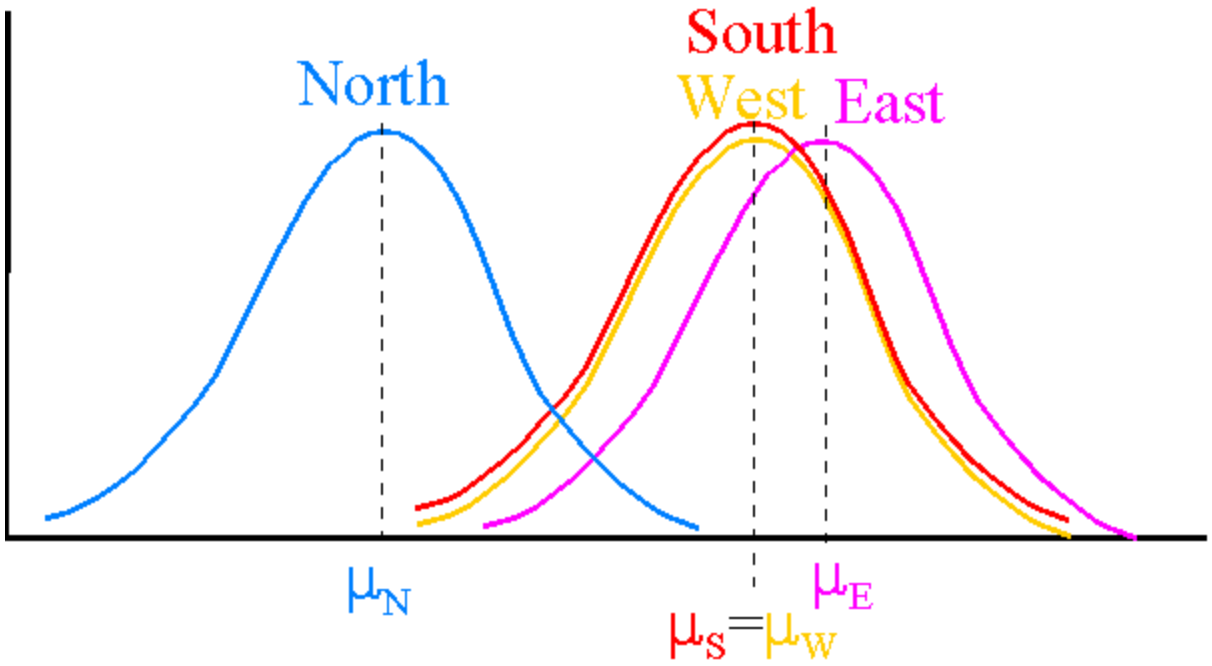
Con I = 3 tendríamos 5 posibilidades:

1. (H_0) Las tres medias iguales
2. (H_1) Dos medias iguales y una distinta
3. (H_1) Las tres medias distintas





ANOVA con I = 2 es matemáticamente equivalente al contraste de la *t* de Student para la igualdad de medias con varianzas iguales



Este gráfico representa las distribuciones ajustadas a los datos de 4 grupos de personas (N del norte, S del sur, E del este y W del oeste) a las que se les pidió que estimasen el precio adecuado de un nuevo producto. Los del norte parece que lo estimaron más barato...

Source of Variation	d.f.	Sums of Squares	Mean Squares	F ratio	Prob>F
Model	3	45	15.00	6.00	0.0061
Error	16	40	2.50		
Total	19	85			

¿Qué proporción de la variabilidad de los datos está explicada porque hay distintos niveles de un factor?

Coeficiente de determinación R^2

$$R^2 = \frac{SCE}{SCT}$$

(A)			
	20	22	24
	19	22	24
	20	22	23
	21	22	25
$\bar{Y}_{i\cdot} =$	20	22	24

(B)			
	45	8	15
	0	30	44
	10	38	2
	25	12	35
$\bar{Y}_{i\cdot} =$	20	22	24

R^2 = 0.89

ANOVA para el ejemplo (A)

	Suma de Cuadrados	g.l.	Varianza	F	$p-$valor
Explicada	32	2	16,000	36	0,000
Residual	4	9	0,444		
Total	36	11			

R^2 = 0.01

ANOVA para el ejemplo (B)

	Suma de Cuadrados	g.l.	Varianza	F	$p-$valor
Explicada	32	2	16,000	0,05	0,951
Residual	2852	9	316,889		
Total	2884	11			

El contraste ANOVA equilibrado (con iguales tamaños de las muestras) es bastante fiable (robusto) al rechazar H_0 incluso con desviaciones pequeñas de los requisitos de igualdad de varianzas o Normalidad.

Si las varianzas son muy diferentes o se detectan serias desviaciones de la Normalidad, se pueden realizar transformaciones de la variable Y que podrían resolver el problema. Por ejemplo tomando el Log Y (si la variabilidad crece con los valores de Y) o alguna potencia de Y.

Otra situación irregular que puede detectarse es la existencia de datos anómalos (*outliers*) que deben detectarse. En este caso habría que estudiar más a fondo dichos datos y su posible causa de anomalía.

Ejemplo* 1

Analysis of differences between the Across Trophic Level System Simulation (ATLSS) High Resolution Topography (HRT) model output and the United States Geological Survey (USGS) High Accuracy Elevation Data (HAED).

by
Scott M. Duke-Sylvester
The Institute for Environmental Modeling
University of Tennessee at Knoxville

*** Todos los ejemplos y sus resultados deben discutirse**

Las condiciones hidrológicas locales son importantes para el comportamiento y la vida de la flora y la fauna.

La elevación del terreno es, a su vez, importante para las condiciones hidrológicas locales.

Tres métodos de medición de la altitud:

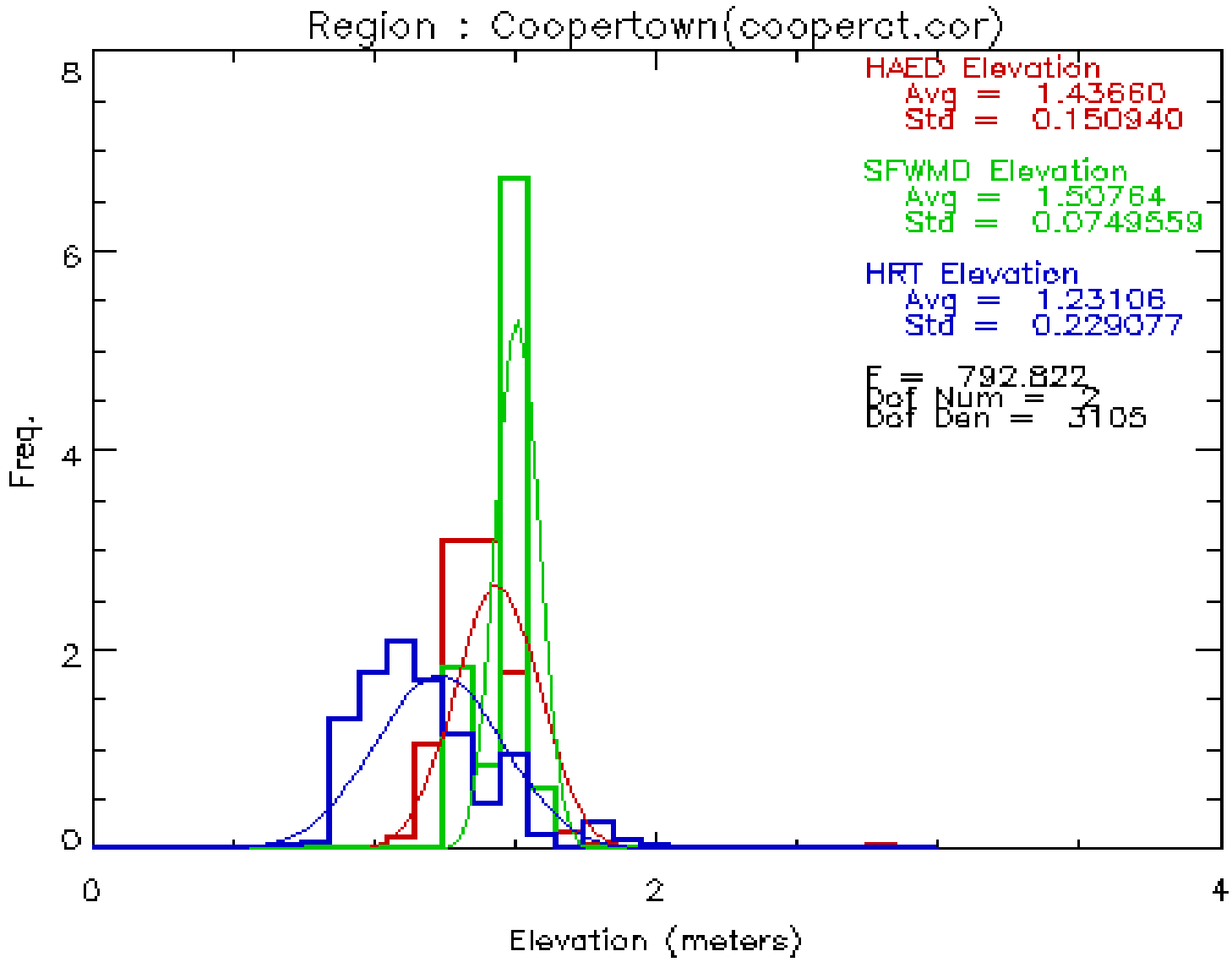
1. **HRT (ATLSS High Resolution Topography)** su precisión y fuentes de error han sido cuestionados.

2. **SFWMD (South Florida Water Management District)** el modelo HRT se basa parcialmente en las medidas proporcionadas por SFWMD.

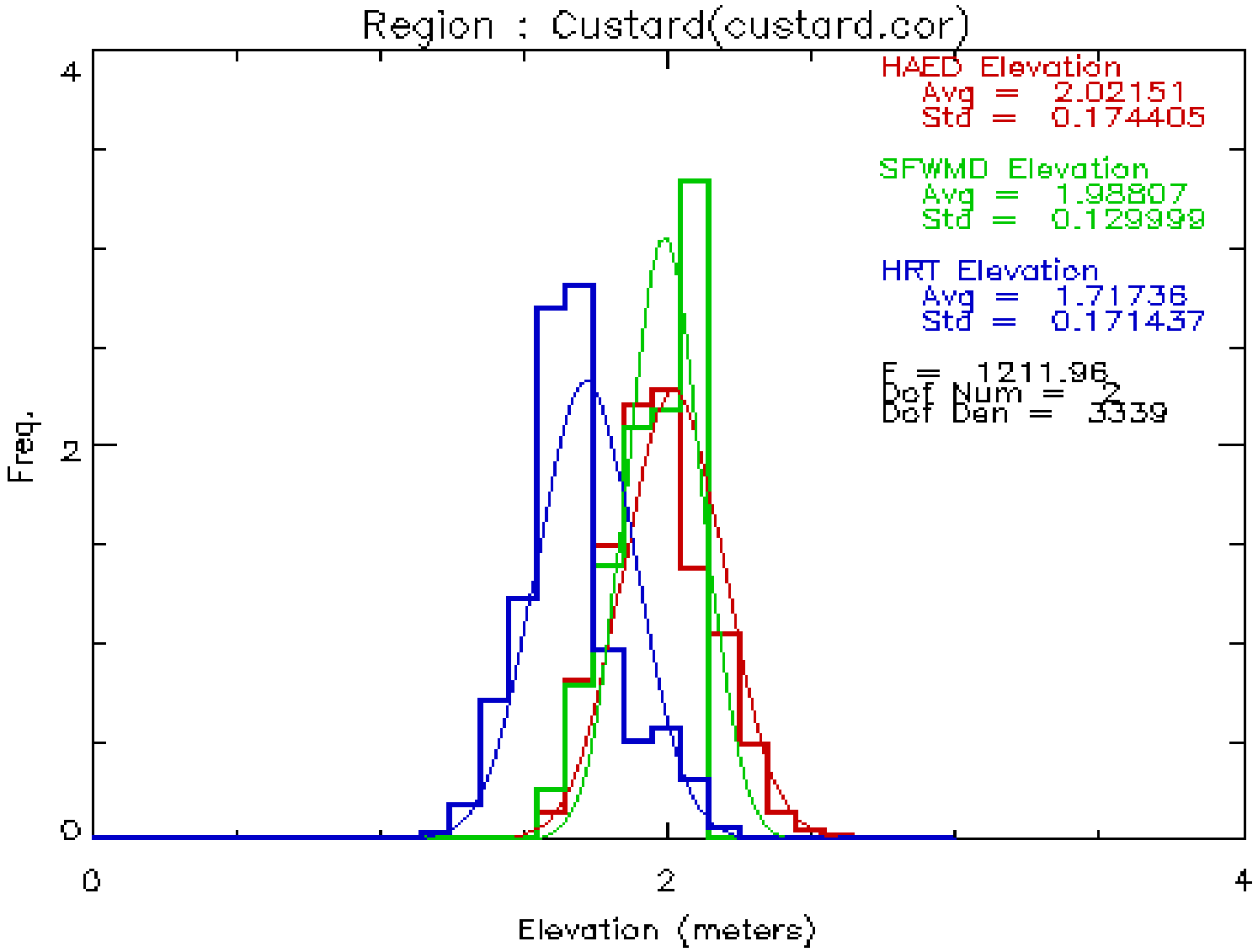
3. **HAED (High Accuracy Elevation Data)** proyecto liderado por el US Geological Survey para obtener medidas precisas de altitud en los Everglades. La técnica utiliza GPS diferencial, que proporciona medidas con una precisión que antes era difícil de conseguir.

El artículo analiza datos, con los tres métodos, de 11 zonas del sur de Florida. A continuación se presentan, para discusión 4 zonas.

Las alturas se agregan en décimas de metro.
Las curvas muestran la distribución Normal ajustada a los datos.



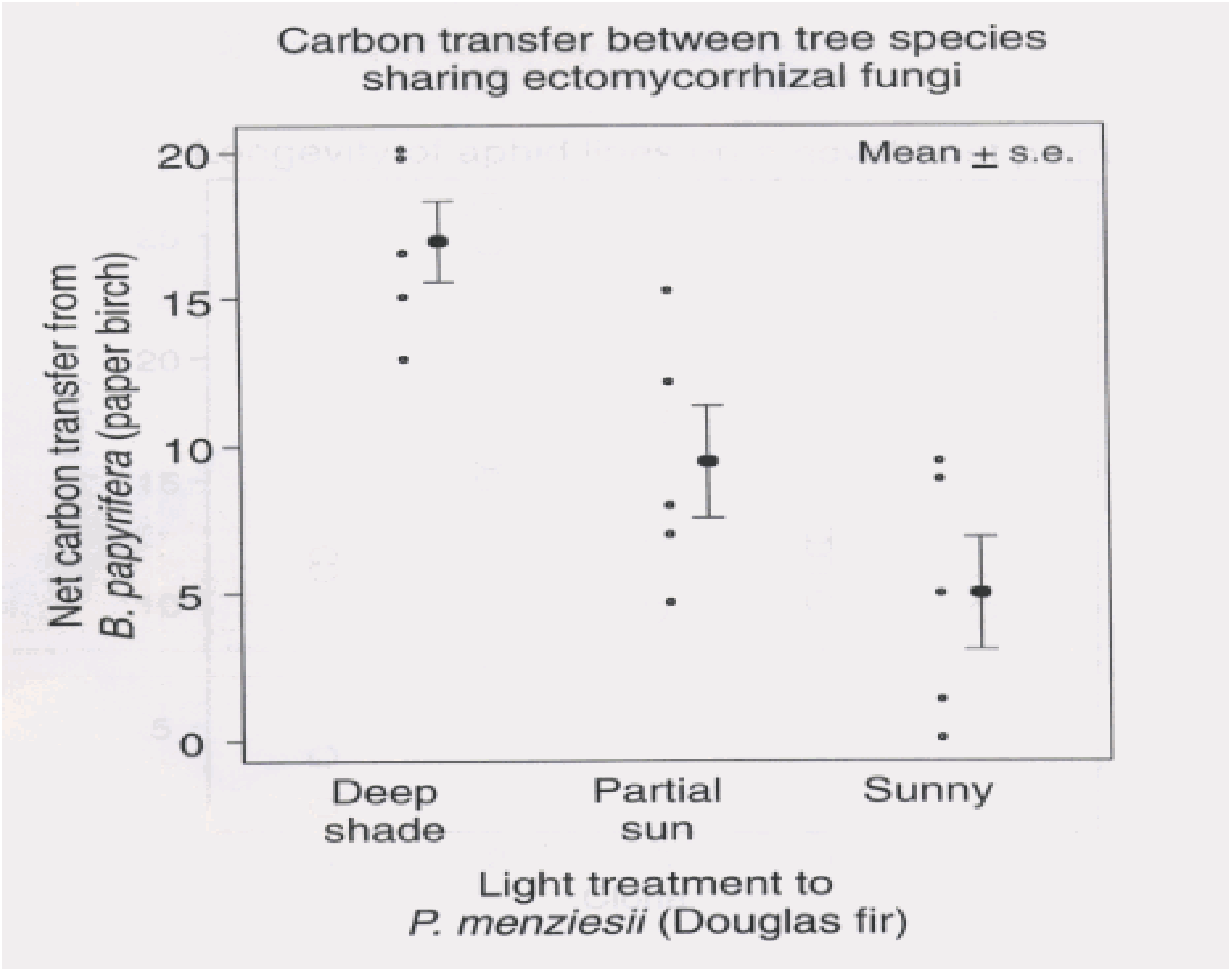
Las alturas se agregan en décimas de metro.
Las curvas muestran la distribución Normal ajustada a los datos.



Ejemplo* 2

Muchos árboles tienen una asociación física con unos hongos llamada mycorrhizae. El árbol proporciona carbono al hongo y el hongo proporciona minerales al árbol.
El micelio vegetativo de estos hongos se extiende lejos por el suelo, poniendo en contacto plantas diferentes, incluso de distintas especies.
Un grupo de investigadores estudiaron si, mediante esta relación, distintos árboles compartían también el carbono.
Para ello, eligieron parejas de árboles, uno de ellos un abedul americano situado al sol y el otro un abeto Douglas con diferentes situaciones (al sol, a la sombra o entre sol y sombra)
Introdujeron C13 y C14 en los abedules y midieron la transferencia neta de carbono a los abetos.
www.zoology.ubc.ca/.../ANOVA/ANOVA.html

*** Todos los ejemplos y sus resultados deben discutirse**


Carbon transfer between tree species
sharing ectomycorrhizal fungi
Mean ± s.e.
Net carbon transfer from
B. papyrifera (paper birch)
20
15
10
5
0
Deep
shade
Partial
sun
Sunny
Light treatment to
P. menziesii (Douglas fir)

Datos	Sombra	Sol y sombra	Sol
	15.1	4.7	8.9
	19.8	12.2	0.1
	13.0	15.3	5.0
	16.6	8.0	9.5
	20.1	7.0	1.4
medias	16.92	9.44	4.98
s_i	3.05	4.26	4.26
n_i	5	5	5

Tabla ANOVA

Source of Variation	SS	df	MS	F
light treatments	364.0	2	182.0	**11.99**
error	182.068	12	15.172	
total	546.0	14		

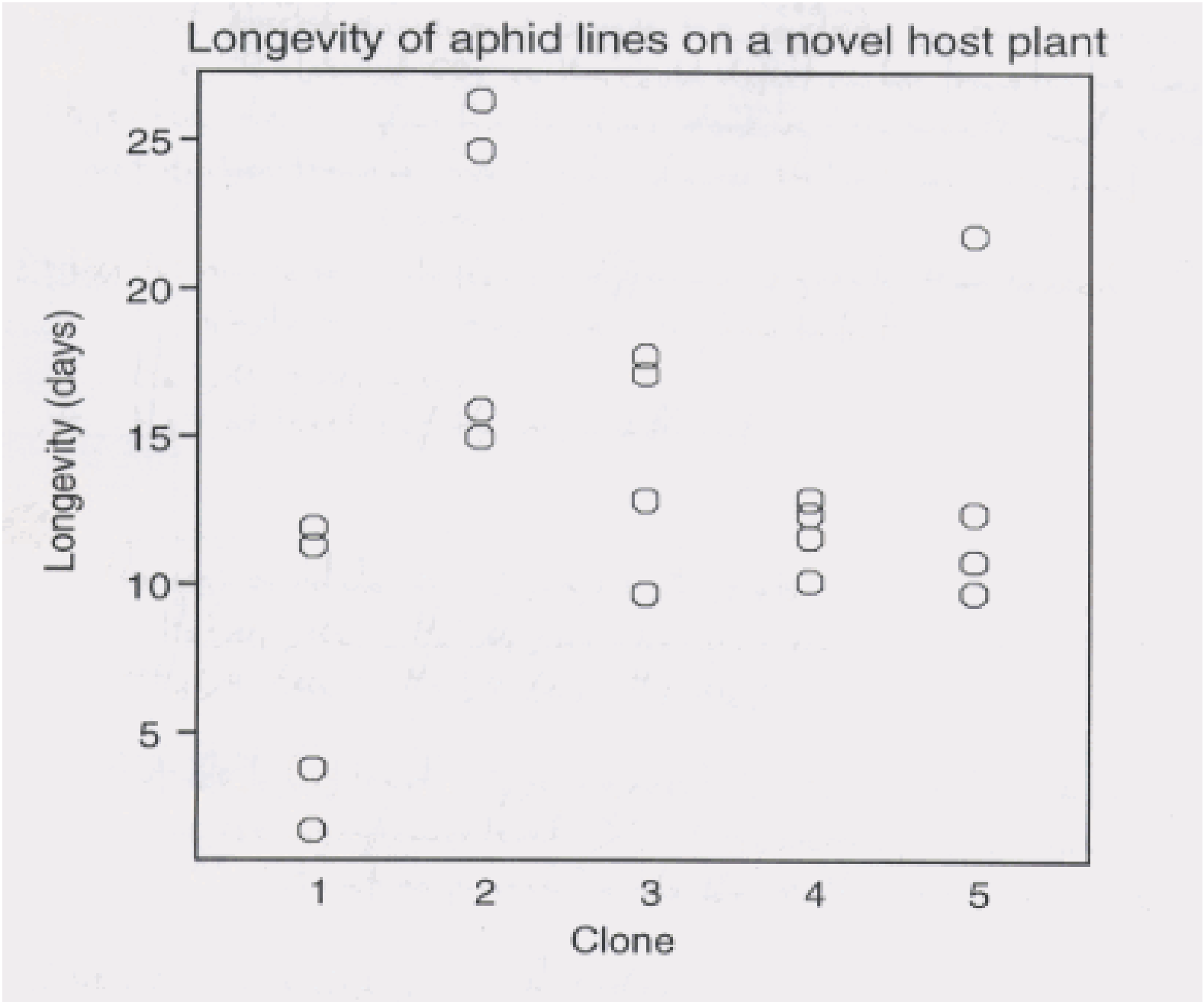
Ejemplo* 3

Una de las cuestiones abiertas en ecología y biología evolutiva es entender los factores que producen cambios evolutivos en una especie debidos al uso de nuevos recursos.

Se llevó a cabo un estudio sobre pulgones del guisante para ver si la habilidad para utilizar un nuevo huésped (alfalfa) tenía relación con variaciones genéticas.
Los investigadores midieron la longevidad de pulgones en alfalfa con 4 individuos en 5 diferentes clones, elegidos al azar en la población natural.

www.zoology.ubc.ca/.../ANOVA/ANOVA.html

*** Todos los ejemplos y sus resultados deben discutirse**


Longevity of aphid lines on a novel host plant
Longevity (days)
25
20
15
10
5
1
2
3
4
5
Clone

Clone	1	2	3	4	5
mean	7.16	20.44	14.34	11.73	13.67
s_i	5.19	5.84	3.78	1.19	5.52
n_i	4	4	4	4	4

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
H_1: At least one of the five families is different from the others.

Source of Variation	SS	df	MS	F
clone	368.55	4	92.139	4.3 $p<0.025$
error	321.76	15	21.45	
total	690.31	19		

$F_{0.05,4,15} = 3.06$

Análisis posteriores al rechazo de H_0

Al rechazar H_0 tenemos evidencia estadística de que al menos una de las μ_i es diferente de las otras pero ¿entre cuales hay diferencia significativa?

Intervalos de confianza para la diferencia de dos de las medias:

$$IC_{1-\alpha}(\mu_i - \mu_j) = \left(\bar{y}_{i.} - \bar{y}_{j.} \pm t_{n-I;\alpha/2} S_R \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\right)$$

Error típico

Contrates de hipótesis sobre dos de las medias:

$$\boldsymbol{H_0 : \mu_i = \mu_j \text{ vs. } H_1 : \mu_i \neq \mu_j}$$

α nivel de significación

$$\Downarrow$$

$$R = \left\{ \left| \frac{\bar{Y}_{i.} - \bar{Y}_{j.}}{\hat{S}_R \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \right| > t_{n-I,\alpha/2} \right\}$$

Equivalente a rechazar H_0 si el cero no está en el intervalo

Comparaciones múltiples:

Pruebas Post hoc: Test de Bonferroni

Si realizamos una comparación (con el mismo nivel α) de todas las posibles parejas de medias la probabilidad de que rechacemos incorrectamente en alguno de los contrastes puede ser muy alta, hasta:

$1- (1- \alpha)^c$ donde c es el número de contrastes que realicemos

por ejemplo si hay cinco niveles del factor, c= 10, si hay 10, c= 45

El test múltiple de Bonferroni fija un nivel de significación total α_T y realiza todos los contrastes de parejas con un $\alpha = \alpha_T / c$

Es importante señalar que puede ocurrir que rechacemos H_0 en ANOVA y no encontremos diferencias entre ningún par de medias con Bonferroni ...

Comparaciones múltiples:

Pruebas Post hoc: otros contrastes

El test de Bonferroni es muy conservador, sobre todo si c es grande. Por ejemplo, si el Factor tiene 5 niveles y fijamos $\alpha_T = 0.05$ tendremos que el α para cada contraste entre dos medias es 0.005.

Otros contrates múltiples:

Tukey (bueno si el diseño es equilibrado)

Scheffé (útil en el caso de tamaños muestrales diferentes, coincide siempre con ANOVA)

Dunnett (si hay un grupo "control")

Duncan

....

....

Ejemplo* 4

ANOVA con SPSS

Se encontraron 24 piezas de cerámica romana en 3 localidades diferentes del Reino Unido:
Llanederyn (L), Island Thorns (I) y Ashley Rails (A).

En cada pieza se midió el porcentaje de óxido de diferentes metales con una técnica de espectrometría de absorción atómica.

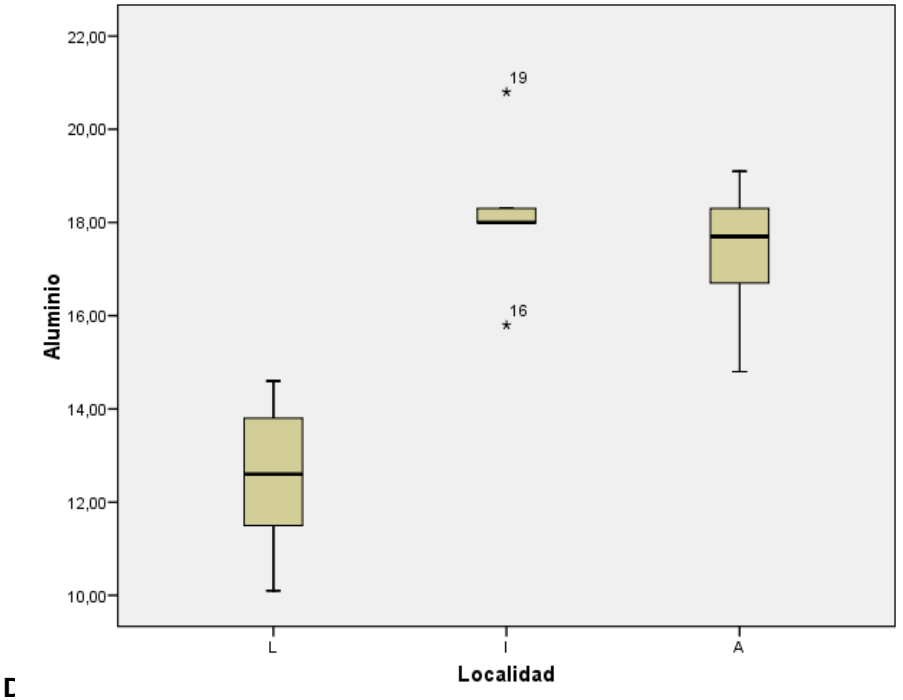
En este ejemplo analizaremos si hay diferencias en el porcentaje de óxido de aluminio en las tres localidades.
El diseño no es equilibrado.

Source: Data and Story Library; from Tubb, A., Parker, A.J. and Nickless, G. (1980), The analysis of Romano-British pottery by atomic absorption spectrophotometry. Archaeometry, 22, 153-171.
Education Queensland

*** Todos los ejemplos y sus resultados deben discutirse**

Análisis descriptivo

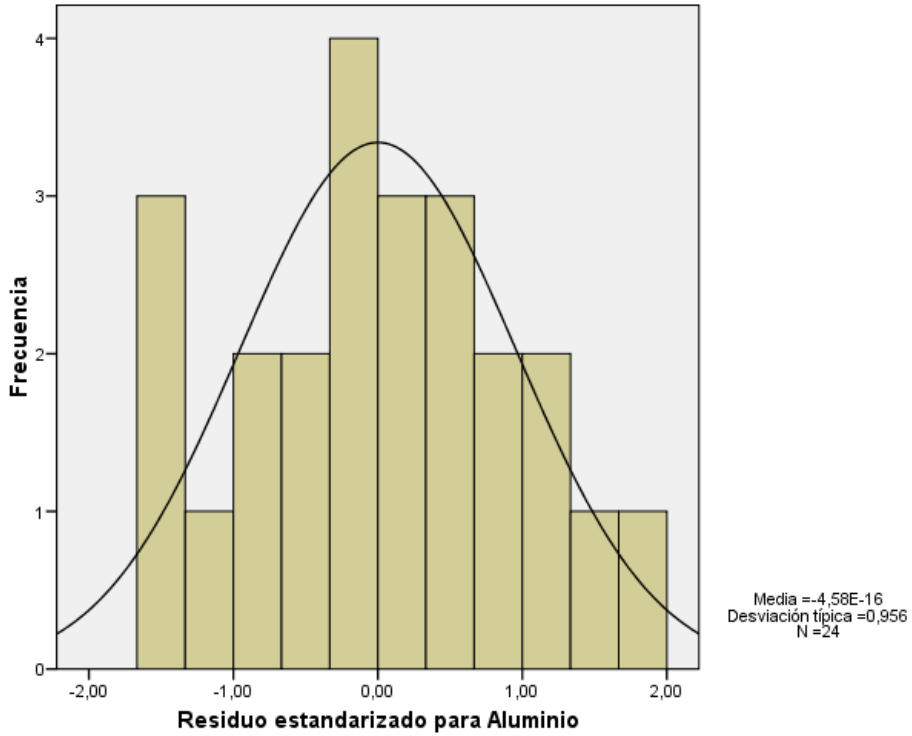
	Datos	
L	**I**	**A**
14,4	18,3	17,7
13,8	15,8	18,3
14,6	18	16,7
11,5	18	14,8
13,8	20,8	19,1
10,9	.	.
10,1	.	.
11,6	.	.
11,1	.	.
13,4	.	.
12,4	.	.
13,1	.	.
12,7	.	.
12,5	.	.

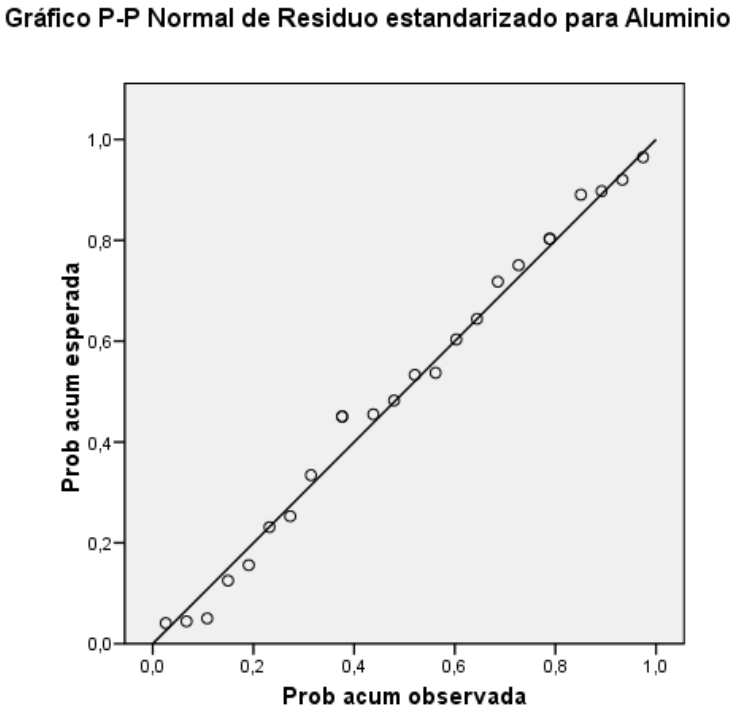


Aluminio

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
L	14	12,5643	1,37707	,36804	11,7692	13,3594	10,10	14,60
I	5	18,1800	1,77539	,79398	15,9756	20,3844	15,80	20,80
A	5	17,3200	1,65892	,74189	15,2602	19,3798	14,80	19,10
Total	24	14,7250	2,99989	,61235	13,4583	15,9917	10,10	20,80

Normalidad e igualdad de varianzas





Prueba de homogeneidad de varianzas

Aluminio

Estadístico de Levene	gl1	gl2	Sig.
,051	2	21	,950

ANOVA

Aluminio

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	158,717	2	79,358	34,526	,000
Intra-grupos	48,268	21	2,298		
Total	206,985	23			

Comparaciones múltiples

Variable dependiente: Aluminio

Bonferroni

(I) Localidad	(J) Localidad	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
L	I	-5,61571*	,78986	,000	-7,6704	-3,5610
	A	-4,75571*	,78986	,000	-6,8104	-2,7010
I	L	5,61571*	,78986	,000	3,5610	7,6704
	A	,86000	,95885	1,000	-1,6343	3,3543
A	L	4,75571*	,78986	,000	2,7010	6,8104
	I	-,86000	,95885	1,000	-3,3543	1,6343

*. La diferencia entre las medias es significativa al nivel .05.

Aceptamos la diferencia, en óxido de aluminio, de la localidad L con A e I

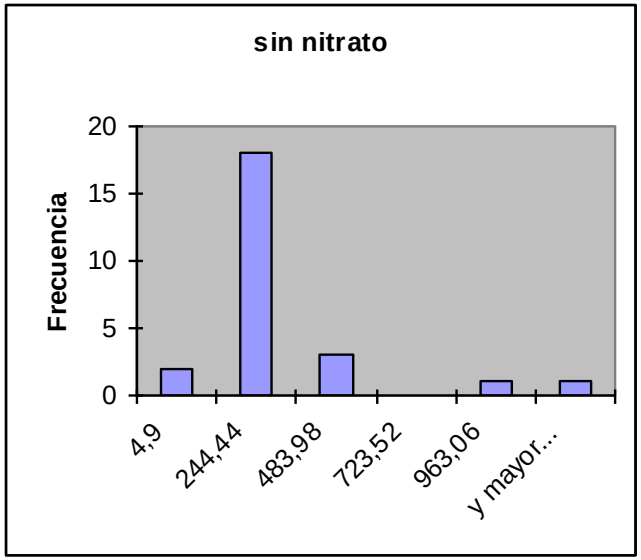
Ejemplo* 5

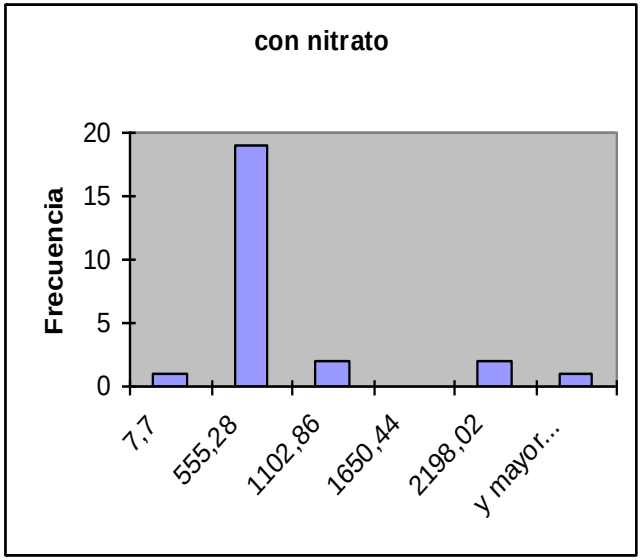
ANOVA con Excel

Se seleccionaron, al azar, 50 nubes.
De ellas, al azar, se sembraron 25 con Nitrato de Plata.
Se midió a continuación la cantidad de lluvia caída de cada una (en pies por acre).
El propósito del experimento era determinar si el sembrado de nitrato de plata incrementa la lluvia.

Reference: Chambers, Cleveland, Kleiner, and Tukey. (1983). Graphical Methods for Data Analysis. Wadsworth International Group, Belmont, CA, 351. Original Source: Simpson, Alsen, and Eden. (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. Technometrics 17, 161-166.
Education Queensland

*** Todos los ejemplos y sus resultados deben discutirse**

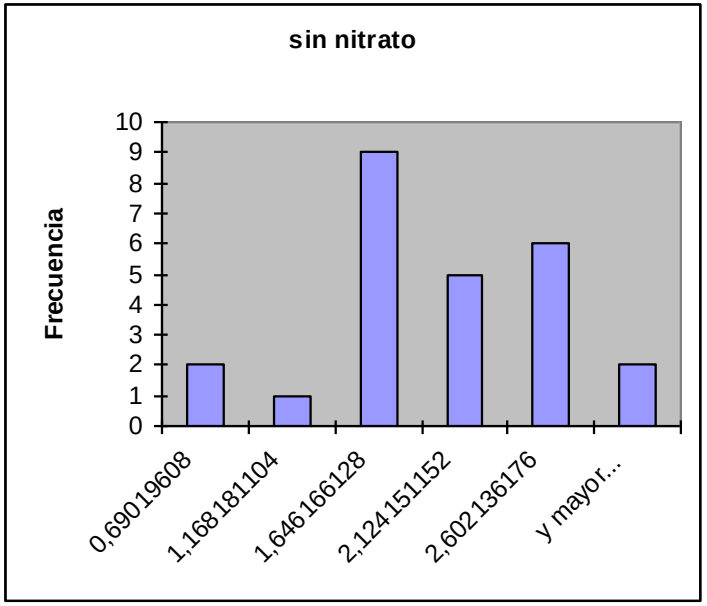


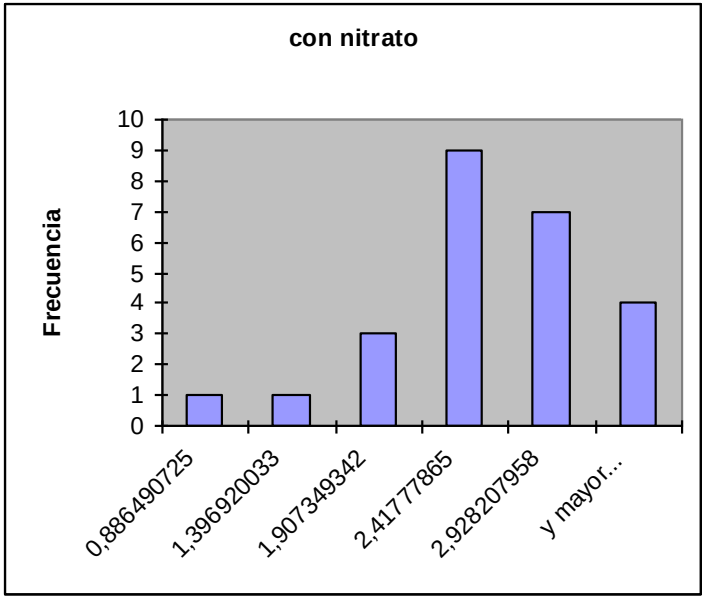


	sin nitrato	*con nitrato*
Media	171,13	459,50
Error típico	56,42	131,58
Mediana	47,30	242,50
Desviación estándar	282,12	657,92
Varianza de la muestra	79591,66	432861,91
Curtosis	7,82	5,74
Coeficiente de asimetría	2,74	2,39
Mínimo	4,90	7,70
Máximo	1202,60	2745,60
Suma	4278,30	11487,50
Cuenta	25	25

¿son aceptables la normalidad y la igualdad de varianzas?

Tomando logaritmos de los datos



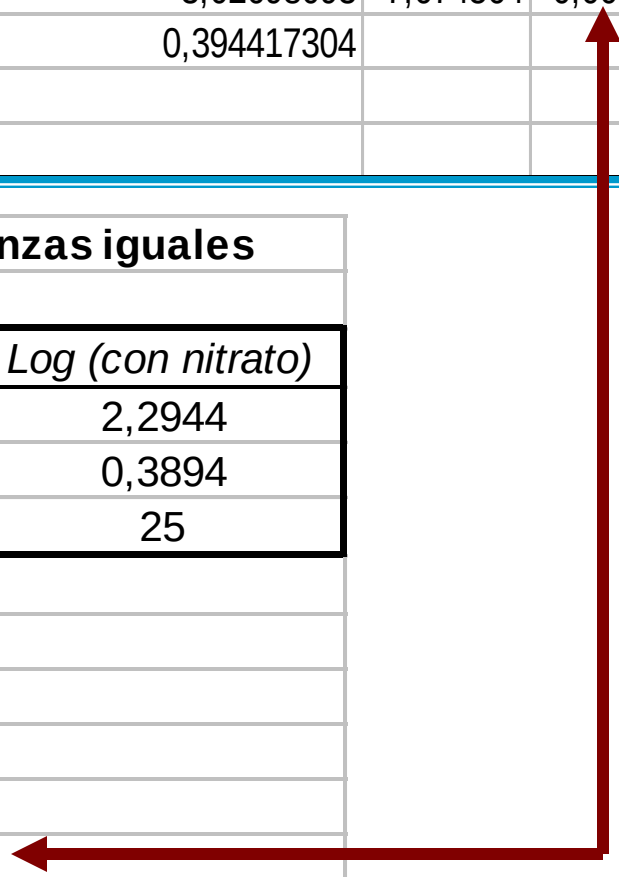


	Log (sin nitrato)	Log (con nitrato)
Media	1,802	2,294
Error típico	0,126	0,125
Mediana	1,675	2,385
Desviación estándar	0,632	0,624
Varianza de la muestra	0,399	0,389
Curtosis	-0,433	0,027
Coeficiente de asimetría	0,230	-0,297
Mínimo	0,690	0,886
Máximo	3,080	3,439
Suma	45,058	57,361
Cuenta	25	25

Ahora parece más aceptable...

ANÁLISIS DE VARIANZA					
fuente de variación	*Suma de cuadrados*	*g.l.*	*Promedio de los cuadrados*	*F*	*p-valor*
Entre grupos	3,02698093	1	3,02698093	7,674564	0,007942
Dentro de los grupos	18,93203057	48	0,394417304		
Total	21,9590115	49			

Prueba t para dos muestras suponiendo varianzas iguales		
	Log (sin nitrato)	*Log (con nitrato)*
Media	1,8023	2,2944
Varianza	0,3995	0,3894
Observaciones	25	25
Varianza agrupada	0,394417	
Grados de libertad	48	
Estadístico t	-2,770300	
P(T<=t) una cola	0,003971	
P(T<=t) dos colas	0,007942	



ANOVA con I = 2 es matemáticamente equivalente al contraste de la *t* de Student para la igualdad de medias con varianzas iguales

Ejemplo* 6

ANOVA con SPSS

100 pacientes con un mismo nivel de depresión diagnosticada se sometieron a un tratamiento con un nuevo fármaco.

Se clasificaron, al azar en 5 grupos de 20 pacientes a los que se les administró diferentes dosis del fármaco (0, 10, 20, 30 y 40 mgr.)

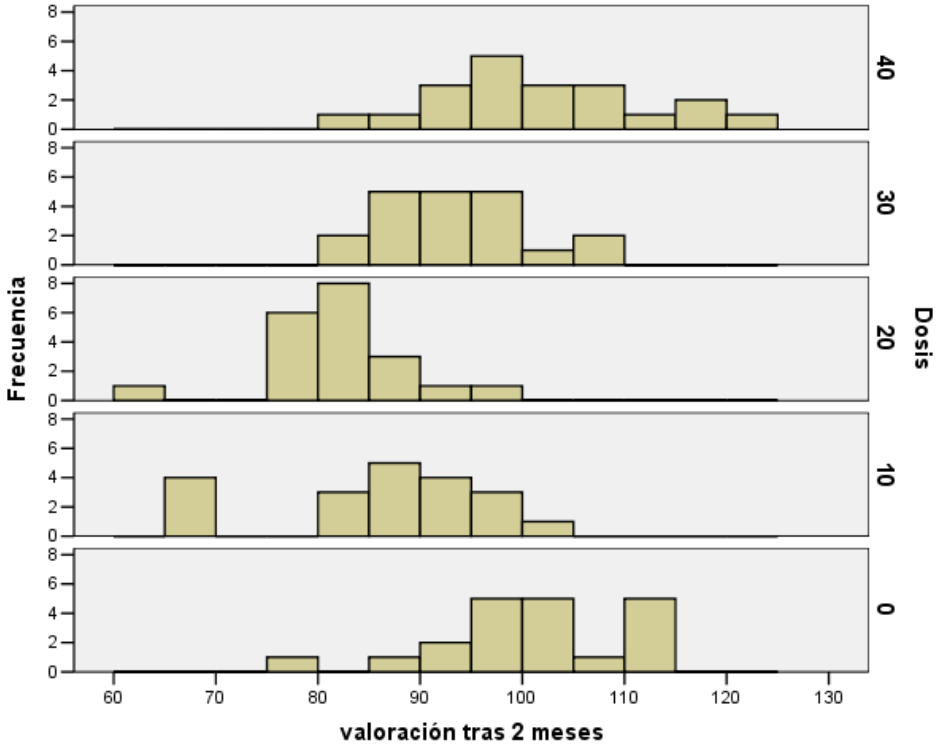
Al cabo de 2 meses de tratamiento se evaluó la situación de la enfermedad.

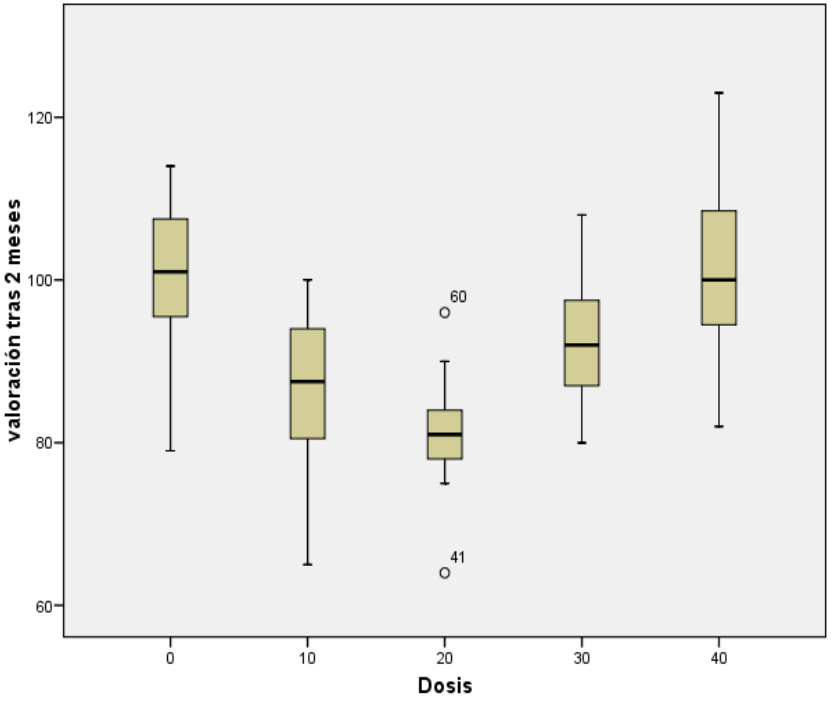
*** Todos los ejemplos y sus resultados deben discutirse**

Análisis descriptivo

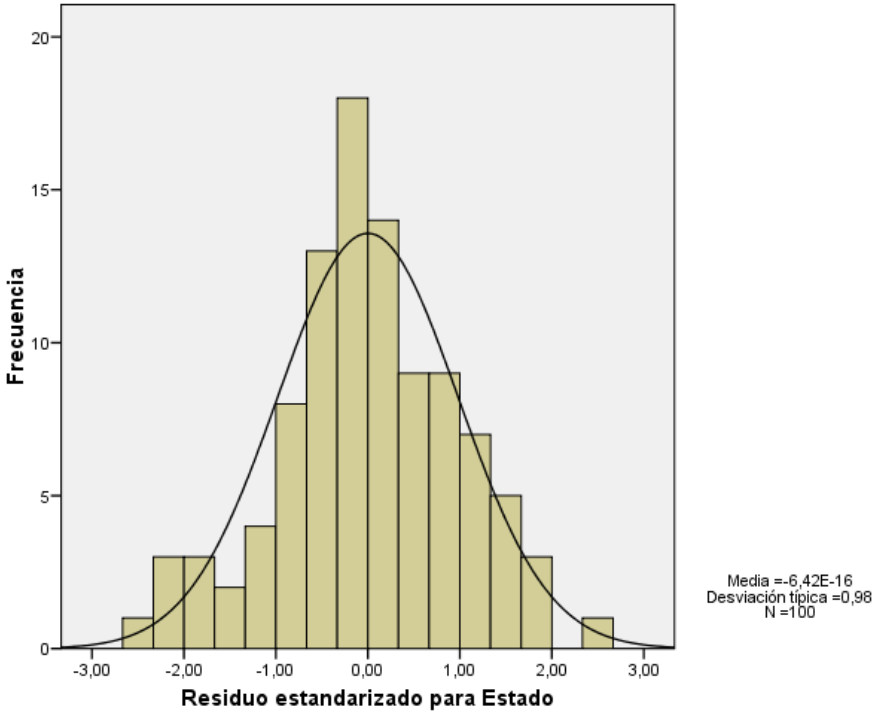
valoración tras 2 meses

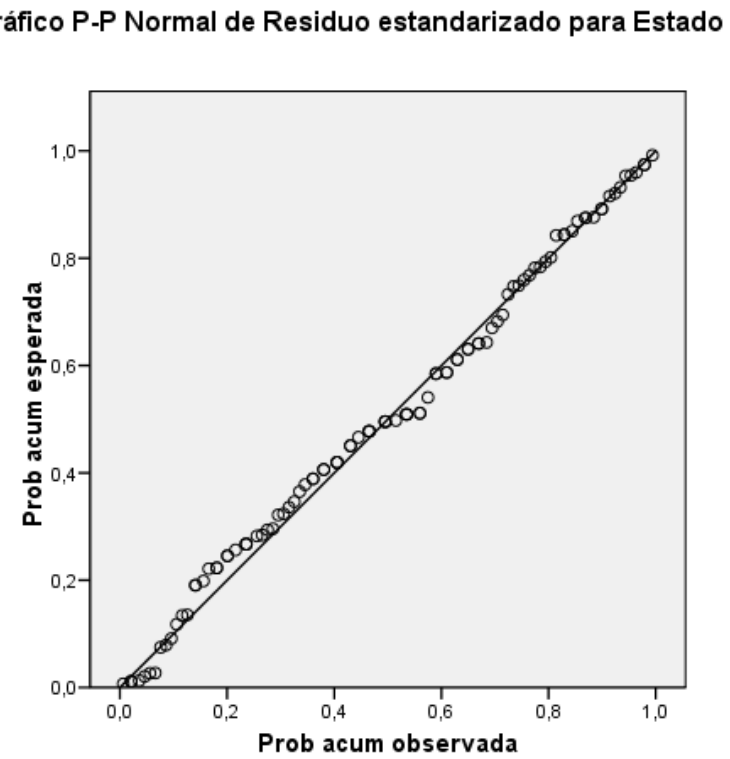
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
0	20	100,80	8,817	1,972	96,67	104,93	79	114
10	20	85,05	11,009	2,462	79,90	90,20	65	100
20	20	81,10	6,601	1,476	78,01	84,19	64	96
30	20	92,50	7,244	1,620	89,11	95,89	80	108
40	20	101,75	10,657	2,383	96,76	106,74	82	123
Total	100	92,24	12,125	1,212	89,83	94,65	64	123





Análisis de los residuos (Normalidad)





Prueba de homogeneidad de varianzas

valoración tras 2 meses

Estadístico de Levene	gl1	gl2	Sig.
2,042	4	95	,095

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_1: al menos una de las dosis es, en media, diferente

ANOVA

valoración tras 2 meses

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	6791,540	4	1697,885	20,779	,000
Intra-grupos	7762,700	95	81,713		
Total	14554,240	99			

Comparaciones múltiples

Variable dependiente: valoración tras 2 meses

	(I) Dosis	(J) Dosis	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
Bonferroni	0	10	15,750*	2,859	,000	7,53	23,97
		20	19,700*	2,859	,000	11,48	27,92
		30	8,300*	2,859	,046	,08	16,52
		40	-,950	2,859	1,000	-9,17	7,27
	10	0	-15,750*	2,859	,000	-23,97	-7,53
		20	3,950	2,859	1,000	-4,27	12,17
		30	-7,450	2,859	,106	-15,67	,77
		40	-16,700*	2,859	,000	-24,92	-8,48
	20	0	-19,700*	2,859	,000	-27,92	-11,48
		10	-3,950	2,859	1,000	-12,17	4,27
		30	-11,400*	2,859	,001	-19,62	-3,18
		40	-20,650*	2,859	,000	-28,87	-12,43
	30	0	-8,300*	2,859	,046	-16,52	-,08
		10	7,450	2,859	,106	-,77	15,67
		20	11,400*	2,859	,001	3,18	19,62
		40	-9,250*	2,859	,017	-17,47	-1,03
	40	0	,950	2,859	1,000	-7,27	9,17
		10	16,700*	2,859	,000	8,48	24,92
		20	20,650*	2,859	,000	12,43	28,87
		30	9,250*	2,859	,017	1,03	17,47
t de Dunnett (bilateral)[a]	10	0	-15,750*	2,859	,000	-22,85	-8,65
	20	0	-19,700*	2,859	,000	-26,80	-12,60
	30	0	-8,300*	2,859	,016	-15,40	-1,20
	40	0	,950	2,859	,992	-6,15	8,05

*. La diferencia entre las medias es significativa al nivel .05.

a. Las pruebas t de Dunnett tratan un grupo como control y lo comparan con todos los demás grupos.