Introducción al análisis estadístico de datos: conceptos básicos (Parte II)

Iniciativas de Investigación y Actividad Creativa Subgraduadas (iINAS)

6 de marzo de 2015

Marta Álvarez, Ph.D.
marta.alvarez1@upr.edu
Instituto de Estadística y Sistemas Computadorizados de Información
Facultad de Administración de Empresas, UPR Río Piedras

Como obtener el programa estadístico SPSS (Statistical Package for the Social Sciences) en la UPRRP

- En DTAA

  Por teléfono:

  787.764.0000 exts. 80400, 83822

  Por Internet:

  http://helpdesk.uprrp.edu:9675/portal/

SPSS

- SPSS tiene dos pantallas: una para los datos y otra para los resultados ("output").

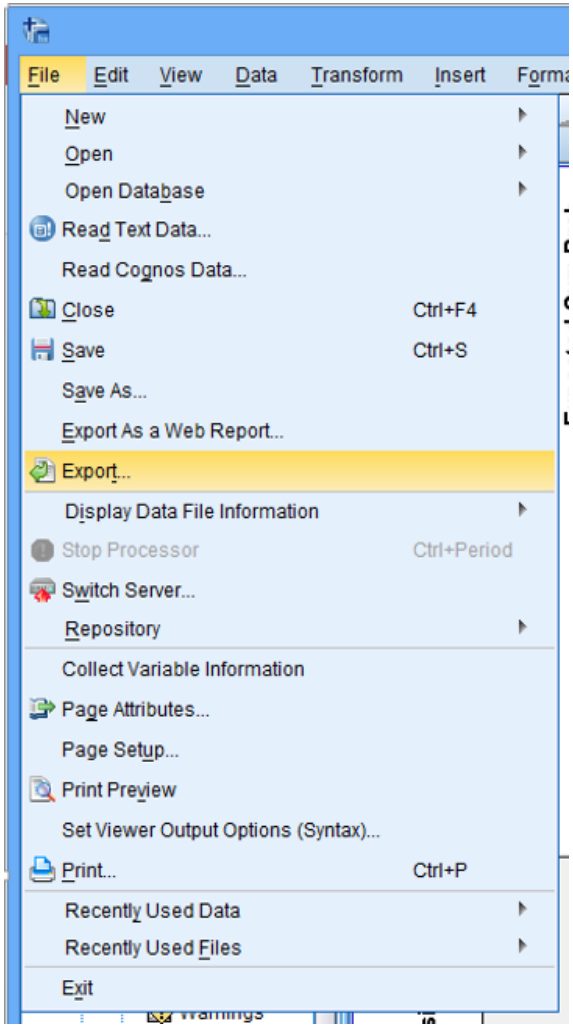
- Es importante no cerrar la pantalla de los resultados sin guardarla ("save").

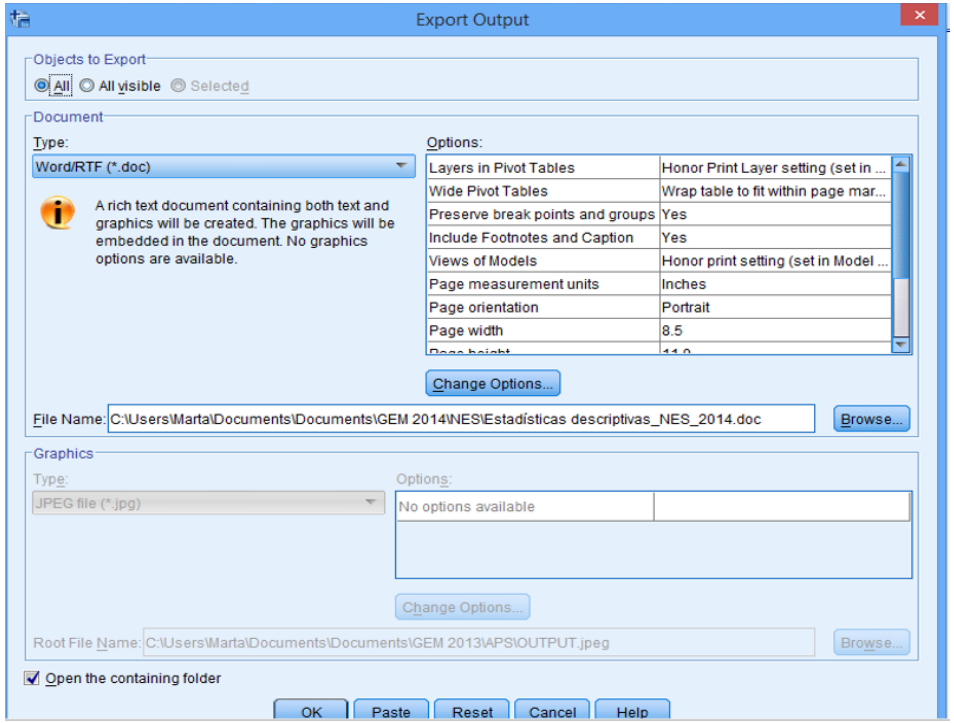
- Puede exportar los resultados a un documento en Word desde la pantalla de resultados, utilizando "Export".

Para exportar los resultados ("output") a un documento en Word

- Estar en la ventana

del "Output". Ir a >File >Export





Temario

- Fundamentos de la inferencia estadística
  - Intervalos de confiabilidad
  - Pruebas de hipótesis
- Inferencias para la media de una y dos poblaciones
- Inferencias para la correlación
- Inferencias para tablas de dos factores
- Regresión lineal simple y múltiple

Datos utilizados: (1) Global Entrepreneurship Monitor (GEM) 2013

- GEM es un estudio académico longitudinal que provee información y medidas armonizadas de:
  - actitudes de la población hacia la creación de empresas y el empresario
  - actividades y características de los individuos que participan en las distintas fases de la creación de empresas
  - Factores del entorno que promueven y obstaculizan el empresarismo
- Esta información permite hacer comparaciones entre países, regiones y niveles de desarrollo económico.
- En el 2013 participaron 70 países, incluyendo Puerto Rico.
- Equipo de Puerto Rico en GEM 2013:
  - Investigadoras: Dra. Marinés Aponte, Dra. Marta Álvarez
  - Gerente de proyecto: Prof. Aida Lozada

(2) Archivo de datos creado para el taller

- Buscar archivo en desktop: “Datos taller iInas_feb 2015.xls”

- Fuentes:
  - Banco Mundial, 2012
  - 2014 Worldwide VAT, GST and Sales Tax Guide
    - EY, http://www.ey.com/GL/en/Home

Estadística inferencial

- Procedimientos para generalizar a la población los hallazgos obtenidos en los datos de una muestra aleatoria de esa población.

- Utilizamos las estadísticas muestrales para hacer inferencias sobre los parámetros de la población.

Estadística paramétrica vs no-paramétrica

- Estadística paramétrica
  - Métodos estadísticos basados en el supuesto de que la distribución de la población estudiada se comporta como una de las distribuciones de probabilidad o funciones de densidad conocidas.

- Estadística no-paramétrica
  - Métodos estadísticos donde no se asume ninguna distribución para la población.

- Dos tipos principales de inferencias:
  - Estimación
  - Prueba de hipótesis

- Hay dos enfoques:

  1. Clásico o frecuentista

     - Los parámetros se consideran valores fijos desconocidos.

  2. Bayesiano

     - Los parámetros se consideran variables y se describen probabilísticamente.

Estimación

- Se utilizan las estadísticas muestrales como estimadores de los parámetros poblacionales.

- Dos tipos de estimación:
  - Estimación puntual
    - El parámetro es estimado por un solo valor.
  - Estimación por intervalo
    - El estimador es un intervalo de posibles valores de un parámetro; se le llama intervalo de confiabilidad y se basa en la teoría de probabilidad antes de tomar la muestra.

Estimación por intervalo

- Un intervalo de confiabilidad para un parámetro tiene la siguiente forma general:

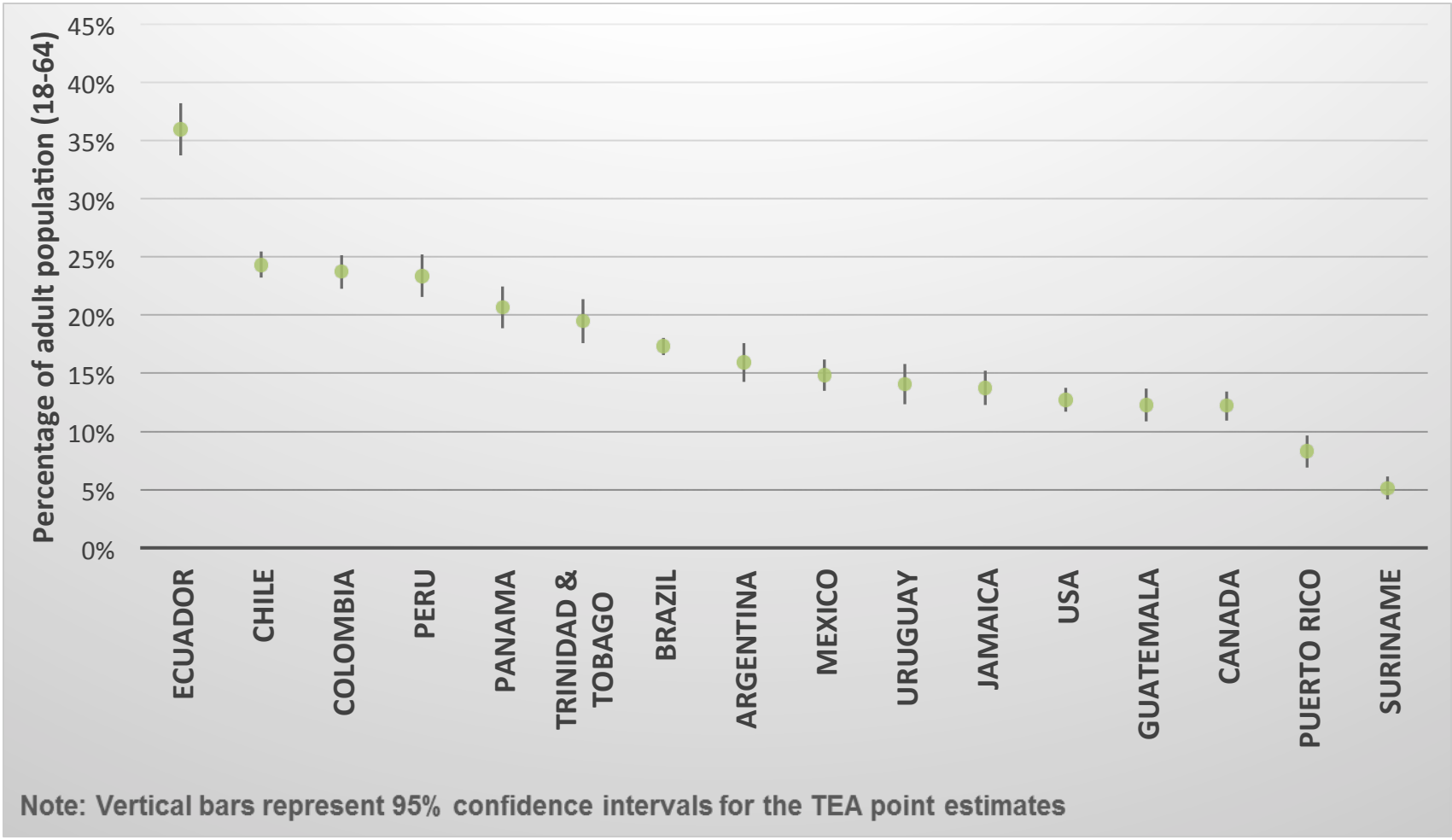
  estimado ± margen de error

- El margen de error nos demuestra la precisión del estimado, basado en la variabilidad del mismo y en el nivel de confiabilidad.

- El nivel de confiabilidad (1-α) es la probabilidad de que el intervalo incluya el verdadero valor del parámetro cuando se toman muestras repetidas del mismo tamaño.

- Los valores de α más utilizados son 0.05, 0.01 y 0.10.

Índice de Actividad Empresarial Temprana (TEA) para los países del Caribe, América Latina y América del Norte participantes en GEM 2013



Inferencias sobre la media poblacional de una variable cuantitativa

Estadística paramétrica: Distribución de muestreo

- La distribución de muestreo de un estimador es la distribución de probabilidad del estimador, si se tomaran muestras repetidas del mismo tamaño. La misma es necesaria para las metodologías paramétricas.

- Si la población tiene una distribución Normal $X \sim N(\mu, \sigma^2)$, entonces la media muestral también se distribuye Normal, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

- Teorema del Límite Central
  - A medida que el tamaño de la muestra aumenta, la distribución de muestreo de la media muestral se aproxima a la distribución Normal.

Intervalo de (1-α)100% de confiabilidad para μ (σ desconocida)

- Caso en el que no conocemos la desviación estándar de la población, σ y la estimamos con la desviación estándar de la muestra, s.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \Rightarrow \quad t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Donde t_{n-1} es la distribución t de estudiante con n-1 grados de libertad.

Intervalo de (1-α)100% de confiabilidad para μ (σ desconocida)

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$$

$$\Rightarrow \quad \left(\bar{x} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}} \right)$$

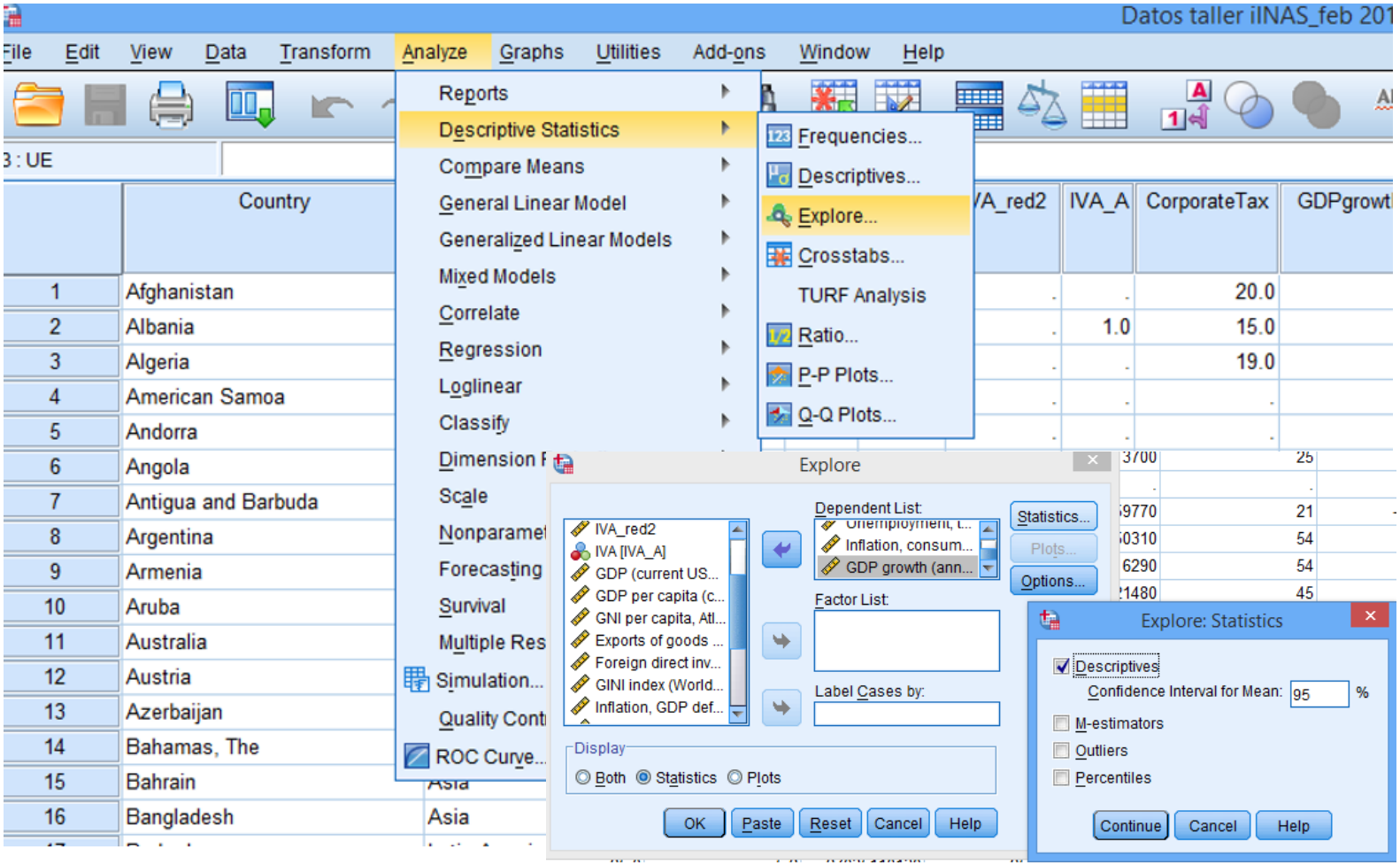
donde $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ es el error estándar de $\bar{x}$.

Margen de error

- El margen de error de un intervalo de confiabilidad se reduce si:
  - se reduce el nivel de confiabilidad
  - la desviación estándar o el error estándar disminuye, ó
  - el tamaño de muestra aumenta.

SPSS

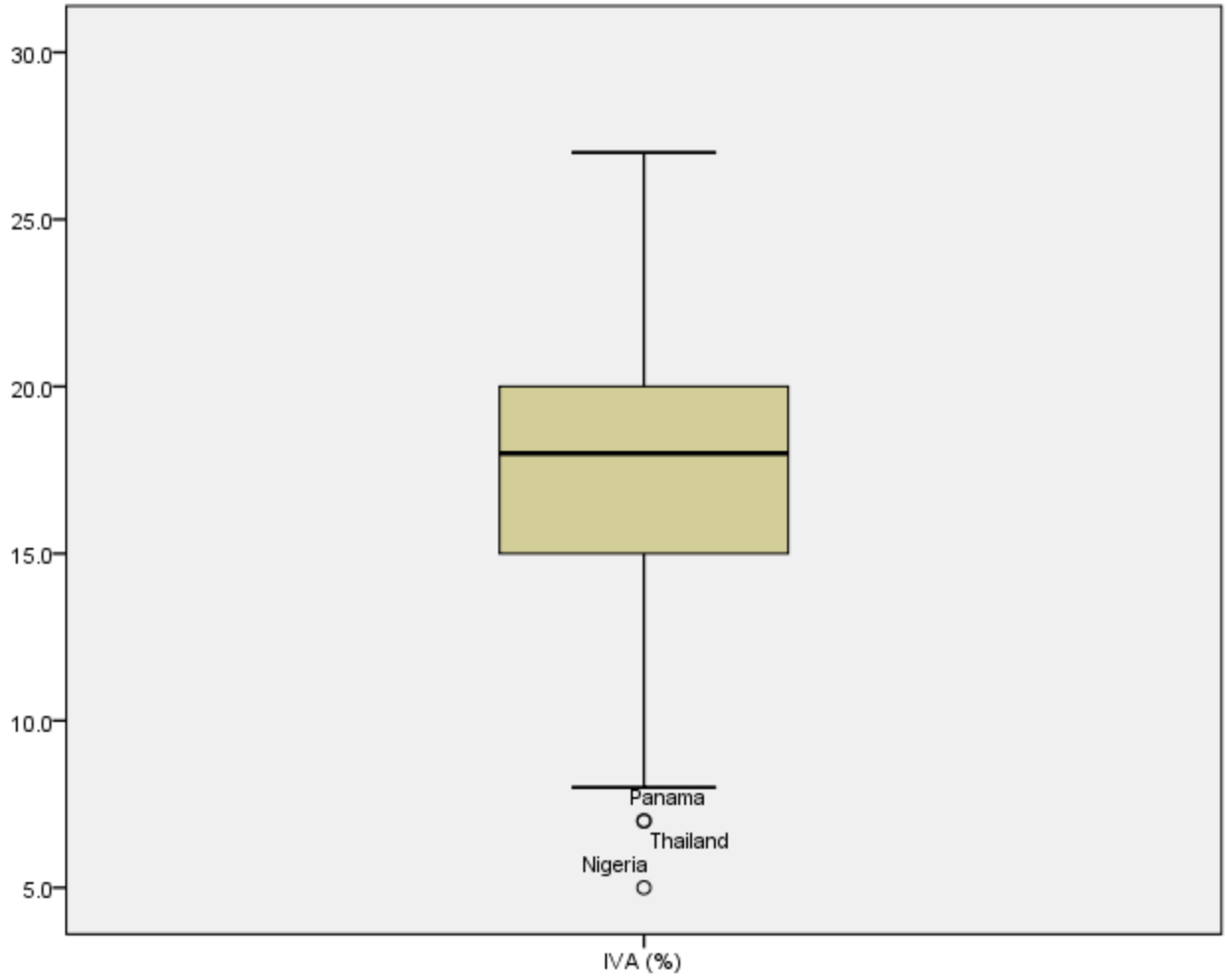
>Analyze >Descriptive statistics >Explore



Intervalo de 95% de confiabilidad para la variable IVA

Descriptives

			Statistic	Std. Error
IVA (%)	Mean		17.538	.4848
	95% Confidence Interval for Mean	Lower Bound	16.575	
		Upper Bound	18.502	
	5% Trimmed Mean		17.671	
	Median		18.000	
	Variance		21.385	
	Std. Deviation		4.6244	
	Minimum		5.0	
	Maximum		27.0	
	Range		22.0	
	Interquartile Range		5.0	
	Skewness		-.408	.253
	Kurtosis		-.054	.500


30.0
25.0
20.0
15.0
10.0
5.0
Panama
Thailand
Nigeria
IVA (%)

Pruebas de hipótesis: idea básica

- Pruebas paramétricas:

  Cuando la hipótesis nula H_0 es cierta, se asume que la población/es tiene cierta distribución de probabilidad. Si el valor de la estadística de prueba tiene una probabilidad baja de ser observado bajo la distribución asumida, entonces decimos que tenemos un resultado significativo de la prueba (o se rechaza la hipótesis nula y se acepta la alterna).

Prueba de hipótesis

- Elementos de una prueba de hipótesis:
  - Hipótesis
    - Hipótesis nula: H_0: $\theta = \theta_0$
    - Hipótesis alterna: H_1: $\theta \neq \theta_0$
      $\theta > \theta_0$
      $\theta < \theta_0$

      La hipótesis alterna representa la pregunta del investigador/a.
  - Estadística de prueba
  - Regla de decisión
    - Valor p ("P-value")
    - Región de rechazo

- La estadística de prueba calculada de los datos de la muestra es comparada con la distribución teórica de la misma.

- Las posibles decisiones en una prueba estadística son:
  - Rechazar la H_0 y aceptar la hipótesis alterna H_1
  - No rechazar la H_0 y concluir que los datos de la muestra no proven suficiente evidencia estadística para apoyar H_1

Posibles errores al tomar una decisión

- Dos tipos de error:
  - Error Tipo I
    - Ocurre cuando se rechaza la hipótesis nula y ésta es cierta
    - P[Error Tipo I] = α
  - Error Tipo II
    - Ocurre cuando no se rechaza la hipótesis nula y ésta es falsa
    - P[Error Tipo II] = β
- La potencia de una prueba (1- β) es la probabilidad de rechazar la hipótesis nula cuando ésta es falsa.
- A α se le conoce como el nivel de significancia de una prueba.

- Regla de decisión:
  - p-value de la prueba
    - Se define como la probabilidad, asumiendo que la hipótesis nula es cierta, de que la estadística de prueba tome un valor tan o más extremo del observado.
    - Es la probabilidad observada de Error Tipo I: probabilidad de rechazar la hipótesis nula cuando ésta es cierta
    - Mientras más pequeño es el P-value, más confiabilidad tenemos de que la hipótesis nula es falsa.
  - Al p-value se le llama la probabilidad de error Tipo I observada, mientras que α es la probabilidad de error Tipo I tolerada.
  - Rechazamos la hipótesis nula y aceptamos la alterna a un nivel de significancia α si el P-value $\leq \alpha$.

Pruebas sobre la media poblacional μ

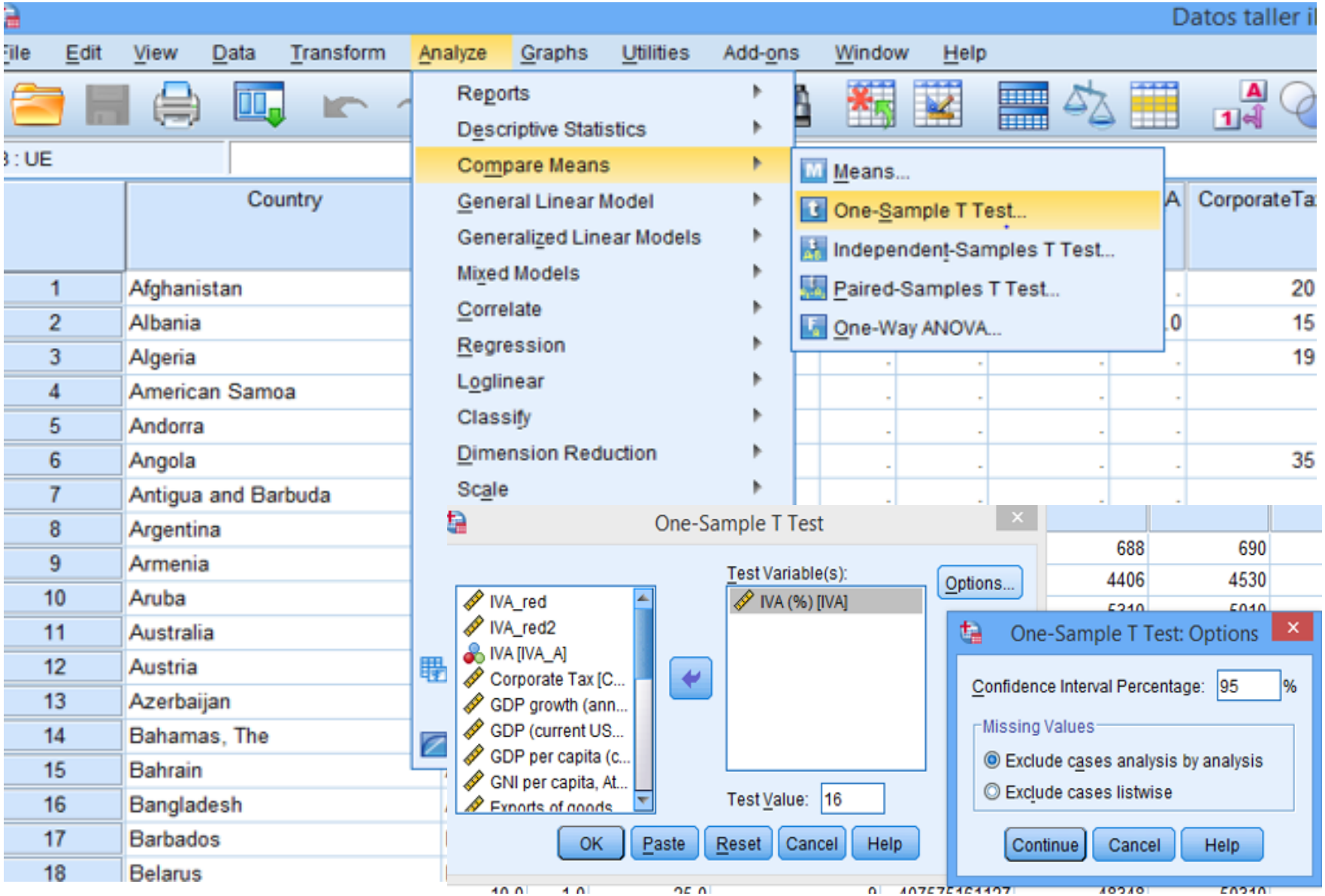
- Prueba de t

  Asumimos que la población que se está estudiando tiene una distribución Normal con media μ y varianza σ^2 (desconocida). Entonces utilizamos la estadística de prueba t.

- Si la distribución no es Normal, esta metodología es bastante robusta para muestras grandes.

$$H_0 : \mu = \mu_0$$

Hipótesis alterna	Estadística de prueba	P-value
$H_1 : \mu \neq \mu_0$		$2P(t_{n-1} \geq \lvert t^* \rvert)$
$H_1 : \mu > \mu_0$	$t^* = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$	$P(t_{n-1} \geq t^*)$
$H_1 : \mu < \mu_0$		$P(t_{n-1} \leq t^*)$


Datos taller il
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
Country
CorporateTa
1 Afghanistan 20
2 Albania 15
3 Algeria 19
4 American Samoa
5 Andorra
6 Angola 35
7 Antigua and Barbuda
8 Argentina
9 Armenia
10 Aruba
11 Australia
12 Austria
13 Azerbaijan
14 Bahamas, The
15 Bahrain
16 Bangladesh
17 Barbados
18 Belarus
688 690
4406 4530
One-Sample T Test
IVA_red
IVA_red2
IVA [IVA_A]
Corporate Tax [C...
GDP growth (ann...
GDP (current US...
GDP per capita (c...
GNI per capita, At...
Exports of goods
Test Variable(s):
IVA (%) [IVA]
Options...
Test Value: 16
OK Paste Reset Cancel Help
One-Sample T Test: Options
Confidence Interval Percentage: 95 %
Missing Values
Exclude cases analysis by analysis
Exclude cases listwise
Continue Cancel Help

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
IVA (%)	91	17.538	4.6244	.4848

One-Sample Test

	Test Value = 16					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
IVA (%)	3.174	90	.002	1.5385	.575	2.502

Estadística de prueba

P-value

$$H_0 : \mu = 16$$
$$H_1 : \mu \neq 16$$

P-value significativo; se rechaza la hipótesis nula y se concluye que la media poblacional del IVA es diferente a 16%.

Comparación de dos medias poblacionales

- Suponga que queremos comparar las medias de dos grupos, donde los grupos se asumen que son muestras aleatorias de diferentes poblaciones y las muestras son independientes entre sí.

Parámetro: $\mu_1 - \mu_2$

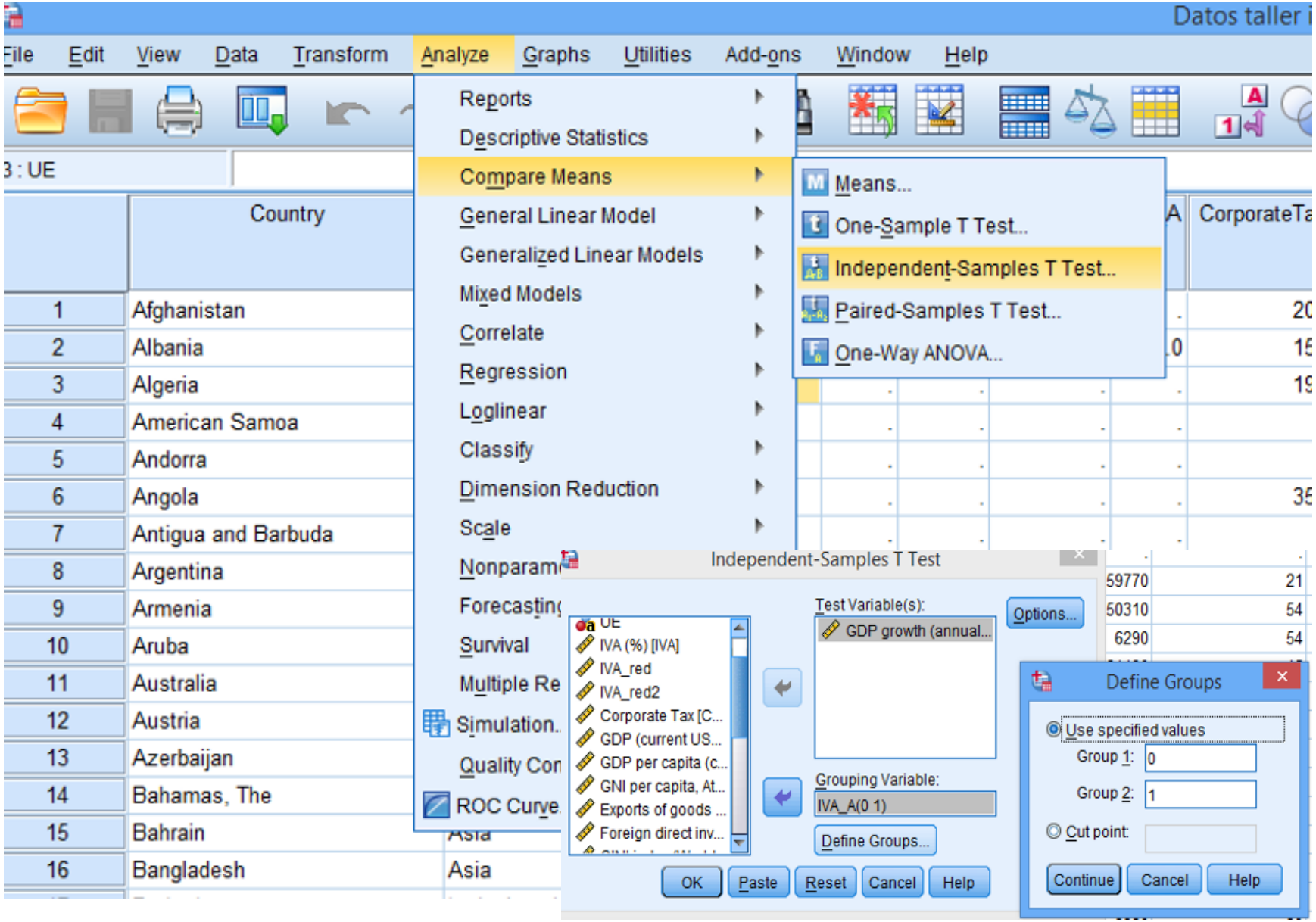
Estimador: $\bar{x}_1 - \bar{x}_2$

Error estándar del estimador: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$$H_0 : \mu_1 - \mu_2 = 0 \quad (\mu_1 = \mu_2)$$

Hipótesis alterna	Estadística de prueba	P-value
$H_1 : \mu_1 - \mu_2 \neq 0$	$t^* = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim t_{gl}$	$2P(t \geq \lvert t^* \rvert)$
$H_1 : \mu_1 - \mu_2 > 0$		$P(t \geq t^*)$
$H_1 : \mu_1 - \mu_2 < 0$		$P(t \leq t^*)$

gl=grados de libertad


Datos taller i
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
3 : UE
Country
1 Afghanistan
2 Albania
3 Algeria
4 American Samoa
5 Andorra
6 Angola
7 Antigua and Barbuda
8 Argentina
9 Armenia
10 Aruba
11 Australia
12 Austria
13 Azerbaijan
14 Bahamas, The
15 Bahrain
16 Bangladesh
Asia
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparam
Forecasting
Survival
Multiple Re
Simulation...
Quality Cor
ROC Curve...
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
CorporateTa
Independent-Samples T Test
UE
IVA (%) [IVA]
IVA_red
IVA_red2
Corporate Tax [C...
GDP (current US...
GDP per capita (c...
GNI per capita, At...
Exports of goods ...
Foreign direct inv...
Test Variable(s):
GDP growth (annual...
Options...
Grouping Variable:
IVA_A(0 1)
Define Groups...
OK Paste Reset Cancel Help
Define Groups
Use specified values
Group 1: 0
Group 2: 1
Cut point:
Continue Cancel Help

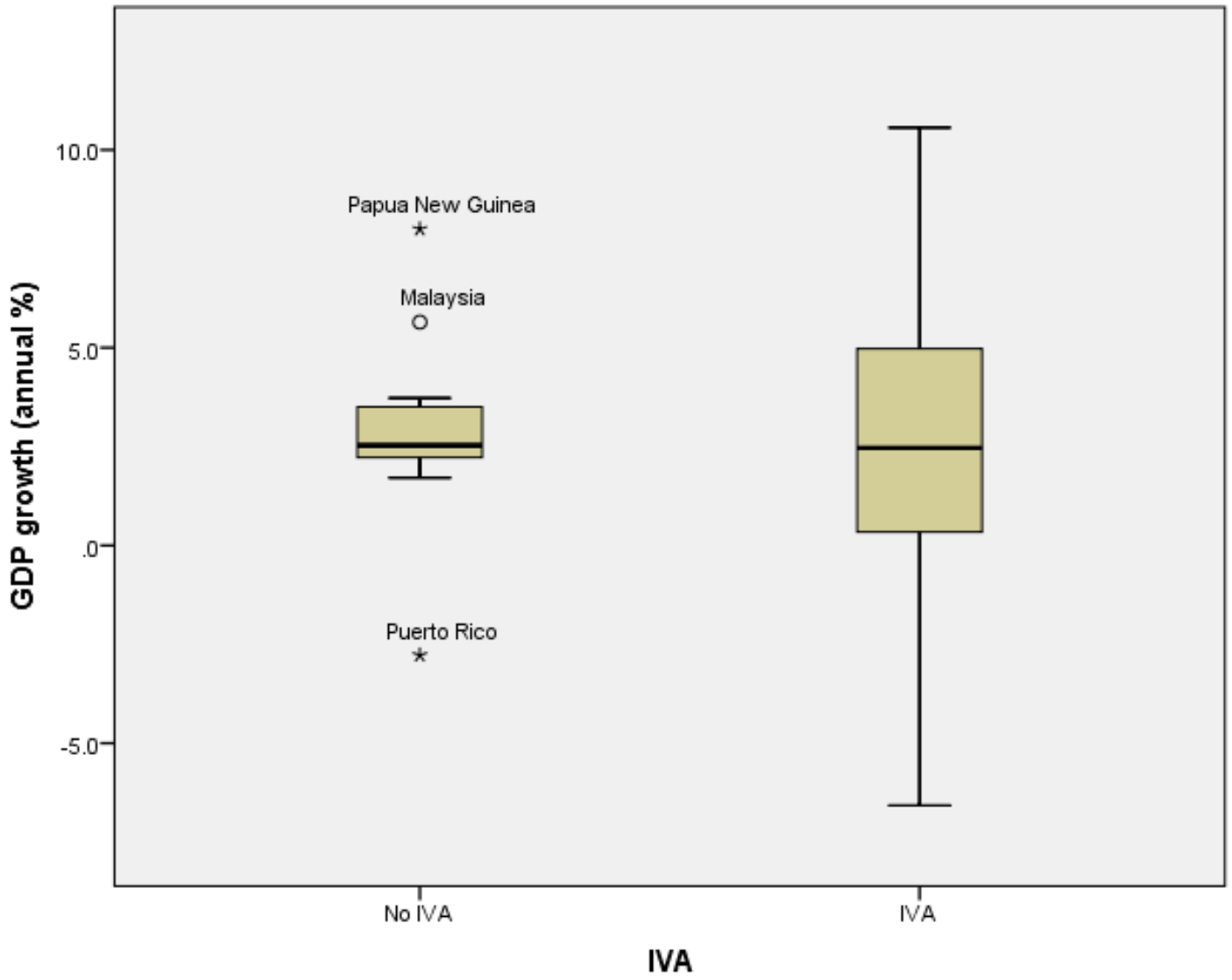
Group Statistics

	IVA	N	Mean	Std. Deviation	Std. Error Mean
GDP growth (annual %)	No IVA	13	2.831	2.4257	.6728
	IVA	91	2.609	3.2531	.3410

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
GDP growth (annual %)	Equal variances assumed	4.094	.046	.236	102	.814	.2220	.9390	-1.6405	2.0845
	Equal variances not assumed			.294	18.793	.772	.2220	.7543	-1.3578	1.8019

P-value no significativo


GDP growth (annual %)
10.0
5.0
.0
-5.0
Papua New Guinea
Malaysia
Puerto Rico
No IVA
IVA
IVA

Pruebas no-paramétricas

- Una muestra: “Wilcoxon signed-rank test”
- Muestras pareadas: aplique pruebas de una muestra a las diferencias entre pares
- Dos muestras independientes: “Wilcoxon rank-sum test”
- Varias muestras independientes: “Kruskall-Wallis test” (prueba paramétrica: ANOVA)

Análisis de correlación

- La correlación es una medida numérica que describe la fuerza y dirección de la relación lineal entre dos variables cuantitativas.
- La correlación de Pearson (o momento-producto) entre dos variables *X* y *Y* se define como:

Población:

$$\rho_{XY} = \frac{E[(X - E(X)(Y - E(Y)]}{\sigma_X \sigma_Y}$$

Estimador muestral:

$$r_{XY} = \frac{\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{s_X s_Y}$$

- La correlación es un número entre -1 y 1.
- El signo de *r* representa la dirección de la relación lineal y la magnitud de *r* representa la fuerza.
- A medida que el valor absoluto de *r* se acerque a 1, $|r| \to 1$, más fuerte es la relación lineal entre las dos variables. Si *r* está cerca de 0, no existe relación lineal entre las variables.

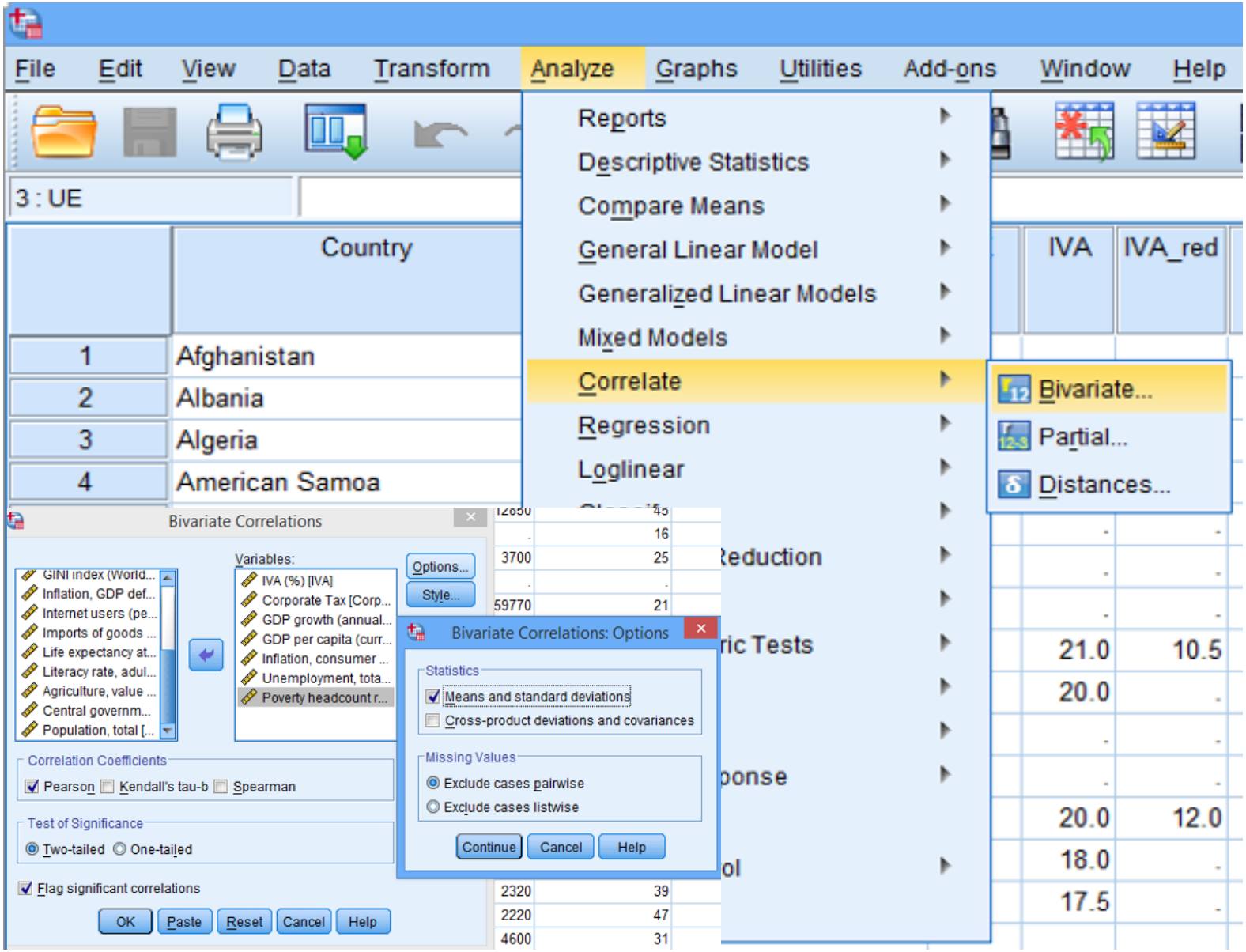
Prueba de hipótesis para la correlación

- Suponga que X y Y son variables aleatorias independientes con una distribución Normal.

- Hipótesis: $H_0 : \rho = 0$ $H_1 : \rho \neq 0$

$\rho > 0$

$\rho < 0$


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
3 : UE
Country
IVA
IVA_red
1 Afghanistan
2 Albania
3 Algeria
4 American Samoa
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Bivariate...
Partial...
Distances...
Bivariate Correlations
Variables:
GINI index (World...
Inflation, GDP def...
Internet users (pe...
Imports of goods ...
Life expectancy at...
Literacy rate, adul...
Agriculture, value ...
Central governm...
Population, total [...
IVA (%) [IVA]
Corporate Tax [Corp...
GDP growth (annual...
GDP per capita (curr...
Inflation, consumer ...
Unemployment, tota...
Poverty headcount r...
Options...
Style...
Correlation Coefficients
Pearson Kendall's tau-b Spearman
Test of Significance
Two-tailed One-tailed
Flag significant correlations
OK Paste Reset Cancel Help
Bivariate Correlations: Options
Statistics
Means and standard deviations
Cross-product deviations and covariances
Missing Values
Exclude cases pairwise
Exclude cases listwise
Continue Cancel Help

Correlations

		IVA (%)	Corporate Tax	GDP growth (annual %)	GDP per capita (current US$)	Inflation, consumer prices (annual %)	Unemployment, total (% of total labor force) (modeled ILO estimate)	Poverty headcount ratio at national poverty lines (% of population)
IVA (%)	Pearson Correlation	1	-.161	-.536**	.322**	-.114	.294**	-.232
	Sig. (2-tailed)		.143	.000	.002	.292	.005	.202
	N	91	84	90	90	88	88	32
Corporate Tax	Pearson Correlation	-.161	1	.011	-.075	.109	-.241**	.398*
	Sig. (2-tailed)	.143		.902	.410	.232	.008	.018
	N	84	130	123	123	122	121	35
GDP growth (annual %)	Pearson Correlation	-.536**	.011	1	-.099	.026	.010	.118
	Sig. (2-tailed)	.000	.902		.174	.732	.902	.450
	N	90	123	189	189	175	169	43
GDP per capita (current US$)	Pearson Correlation	.322**	-.075	-.099	1	-.259**	-.152*	-.410**
	Sig. (2-tailed)	.002	.410	.174		.001	.049	.006
	N	90	123	189	189	175	169	43
Inflation, consumer prices (annual %)	Pearson Correlation	-.114	.109	.026	-.259**	1	-.027	-.067
	Sig. (2-tailed)	.292	.232	.732	.001		.735	.670
	N	88	122	175	175	179	164	43
Unemployment, total (% of total labor force) (modeled ILO estimate)	Pearson Correlation	.294**	-.241**	.010	-.152*	-.027	1	-.076
	Sig. (2-tailed)	.005	.008	.902	.049	.735		.627
	N	88	121	169	169	164	174	43
Poverty headcount ratio at national poverty lines (% of population)	Pearson Correlation	-.232	.398*	.118	-.410**	-.067	-.076	1
	Sig. (2-tailed)	.202	.018	.450	.006	.670	.627	
	N	32	35	43	43	43	43	43

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Inferencias para tablas de dos factores o variables cualitativas

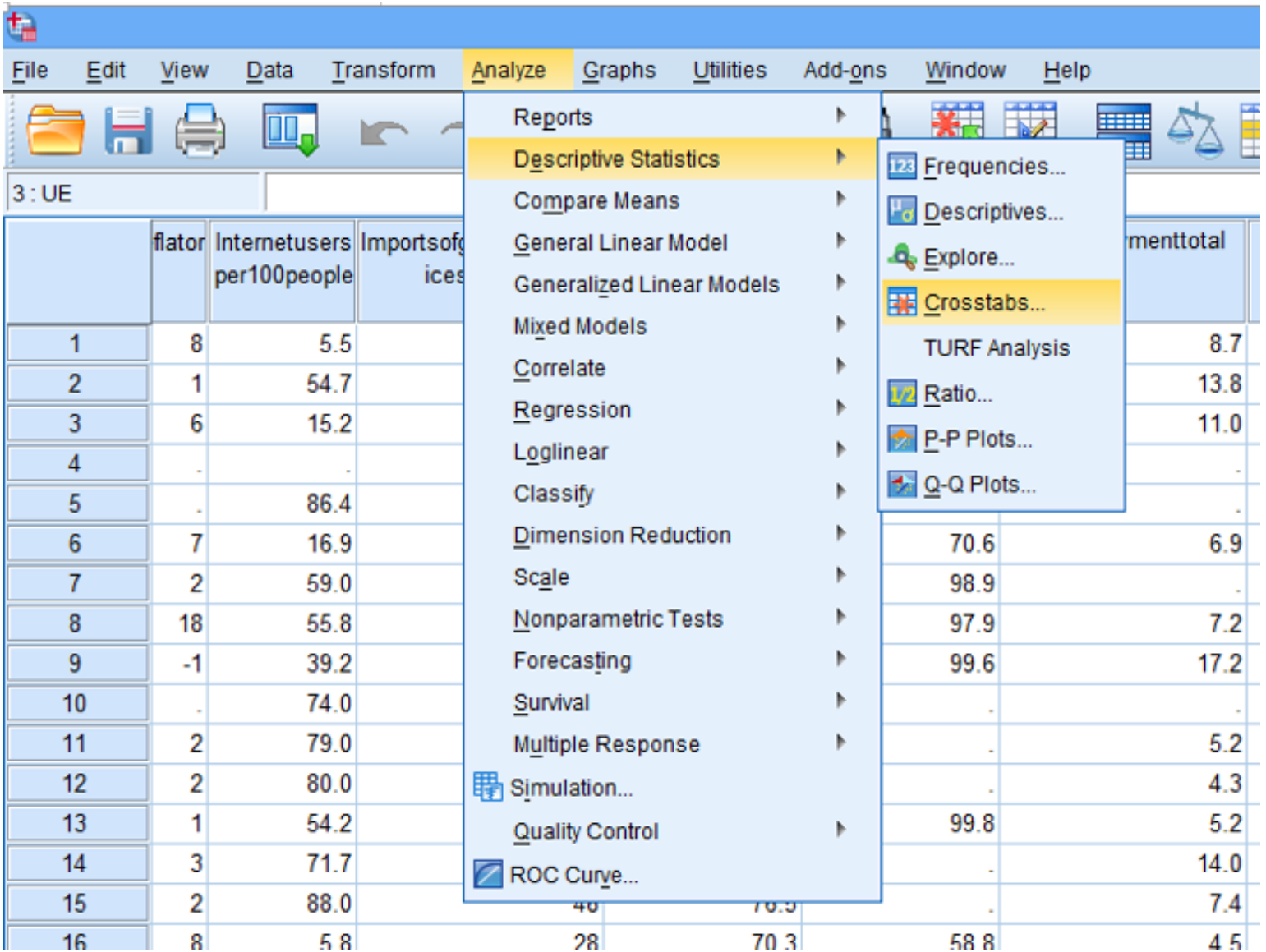
- Suponga que tenemos dos variables cualitativas (factores); los datos serían el número de observaciones (frecuencias) en cada combinación de las variables.
- "Two-way contingency table"
- Nos interesa investigar si existe asociación entre las dos variables.
- La estadística usualmente utilizada para el análisis de datos categóricos es la Ji-cuadrada ("Chi-square").

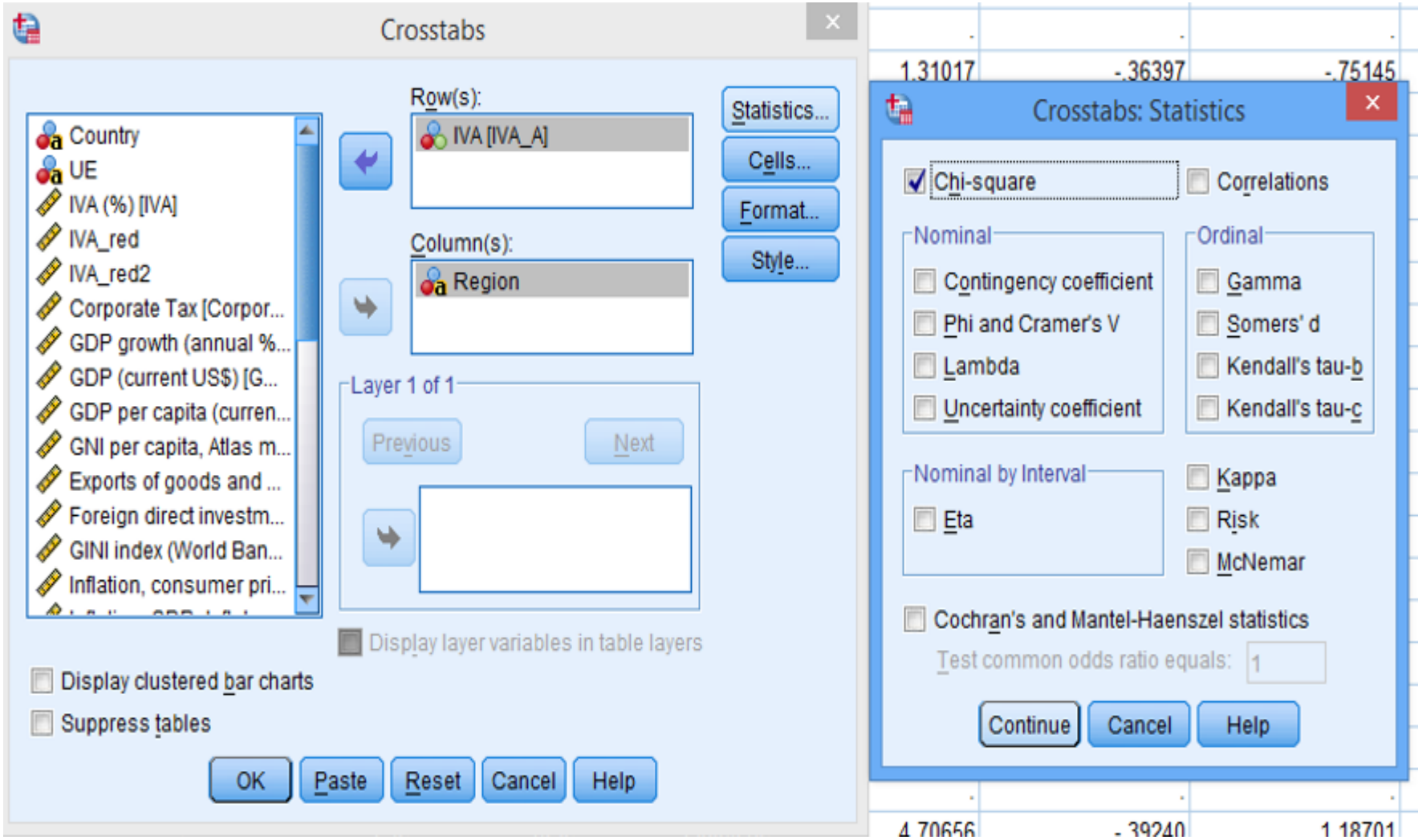
Prueba de asociación con la Ji-cuadrada ("Chi-square")

H_0: No existe asociación entre las dos variables (son independientes)

H_1: Existe asociación entre las dos variables (no son independientes)

- Nota: Una regla razonable para que la prueba anterior sea válida es que todas las frecuencias esperadas deben de ser por lo menos 2; o que no más del 20% de las categorías tengan frecuencias esperadas menor de 5.


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
3 : UE
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Simulation...
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
TURF Analysis
Ratio...
P-P Plots...
Q-Q Plots...
flator
Internetusers per100people
Importsofg ices
menttotal


Crosstabs
Country
UE
IVA (%) [IVA]
IVA_red
IVA_red2
Corporate Tax [Corpor...
GDP growth (annual %...
GDP (current US$) [G...
GDP per capita (curren...
GNI per capita, Atlas m...
Exports of goods and ...
Foreign direct investm...
GINI index (World Ban...
Inflation, consumer pri...
Row(s):
IVA [IVA_A]
Column(s):
Region
Layer 1 of 1
Previous
Next
Statistics...
Cells...
Format...
Style...
Display layer variables in table layers
Display clustered bar charts
Suppress tables
OK
Paste
Reset
Cancel
Help
1.31017
-.36397
-.75145
Crosstabs: Statistics
Chi-square
Correlations
Nominal
Contingency coefficient
Phi and Cramer's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals:
1
Continue
Cancel
Help
4.70656
-.39240
1.18701

IVA * Region Crosstabulation

			Region							Total
				Africa	Asia	Europe	Latin America and Caribbean	North America	Oceania	
IVA	No IVA	Count	0	1	6	0	5	2	3	17
		% within IVA	0.0%	5.9%	35.3%	0.0%	29.4%	11.8%	17.6%	100.0%
		% within Region	0.0%	6.3%	28.6%	0.0%	18.5%	100.0%	100.0%	15.6%
		% of Total	0.0%	0.9%	5.5%	0.0%	4.6%	1.8%	2.8%	15.6%
	IVA	Count	3	15	15	37	22	0	0	92
		% within IVA	3.3%	16.3%	16.3%	40.2%	23.9%	0.0%	0.0%	100.0%
		% within Region	100.0%	93.8%	71.4%	100.0%	81.5%	0.0%	0.0%	84.4%
		% of Total	2.8%	13.8%	13.8%	33.9%	20.2%	0.0%	0.0%	84.4%
Total		Count	3	16	21	37	27	2	3	109
		% within IVA	2.8%	14.7%	19.3%	33.9%	24.8%	1.8%	2.8%	100.0%
		% within Region	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	2.8%	14.7%	19.3%	33.9%	24.8%	1.8%	2.8%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	38.373[a]	6	.000
Likelihood Ratio	35.892	6	.000
N of Valid Cases	109		

a. 9 cells (64.3%) have expected count less than 5. The minimum expected count is .31.

Modelos lineales

- Modelos de la forma:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

donde μ_i representa el valor central de la distribución de las observaciones y ε_{ij} representa el error aleatorio.

- Nos interesa hacer inferencias sobre la variable de respuesta Y.

- Se le llaman modelos lineales ya que μ_i es lineal en los parámetros y el error es aditivo.

Suposiciones sobre los errores

- $E[\varepsilon_i] = 0$; el error tiene media 0.
- $Var[\varepsilon_i] = \sigma^2$; la varianza de los errores es σ^2 , constante para todos los errores.
- Errores son independientes entre sí.
- Para hacer inferencias se asume que

$$\varepsilon_i \sim N\left(0, \sigma^2\right).$$

Regresión lineal simple (RLS)

El modelo de regresión lineal simple es:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

donde:

y_i es el valor de la variable de respuesta para la observación *i*,

β_0 y β_1 son parámetros y se les llama los coeficientes de regresión,

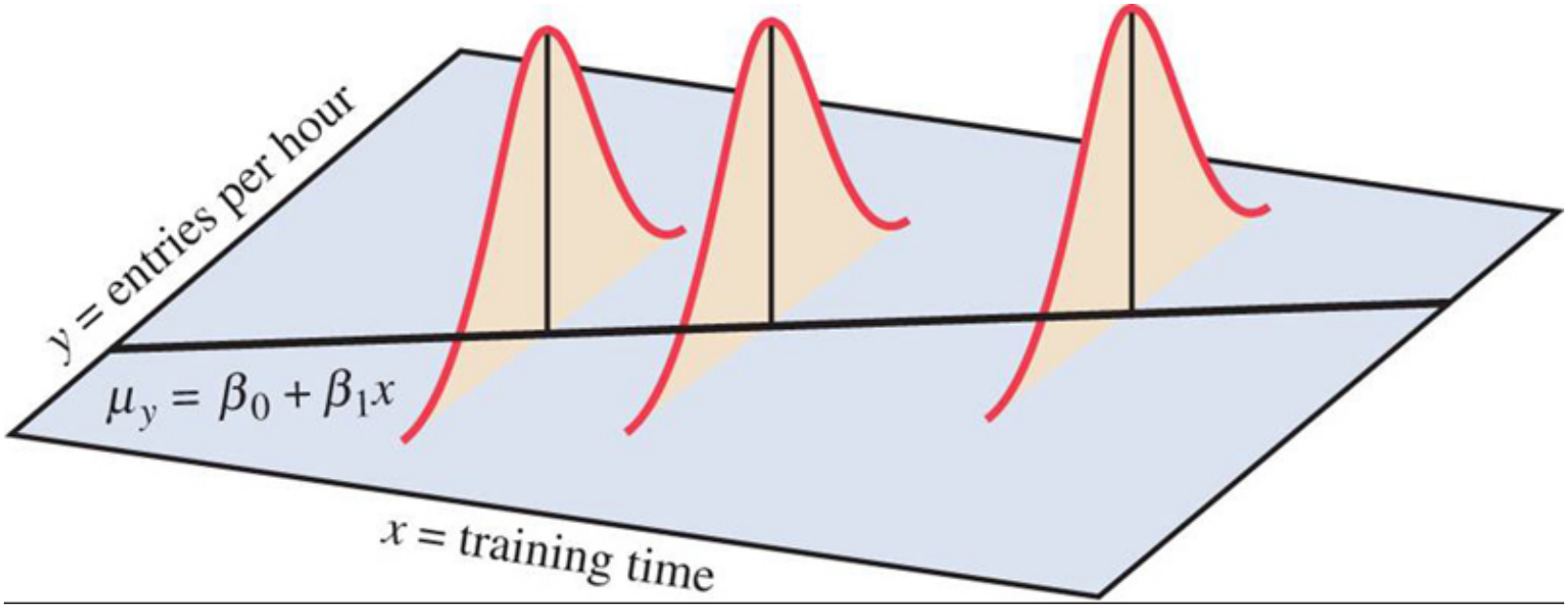
x_i es el valor de la variable predictora (se asume fija) para la observación *i*,

ε_i es el término de error aleatorio tal que $\varepsilon_i \sim N\left(0, \sigma^2\right)$, y son independientes entre sí.

- Al asumir que los errores son independientes e idénticamente distribuidos (iid), la variable aleatoria Y tiene las siguientes propiedades:

  - $\mu_{Y_i} = E[Y_i] = \beta_0 + \beta_1 x_i$
  - $Var[Y_i] = \sigma^2$
  - $Y_i's$ son independientes
  - $Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$

Ejemplo: Moore *et.al.,* Figura 10.2, pág. 572



Interpretación de β_0 y β_1

- β_0 : Intercepto en Y de la línea de regresión. Si el modelo incluye x=0, β_0 es la media de de Y cuando x=0.

- β_1 : es la pendiente de la línea de regresión; indica el cambio en la media de Y cuando X aumenta una unidad.

Estimación de β_0 y β_1

- Como no conocemos los valores de los coeficientes de regresión, los tenemos que estimar de los datos.
- Los datos pueden ser producto de estudios observacionales o experimentales.
- Los datos consisten de n observaciones de la variable predictora X y n observaciones de la variable de respuesta Y:

$$(x_i, y_i), \quad i = 1, \ldots, n.$$

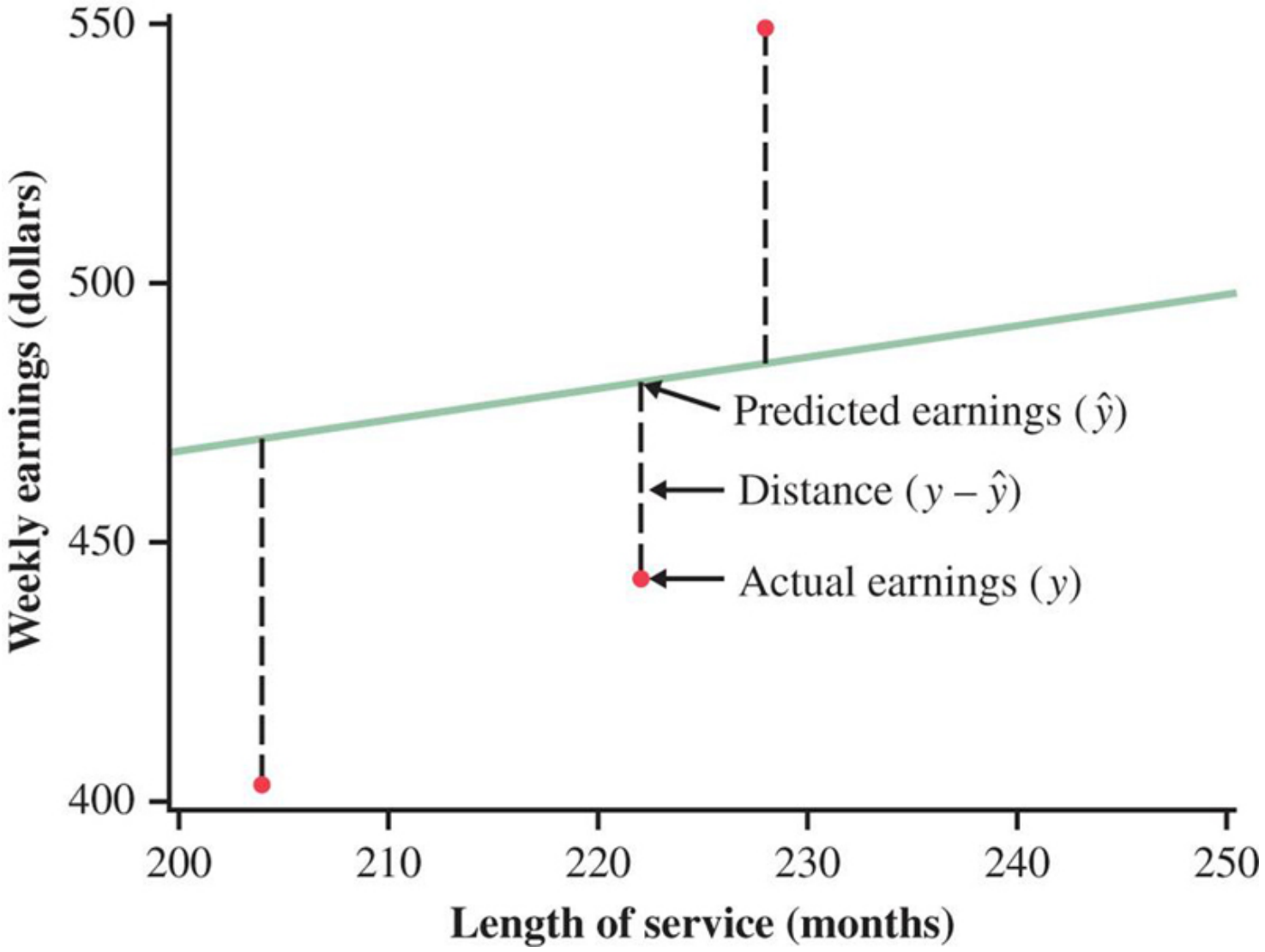
Línea o función de regresión estimada:

$$\hat{y}_i = b_0 + b_1 x_i$$

Residuales (errores observados):

$$e_i = y_i - \hat{y}_i$$

Moore *et.al.*, Figura 10.4, pág. 574



Intervalos de (1-α)100% de confiabilidad

- Para β_1 : $b_1 \pm t_{\frac{\alpha}{2},n-2} s_{b_1}$

- Para β_0 : $b_0 \pm t_{\frac{\alpha}{2},n-2} s_{b_0}$

Prueba de hipótesis para los coeficientes de regresión

- Hipótesis: $H_0 : \beta_i = 0$ $H_1 : \beta_i \neq 0$

  $\beta_i > 0$

  $\beta_i < 0$

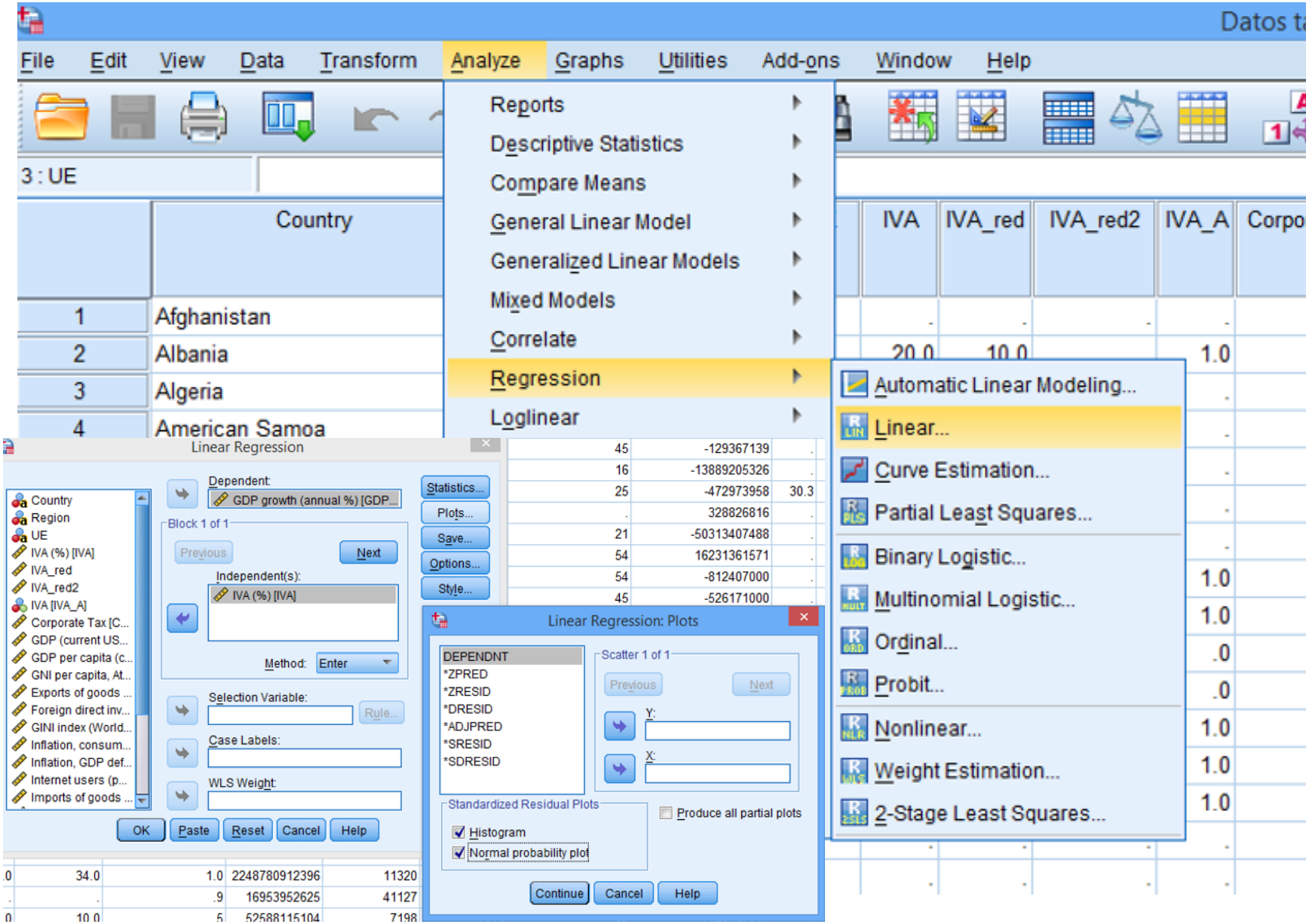
- Estadística de prueba: $t^* = \dfrac{b_i - 0}{s_{b_i}}$

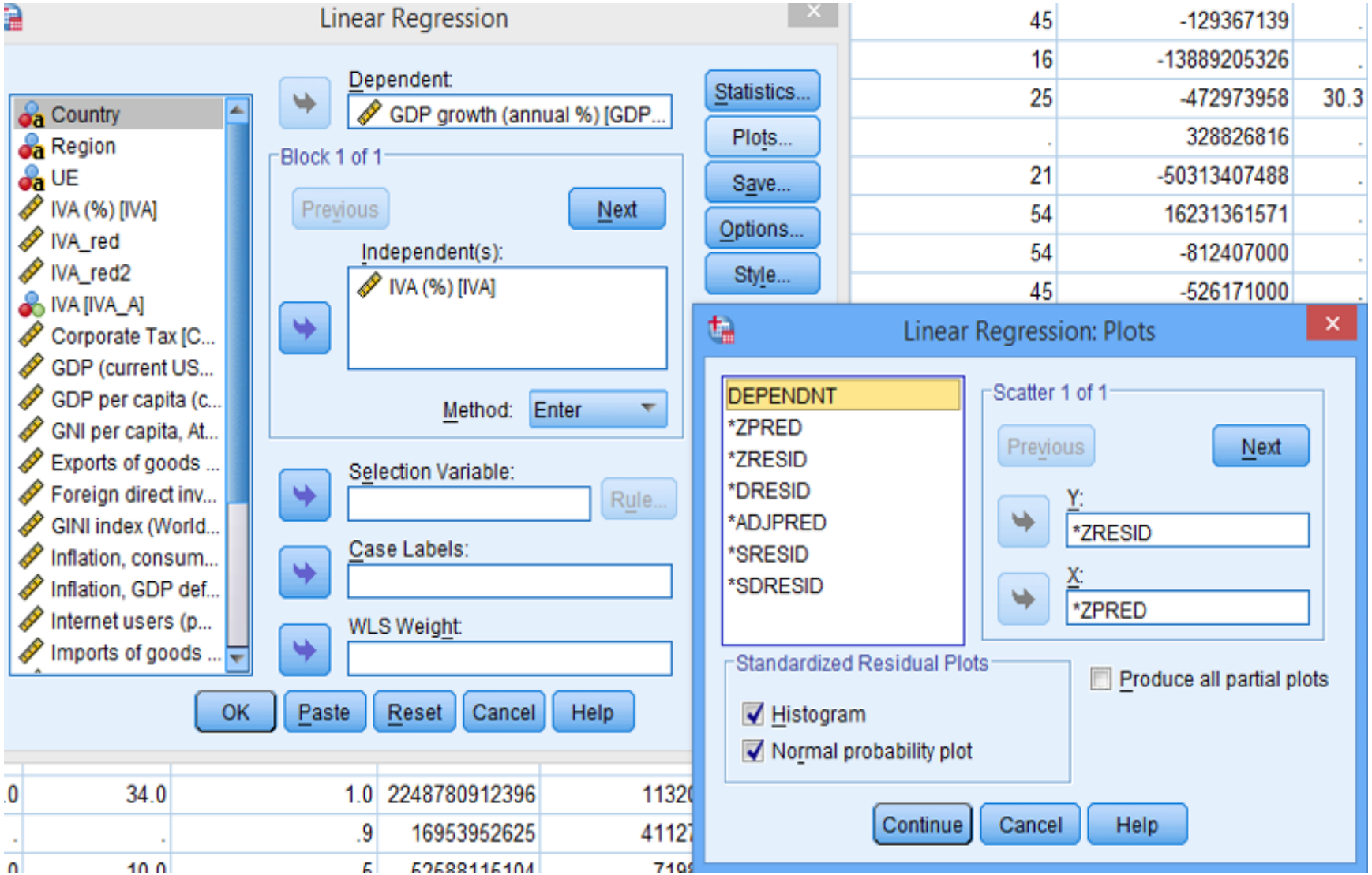
- Bajo la hipótesis nula, $t^* \sim t_{n-2}$.

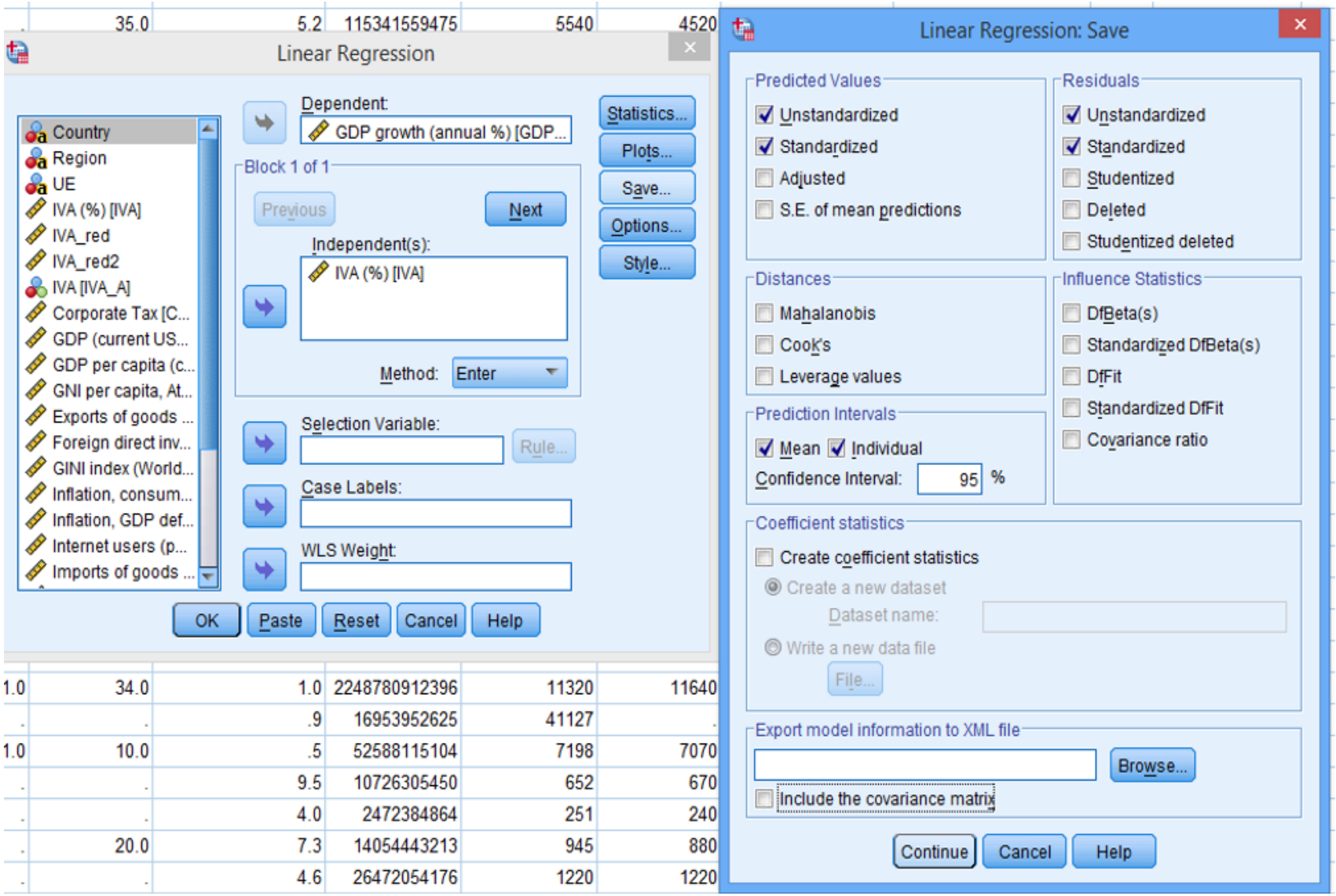
- Nota: para probar si existe una relación lineal entre las variables X y Y, las hipótesis son las siguientes:

$H_0 : \beta_1 = 0 \quad \rightarrow$ No existe relación lineal entre X y Y

$H_1 : \beta_1 \neq 0 \quad \rightarrow$ Existe relación lineal entre X y Y


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
3 : UE
Country
IVA
IVA_red
IVA_red2
IVA_A
Corpo
1 Afghanistan
2 Albania
3 Algeria
4 American Samoa
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Linear Regression
Dependent:
GDP growth (annual %) [GDP...
Block 1 of 1
Previous
Next
Independent(s):
IVA (%) [IVA]
Method: Enter
Selection Variable:
Case Labels:
WLS Weight:
Rule...
Statistics...
Plots...
Save...
Options...
Style...
OK
Paste
Reset
Cancel
Help
Country
Region
UE
IVA (%) [IVA]
IVA_red
IVA_red2
IVA [IVA_A]
Corporate Tax [C...
GDP (current US...
GDP per capita (c...
GNI per capita, At...
Exports of goods ...
Foreign direct inv...
GINI index (World...
Inflation, consum...
Inflation, GDP def...
Internet users (p...
Imports of goods ...
Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
X:
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Continue
Cancel
Help


Linear Regression
Country
Region
UE
IVA (%) [IVA]
IVA_red
IVA_red2
IVA [IVA_A]
Corporate Tax [C...
GDP (current US...
GDP per capita (c...
GNI per capita, At...
Exports of goods ...
Foreign direct inv...
GINI index (World...
Inflation, consum...
Inflation, GDP def...
Internet users (p...
Imports of goods ...
Dependent:
GDP growth (annual %) [GDP...
Block 1 of 1
Previous
Next
Independent(s):
IVA (%) [IVA]
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
OK
Paste
Reset
Cancel
Help
45 -129367139 .
16 -13889205326 .
25 -472973958 30.3
. 328826816 .
21 -50313407488 .
54 16231361571 .
54 -812407000 .
45 -526171000 .
Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
*ZRESID
X:
*ZPRED
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Continue
Cancel
Help
.0 34.0 1.0 2248780912396 11320
. . .9 16953952625 41127
0 10.0 5 52588115104 7198



	35.0	5.2	115341559475	5540	4520
1.0	34.0	1.0	2248780912396	11320	11640
.	.	.9	16953952625	41127	.
1.0	10.0	.5	52588115104	7198	7070
.	.	9.5	10726305450	652	670
.	.	4.0	2472384864	251	240
.	20.0	7.3	14054443213	945	880
.	.	4.6	26472054176	1220	1220

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.536[a]	.288	.280	2.7730

a. Predictors: (Constant), IVA (%)

b. Dependent Variable: GDP growth (annual %)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	273.218	1	273.218	35.530	.000[b]
	Residual	676.692	88	7.690		
	Total	949.910	89			

a. Dependent Variable: GDP growth (annual %)

b. Predictors: (Constant), IVA (%)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	9.235	1.147		8.055	.000	6.957	11.514
	IVA (%)	-.377	.063	-.536	-5.961	.000	-.503	-.252

a. Dependent Variable: GDP growth (annual %)

Enfoque de Análisis de Varianza

- Idea básica:
  - Se particiona la variabilidad total de la variable de respuesta Y en dos fuentes:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

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

$$\Rightarrow \qquad SST = SSR + SSE$$

- SST: Suma de cuadrados total
  - Variabilidad de Y sin tomar en cuenta la variable preditora X.
- SSR: Suma de cuadrados de la regresión
  - Variabilidad de Y asociada a la variable predictora X.
- SSE: Suma de cuadrados del error
  - Variabilidad de Y que no es explicada por la regresión.

Tabla de ANOVA

Fuentes de variación	**Grados de libertad**	**SS**	**MS**	**F**
Regresión	1	SSR= $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSR=SSR/1	F*=MSR/MSE
Error	n-2	SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE=SSE/(n-2)	
Total	n-1	SST= $\sum_{i=1}^{n}(y_i - \bar{y})^2$		

Prueba de F – existencia de relación lineal

- Hipótesis: $H_0 : \beta_1 = 0$

  $H_1 : \beta_1 \neq 0$

- Estadística de prueba: $F^* = \dfrac{MSR}{MSE} \sim F_{1,n-2}$

- Regla de decisión: Rechazo H_0 si $F^* > F_{\alpha;1,n-2}$.

Coeficiente de determinación

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \qquad 0 \leq r^2 \leq 1$$

- Este coeficiente mide la reducción proporcional de la variabilidad total asociada con el uso de la variable predictora X; o puede también ser interpretado como la proporción de la variabilidad total que es explicada por el modelo.

- Valores de r^2 cerca de 1 implica que el modelo provee un buen ajuste ("good fit") a los datos.

Diagnósticos

- El análisis con el modelo de regresión lineal simple está basado en cuatro supuestos básicos:
  - La relación entre las variables es lineal.
  - Los errores tienen una distribución Normal.
  - La varianza de los errores es constante. (homocedasticidad)
  - Los errores son independientes.

- Antes de hacer cualquier inferencia con el modelo es necesario verificar que estos supuestos se cumplan.

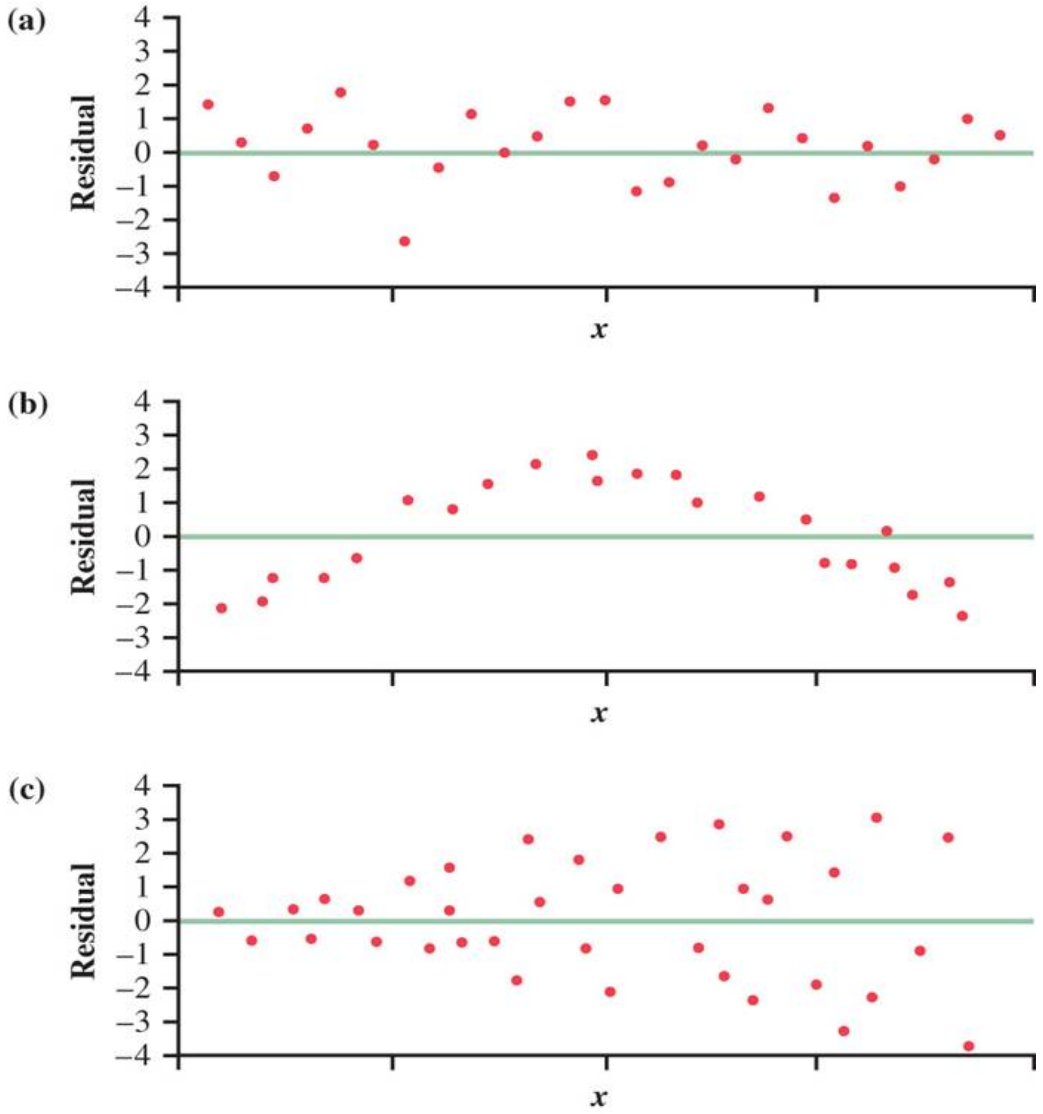
Gráficas de residuales

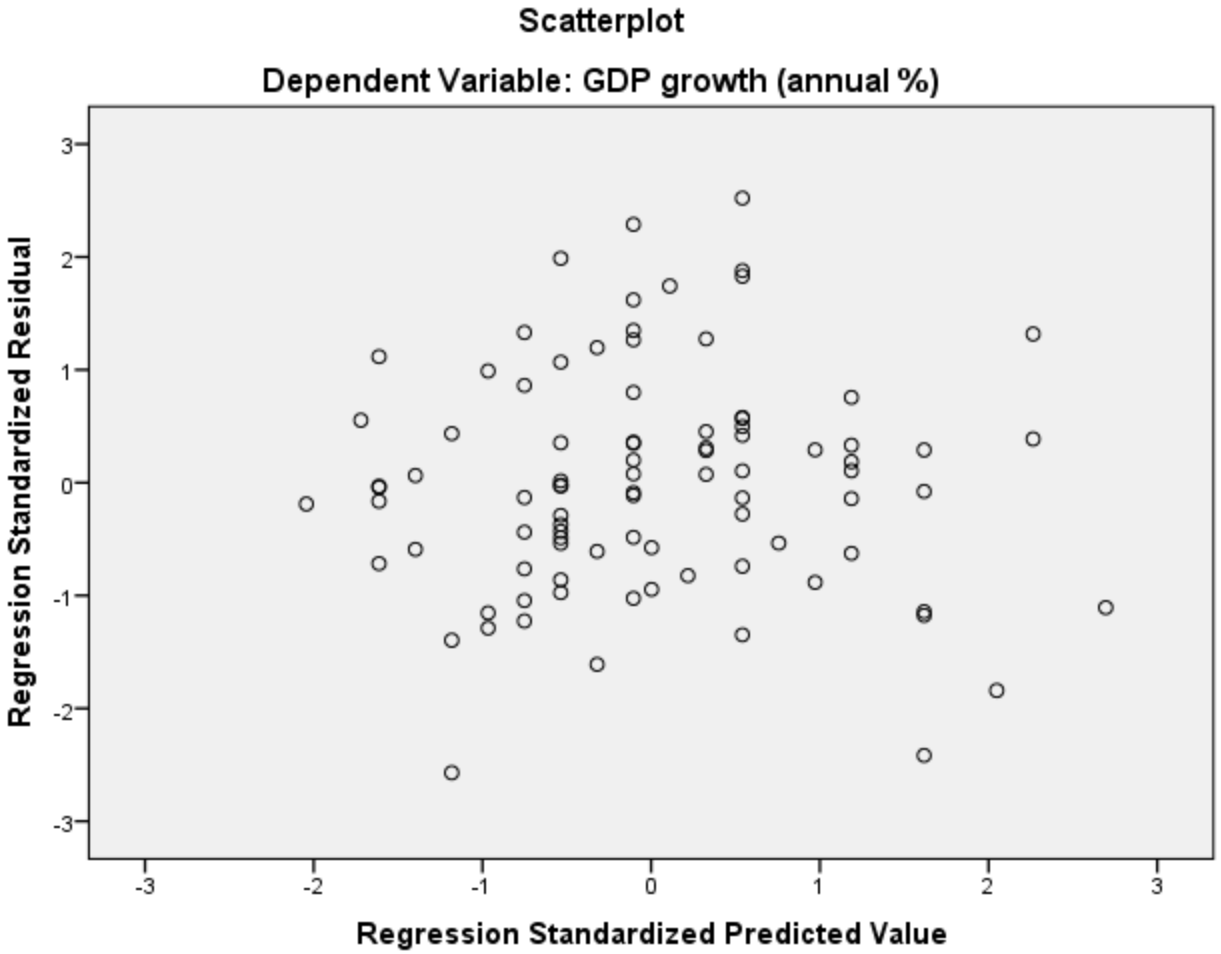
- Residuales: $e_i = y_i - \hat{y}_i$
- Residuales estandarizados: $e_i^* = \dfrac{e_i}{\sqrt{MSE\,(1-h_i)}}$

  h_i es una medida de la influencia de la observación i en el modelo de acuerdo a su valor de x.

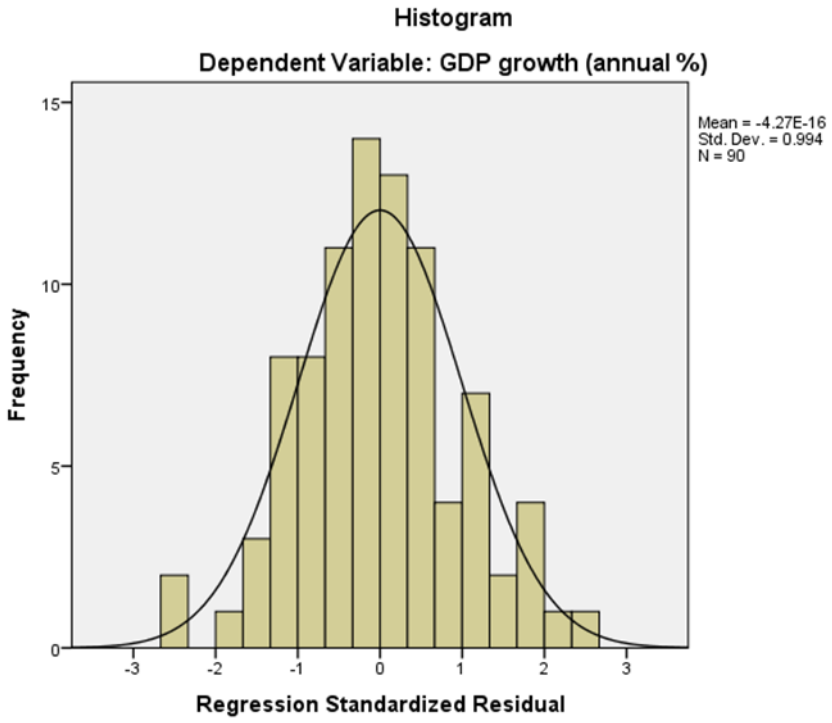
- Gráficas diagnósticas:
  - Gráfica de residuales vs valores ajustados (o vs X)
    - Podemos estudiar si la relación entre las variables es lineal.
    - Podemos estudiar si las varianzas de los errores son constantes.
    - En esta gráfica, no se violan las suposiciones anteriormente mencionadas si los residuales no tienen ningún tipo de patrón.
  - "Normal probability plot"
    - Podemos estudiar si los errores se distribuyen Normal (si la gráfica es una línea ascendente).

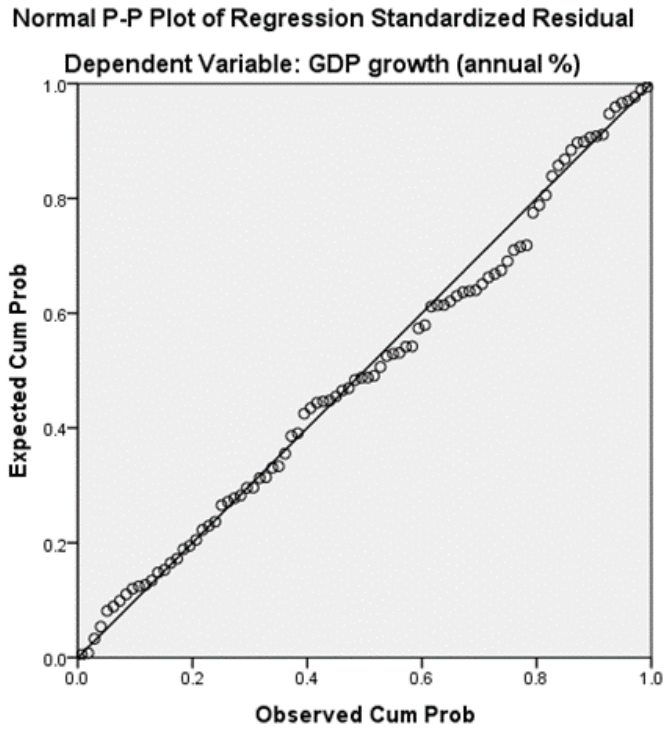
Moore *et.al.*, Figura 2.18, pág. 129




Scatterplot
Dependent Variable: GDP growth (annual %)
Regression Standardized Residual
Regression Standardized Predicted Value
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Gráficas de residuales





Prueba de normalidad de Shapiro-Wilks

H_0: Residuales se distribuyen Normal

H_1: Residuales no se distribuyen Normal

>Analyze >Descriptive Statistics >Explore

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.070	90	.200*	.990	90	.710

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- El p-value no es significativo, por lo que no rechazamos la hipótesis nula y concluimos que no hay evidencia de que los residuales no se distribuyan Normal.

Prueba de Breusch-Pagan

- Prueba para verificar si la varianza de los errores es constante.

- H_0: varianza de los errores es constante

  H_1: varianza de los errores no es constante

Medidas remediales

- Si el modelo de regresión lineal simple no es apropiado para describir los datos, tenemos dos alternativas:
  - Desarrollamos otro modelo,
  - Utilizamos algún tipo de transformación en los datos de manera que el modelo de regresión es apropiado para los datos transformados.

- Si la varianza no es constante, pero varía de forma sistemática, se puede considerar transformar los datos o utilizar “Weighted least squares” para estimar los parámetros.
- Frecuentemente se observa el caso donde las varianzas de los errores no son constantes y los errores no se distribuyen Normal.
- Algunas transformaciones de Y muy utilizadas para resolver el punto anterior son:

$$\sqrt{Y}, \log_{10} Y, 1/Y$$

Regresión múltiple

El modelo de regresión lineal múltiple es:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \ , \quad i = 1, \ldots, n$$

donde:

y_i es el valor de la variable de respuesta para la observación *i*,

$\beta_o, \beta_1, \beta_2, \ldots, \beta_p$ son parámetros y se les llama los coeficientes de regresión,

$x_{i1}, x_{i2}, \ldots, x_{ip}$ son las variables predictoras (se asumen fijas) para la observación *i*,

ε_i es el término de error aleatorio tal que $\varepsilon_i \sim N\left(0, \sigma^2\right)$, y son independientes entre sí.

- Al asumir que los errores son independientes e idénticamente distribuidos (iid), la variable aleatoria Y tiene las siguientes propiedades:

- $\mu_{Y_i} = E[Y_i] = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$
- $Var[Y_i] = \sigma^2$
- $Y_i's$ son independientes
- $Y_i \sim N\left(\beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \sigma^2\right)$

Prueba de hipótesis para los coeficientes de regresión

- Hipótesis: $H_0 : \beta_i = 0$ $\qquad$ $H_1 : \beta_i \neq 0$

- Estadística de prueba: $t^* = \dfrac{b_i - 0}{s_{b_i}}$

- Bajo la hipótesis nula, $t^* \sim t_{n-p-1}$.

Prueba de F

- Hipótesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1 : \beta_j \neq 0, \text{ para algún } j = 1, \ldots, p$$

- Estadística de prueba:

$$F^* = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \sim F_{p,n-p-1}$$

Tabla de Análisis de Varianza

Fuente de variación	Grados de libertad	SS	MS	F
Regresión	p	SSR	MSR=SSR/p	MSR/MSE
Error	n-(p+1)	SSE	MSE=SSE/(n-p-1)	
Total	n-1	SST		

Intervalos de (1-α)100% de confiabilidad para los coeficientes de regresión

Intervalo de (1-α) 100% de confiabilidad para β_i :

$$b_i \pm t_{\frac{\alpha}{2},n-p-1} s_{b_i}$$

- Interpretación de b_i: estima el cambio en la variable de respuesta Y cuando la variable X_i aumenta una unidad y las demás variables predictoras se mantienen constante.

Coeficiente de determinación múltiple R^2

- Al igual que en regresión lineal simple, podemos definir el coeficiente de determinación múltiple como

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \qquad 0 \leq R^2 \leq 1$$

- Este coeficiente mide la reducción proporcional de la variabilidad total asociada con el uso de las variables predictoras $X_1,\ldots,X_p$; o puede también ser interpretado como la proporción de la variabilidad de Y que es explicada por su relación con las variables predictoras (modelo).

- El problema con el coeficiente de regresión múltiple es que cada vez que se añade una variable al modelo, aumenta la suma de cuadrados de regresión (SSR) y por ende R^2 aumenta. Esto dificulta comparar modelos en términos de R^2.
- Para resolver ese problema se define el coeficiente de regresión múltiple ajustado para comparar modelos:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}, \qquad 0 \leq R^2_{adj} \leq 1$$

Selección de variables predictoras

- Existen una serie de algoritmos computacionales de búsqueda automática para seleccionar el "mejor" grupo de variables predictoras, entre ellos:
  - "Stepwise"
    - El algoritmo va seleccionando /eliminando variables en cada paso.
  - "Forward selection"
    - El algoritmo empieza sin variables predictoras y va añadiendo variables una a una.
  - "Backward elimination"
    - El algoritmo empieza con todas las posibles variables predictoras y va eliminando variables una a una.