

AYUDANTÍA MÉTODOS CUANTITATIVOS

Axel Valenzuela - 2020

Estadística en investigación

- La evidencia empírica ayuda a la comprensión de la realidad social.
- La estadística es útil para procesar la evidencia, resumirla, entenderla y sacar conclusiones a partir de la interpretación de datos.
- Una mala interpretación de datos puede tener causas negativas en la población.
- Permite análisis profundos obteniendo hallazgos que no son siempre evidentes ante análisis superficiales.

Tipos de Estadística

- **Estadística Descriptiva**: Es aquella que se encarga de describir y analizar un conjunto de datos.
- **Estadística inferencial**: Es aquella que se encarga de sacar conclusiones a partir del análisis de los datos extraídos.

¿Qué es una variable?

- Son representaciones de alguna unidad observada, pueden representar a aquello que varía o que está sujeto a algún tipo de cambio.
- Pueden clasificarse:

Variable dependiente: Son aquellas que su valor depende de otra(s) variable(s). Los valores que tienen estas variables están supeditados a los que tengan las variables independientes.

Variable independiente: Son aquellas que no dependen de ninguna otra variable. Se le pueden asignar valores sin tener en cuenta otras variables.

Las variables independientes buscan explicar a la variable dependiente.

Tipos de variables: Cualitativas

También llamadas **Categóricas**. Corresponde a aquellas variables que refieren a características o bien, cualidades, **que no pueden ser medidas**. Establecen una condición. Se clasifican en:

- **Nominales**:. Se caracterizan porque **no pueden ordenarse ni establecer jerarquías entre ellas**, tampoco se pueden realizarse operaciones aritméticas con ellas.

Ejemplo: Estado civil, sexo, preferencia política.

- **Ordinales**: Son aquellas que clasifican de manera que **se puede establecer un orden**. **Ejemplo:**

Nivel de estudios → 1. Ed. Básica

2. Ed. Media

3. Ed superior

Tipos de variables: Cuantitativas

También llamadas **Numéricas**. Corresponde a aquellas variables que **se expresan mediante una unidad de medida** y se caracterizan porque se puede realizar operaciones con ellas. Establecen una medida. Se clasifican en:

- **Discreta**: Son aquellas que representan cantidades expresadas solo **por números enteros**. Se realiza un proceso de conteo.

Ejemplo: Número de hijos, cantidad de alumnos por sala, entre otros.

- **Continua**: Son aquellas que representan cantidades que pueden ser expresadas por **números infinitos**. Se realiza proceso de medición

Ejemplos: Estatura, peso, temperatura, entre otros.

Resumen tipo de variables

Escala	Característica	Ejemplo
Nominal	No establece jerarquías.	Sexo Color de pelo
Ordinal	Establece jerarquías	Nivel educacional Nivel socioeconómico
Discreta	Se cuenta (números enteros).	Número de hijos Cantidad de autos por hogar.
Continua	Se mide (números infinitos).	Peso Talla

Variable Dependiente	Variable Independiente	Prueba
Dicotómica	Dicotómica	z, Chi-2, Fisher
Dicotómica	Nominal	z, Chi-2, Metha-Patel
Dicotómica	Ordinal	Mann-Whitney / Wilcoxon
Dicotómica	Cuantitativa	Regresión Logística
Nominal	Dicotómica	z, Chi-2, Metha-Patel
Nominal	Cuantitativa	Regresión Discriminante
Ordinal	Nominal	Kruskal-Wallis
Ordinal	Cuantitativa	Regresión Odds Proporcional
Cuantitativa	Dicotómica	t de Student
Cuantitativa	Nominal	ANOVA
Cuantitativa	Cuantitativa	Regresión Lineal, correlaciones

TIPOS DE VARIABLES		DEPENDIENTE		
		CUALITATIVA (2 grupos)	CUALITATIVA (> 2 grupos)	CUANTITATIVA
I N D E P E N D I E N T E	CUALITATIVA (2 grupos)	Chi cuadrado	Chi cuadrado	T de Student
	CUALITATIVA (> 2 grupos)	Chi cuadrado	Chi cuadrado	Anova
	CUANTITATIVA	Regresión logística	Regresión logística	Regresión lineal Correlaciones: - Pearson - Spearman

Base de datos

Variable

ID	Edad	Sexo	Preferencia política
1	24	1	2
2	18	2	1
3	42	2	5
4	66	2	3
5	27	1	4

Casos

Valor número asociado a una etiqueta o un significado

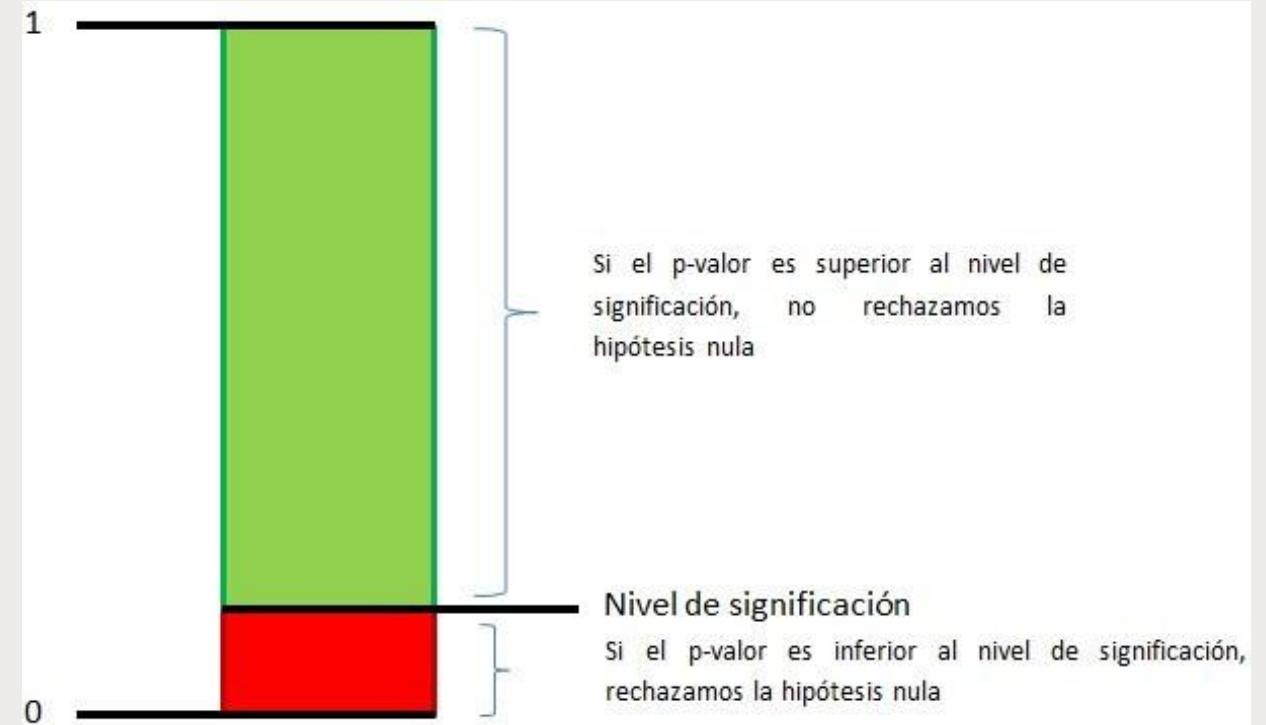
Prueba de hipótesis

		Ho	
		Verdadera	Falsa (H1 verdadera)
Ho	Aceptar	Decisión correcta $1 - \alpha$	Decisión incorrecta <u>Error de tipo II</u> β
	Rechazar (Aceptar H1)	Decisión incorrecta <u>Error de tipo I</u> α	Decisión correcta $1 - \beta$

- El **error tipo I**, también llamado alfa (α), se comete al rechazar la hipótesis nula (H_0) siendo esta verdadera.
- El **error tipo II** o beta (β), se comete al **aceptar** la hipótesis nula (H_0) siendo esta falsa. Es decir, la probabilidad de cometer un error tipo II es beta (β), y depende de la potencia de la prueba ($1-\beta$).

Prueba de hipótesis

1. Se debe identificar variables dependiente e independiente(s) a estudiar.
2. Se establecen las hipótesis
3. Se realiza la prueba estadística
4. Se interpreta el resultado (No se rechaza H_0 o, se rechaza H_0 y se acepta H_1).



Por lo general la hipótesis nula incluye un no (o un desigual a) en su enunciado.

Ejemplo:

H_0 : Las variables sexo e ingreso autónomo no están relacionadas (son independientes).

H_1 : Las variables sexo e ingreso autónomo sí están relacionadas.



TABLAS DE CONTIGENCIA Y CRUZADAS



Tablas

- Son distribuciones de frecuencia de **dos o mas variables categóricas**. También puede conceptualizarse como una clasificación cruzada de los valores (o categorías) de dos o mas variables mostrándose el número de observaciones en cada celda reportada.

Comando	Comando abreviado	Uso	Utilidad
Tabulate	tab	tabulate variable tabulate variable variable2	Describe propiedades de las variables que están en la base de datos.
Tabstat		tabstat <i>variable</i>	Describe propiedades de las variables que están en la base de datos y entrega estadísticos asociados a estas variables.

Opcionales comando tab con dos variables

options	Description
Main	
<u>chi2</u>	report Pearson's chi-squared
<u>exact</u> [(#)]	report Fisher's exact test
<u>gamma</u>	report Goodman and Kruskal's gamma
<u>lrchi2</u>	report likelihood-ratio chi-squared
<u>taub</u>	report Kendall's tau-b
<u>V</u>	report Cramér's V
<u>cchi2</u>	report Pearson's chi-squared in each cell
<u>column</u>	report relative frequency within its column of each cell
<u>row</u>	report relative frequency within its row of each cell
<u>clrchi2</u>	report likelihood-ratio chi-squared in each cell
<u>cell</u>	report the relative frequency of each cell
<u>expected</u>	report expected frequency in each cell
<u>nofreq</u>	do not display frequencies
<u>row</u> sort	list rows in order of observed frequency
<u>col</u> sort	list columns in order of observed frequency
<u>missing</u>	treat missing values like other values
<u>wrap</u>	do not wrap wide tables
[no]key	report/suppress cell contents key
<u>no</u> label	display numeric codes rather than value labels
<u>no</u> log	do not display enumeration log for Fisher's exact test
* <u>firstonly</u>	show only tables that include the first variable in varlist

Interpretación Tablas de Contingencia

Número de observaciones (frecuencia)

Porcentaje relativo

Porcentaje acumulado

tab sexo

Sexo	Freq.	Percent	Cum.
Hombre	103,322	47.74	47.74
Mujer	113,117	52.26	100.00
Total	216,439	100.00	

. tab sexo if ecivil==5

Sexo	Freq.	Percent	Cum.
Hombre	3,168	35.65	35.65
Mujer	5,718	64.35	100.00
Total	8,886	100.00	

Tabla de contingencia de una variable

Interpretaciones:

- Mayoría de personas encuestadas son mujer (113,117 > 103,322).
- La cantidad de hombres encuestados representa un 47.74%.

Tabla de contingencia de una variable con filtro if. Ecivil==5 representa a separados/as.

Interpretaciones:

- Mayoría de personas separados/as encuestadas son mujeres (5,718 > 3,168).
- La cantidad de mujeres separadas encuestadas representa el 64.35%

```
. tab ecivil sexo
```

Estado civil	Sexo		Total
	Hombre	Mujer	
Casado(a)	29,795	29,803	59,598
Conviviente o pareja	13,776	13,786	27,562
Conviviente civil (co	186	180	366
Anulado(a)	66	156	222
Separado(a)	3,168	5,718	8,886
Divorciado (a)	940	2,105	3,045
Viudo(a)	2,256	8,370	10,626
Soltero(a)	53,135	52,999	106,134
Total	103,322	113,117	216,439

Tabla de contingencia de dos variables.

Interpretaciones:

- Comparación entre hombre y mujer según Estado Civil.

```
. tab ecivil sexo, col
```

Estado civil	Sexo		Total
	Hombre	Mujer	
Casado(a)	29,795 28.84	29,803 26.35	59,598 27.54
Conviviente o pareja	13,776 13.33	13,786 12.19	27,562 12.73
Conviviente civil (co	186 0.18	180 0.16	366 0.17
Anulado(a)	66 0.06	156 0.14	222 0.10
Separado(a)	3,168 3.07	5,718 5.05	8,886 4.11
Divorciado (a)	940 0.91	2,105 1.86	3,045 1.41
Viudo(a)	2,256 2.18	8,370 7.40	10,626 4.91
Soltero(a)	53,135 51.43	52,999 46.85	106,134 49.04
Total	103,322 100.00	113,117 100.00	216,439 100.00

Tabla de contingencia de dos variables y porcentajes según columna.

Interpretaciones:

- Se pueden hacer interpretaciones con los porcentajes y número de observaciones entre hombres y mujeres según Estado Civil.

```
. tab ecivil sexo, row
```

Estado civil	Sexo		Total
	Hombre	Mujer	
Casado(a)	29,795 49.99	29,803 50.01	59,598 100.00
Conviviente o pareja	13,776 49.98	13,786 50.02	27,562 100.00
Conviviente civil (co	186 50.82	180 49.18	366 100.00
Anulado(a)	66 29.73	156 70.27	222 100.00
Separado(a)	3,168 35.65	5,718 64.35	8,886 100.00
Divorciado (a)	940 30.87	2,105 69.13	3,045 100.00
Viudo(a)	2,256 21.23	8,370 78.77	10,626 100.00
Soltero(a)	53,135 50.06	52,999 49.94	106,134 100.00
Total	103,322 47.74	113,117 52.26	216,439 100.00

Interpretaciones:

- Se pueden hacer interpretaciones con los porcentajes y número de observaciones entre hombres y mujeres según Estado Civil.

Tabla de contingencia de dos variables y porcentajes según filas.

```
. tab sexo ecivil, cel
```

Key
frequency
cell percentage

Sexo	Estado civil								Total
	Casado(a)	Convivien	Convivien	Anulado(a)	Separado(a)	Divorciad	Viudo(a)	Soltero(a)	
Hombre	29,795 13.77	13,776 6.36	186 0.09	66 0.03	3,168 1.46	940 0.43	2,256 1.04	53,135 24.55	103,322 47.74
Mujer	29,803 13.77	13,786 6.37	180 0.08	156 0.07	5,718 2.64	2,105 0.97	8,370 3.87	52,999 24.49	113,117 52.26
Total	59,598 27.54	27,562 12.73	366 0.17	222 0.10	8,886 4.11	3,045 1.41	10,626 4.91	106,134 49.04	216,439 100.00

Tabla de contingencia de dos variables y porcentajes para columnas y filas.

Interpretaciones:

- Se pueden hacer interpretaciones con los porcentajes y número de observaciones entre hombres y mujeres según Estado Civil.

A tener en cuenta...

- **Coeficiente de Cramer:** Toma valores entre 0 y 1. Cuanto más próximo a cero se encuentre, más independientes serán las variables; cuanto más próximo a uno sea el número, más asociadas estarán las variables que se estudien.

Coeficiente V de Cramer	Significado
0 y 0.2	Asociación nula
0.2	Asociación débil
0.2 y 0.6	Asociación moderada
0.6 y 0.9999	Asociación fuerte
1	Asociación perfecta

- **Prueba chi cuadrado:** Una prueba de chi-cuadrada es una prueba de hipótesis que compara la distribución observada de los datos con una distribución esperada de los datos. Se deben establecer hipótesis, comparar el valor Pr obtenido con el nivel de significancia y luego aceptar o rechazar la Hipótesis nula en base a los valores obtenidos.

```
tab Winner17 Nation, col nofreq all
```

2017 Winning Party	Nation				Total
	England	Norther..	Scotland	Wales	
Con	55.53	0.00	22.03	20.00	48.77
DUP	0.00	55.56	0.00	0.00	1.54
Green	0.19	0.00	0.00	0.00	0.15
Ind	0.00	5.56	0.00	0.00	0.15
LD	1.50	0.00	6.78	0.00	1.85
Lab	42.59	0.00	11.86	70.00	40.31
PC	0.00	0.00	0.00	10.00	0.62
SF	0.00	38.89	0.00	0.00	1.08
SNP	0.00	0.00	59.32	0.00	5.38
Spk	0.19	0.00	0.00	0.00	0.15
Total	100.00	100.00	100.00	100.00	100.00

```
Pearson chi2(27) = 1.1e+03 Pr = 0.000
```

```
likelihood-ratio chi2(27) = 408.0405 Pr = 0.000
```

```
Cramér's V = 0.7548
```

```
gamma = 0.5869 ASE = 0.051
```

```
Kendall's tau-b = 0.3336 ASE = 0.034
```

Variables:

Winner17: Partido ganador elecciones 2017

Nation: País del Reino Unido

Se obtiene que:

Chi2= 0.000

Cramer=0.7548

Se puede interpretar que:

- Existen diferencias entre las variables que son significativas $Pr=0.000<0.05$.
- El grado de asociación entre estas variables es fuerte ($V=0.7458$).
- Debido a que $Pr = 0.000<0.005$, se rechaza la Hipótesis nula y se acepta la hipótesis alternativa.

H0: Las variables Winner17 y Nation no están relacionadas(son independientes).

H1: Las variables Winner17 y Nation sí están relacionadas.

Resumen

Tabla (tab)	
Tipos de Variables	Dos o más Cuantitativas(categóricas)
Análisis	V de Cramer y Chi2
Comando(s) Stata	tab

Tabla (tabstat)	
Tipos de Variables	Dos o más Cuantitativas(categóricas)
Análisis	Prob > F – Prob >chi2
Comando(s) Stata	tabstat



DIFERENCIAS DE MEDIAS



A tener en cuenta...

- **Análisis de Varianzas (ANOVA):** Sirve para evaluar si las diferencias que se observan entre las medias son significativas, cuando la variable categórica presenta más de dos valores, lo podemos realizar mediante el ANOVA de un factor. Se prueba la hipótesis nula que “todos los tratamientos son iguales, contra la hipótesis alternativa que “al menos uno de los tratamientos es distinto a los demás”.
- **Prueba de Barlett:** Mediante esta prueba se asume que las varianzas son homogéneas. Es decir, este es uno de los extraños casos en que debemos aceptar una H_0 y no rechazarla, es decir, no asumir que un resultado de $p < \alpha$ es lo que buscamos. Se espera que se cumpla su hipótesis nula o H_0 .
- **Prueba de Tukey:** Esta prueba es especial para varianzas heterogéneas. Esta es una prueba de contraste entre las distintas categorías de la variable, y su aplicación muestra que las diferencias solo son significativas para los casos en que $p < 0.05$.

```
. tabstat ingautonomo, by( sexo) stats(mean median sd min max)
```

Summary for variables: ingautonomo
by categories of: sexo (Sexo)

sexo	mean	p50	sd	min	max
Hombre	564333.9	350000	1022775	83	8.47e+07
Mujer	396815	270167	604536.1	42	6.43e+07
Total	485739.1	300000	856581.8	42	8.47e+07

Primer paso:

Realizar tabla descriptiva para observar las medias de las categorías de la variable.

Segundo paso:

Se realiza análisis estadístico correspondiente.

Tercer paso:

Se interpretan los outputs.

Cuarto paso

Se realiza la interpretación.

Variable continua

Variable categórica

```
. oneway ingautonomo sexo, tabulate
```

Sexo	Summary of Ingreso autonomo		
	Mean	Std. Dev.	Freq.
Hombre	564333.92	1022774.8	61,175
Mujer	396814.98	604536.11	54,069
Total	485739.1	856581.82	115,244

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	8.0544e+14	1	8.0544e+14	1108.27	0.0000
Within groups	8.3752e+16	115242	7.2675e+11		
Total	8.4558e+16	115243	7.3373e+11		
Bartlett's test for equal variances: chi2(1) = 1.5e+04 Prob>chi2 = 0.000					

Variables:

ingautonomo: Ingreso monetario (\$).

sexo: Sexo (Hombre/Mujer).

Se obtiene que:

Prob>F:0.000

Se puede interpretar que:

- Existen diferencias significativas (Prob>F:0.000).
- Debido a que $Pr = 0.000 < 0.005$, se rechaza la Hipótesis nula y se acepta la hipótesis alternativa.
- Debido a que se rechaza la hipótesis nula, el análisis **no es válido** y por tanto se debe efectuar una **Prueba de Tukey**.

H0: Las variables sexo e ingreso autónomo no están relacionadas (son independientes).

H1: Las variables sexo e ingreso autónomo sí están relacionadas.

Variable continua

Variable categórica

```
. pwmean ingautonomo, over(sexo) mcompare(tukey) effects
```

Pairwise comparisons of means with equal variances

over : sexo

note: option tukey ignored since there is only one comparison

ingautonomo	Contrast	Std. Err.	t	Unadjusted	Unadjusted	
				P> t	[95% Conf. Interval]	
sexo						
Mujer vs Hombre	-167518.9	5031.998	-33.29	0.000	-177381.6	-157656.3

Se obtiene que:

P>|t|:0.000

Contraste: -167.518,9

Se puede interpretar que:

- Se obtiene que las diferencias entre ingresos entre mujeres y hombres son estadísticamente significativas (P>|t|:0.000).
- Mujeres ganan 167.518,9 menos que los hombres.

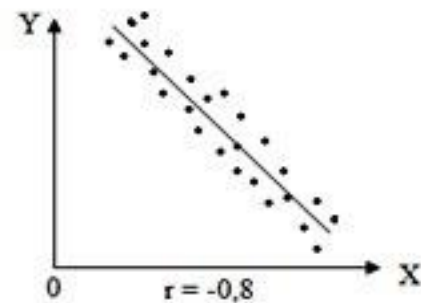
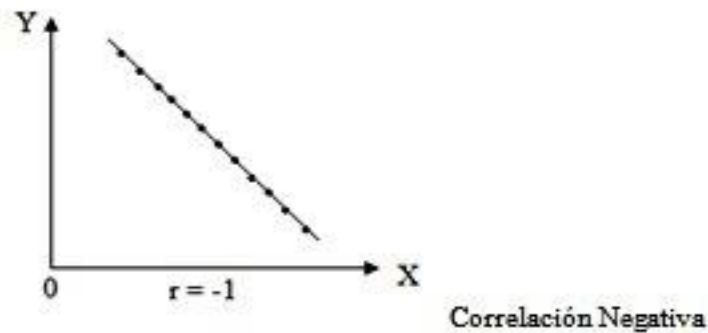
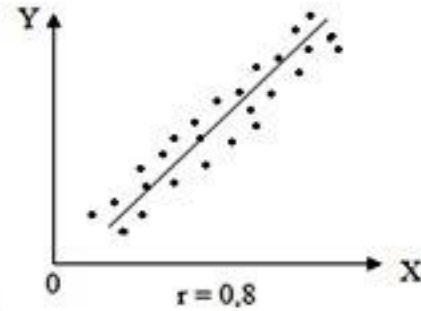
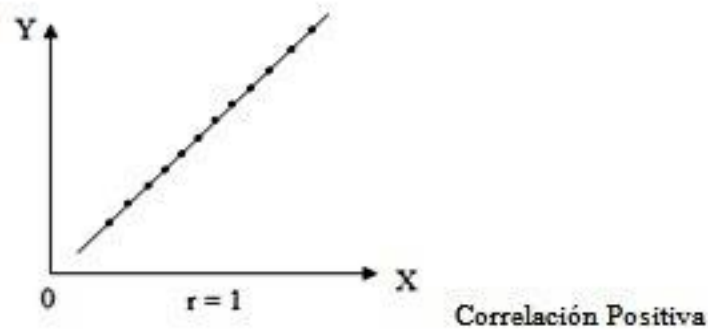
Resumen

ANOVA	
Tipos de Variables	Continua (dependiente) y categórica
Análisis a considerar	Prob > F – Prob >chi2
Comando(s) Stata	oneway
Observaciones	

Prueba Tukey	
Tipos de Variables	Continua (dependiente) y categórica
Análisis a considerar	$P > t $ - Contrast
Comando(s) Stata	pwmean var, over(var2) mcompare(tukey) effects
Observaciones	Varianzas heterogéneas. Esta prueba no es apropiada cuando existen grandes diferencias en el tamaño de la muestra por tratamiento



CORRELACIONES



- Las correlaciones sirven para establecer relaciones lineales entre variable. Estas se pueden observar mediante un gráfico de dispersión.
- Se pueden clasificar según su **coeficiente de correlación(r)**.
- Los coeficientes de correlación pueden tomar valor entre **-1 y 1**.
- Un coeficiente de correlación de $r=1$ se considera **relación lineal perfecta positiva**, en tanto si fuese $r=-1$ se considera como **relación lineal perfecta negativa**.
- Puede considerarse que existe una **relación lineal fuerte** en caso que esta tome un valor $r=\pm 0,8$, siendo negativa o positiva según el signo del caso.
- En tanto para $r=0$, se considera que **no existe correlación lineal**.

Rango		Relacion Lineal
$\pm 0,96$	$\pm 1,0$	Perfecta
$\pm 0,85$	$\pm 0,95$	Fuerte
$\pm 0,70$	$\pm 0,84$	Significativa
$\pm 0,50$	$\pm 0,69$	Moderada
$\pm 0,20$	$\pm 0,49$	Débil
$\pm 0,10$	$\pm 0,19$	Muy Débil
$\pm 0,09$	$\pm 0,0$	Nula

Correlación en Stata

Comando para
calcular correlación
mediante Pearson

Variables estudiadas

Realiza en cálculo en base
a una significancia del 95%

```
. pwcorr ingautonomo esc, ob sig star (.05)
```

	ingaut~o	esc
ingautonomo	1.0000	1.2e+05
esc	0.3003*	1.0000
	0.0000	1.1e+05
		1.7e+05

Coeficiente de
correlación.

Valor P

Se obtiene que:

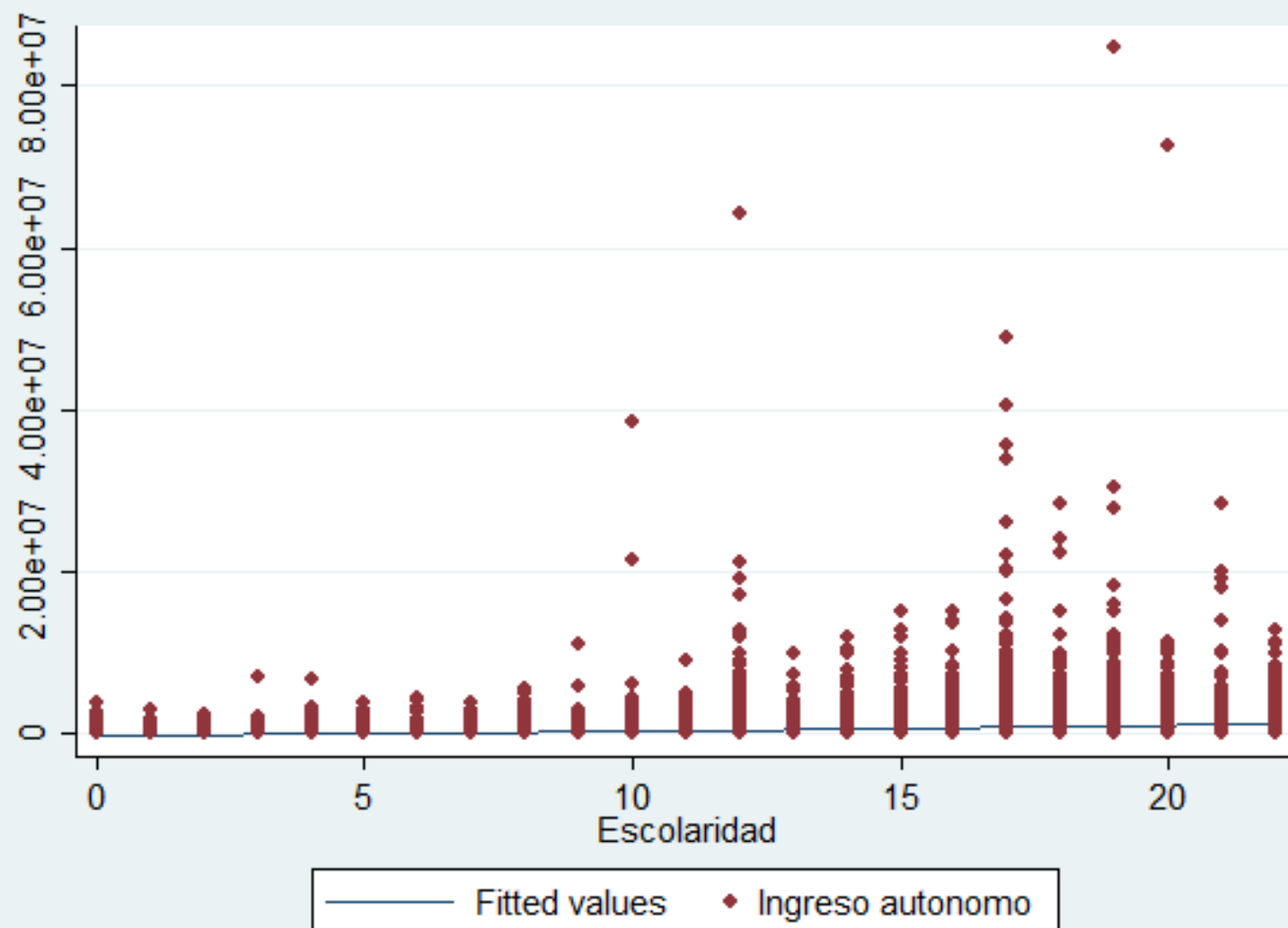
R=0,3003

P=0.000

Se puede interpretar que:

- El modelo es estadísticamente significativo (P=0.000).
- No existe correlación lineal entre las variables Ingreso autónomo (ingautonomo) y Nivel educacional (esc) para un 95% de nivel de confianza.

```
. twoway(lfit ingautonomo esc) (scatter ingautonomo esc)
```



Resumen

Correlación	
Tipos de Variables	Cuantitativas
Análisis a considerar	Coeficiente de correlación y Valor P.
Comando(s) Stata	pwcorr
Observaciones	



REGRESIONES



Regresión Lineal

- La regresión lineal es una técnica que nos permite modelar respecto del comportamiento de una determinada variable que nos interesa estudiar, a partir del cambio que produce en esta una determinada variable.

- Ecuación simple: $Y = B_0 + B_1X + e$

Diagram illustrating the components of the simple linear regression equation $Y = B_0 + B_1X + e$:

- Y : Variable dependiente
- B_0 : Constante(intercepto)
- B_1X : Variable independiente
- e : Error

- Ecuación múltiple: $Y = B_0 + B_1X + B_2X + B_3X... + e$

¿Qué debo considerar para interpretar una regresión?

```
. reg ingautonomo edad esc sexo
```

Source	SS	df	MS	Number of obs	=	113,385
Model	9.5352e+15	3	3.1784e+15	F(3, 113381)	=	4836.90
Residual	7.4505e+16	113,381	6.5712e+11	Prob > F	=	0.0000
				R-squared	=	0.1135
				Adj R-squared	=	0.1134
Total	8.4040e+16	113,384	7.4120e+11	Root MSE	=	8.1e+05

ingautonomo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edad	5801.846	151.0424	38.41	0.000	5505.805	6097.887
esc	70588.79	615.8857	114.61	0.000	69381.66	71795.92
sexo	-194744.3	4830.895	-40.31	0.000	-204212.8	-185275.8
_cons	-283771.6	13741.2	-20.65	0.000	-310704.1	-256839.1

- Prob > F
- R-squared
- Adj R-Squared
- P>|t|

```
. reg ingautonomo edad esc sexo
```

Source	SS	df	MS	Number of obs	=	113,385
Model	9.5352e+15	3	3.1784e+15	F(3, 113381)	=	4836.90
Residual	7.4505e+16	113,381	6.5712e+11	Prob > F	=	0.0000
				R-squared	=	0.1135
				Adj R-squared	=	0.1134
Total	8.4040e+16	113,384	7.4120e+11	Root MSE	=	8.1e+05

ingautonomo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edad	5801.846	151.0424	38.41	0.000	5505.805	6097.887
esc	70588.79	615.8857	114.61	0.000	69381.66	71795.92
sexo	-194744.3	4830.895	-40.31	0.000	-204212.8	-185275.8
_cons	-283771.6	13741.2	-20.65	0.000	-310704.1	-256839.1

Se obtiene que:

Prob>F= 0.0000

R-squared=0.1135

Sobre el modelo:

- El modelo (variables en su conjunto) explica en un **11%** los ingresos autónomos($0,1135 \times 100$).
- Debido a que $P < 0.0000$, se rechaza hipótesis nula y se **acepta hipótesis alternativa** en que al menos uno de los parámetros incluidos es distinto de cero o significativo.
- Debido a que en todas las variables en la prueba T son inferior a 0.05, **se considera que todas son significativas estadísticamente.**

Se puede interpretar que:

- Por cada año de edad que aumenta, el ingreso aumenta en 5801 pesos.
- Por cada año de escolaridad el ingreso aumenta en 70.588 pesos.
- Las mujeres tienen un ingreso menor de 194.744 pesos respecto de los hombres

• El modelo quedaría de esta manera:

Ingreso Autónomo: $-283771,6 + (5801,846 \times \text{años de edad}) + (70.588,79 \times \text{años de escolaridad}) - (194744,3 \times 1/0)$. (1 en caso de mujer, 0 en caso de hombre)

Resumen

Regresión Lineal	
Tipos de Variables	Cuantitativas
Análisis a considerar	Prob > F R-squared Adj R-Squared P> t
Comando(s) Stata	reg



ANÁLISIS FACTORIAL

¿Qué es Análisis Factorial?

- Se utiliza para analizar interrelaciones entre un **número elevado** de variables cuantitativas explicando dichas interrelaciones en términos de un **número menor de variables** que se denominan **factores** o componentes principales.
- Su finalidad principal no es el análisis de relaciones causales, sino la **agrupación de variables**, en función de la variabilidad que cada variable comparte con otras variables (varianza o covarianza).
- El Análisis Factorial es, por tanto, una técnica de reducción de la dimensionalidad de los datos. Su propósito último consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos.
- Nos serviría para la creación de indicadores, principalmente bajo variables de tipo cuantitativas.

Interpretar Análisis Factorial

- **Retained factors:** Indica la cantidad de factores “retenidos”, es decir los que serían útiles para la creación del indicador.
- **Eigenvalue:** Su valor debe estar por sobre 1 para que se “retenga” el factor y se considere útil para la creación del indicador.
- **Proportion:** Indica la proporción en que las variables explican lo que se desea medir. Se puede tomar en términos porcentuales. “El factor acumula el x% de varianza explicada”.
- **KMO**(Kaiser, Meyer y Olkin): El índice KMO se utiliza para comparar las magnitudes de los coeficientes de correlación parcial, de forma que cuánto más pequeño sea su valor, mayor será el valor de los coeficientes de correlación parciales $ij(p)$ y, en consecuencia, menos apropiado es realizar un Análisis Factorial. Entre 0 y 1.

0,90 > KMO	Muy bueno
0,90 > KMO > 0,80	Bueno
0,80 > KMO > 0,70	Aceptable
0,70 > KMO > 0,60	Mediocre o regular
0,60 > KMO > 0,50	Malo
0,50 > KMO	Inaceptable o muy malo ³

```
. factor v2x_polyarchy v2x_libdem v2x_delibdem v2x_partipdem v2x_egaldem, pcf
(obs=18,348)
```

Factor analysis/correlation

Method: principal-component factors

Rotation: (unrotated)

Number of obs = 18,348

Retained factors = 1

Number of params = 5

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.86387	4.80909	0.9728	0.9728
Factor2	0.05478	0.02335	0.0110	0.9837
Factor3	0.03142	0.00355	0.0063	0.9900
Factor4	0.02787	0.00581	0.0056	0.9956
Factor5	0.02206	.	0.0044	1.0000

LR test: independent vs. saturated: $\chi^2(10) = 2.2e+05$ Prob> $\chi^2 = 0.0000$

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Uniqueness
v2x_polyar~y	0.9869	0.0261
v2x_libdem	0.9904	0.0192
v2x_delibdem	0.9893	0.0214
v2x_partip~m	0.9857	0.0284
v2x_egaldem	0.9792	0.0411

El análisis factorial de las variables arroja que mediante un factor “Factor 1” es correcto crear un indicador. Esto por su valor Eigenvalue (4.86387) y por su valor Proportion (0,9728=97,7%).

El factor 1 explica la variabilidad conjunta de las variables en un 97%.

Las demás combinaciones (Factor 2, 3, 4 y 5) no son útiles estadísticamente para crear el indicador debido a sus valores Eigenvalue y Proportion.

```
. estat kmo
```

Kaiser-Meyer-Olkin measure of sampling adequacy

Variable	kmo
v2x_polyar~y	0.9290
v2x_libdem	0.9122
v2x_delibdem	0.9231
v2x_partip~m	0.9352
v2x_egalDEM	0.9563
Overall	0.9308

La prueba KMO analiza cada variable que forma parte del factor. En este caso al estar todas sobre 0.92(y 0,9308 en su promedio), son variables apropiadas para realizar construir el indicador.

Resumen

Análisis factorial	
Tipos de Variables	Variables cuantitativas
Análisis a considerar	Eigenvalue Proportion Retained factors
Comando(s) Stata	factor
Observaciones	No exige ninguna distinción entre variables dependientes e independientes.



OTRAS TABLAS

```

1._at      : edad      =      18
2._at      : edad      =      28
3._at      : edad      =      38
4._at      : edad      =      48
5._at      : edad      =      58
6._at      : edad      =      68
7._at      : edad      =      78

```

	Delta-method					
	Margin	Std. Err.	t	P> t	[95% Conf. Interval]	
_at						
1	397357.3	4721.193	84.16	0.000	388103.8	406610.7
2	427244.9	3583.598	119.22	0.000	420221.1	434268.7
3	457132.6	2690.653	169.90	0.000	451859	462406.3
4	487020.3	2341.362	208.01	0.000	482431.3	491609.4
5	516908	2751.229	187.88	0.000	511515.7	522300.4
6	546795.7	3674.435	148.81	0.000	539593.9	553997.5
7	576683.4	4824.791	119.53	0.000	567226.9	586139.9

Comando **margins** realiza estimaciones en una variable. En este caso realizó estimación de ingreso para 18 años(1), 28 años(2), 38 años(3), 48 años(4), 58 años(5), 68 años(6) y 78 años(7).

Es decir para 18 años estima un ingreso de 397357.3 pesos. En tanto para 58 años, estimó un ingreso 516906.

Duda Sesión

- ¿Coeficiente de V de Cramer y prueba χ^2 son excluyentes?, ¿Puede darse un caso con bajo de asociatividad (V de Cramer) y diferencias significativas (χ^2)?.

No son excluyentes, puede darse el caso en que existe una baja asociatividad con diferencias estadísticamente significativas en la prueba χ^2 . Aunque es poco probable que la asociatividad sea nula (0), pero sí puede darse el caso en que esta asociatividad sea débil junto con la otra condición. (Consulté la duda con el profe)