

Universidad Católica del Uruguay
Facultad de Ciencias Humanas

Autotexto

Serie Estadística

Comparación de medias:

ANOVA



Universidad
Católica

DAMASO A. LARRAÑAGA • URUGUAY

Laboratorio Metodológico

Versión original:

M. J. Álvarez – J. Bogliaccini – D. Gelber – F. Rodríguez

Autores revisión 2008:

A. de León – M. Dodel – C. Rafaniello

Módulo de Práctica de Análisis

Serie Estadística

Tema: Comparación de medias

Descripción

Este documento muestra cómo se puede analizar la asociación entre una variable categórica y una interval a través de un conjunto de procedimientos diferentes:

- Una **comparación de medias**, de utilidad básicamente **descriptiva**
- Un **análisis de varianza** de una vía, que puede aplicarse a **variables categóricas dicotómicas o politómicas**

Como ejemplo para esta Guía, utilizaremos la Encuesta Continua de Hogares (ECH) que realiza el Instituto Nacional de Estadística (INE). Dicha encuesta se practica en todo el país urbano de forma continua, es decir a lo largo de todo el año, desde hace más de 20 años consecutivos, proporcionando información para el cálculo de indicadores de la actividad laboral, de los ingresos de las personas y los hogares, sobre los años de educación alcanzados por las personas, etc. Para mayor información acerca de la metodología de dicha encuesta puede consultarse:

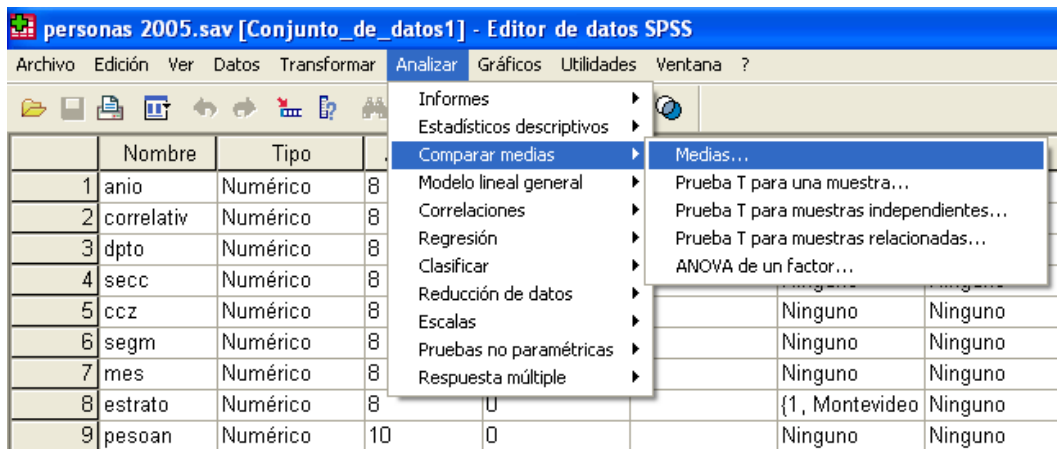
<http://www.ine.gub.uy/biblioteca/metodologias/ech/metodologiaech.htm>

A modo de ejemplo analizaremos la asociación entre la variable interval ingreso por todos los conceptos y la variable ordinal nivel educativo.

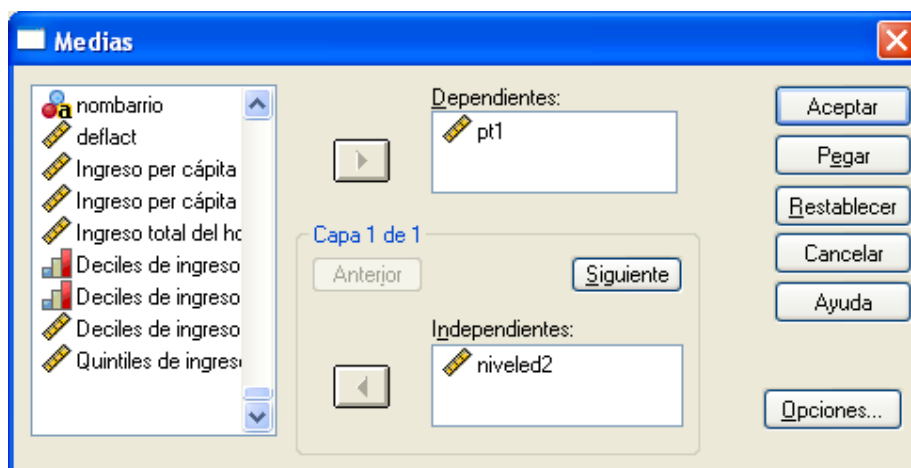
Para el caso de los procedimientos que utilizaremos en esta guía es necesario recodificar¹ la variable nivel educativo agrupando las categorías: “sin instrucción” junto con “primaria”, “secundaria hasta 3 años” junto con “secundaria de 4 a 6 años” y junto a “enseñanza técnica”, por último “magistrado o profesorado” junto con “universidad o similar”

¹ Ver autotexto Recodificar y calcular

¿Cuál es la media de ingreso para los diferentes niveles educativos?



El procedimiento de comparación de medias llama "dependientes" a las variables intervalales de las que se calcula la media, e "independientes" a las variables, generalmente categóricas, que definen los grupos para los cuales se calculan las medias de las variables intervalales:



Para que el procedimiento no resulte meramente descriptivo, a través de las "Opciones" podemos ordenar una tabla de análisis de varianza, que proporciona una prueba de significación.



Todo lo anterior resulta en la siguiente sintaxis, que una vez ejecutada genera los resultados que se presentan en la página siguiente:

```
MEANS
  TABLES=pt1 BY niveled2
  /CELLS MEAN COUNT STDDEV
  /STATISTICS ANOVA .
```

Report

Ingresos por todos los conceptos

Nivel educativo	Mean	N	Std. Deviation
Primaria o menos	2339.65	1125465	3555.67
Secundaria hasta 3 años	4244.44	970477	6435.60
Universidad o similar	11184.21	269550	15944.25
Total	4128.97	2365492	7692.62

En la tabla anterior podemos observar que la media de ingresos es mayor a medida de que aumenta el nivel educativo. ¿Será significativa la diferencia entre las medias según nivel educativo? La tabla de análisis de varianza me permite responder a esta pregunta.

El análisis de varianza

El análisis de varianza, como su nombre lo sugiere, separa o particiona la varianza en dos partes:

- La varianza de ingresos entre los grupos (between groups) o inter-grupos, en este caso entre los distintos niveles educativos.
- La varianza dentro de cada uno de los grupos (within groups) o intra-grupos. En este caso refiere a la distribución de ingresos al interior de cada grupo de nivel educativo.

Luego compara qué proporción de la varianza es varianza explicada y no explicada. Cuanto mayor sea el peso de la varianza explicada en el total, mayor será la certeza de que las medias son significativamente distintas. Dicho en otros términos, cuanto más homogéneos sean los grupos al interior, y más heterogéneos sean entre sí, mayor seguridad tendré de que las variables están asociadas.

Como el lector recordará, la varianza se calcula en base a:

a) las diferencias de los valores individuales con respecto a la media;

$$x_i - \bar{x}$$

b) elevadas al cuadrado (para evitar que las diferencias negativas compensen a las positivas y el resultado final de la suma sea igual a 0);

$$(x_i - \bar{x})^2$$

c) sumadas;

$$\sum (x_i - \bar{x})^2$$

d) y divididas sobre el número de casos de la muestra menos 1. La fórmula de cálculo sería la siguiente:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

La cantidad obtenida en el paso c) recibe el nombre de "suma de cuadrados" (de las desviaciones con respecto a la media). Si conocemos el dato de la varianza y el número de casos, podemos calcularla.

La tabla de análisis de varianza muestra esta "suma de cuadrados" total, la cual luego particiona entre a) la que puede atribuirse al nivel educativo (entre los grupos) y b) la que no puede atribuirse a dicha variable (varianza dentro de los grupos). En nuestro ejemplo, el valor de la primera es 1,70E+13 y el de la segunda es 1,23E+14.

A su vez, el procedimiento tiene en cuenta el número de grados de libertad (*degrees of freedom*, abreviados *df*) de cada uno de los componentes de la suma de cuadrados.

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
Ingresos por todos los conceptos * Nivel educativo recodificado	Between Groups (Combined)	1.70E+13	2	8.517E+12	163861.0	.000
	Within Groups	1.23E+14	2365489	51975583.58		
	Total	1.40E+14	2365491			

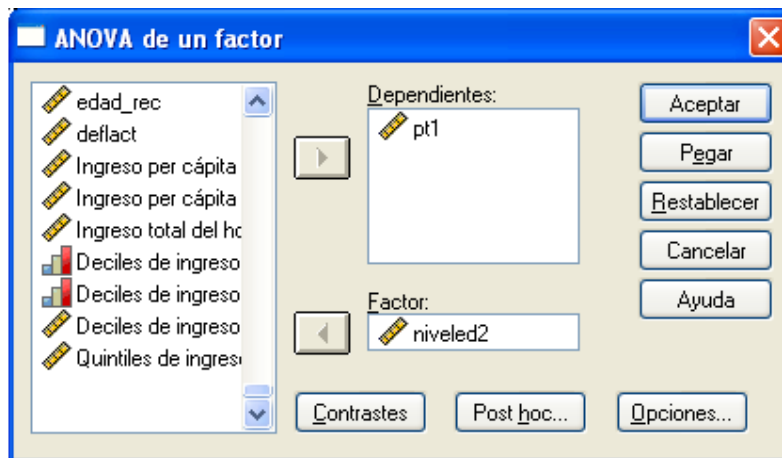
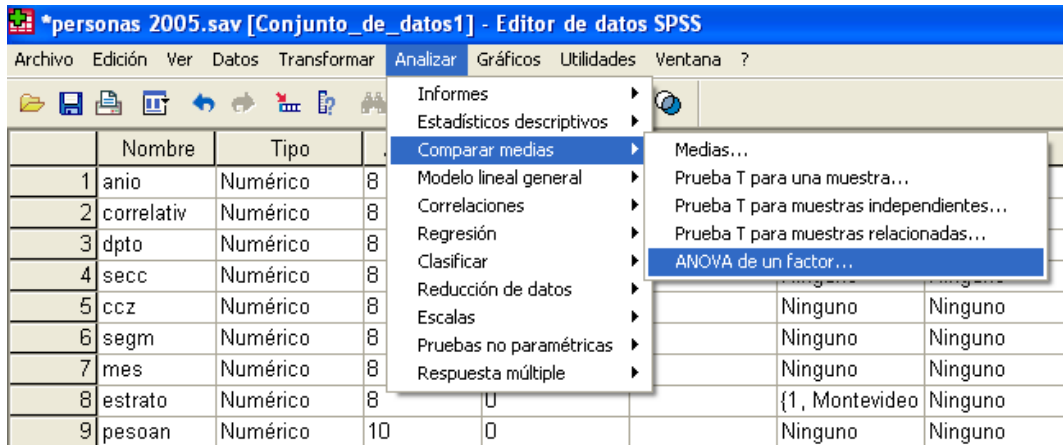
La razón de F se calcula dividiendo el cuadrado medio de la varianza explicada por el de la varianza no explicada. Como en toda prueba de hipótesis, éste valor empírico de F se compara con el valor teórico para cierto nivel de confianza (de una tabla de distribución F).

En este caso, la probabilidad de que las varianzas entre los grupos sean tan distintas por producto solo del azar es muy baja (0.000). El valor p es menor que el valor alfa 0.05 o 0.01. Por lo tanto, rechazamos la hipótesis nula: igualdad de varianzas.

Cuando la significación asociada a F es > a 0.05 no podemos afirmar que existan diferencias entre las medias de los grupos o categorías de comparación.

Pero aún nos queda saber algo....cuál o cuáles medias son las que se distinguen del resto.

Hasta ahora solo sabemos que por lo menos alguno de los grupos es diferente a los otros, pero no sabemos cuál o cuáles. Para eso necesitamos ir más allá de F...necesitamos un **test de post-estimación, o post hoc**. Hay varios



```

ONEWAY
  pt1 BY niveled2
/MISSING ANALYSIS
/POSTHOC = LSD ALPHA(.05).

```

El DMS es uno de los posibles tests que podemos usar. Si elegimos el LSD (least significant difference) o DMS (diferencia mínima significativa), obtendremos un output como el siguiente: que compara cada grupo con los otros dos.

Multiple Comparisons

Dependent Variable: Ingresos por todos los conceptos

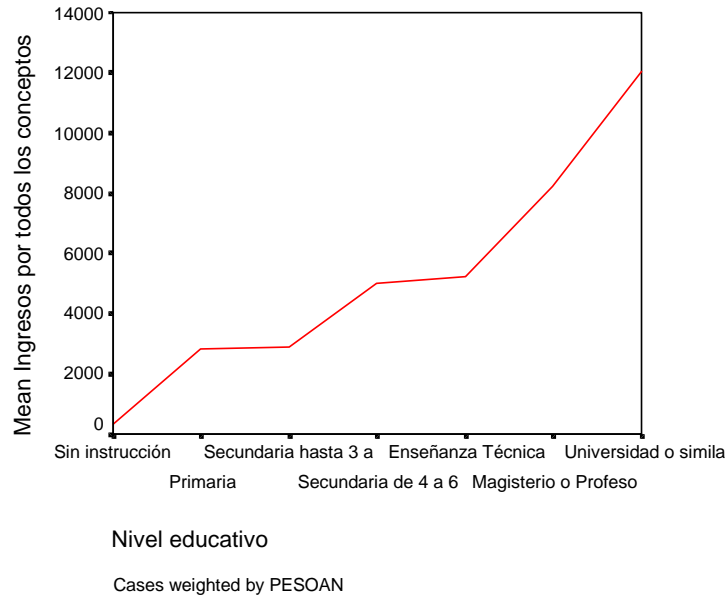
LSD

(I) Nivel educativo recodificado	(J) Nivel educativo recodificado	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Primaria o menos	Secundaria hasta 3 años	-1904.79*	9.99	.000	-1924.37	-1885.22
	Universidad o similar	-8844.56*	15.46	.000	-8874.86	-8814.26
Secundaria hasta 3 años	Primaria o menos	1904.79*	9.99	.000	1885.22	1924.37
	Universidad o similar	-6939.77*	15.70	.000	-6970.53	-6909.01
Universidad o similar	Primaria o menos	8844.56*	15.46	.000	8814.26	8874.86
	Secundaria hasta 3 años	6939.77*	15.70	.000	6909.01	6970.53

*. The mean difference is significant at the .05 level.

En este caso, todos los grupos son significativamente diferentes respecto a los otros. Si no fuera así, la significación de alguna de las comparaciones sería mayor a 0.05.

Podemos pedir un gráfico de las medias, para poder visualizar las diferencias entre las medias de los distintos grupos.



Observando el cuadro se puede concluir que las distintas categorías educativas tienen diferencias significativas en sus medias de ingreso.

Del mismo modo, si se observa el gráfico se puede determinar que a medida que aumenta el nivel educativo, aumenta la media de ingresos.

Ejercicios:

1. Reproduzca la remodificación de nivel educativo de este autotexto para la ECH más reciente disponible en el laboratorio y replique el análisis. Describa similitudes y diferencias en el análisis.

2. Formule una hipótesis de comparación de medias entre más de dos grupos para dos variables que usted quiera de la ECH (recuerde que una de las variables deberá ser de tipo interval- la dependiente- y otra de tipo nominal –la independiente o la que define los grupos. Esta última tendrá que tener en lo posible más de dos categorías para poder usar los conocimientos específicos adquiridos en este autotexto. Formule la hipótesis alternativa estadísticamente y en prosa. Utilice ANOVA para probarla. Escriba su conclusión.

3. Utilizando la base de datos de la OEA “La Niñez y las Familias en la Americas”

1. Recodifique la variable “idh” en tres categorías donde la primera abarca los países con menos de 0,653, la segunda son los países entre valores entre 0,654 y 0,776 y finalmente la tercer categoría los mayores a 0,777.

2. Una vez recodificada la variable, analice el cuadro resultante de la comparación de medias ANOVA, entre la variable de idh recodificada y el porcentaje de metas logradas por país (metpor).
3. A la luz de los resultados del punto 2 vuelva a realizar la comparación de medias, esta vez introduciendo en el análisis el test de comparación de medias al interior de la variable (LSD)

Bibliografía de referencia

- Blalock, H. (1966) Estadística Social. FCE. México.
 - Peña, D. Romo, J - Introducción a la Estadística para las Ciencias Sociales. Mc Graw Hill 1997.
 - Mason y Lind - Estadística para administración y economía. Alfaomega 1998. México, D.F. 8ª edición.
-