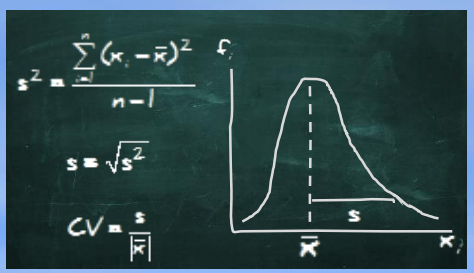
# METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli




UAB
Universitat Autònoma de Barcelona

# METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

# Índice general

# Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

## PARTE III. ANÁLISIS

## Capítulo III.8
## Análisis de varianza

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2016). Análisis de varianza. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. 1ª edición. Edición digital: http://ddd.uab.cat/record/163568.

Capítulo acabado de redactar en septiembre de 2016

# Índice de contenidos

PARTE III
Capítulo 8

# Análisis de varianza

En los dos capítulos anteriores hemos tratado los procedimientos de análisis de la relación entre variables cualitativas, a través de tablas de contingencia y el log-lineal. En este capítulo y en el siguiente siguientes veremos cómo analizar la relación entre variables donde intervienen variables cuantitativas. Con el título de análisis de varianza recogemos diversas técnicas de análisis que nos permiten relacionar variables cuantitativas con variables cualitativas a partir de relaciones de dependencia. Con el análisis de regresión, en el próximo capítulo, veremos cómo relacionar entre sí variables cuantitativas a partir también de modelos explicativos.

En este tema se proporciona una presentación del análisis de varianza con un carácter introductorio para conocer sus características principales y orientar un seguimiento posterior. En esta introducción se comienza con los conceptos más básicos que se derivan de la prueba estadística del contrate de una media, que vimos en el Capítulo III.4, y de la prueba estadística de la diferencia entre dos medias que están en la base de la visión más amplia que implica la técnica de análisis de varianza donde se comparan más de dos y se plantean aspectos adicionales sobre los modelos de relación entre las variables. Tras presentar una panorámica de los distintos diseños de análisis que se pueden formalizar, nos centraremos en el análisis de varianza unifactorial y multifactorial.

Tanto el análisis de varianza como el análisis de regresión son métodos basados en el modelo estadístico de relaciones lineales entre variables, de donde se derivan también otras técnicas de análisis, como el análisis log-lineal que hemos visto en el tema anterior. En este sentido comparten planteamientos similares de un mismo modelo lineal con una formulación específica. Ya comentamos que la perspectiva de los modelos lineales generalizados concibe a estas distintas técnicas como casos particulares de un mismo modelo general. El modelo lineal general (*Ecuación 1*) plantea la existencia de una variable dependiente $Y_i$ que es expresada como una función lineal de un número de términos que representan a las variables independientes $X_j$, cada una de las cuales se multiplica por un coeficiente $\beta_j$, que valora la importancia del efecto de cada variable independiente sobre la variable dependiente. Además cabe considerar un término de error $\varepsilon_i$ que representa los efectos de todas las fuentes aleatorias de variación que no se han tenido en cuenta o que son desconocidas.

Así, el modelo lineal se expresa de la siguiente forma:

$$Y_i = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

*Ecuación 1*

*Variable dependiente* — *Constante* — *Variables independientes* — *Error*

El análisis de varianza, también denominado ANOVA como abreviación de *ANalysis Of VAriance*, está destinado a examinar la relación de dependencia entre variables cuantitativas en relación a variables cualitativas. El modelo de la varianza es un análisis dependencia entre dos conjuntos de variables: la o las variables dependientes, que se consideran como explicadas, se miden en una escala cuantitativa, mientras que las variables independientes son cualitativas, si bien también existe la posibilidad de introducir, como veremos, variables independientes de control cuantitativas. Se trata de un método de análisis estadístico que permite probar hipótesis sobre si existen diferencias significativas de una característica observada, o varias de ellas, medidas con variables cuantitativas, entre los diferentes grupos formados a partir de las categorías de una o de más variables cualitativas.

En un análisis de varianza donde se consideran uno o más factores para explicar la variabilidad de una variable dependiente podríamos responder por ejemplo a cuestiones, expresadas en términos de hipótesis, como las siguientes: ¿existen diferencias significativas entre los ingresos de hombres y mujeres?, ¿o entre clases sociales altas y bajas?, ¿dependen del nivel educativo? ¿del origen social de los padres? ¿de la antigüedad en el mercado de trabajo? ¿es diferente entre sectores productivos o entre zonas geográficas? Si analizáramos el nivel de consumo de un determinado producto nos podríamos preguntar si está determinado por la clase social, por el sexo, por la edad, por el nivel de ingresos, por el nivel educativo. Si estudiamos resultados académicos de los estudiantes/as universitarios nos podemos interrogar si éstos dependen de si trabajan o no, de su origen social, del lugar donde estudiaron, del área de conocimiento, del género, de la motivación. En todos estos casos podemos hacer uso del análisis de varianza para dar respuesta a nuestras hipótesis.

La aplicación de este método de análisis responde a estas preguntas en base a la comparación de las medias observadas de la variable cuantitativa, por el conjunto de la muestra y para cada grupo definido por la o variables independientes consideradas en el modelo, y en una prueba de hipótesis basada en el cálculo de la varianza entre estas medias. Miraremos hasta qué punto las medias calculadas en cada grupo se dispersan, se alejan de la media global y nos permiten hablar de diferencias significativas entre los grupos, o, por el contrario, hasta qué punto esta variabilidad no es suficientemente importante para afirmar que los grupos se comportan de manera diferente. Por tanto, se trata de ver, según un modelo de hipótesis, cuáles son los factores o los determinantes de la variabilidad de los ingresos, del consumo, de los resultados académicos; de si las variables independientes como el sexo, la clase social social, el nivel educativo, etc., consideradas por separado (análisis de un solo factor) o conjuntamente (análisis multifactorial), son fuentes de la variación de la variable dependiente, explican la existencia de las variaciones, de las diferencias.

Introduciremos a continuación algunos conceptos generales del análisis de varianza que son característicos de la terminología utilizada con este tipo de método y en los distintos diseños de análisis que se pueden plantear, los cuales tienen como referente a los diseños experimentales. A continuación, introduciremos el ejercicio básico y fundamental de un análisis descriptivo de comparación de las medias entre los distintos grupos. Veremos que no es más que reproducir el análisis descriptivo de una variable cuantitativa, como vimos en el Capítulo III.3, repetido dentro de cada grupo y luego comparados entre sí. Junto al aspecto descriptivo introduciremos el aspecto inferencial, primero comparando solamente dos medias y aplicando la prueba estadística correspondiente. Luego generalizando este tipo de análisis a una técnica de análisis de dependencia como es el análisis de varianza, en su versión más sencilla, análisis de varianza unifactorial, y en un modelo más complejo: el análisis de varianza multifactorial, donde se consideran dos o más variables independientes.

## 1. Conceptos generales y terminología

Es habitual denominar a la variable dependiente como variable explicada o resultado (*outcome*) mientras que a cada una las variables independientes se la denomina factor, es decir, es cada una de las "causas" o "factores" que explican la heterogeneidad (variabilidad) de la variable dependiente, según establecemos en una hipótesis de nuestro modelo de análisis. Cada valor o categoría de la variable independiente se denomina tratamiento o nivel del factor configurando los grupos de individuos que son los que motivan el comportamiento diferenciado de la variable dependiente.

Si las variables independientes determinan, afectan, a la dependiente la medida de la influencia de cada una de ellas se denomina efecto ($\alpha_j$) y la importancia de su magnitud como tamaño del efecto. Se suele utilizar esta expresión: $\alpha_j = \mu_j - \mu$ para referirse al efecto de la categoría *j* de una variable independiente como la diferencia entre la media global (los ingresos de toda la muestra) y la media del grupo (del grupo de mujeres, por ejemplo).

Cuando se consideran varias variables independientes cada una de ellas tiene un efecto principal, es decir, un efecto individual, pero pueden existir efectos adicionales como resultado de la interacción entre los factores o variables independientes.

La o las variables factor se pueden caracterizar por ser de efectos fijos, es decir, se incluyen o interesan los niveles de la variable independiente sobre los que se desea extraer conclusiones. El investigador/a fija o controla los niveles (en un contexto experimental) o bien vienen dados por las características del factor y se consideran todos (por ejemplo, si se considera la variable sexo, los niveles son varón y mujer) o elije algunos de ellos, los que se consideran relevantes para el estudio (por ejemplo, los municipios del ámbito metropolitano). También pueden ser efectos aleatorios, en este caso los niveles de un factor aleatorio son una muestra aleatoria de los posibles niveles, son tantos los valores posibles que se eligen unos pocos y luego se extrapolan los resultados como representativos de los demás (por ejemplo, se seleccionan unos pocos municipios al azar). A los modelos que emplean efectos aleatorios se les denomina como modelo II, frente al modelo I que sería de efectos fijos.

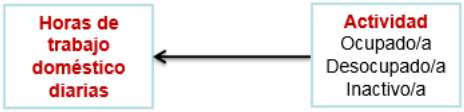
Además, se pueden considerar covariables, es decir, variables predictoras cuantitativas que se relacionan con la variable dependiente (o respuesta) y que se consideran con una función de control estadístico de la variable dependiente. Este planteamiento corresponde al llamado análisis de covarianza (ANCOVA), y se trata de contrastar si los resultados de una ANOVA se ven alterados cuando se introduce la variable de control.

Se pueden contrastar tanto los modelos equilibrados o balanceados, cuando se tiene igual número de casos por grupo o celda, definida por cada valor de la variable independiente o combinación de valores de varios factores. Si es diferentes los modelos son no equilibrados.

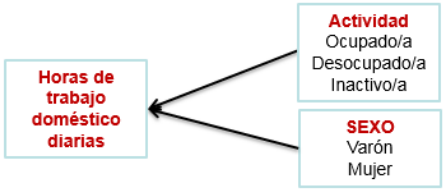
Los modelos de un análisis de varianza intentan ser explicativos. La parte no explicada por nuestro modelo a partir de nuestros datos muestrales constituirá el error ($\varepsilon_i$) que pueden deberse a diversas razones: errores de medida, efectos aleatorios, variables relevantes no incluidas en el modelo, etc.

Los modelos o diseños particulares que se pueden contemplar en un análisis de varianza los esquematizamos y ejemplificamos seguidamente.

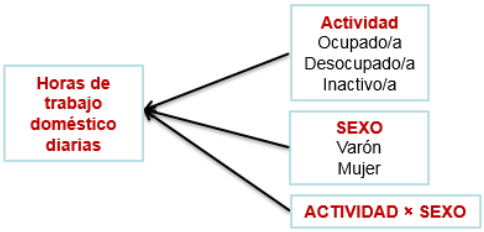
1) ANOVA unifactorial: se considera una variable dependiente cuantitativa (univariable) y una variable independiente cualitativa (unifactorial, de una sola vía: *oneway*) entre sujetos (diseño de muestras independientes).



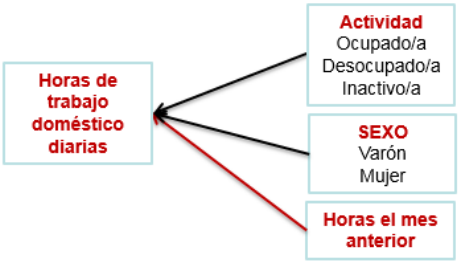

2) ANOVA multifactorial: se considera una variable dependiente cuantitativa (univariable) y dos o más variables independientes cualitativas (multifactorial, de dos o más vías) entre sujetos (diseño de muestras independientes).
El diseño puede ser sin interacción:



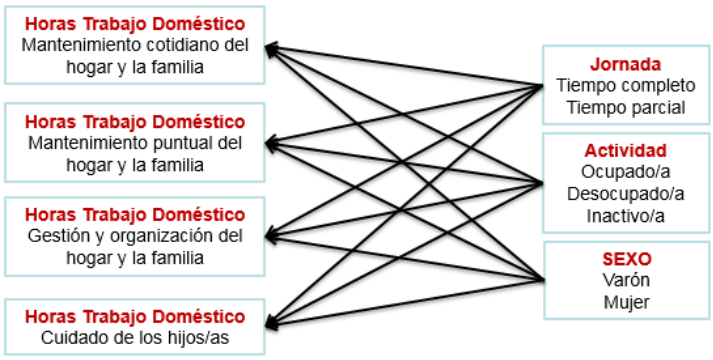

O con interacción de las variables independientes:



3) Análisis de covarianza (ANCOVA): se considera una variable dependiente cuantitativa (univariable) y dos o más variables independientes cualitativas (multifactorial, de dos o más vías) con una variable independiente adicional de control medida a nivel cuantitativo y correlacionada con la variable dependiente, entre sujetos (diseño de muestras independientes). El diseño puede ser con o sin interacción.




4) Análisis multivariable de varianza (MANOVA): se considera más de una variable dependiente cuantitativa (multivariable) y dos o más variables independientes cualitativas (multifactorial, de dos o más vías) sin una variable independiente adicional de control (MANOVA) o con ella (MANCOVA), entre sujetos (diseño de muestras independientes). El diseño puede ser con o sin interacción.




Otros diseños en lugar de tener muestras independientes (inter sujetos) consideran muestras apareadas o relacionadas (intra sujetos), es decir, se disponen, para los mismos individuos, de diversas medidas o variables de las que se obtienen las medias. El diseño se denomina de medidas repetidas: podemos medir la misma característica a los mismos sujetos en diversos momentos en el tiempo, por ejemplo, la valoración de un producto de consumo a lo largo de cuatro trimestres, o sería también el caso de esa valoración realizada por los mismos sujetos de cuatro productos distintos. Si la valoración de los productos se realizara sobre, por ejemplo, cuatro grupos de sujetos distintos elegidos al azar, el diseño de denominaría completamente aleatorizado.

Cuando se realiza un análisis de varianza donde las variables independientes están anidadas, es decir, se relacionan o están estructurados de manera jerárquica el diseño de denomina de componentes de la varianza. En esta situación las muestras de cada factor son tomadas del interior de las muestras del factor inmediatamente superior a él.

Cuando se consideran factores no relacionados entre ellos el diseño se llama ortogonal; es no ortogonal cuando las variables independientes están asociadas, interaccionan.

Tras este repaso por los distintos elementos que caracterizan a un diseño con el que realizar un análisis de varianza nos centraremos en las páginas siguientes en el ANOVA unifactorial y multifactorial. Como comentábamos al inicio, la formulación del análisis de varianza es una extensión o una generalización de la prueba de la diferencia entre dos medias, la cual es a la vez una extensión de la prueba estadística de una sola media que se vio en el Capítulo III.4 (apartado 2.5). Antes de pasar a presentar ese contenido presentaremos en el apartado siguiente el necesario análisis descriptivo preliminar que se deriva de la comparación de medias a través de tablas y gráficos.

## 2. Análisis descriptivo de la comparación de medias

En el análisis de la relación entre una variable cuantitativa y otra (u otras) cualitativa(s) podemos emplear tablas de medias y representaciones gráficas, como los gráficos de medias o los diagramas de caja. Se trata de evaluar la información en un análisis exploratorio del comportamiento de los datos y descriptivo para observar inicialmente si existen diferencias entre los grupos en cuanto a la distribución de la variable cuantitativa de interés y en cuanto a los descriptivos más relevantes (posición, dispersión o forma). Es una primera forma de evidenciar los resultados obtenidos con nuestros datos y dar cuenta del contenido de la hipótesis de relación entre las variables.

Por ejemplo en el análisis de la relación entre el número de horas dedicadas semanalmente a las tareas del hogar según la edad o el estado civil, con datos de la *Encuesta de Condiciones de Vida de Catalunya* del año 2006, se observa el comportamiento del promedio de horas dentro de cada grupo de edad (Tabla III.8.2Tabla III.8.2) y de cada estado civil (Tabla III.8.1), constatando que son valores que difieren entre los grupos.

Tabla III.8.1. Horas dedicadas a las tareas del hogar según la edad del entrevistado/a

| Edad | Media | Frecuencia | Desviación típica |
|---|---|---|---|
| 1 16-25 | 6,25 | 1.231 | 7,663 |
| 2 26-35 | 12,79 | 2.138 | 12,810 |
| 3 36-45 | 16,36 | 1.911 | 13,542 |
| 4 46-55 | 17,49 | 1.512 | 15,093 |
| 5 56-65 | 17,90 | 1.334 | 15,793 |
| 6 66-75 | 18,94 | 1.003 | 14,945 |
| 7 76-85 | 15,74 | 721 | 13,489 |
| 8 Más de 85 | 8,10 | 181 | 12,180 |
| Total | 14,80 | 10.031 | 14,066 |

Fuente: Enquesta de Condicions de Vida i Hàbits de la Població de Catalunya 2006.

A medida que aumenta la edad se incrementa la dedicación a las tareas del hogar pasando de una media 6,2 entre las personas de 16 a 25 años a las 18,9 de la franja entre 66 y 75, bajando finalmente en los dos últimos tramos etarios. En el caso del estado civil se pasa de 8,5 horas de las personas solteras a las 17,2 cuando se está casado

o las 18,8 horas en el caso de los viudos/as. Los separados y divorciado bajan algo respecto de estos valores medios.

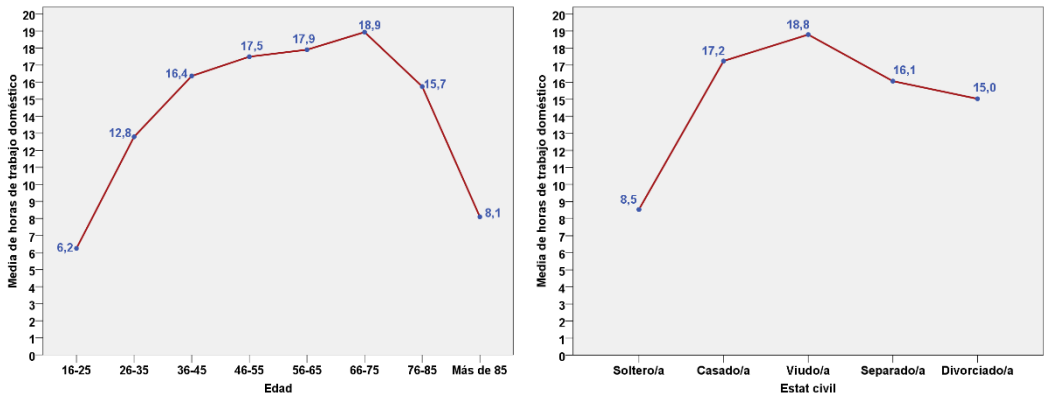
Tabla III.8.2. Horas dedicadas a las tareas del hogar según el estado civil del entrevistado/a

| Estado civil | Media | Frecuencia | Desviación típica |
|---|---|---|---|
| 1 Soltero/a | 8,54 | 2.873 | 9,477 |
| 2 Casado/da | 17,24 | 5.944 | 15,075 |
| 3 Viudo/a | 18,79 | 778 | 14,498 |
| 4 Separado/a | 16,06 | 256 | 11,229 |
| 5 Divorciado/a | 15,03 | 181 | 10,451 |
| Total | 14,80 | 10.031 | 14,066 |

Fuente: Enquesta de Condicions de Vida i Hàbits de la Població de Catalunya 2006.

Estas tendencias que muestran los datos se pueden representar gráficamente a través de un gráfico de líneas con los valores de las medias, como aparece en el Gráfico III.8.1.

Gráfico III.8.1. Gráficos de medias de las Tablas III.8.1 y III.8.2




En este tipo de análisis se pueden calcular los distintos estadísticos característicos de un ejercicio de lectura descriptivo y exploratorio, como se muestra en la Tabla III.8.3, en este caso comparando la variable de horas dedicadas a las tareas del hogar según el sexo. Los datos evidencian la desigual distribución del tiempo y las actividades del trabajo reproductivo entre ambos sexos. Mientras que los varones dedican una media de 8 horas a la semana a las tareas domésticas las mujeres casi lo triplican con una media de 21 horas semanales. Vemos además que la desviación típica o estándar es el doble en el caso de las mujeres, poniendo de manifiesto el mayor nivel de dispersión de los datos, entre valores bajos y altos alrededor de la media. Los varones tienden a concentrarse más, en los valores inferiores en particular. Este rasgo se puede apreciar más claramente en el diagrama de caja del **¡Error! No se encuentra el origen de la referencia.**. Las horas de dedicación de las mujeres se distribuyen a lo largo de un rango de valores mayor, hasta un máximo declarado de 98 horas. En el caso de los varones se concentran en valores bajos de forma que una dedicación de 35 horas

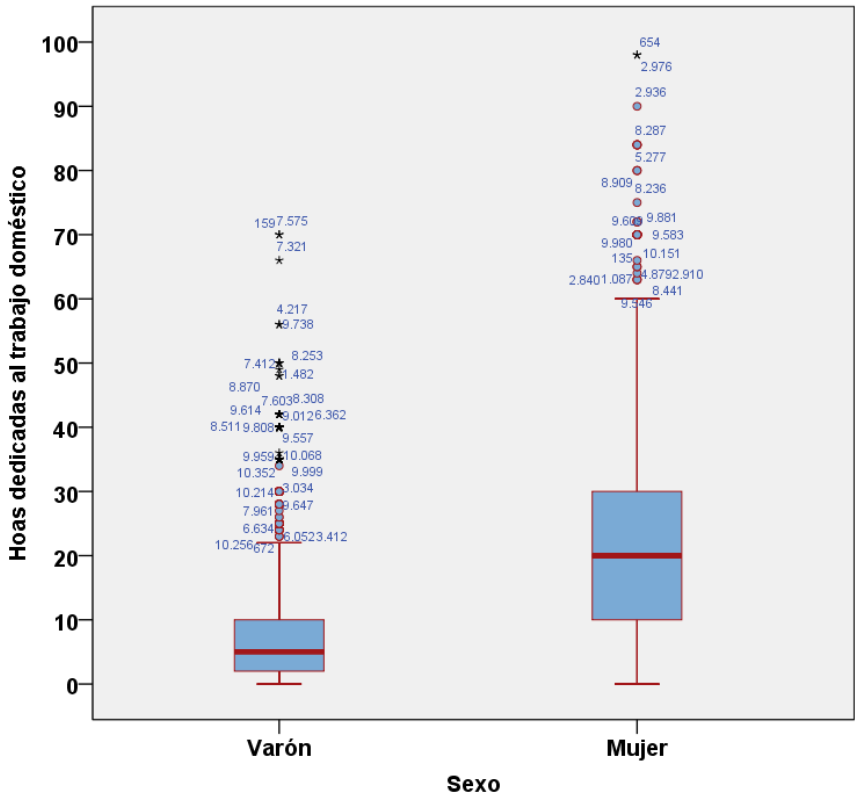
semanales se convierte en un valor extremo severo (identificado con el asterisco *), situación nada extraña o extrema en el caso de una mujer.

Tabla III.8.3. Estadísticos de las horas dedicadas a las tareas del hogar según el sexo

| | | Sexo | | | |
|---|---|---|---|---|---|
| | | 1 Varón | | 2 Mujer | |
| | | Estadístico | Error estándar | Estadístico | Error estándar |
| Media | | 7,97 | 0,118 | 21,35 | 0,214 |
| 95% de intervalo de confianza para la media | Límite inferior | 7,74 | | 20,94 | |
| | Límite superior | 8,20 | | 21,77 | |
| Media recortada al 5% | | 7,03 | | 20,40 | |
| Mediana | | 5,00 | | 20,00 | |
| Varianza | | 68,657 | | 234,156 | |
| Desviación típica | | 8,286 | | 15,302 | |
| Mínimo | | 0 | | 0 | |
| Máximo | | 70 | | 98 | |
| Rango | | 70 | | 98 | |
| Rango intercuartil | | 8 | | 20 | |
| Asimetría | | 1,868 | 0,035 | 0,885 | 0,034 |
| Curtosis | | 5,003 | 0,070 | 0,761 | 0,068 |

Fuente: Enquesta de Condicions de Vida i Hàbits de la Població de Catalunya 2006.

Gráfico III.8.2. Diagrama de caja de las horas dedicadas a las tareas del hogar según el sexo



Consideremos un segundo ejemplo relacionando los ingresos individuales, de toda la población (mayores de 18 años), con el nivel de estudios, según los datos del estudio del CIS 3041.[1]. Los datos de la tabla de frecuencias (Tabla III.8.4) y del histograma (Gráfico III.8.3) evidencias la concentración de la población en los niveles inferiores, con una media global de unos 760€ mensuales, y una desviación típica de 765.

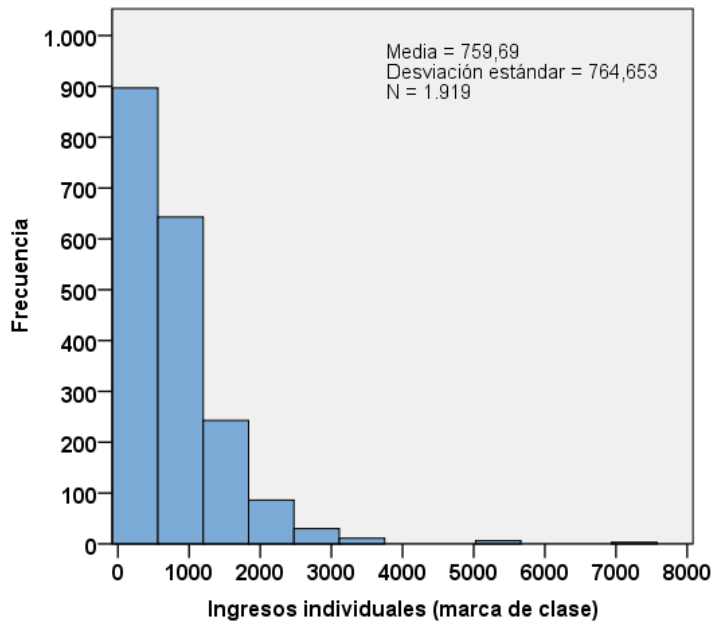
Tabla III.8.4. Distribución de frecuencias de los ingresos individuales

P46m Ingresos individuales (marca de clase)

| | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | 0 | 484 | 19,5 | 25,2 | 25,2 |
| | 150 | 75 | 3,0 | 3,9 | 29,1 |
| | 450 | 338 | 13,6 | 17,6 | 46,7 |
| | 750 | 340 | 13,7 | 17,7 | 64,5 |
| | 1050 | 303 | 12,2 | 15,8 | 80,3 |
| | 1500 | 243 | 9,8 | 12,7 | 92,9 |
| | 2100 | 86 | 3,5 | 4,5 | 97,4 |
| | 2700 | 30 | 1,2 | 1,6 | 99,0 |
| | 3750 | 11 | 0,4 | 0,6 | 99,5 |
| | 5250 | 6 | 0,2 | 0,3 | 99,8 |
| | 7500 | 3 | 0,1 | 0,2 | 100,0 |
| | Total | 1919 | 77,4 | 100,0 | |
| Perdidos | 9999 NC | 561 | 22,6 | | |
| Total | | 2480 | 100,0 | | |

Fuente: Centro de Investigaciones Sociológicas, Estudio 3041.

Gráfico III.8.3. Histograma de los ingresos individuales




[1] Estos resultados se pueden reproducir a partir del archivo de sintaxis AVA-Ingresos.sps de la página web.

Si analizamos la distribución de ingresos en función del nivel de estudios (Tabla III.8.5 y Gráfico III.8.4) vemos cómo se pone de manifiesto que un mayor nivel educativo comporta un aumento de los ingresos, sobre todo si se alcanza el nivel universitario. Las diferencias entre los cinco primeros niveles de titulación muestran un orden donde a cada nivel educativo que aumenta se da un un incremento de la media de ingresos (equiparándose la formación secundaria 2ª etapa con la formación profesional), pero no son diferencias tan destacadas como cuando consideramos el nivel de estudios superiores, donde la media se dispara a 1317€.

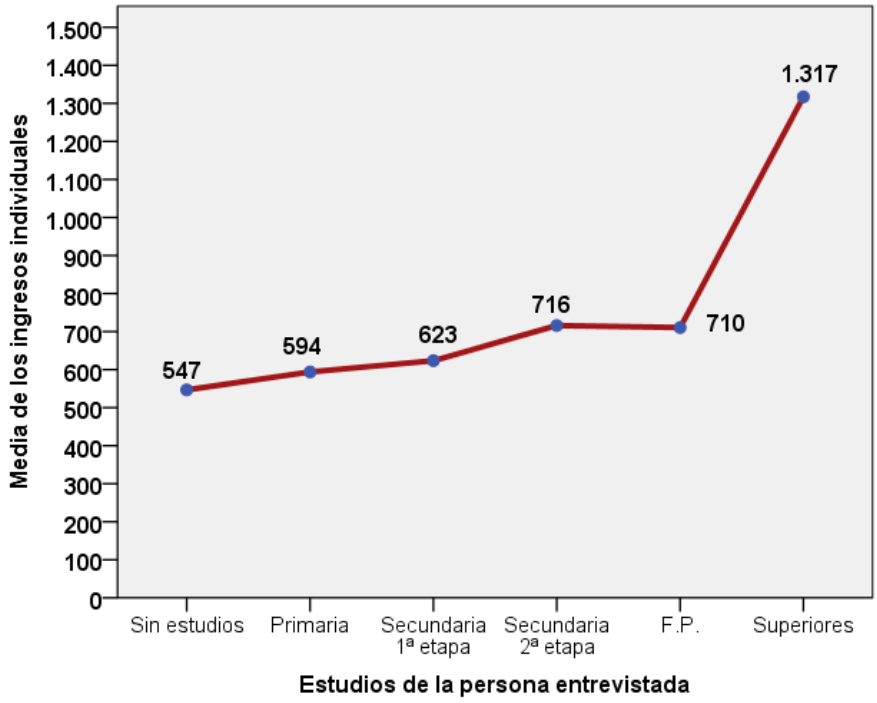
Tabla III.8.5. Ingresos individuales según el nivel de estudios

P46m Ingresos individuales (marca de clase)

| ESTUDIOS Estudios de la persona entrevistada | Media | Frecuencia | Desviación estándar |
|---|---|---|---|
| 1 Sin estudios | 546,73 | 107 | 375,567 |
| 2 Primaria | 593,86 | 391 | 473,937 |
| 3 Secundaria 1ª etapa | 623,39 | 513 | 635,066 |
| 4 Secundaria 2ª etapa | 716,02 | 256 | 716,786 |
| 5 F.P. | 710,38 | 318 | 737,345 |
| 6 Superiores | 1317,02 | 332 | 1070,260 |
| Total | 760,02 | 1917 | 764,925 |

Fuente: Centro de Investigaciones Sociológicas, Estudio 3041.

Gráfico III.8.4. Ingresos individuales medios según el nivel de estudios



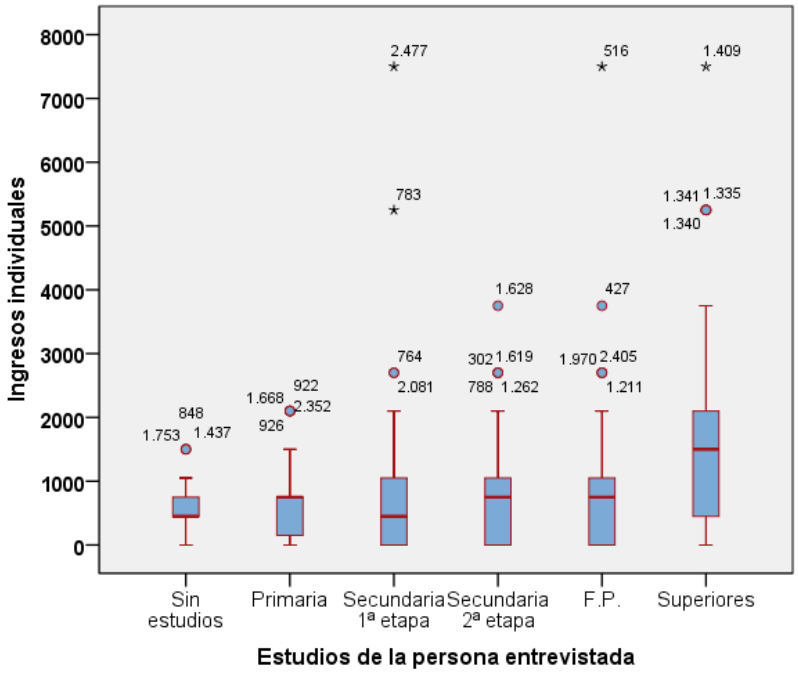
Esta lectura de la relación entre ingresos y estudios se puede complementar con la información de los descripctivos de la Tabla III.8.6 y del diagrama de caja del Gráfico III.8.5.

Tabla III.8.6. Descriptivos de los ingresos individuales según el nivel de estudios

| Descriptivo de P46m Ingresos individuales (marca de clase) | | ESTUDIOS Estudios de la persona entrevistada | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 Sin estudios | 2 Primaria | 3 Secundaria 1ª etapa | 4 Secundaria 2ª etapa | 5 F.P. | 6 Superiores |
| Media | Estadístico | 546,7 | 593,9 | 623,4 | 716,0 | 710,4 | 1317,0 |
| | Error estándar | 36,3 | 24,0 | 28,0 | 44,8 | 41,3 | 58,7 |
| 95% de intervalo de confianza para la media | | | | | | | |
| Límite inferior | Estadístico | 474,7 | 546,7 | 568,3 | 627,8 | 629,0 | 1201,5 |
| Límite superior | Estadístico | 618,7 | 641,0 | 678,5 | 804,2 | 791,7 | 1432,6 |
| Media recortada al 5% | Estadístico | 535,12 | 535,1 | 568,0 | 575,2 | 661,3 | 645,7 |
| Mediana | Estadístico | 450,00 | 450,0 | 750,0 | 450,0 | 750,0 | 750,0 |
| Varianza | Estadístico | 141.050,5 | 141.050,5 | 224.616,1 | 403.309,2 | 513.781,7 | 543.677,5 |
| Desviación estándar | Estadístico | 375,6 | 375,6 | 473,9 | 635,1 | 716,8 | 737,3 |
| Mínimo | Estadístico | 0 | 0 | 0 | 0 | 0 | 0 |
| Máximo | Estadístico | 1.500 | 2.100 | 7.500 | 3.750 | 7.500 | 7.500 |
| Rango | Estadístico | 1.500 | 2.100 | 7.500 | 3.750 | 7.500 | 7.500 |
| Rango intercuartil | Estadístico | 300 | 600 | 1.050 | 1.050 | 1.050 | 1.650 |
| Asimetría | Estadístico | 0,123 | 0,520 | 3,554 | 0,900 | 2,998 | 1,471 |
| | Error estándar | 0,234 | 0,123 | 0,108 | 0,152 | 0,137 | 0,134 |
| Curtosis | Estadístico | -0,358 | 0,038 | 31,157 | 0,621 | 22,354 | 4,619 |
| | Error estándar | 0,463 | 0,246 | 0,215 | 0,303 | 0,273 | 0,267 |

Fuente: Centro de Investigaciones Sociológicas, Estudio 3041.

Gráfico III.8.5. Diagramas de caja de los ingresos individuales según el nivel de estudios



## 3. Prueba estadística del contraste de una media

Recordaremos algunas de las ideas y cálculos que presentamos en el Capítulo III.4. La prueba estadística de una media permite generalizar a la población el valor obtenido en una muestra: se calcula el estadístico muestral de la media $\bar{x}$ que estima el parámetro poblacional de la media $\mu$. Cada estimación del valor verdadero (y desconocido) del parámetro a partir de los estadísticos de la muestra tiene asociado un margen de error respecto del valor real y desconocido del parámetro. El margen de error se define como la diferencia máxima entre la estimación puntual obtenida en la muestra y el valor verdadero del parámetro poblacional, $\bar{x}-\mu$. Cada posible estimación de la media obtenida en cada posible muestra genera un resultado distinto, pero sabemos que todos estos posibles valores que intentan estimar el valor verdadero del parámetro forman lo que se llama una distribución muestral del estadístico. A partir del conocimiento general de esta forma de la distribución del estadístico de la media se puede determinar, con una certeza suficiente, el margen de error que cometemos al hacer estimaciones muestrales. Según el teorema central del límite sabemos que, teniendo en cuenta todas posibles medias $\bar{x}_i$ obtenidas en todas las posibles muestras de tamaño $n$ de una población $N$, a medida que aumenta el tamaño de la muestra y con independencia de la distribución de la variable en la población, las medias muestrales siguen aproximadamente una distribución normal de media $\mu$, la media poblacional, y varianza $\sigma^2/n$, la varianza poblacional de la variable dividida por el tamaño de la muestra.

Se denomina error típico a la desviación del estadístico: $\sigma/\sqrt{n}$, y se considera que la media sigue una distribución normal $N\left(\mu, \sigma/\sqrt{n}\right)$ si n ≥ 30.

Fijado un nivel de confianza, habitualmente el 95,5%, es decir, con $z=2$, el intervalo de confianza de la media es $\bar{x} \pm z \cdot \sigma/\sqrt{n}$, siendo $e = z \cdot \sigma/\sqrt{n}$ el error muestral. Cuando la desviación típica de la variable en la población $\sigma$ es desconocida se sustituye por la desviación típica muestral $s$, y la distribución normal del estadístico se sustituye por la distribución de la t de Student. Así la expresión anterior queda como $\bar{x} \pm t \cdot s/\sqrt{n}$.

Veamos un ejemplo. Consideremos que los ingresos medios de un grupo social de 2.500 personas que han sido entrevistadas en una encuesta de condiciones de vida da como resultado 1.500 € y que la desviación típica muestral es de 1000 €. Si tenemos en cuenta un nivel de confianza del 95,5%[2], el error muestral que cometemos en nuestra estimación es $e = t \cdot s/\sqrt{n} = 2 \times 1500/\sqrt{2500} = 60$. Por tanto, el intervalo donde se encontrará la media poblacional, a partir de $\bar{x} \pm e = 1500 \pm 60$, es (1440,1560).

[2] Cuando la muestra es grande como en este caso con n=2500, la distribución de la t de student alcanza los valores de la normal. En este caso el valor de la $t$ es igual al de la $z$, con un nivel de confianza del 95,5% $z=t=2$. Algunos de los valores teóricos de la distribución de la t de student se presentan en una tabla en el anexo de este capítulo. Para realizar la lectura de la tabla y elegir el valor teórico en cada caso se determina primero el nivel de significación (el complementario del de confianza), por ejemplo del 5%, o bien del 0,05. Si el contraste se realiza a dos colas como en este caso se considera el valor de probabilidad del 0,025. Así determinamos qué columna considerar. Las filas de la tabla se refieren a los grados de libertad, en este caso son n-1=2500-1, un valor muy alto equivalente a un valor infinito, que se localiza en la última fila de la tabla. Podemos observar como el cruce de la fila y la columna citadas permite obtener el valor de 1,96.

Podemos reproducir este ejercicio con los datos de la matriz de datos CIS3041. Los resultados de la denominada prueba de una muestra (*one samples test*) considerando la variable de ingresos del hogar P45m son los de la Tabla III.8.7, utilizando el software R y Deducer.

Tabla III.8.7. Resultados de la prueba de una media con R

**Descriptive Statistics**

| | Mean | St. Deviation | Valid N |
|---|---|---|---|
| P45m | 1500.18 | 1078.10 | 1706 |

**One-Sample Test**

**Method: One Sample t-test**

| | mean of x | 95% CI Lower | 95% CI Upper | t | df | p-value |
|---|---|---|---|---|---|---|
| P45m | 1500.18 | 1448.98 | 1551.37 | 57.47 | 1705.00 | <0.001 |

**Notes:**

HA: two.sided
H0: mean=0

En este caso a partir de una media muestral de 1500,18 € y una desviación típica de 1078,10 € se estima que el valor poblacional de la media se encuentra en el intervalo (1448,98 , 1551,37).

Además de determinar el margen de error y el intervalo de confianza de la estimación se pueden realizar pruebas de hipótesis de carácter univariable donde se testea si la media en la población es igual a un determinado valor que hemos obtenido en la muestra. Remitimos al lector/a a la presentación realizada en el apartado 2.5 del Capítulo III.4.

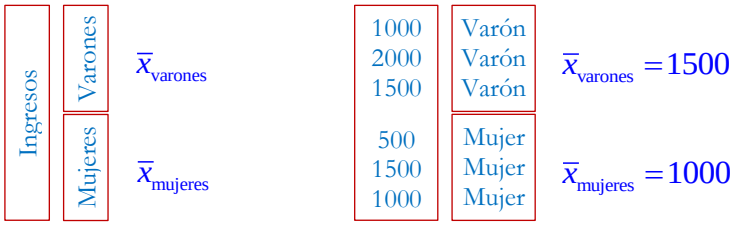
## 4. Prueba estadística del contraste de dos medias

El contraste de dos medias se inscribe en una situación de relación entre dos variables, una cuantitativa de la que se calculan las medias y una cualitativa que define dos grupos que son contrastados y en el interior de los cuales se calculan esas medias. Consiste pues en comparar los resultados obtenidos en la estimación de dos medias de una variable cuantitativa a partir de dos grupos de casos obtenidos en muestras aleatorias y ver si existen diferencias entre ambas determinando si éstas son significativas o las diferencias simplemente se deben al azar. El resultado de esta prueba es similar al que se obtendrá en un análisis de varianza cuando la variable independiente que define los grupos sólo tiene dos valores. Pero en este caso se trata sencillamente de una prueba estadística, la denominada prueba de la t (*t-test*) y en el caso del análisis de varianza se trata de toda una técnica de análisis que permite contratar diferentes tipos de diseños de análisis.

La prueba estadística del contraste de dos medias formula las hipótesis de si éstas son iguales o diferentes. Estas medias se pueden obtener a partir de diseños de análisis diferentes que mostramos a continuación.

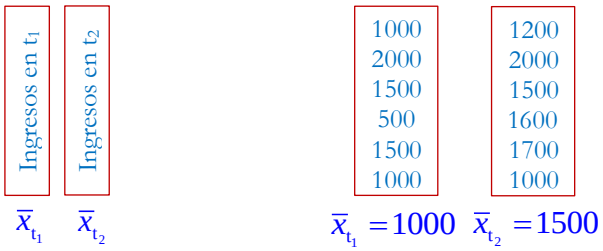
1) Diseño de muestras independientes. En este caso comparamos dos muestras que se obtienen de forma independiente y en donde podemos encontrarnos con dos situaciones que trataremos de la forma similar:
   - O bien se trata de un diseño estrictamente de dos muestras aleatorias independientes con $n_1$ y $n_2$ casos de dos poblaciones diferentes $P_1$ y $P_2$ de donde compramos las dos medias.
   - O bien se trata de un diseño de dos subpoblaciones $P_1$ y $P_2$ de una sola población $P$ y disponemos de dos submuestras aleatorias e independientes con $n_1$ y $n_2$ casos en donde obtenemos las dos medias que se comparan.

   Es un diseño que se puede identificar con la expresión entre-sujetos. Por ejemplo, consideramos la variable cuantitativa ingresos y comparamos los ingresos medios según el sexo: calculamos la media de los ingresos de varones y de mujeres por separado y las compramos entre sí. Gráficamente se podría representar así:




   Este ejemplo corresponde a la segunda situación reseñada. En el primer caso se podría considerar por ejemplo dos grupos de individuos separados de entrada como sería el caso de las notas obtenidas por los dos grupos de una misma asignatura.

2) Diseño de muestras apareadas. En este caso disponemos de la misma muestra aleatoria $n$ y lo que hacemos es comparar dos medias de dos variables, ambas evaluadas para todos los casos pues los datos están emparejados. Es un diseño que se puede identificar con la expresión intra-sujetos. Por ejemplo, si comparamos los ingresos de los mismos individuos en dos momentos en el tiempo $t_1$ y $t_2$, dispondremos de dos medias y contrastamos si son significativamente diferentes, si han cambiado en el tiempo entre $t_1$ y $t_2$. Gráficamente se puede representar así:



## 4.1. Muestras independientes

Trataremos en primer lugar la situación del contraste de dos medias con muestras independientes. Partimos de $n_1$ y $n_2$ elementos de dos poblaciones, con medias poblacionales $\mu_1$ y $\mu_2$, y con varianzas poblacionales $\sigma_1^2$ y $\sigma_2^2$.

Para realizar la prueba estadística establecemos cuatro pasos como en toda prueba de contraste de hipótesis. Presentamos estos pasos de forma genérica y a continuación los ejemplificaremos con datos reales.

1. Formulación de las hipótesis
   Se formulan la hipótesis nula ($H_0$) y la alternativa ($H_A$) siguientes:

   $H_0$: Las medias poblacionales son iguales, $\mu_1 = \mu_2$, o bien, $\mu_1 - \mu_2 = 0$
   $H_A$: Las medias poblacionales son distintas, $\mu_1 \neq \mu_2$, o bien, $\mu_1 - \mu_2 \neq 0$

2. Cálculo del valor del estadístico muestral
   El estadístico es la diferencia entre las dos medias que se obtiene en la muestra, $\bar{x}_1$ y $\bar{x}_2$, y se trata de determinar si $\bar{x}_1 - \bar{x}_2$ es significativamente diferente de 0. Para valorar la diferencia debemos conocer la forma de la distribución muestral del estadístico: la diferencia de medias dividida por su error típico ($s_{\bar{x}_1-\bar{x}_2}$) se distribuye según una t de student, si la muestra es suficientemente grande (superior a 30 casos). Obtenemos así el estadístico $t$:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}} \qquad \textit{Ecuación 2}$$

   Este cálculo es diferente dependiendo de si las varianzas son iguales (homoscedasticidad) o son diferentes (heteroscedasticidad). Para determinar en qué situación nos encontramos se requerirá realizar una prueba estadística previa que lo establezca: la prueba estadística de igualdad de varianzas. Posteriormente veremos la prueba de Levene de homogeneidad de varianzas.

   Si tenemos varianzas iguales, el error típico del estadístico, utilizando la varianza muestral, es:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1 \cdot n_2}} \qquad \textit{Ecuación 3}$$

   En este caso el estadístico $t$ se distribuye con $\nu$ grados de libertad, valor que se obtiene de calcular $\nu = n_1 + n_2 - 2$.

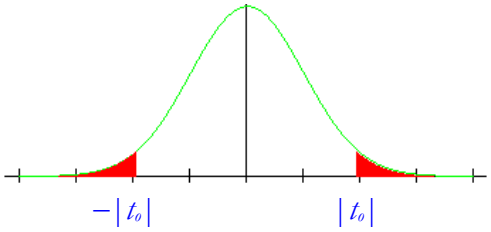
   Si tenemos varianzas diferentes, el error típico del estadístico es:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}} \qquad \textit{Ecuación 4}$$

En este caso el estadístico $t$ se distribuye con $\nu$ grados de libertad, valor que se obtiene de calcular $\upsilon = \dfrac{\left(s_1^2/n_1 - 1 + s_2^2/n_2 - 1\right)^2}{\left(s_1^2/n_1 - 1\right)^2 \cdot \left(\dfrac{1}{n_1 + 1}\right) + \left(s_2^2/n_2 - 1\right)^2 \cdot \left(\dfrac{1}{n_2 + 1}\right)}$.

3. Determinación de la significación

   A partir del estadístico determinamos la probabilidad de obtener una diferencia como la muestral (o mayor, en valor absoluto), $Pr(|t| \geq |t_o| = \alpha$, suponiendo que en la población las medias son iguales, es decir, que se da la hipótesis nula, y valoraremos si es suficientemente significativa esta probabilidad. El valor concreto del estadístico $t$ se expresa por $t_o$, y no sabemos a priori si esta diferencia será positiva o negativa, por lo que consideraremos en valor absoluto $|t_o|$, y realizaremos un contraste a dos colas:




4. Decisión sobre la significación del estadístico

   El contraste, finalmente, se realiza fijando un nivel de significación a partir del cual determinar si se puede rechazar o no la hipótesis nula. Como es habitual se considera el valor de significación $\alpha$=0,05, y la decisión se formaliza de la siguiente manera:

   Si $Pr(t_o) \geq \alpha$ aceptamos la hipótesis nula, las medias son iguales.
   Si $Pr(t_o) < \alpha$ rechazamos la hipótesis nula, las medias son diferentes.

Antes de ver el ejemplo de aplicación formularemos la prueba que establece si las varianzas son iguales o diferentes, la denominada prueba de Levene.[3]. De nuevo presentamos la prueba en cuatro pasos:

1. Formulación de las hipótesis

   $H_0$: Las varianzas poblacionales son iguales (homoscedasticidad): $\sigma_1^2 = \sigma_2^2$

   $H_A$: Las varianzas poblacionales son distintas (heteroscedasticidad): $\sigma_1^2 \neq \sigma_2^2$

2. Cálculo del valor del estadístico muestral

   En este caso se realiza un cálculo de varianzas intergrupos e intragrupos que de lugar al estadístico F de Fisher-Snedecor[4].

[3] Otras pruebas estadísticas para determinar la igualdad de varianzas son la *C de Cochran*, la *F máximos de Hartley* o la *F de Bartlett-Box*.
[4] No detallamos las fórmulas de cálculo, las veremos calculadas por el software estadístico.

3. Determinación de la significación
   Se estima la probabilidad asociada al estadístico a partir del valor concreto $F_o$ del estadístico $F$.

4. Decisión sobre la significación del estadístico
   Tomando el valor de significación $\alpha=0,05$, la decisión se formaliza de la siguiente manera[5]:
   Si $Pr(F_o) \geq \alpha$ aceptamos la hipótesis nula, las varianzas son iguales.
   Si $Pr(F_o) < \alpha$ rechazamos la hipótesis nula, las varianzas son diferentes.

Aplicaremos la prueba de la diferencia entre dos medias a la determinación de si las medias de ingresos individuales de varones y de mujeres son iguales o, como es previsible, son diferentes. A partir del estudio del CIS 3041 consideramos las variables P46m y P31. Aplicamos la prueba de la t con el software SPSS y obtenemos los resultados de la Tabla III.8.8.

Tabla III.8.8. Resultados de la prueba del contraste de dos medias

**Estadísticos de grupo**

| | P31 Sexo | N | Media | Desviación estándar | Media de error estándar |
|---|---|---|---|---|---|
| P46m Ingresos individuales (marca de clase) | 1 Hombre | 923 | 986,46 | 852,507 | 28,061 |
| | 2 Mujer | 996 | 549,55 | 601,437 | 19,057 |

**Prueba de muestras independientes**

| | Prueba de Levene de calidad de varianzas | | Prueba t para la igualdad de medias | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| P46m Ingresos individuales (marca de clase) | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Diferencia de error estándar | 95% de intervalo de confianza de la diferencia | |
| | | | | | | | | Inferior | Superior |
| Se asumen varianzas iguales | 28,331 | 0,000 | 13,046 | 1917 | 0,000 | 436,909 | 33,490 | 371,229 | 502,590 |
| No se asumen varianzas iguales | | | 12,881 | 1644,5 | 0,000 | 436,909 | 33,920 | 370,378 | 503,440 |

La media de ingresos de los varones es de 986€ frente a los 550€ de las mujeres, una diferencia de medias de 437€ ¿Esta diferencia observada en la muestra es estadísticamente significativa? Miremos en primer lugar si debemos tener en cuenta la situación de varianzas iguales o diferentes. El resultado de la prueba muestra en la tabla unas primeras columnas donde se lee el resultado de la prueba de Levene, como la significación es 0,000 menor de 0,05, se concluye que es significativa, es decir, que

[5] La prueba utiliza dos parámetros de grados de libertad: $\nu_1 = k-1$ y $\nu_2 = \sum_{j=1}^{k} n_j - 1$, donde $k$ es el número de grupos y $n_j$ el números de casos de cada grupo. Se puede consultar la tabla de distribución teórica en el anexo.

rechazamos la hipótesis nula (igualdad de varianzas) y aceptamos la alternativa (no igualdad de varianzas). Este resultado supone mirar la segunda línea de la tabla de la prueba de la t. Considerando un nivel de confianza del 95% y con un estadístico t calculado de 12,881, la probabilidad asociada al estadístico (con 1.644,5 grados de libertad) es de 0,000, es decir, es significativa, menor de 0,05, y se concluye el rechazo de la hipótesis nula (igualdad de medias) y se acepta la alternativa (medias diferentes). En conclusión, pues, establecemos estadísticamente que los ingresos de los varones (986€) difieren significativamente de los de las mujeres (550€), la diferencia de medias de 437€ se puede extrapolar a la población.

### 4.2. Muestras relacionadas

Las muestras relacionadas implican un diseño diferente pero que supone una prueba estadística de similares características pues suponen igualmente la comparación de dos medias. En este caso las dos medias que se comparan se calculan para todos los casos en dos variables diferentes, obteniendo así dos medias referidas a los mismos individuos, pero de mediciones distintas. Podríamos considerar la media de los ingresos obtenidos por una muestra de trabajadores/as de una multinacional en el año 2016 y compararla, por ejemplo, con la media de ingresos que tenían justo antes del inicio de la crisis en 2007. Obtendríamos dos medias, y una diferencia entre ambas cuya prueba de significación se formularía en los mismos términos que acabamos de ver para un diseño de muestras independientes.

## 5. El análisis de varianza unifactorial

El análisis de varianza unifactorial tiene como objetivo mostrar hasta qué punto la variación de una variable o factor de naturaleza cualitativa $X$ influye o explica ("causa") la variación de otro variable de naturaleza cuantitativa $Y$. Es decir, cómo los diferentes valores de la variable $X$, que configuran los llamados grupos factoriales (los $k$ grupos[6]: $X_1, X_2,..., X_k$), introducen variaciones en los valores correspondientes de las medias de $Y$, considerando las medias poblacionales $(\mu_1, \mu_2,...,\mu_k)$.

En términos muestrales obtenemos medias muestrales de los datos observados a partir de ambas variables, $y$ y $x$. La información en este tipo de análisis se puede disponer como se representa en la Tabla III.8.9.

El conjunto de valores de la variable dependiente $y_{ij}$ , es decir, de cada individuo $i$ que pertenece a un grupo $j$, se distribuye dentro cada grupo que configura la variable independiente o factor. En el interior de cada uno de los $k$ grupos determinamos el número de casos, la media y la desviación.

[6] La terminología del análisis de varianza emplea las expresiones factor y niveles para identificar a la variable independiente y a los grupos factoriales.

Tabla III.8.9. Tabla de datos de un análisis de varianza

| | $x_j$: variable independiente (grupos) | | | Total |
|---|---|---|---|---|
| | $j=1$ | … | $j=k$ | |
| Valores de la variable dependiente $y_{ij}$ | $y_{11}$<br>$y_{21}$<br>:<br>$y_{n_1 1}$ | …<br>…<br>:<br>… | $y_{1k}$<br>$y_{2k}$<br>:<br>$y_{n_k k}$ | $y_{1+}$<br>$y_{2+}$<br>:<br>$y_{n_k +}$ |
| Casos | $n_1$ | … | $n_k$ | $n$ |
| Media | $\bar{y}_{+1}$ | … | $\bar{y}_{+k}$ | $\bar{y}$ |
| Desviación | $s_1$ | … | $s_k$ | $s$ |

Trabajaremos con un ejemplo sencillo, el que se presenta en la Tabla III.8.10, donde se relacionan la valoración de un producto de consumo (en una escala de 1 a 9) con la clase social, con tres grupos sociales: alta, media y baja.

Tabla III.8.10. Tabla de datos del ejemplo de Valoración del producto de consumo según la Clase social

| | Clase social | | | Total |
|---|---|---|---|---|
| | Baja | Media | Alta | |
| Valoración de un producto de consumo (de 1 a 9) $y_{ij}$ | 5 6 9 6 8 5<br>7 7 6 9 5 7<br>5 6 8 5 7 6 | 6 7 4 4 5 5<br>6 6 3 4 4 3 | 1 3 2 1 4 3 | 5 6 9 6 8 5 7 7 6<br>9 5 7 5 6 8 5 7 6<br>6 7 4 4 5 5 6 6 3<br>4 4 3 1 3 2 1 4 3 |
| Casos | $n_1$=18 | $n_2$=12 | $n_3$=6 | n=36 |
| Media | $\bar{y}_{+1}$=6,5 | $\bar{y}_{+2}$=4,75 | $\bar{y}_{+3}$=2,33 | $\bar{y}$=5,22 |
| Desviación | $s_1$=1,34 | $s_2$=1,29 | $s_2$=1,22 | s=1,99 |

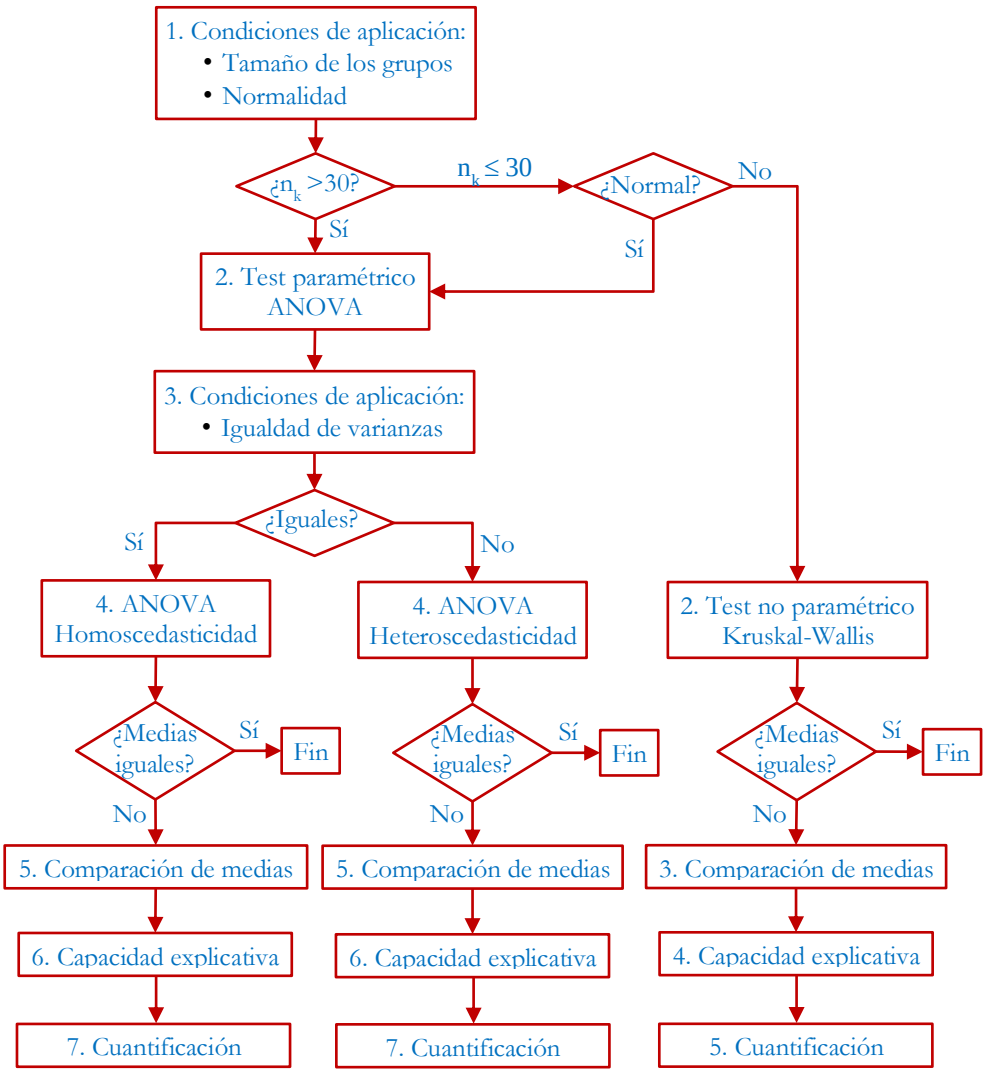
La valoración media del producto realizada por el conjunto de la muestra, de los 36 casos, es de 5,22. Este valor total vemos que varía según la clase social: la clase baja lo valora mejor (6,5) que la clase media (4,75) y que la clase alta (2,33). También podemos observar como las desviaciones típicas dentro de cada clase son valores similares: 1,34, 1,29 y 1,22.

Con este tipo de planteamiento vemos que el análisis de varianza presenta un cierto parecido con el análisis de diferencias de medias, no obstante, tiene especificidades notables. Básicamente estas diferencias provienen del hecho de considerar la existencia de un modelo explicativo entre dos variables, una, que la declaramos como dependiente, la $Y$, y otra como variable independiente, la $X$, siendo ésta la variable que determina el número $k$ de grupos y el número de medias que se comparan. Esto significa que no sólo podremos comprobar si existen diferencias entre las medias de los diferentes grupos de una muestra: $\bar{Y}_1, \bar{Y}_2, \cdots, \bar{Y}_k$, entre cuáles y si son significativamente diferentes, sino que también podremos establecer qué parte de la varianza de $Y$ está explicada por $X$. Por tanto, el análisis de varianza contiene tres objetivos esenciales:

1) El cálculo de las medias y de sus diferencias.
2) El cálculo de su significación, como veremos, a partir de la comparación de varianzas internas y externas, intra/inter, de la variable dependiente *Y* para todos los valores de la variable *X*.
3) Adicionalmente se calcula la parte explicada de la variable dependiente explicada por la variable independiente o factor.

Para conseguir estos objetivos se deben establecer determinadas condiciones en los datos. Las veremos en primer lugar en el apartado siguiente. Antes de acabar este apartado introductorio presentaremos el proceso general de un análisis de varianza que se puede estructurar mediante el esquema que se representa en el Gráfico III.8.6.

Gráfico III.8.6. Proceso de análisis del ANOVA




En función del tamaño de los grupos y de la condición de normalidad aplicaremos la prueba paramétrica del análisis de varianza o bien una prueba no paramétrica en el caso de grupo pequeños, de menos de 30 casos, que no siguen una distribución normal. En el caso de poder aplicar un modelo ANOVA debemos considerar de forma

diferenciada los casos en que se da igualdad de varianzas (homoscedasticidad) de los casos en que no son iguales (heteroscedasticidad). Finalmente, en todas las situaciones, se trata de determinar si las medias son iguales o diferentes. Si son iguales el análisis concluye sin poder validar la influencia de la variable independiente sobre la dependiente. Si son diferentes determinaremos entre qué categorías se dan las diferencias y valoraremos la influencia que tiene la variable independiente sobre la dependiente estableciendo su capacidad explicativa.

Así pues, visto el planteamiento general de un modelo de análisis de varianza procederemos a continuación a dar cuenta de:

- Las condiciones de aplicación: normalidad, homoscedasticidad, independencia.
- La formulación del modelo de análisis unifactorial estableciendo la hipótesis nula y alternativa.
- La validación de dicho modelo con los conceptos de varianza intragrupos y entregrupos y la construcción de la tabla ANOVA que permite la contrastación estadística de las hipótesis.
- La comparación entre grupos para determinar dónde están las fuentes de variación y contrastar entre qué grupos se producen las diferencias.
- La determinación de la fuerza de la relación o capacidad explicativa del modelo con el estadístico eta cuadrado.
- La interpretación de los resultados en relación al modelo de análisis.

## 5.1. Condiciones de aplicación

Como modelo y prueba estadística los datos deben cumplir diversos supuestos:

1) Que las muestras sean aleatorias e independientes en cada grupo factorial.
2) Que el modelo incluya todas las variables independientes relevantes (complitud).
3) Que la variable dependiente sea cuantitativa. Siendo la(s) variable(s) independiente(s) cualitativa(s).
4) Que se dé una relación lineal, la suma ponderada de las variables independientes sea la variable dependiente (aditividad).
5) Que la distribución de la variable dependiente sea normal en cada grupo factorial (normalidad de los errores).
6) Que las varianzas de la variable dependiente en cada grupo de la población sean iguales (supuesto de homoscedasticidad): $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$.

Aunque el análisis de varianza fue ideado para condiciones experimentales, puede extenderse a situaciones de cuasi-experimentalidad. En este caso la muestra es de toda la población y no una para cada grupo factorial. El análisis de varianza se considera una prueba estadística robusta, por lo tanto, el incumplimiento de alguno de los supuestos no debe entenderse como totalmente restrictivo en la aplicación de la prueba, ni el análisis resulta invalidado. No obstante, hay que ser conscientes de los incumplimientos.

Para determinar la condición de normalidad se aplica una prueba estadística que la determina como la prueba de Kolmogorov-Smirnov, u otras como las de Lilliefors, Anderson-Darling o Shapiro-Wilk. El no cumplimiento de esta condición no afecta

gravemente al contraste de la $F$ si el tamaño de la muestra es suficientemente grande en cada grupo: si $n>30$ dentro de cada grupo factorial, por el Teorema Central del Límite, se garantiza la robustez del análisis. Si $n\leq 30$ se tiene que verificar con la prueba estadística de normalidad de Shapiro-Wilk.

En caso de incumplimiento se puede aplicar una transformación de los datos (logaritmo, función inversa,...) que corrige el efecto no deseado en las estimaciones.

Para determinar la existencia de normalidad podemos optar por realizar la prueba de Kolmogorov-Smirnov (K-S), se trata de un test no paramétrico que determina la bondad de ajuste de dos distribuciones: la muestral y una distribución de probabilidad. En este caso el test se adapta para establecer la normalidad. Se realiza de la forma siguiente:

1. Formulación de las hipótesis
   $H_0$: La distribución es normal
   $H_A$: La distribución no es normal
2. Cálculo del valor del estadístico muestral
   Se realiza el cálculo del estadístico $D$ basado en la comparación de las distribuciones acumuladas observada y teórica.
3. Determinación de la significación
   Se estima la probabilidad asociada al estadístico, con sus correspondientes grados de libertad, a partir de las tablas de distribución de Kolmogorov-Smirnov.
4. Decisión sobre la significación del estadístico
   Tomando el valor de significación $\alpha=0,05$, la decisión se formaliza de la siguiente manera:
   Si $Pr(D_0) \geq \alpha$ aceptamos la hipótesis nula, existe normalidad.
   Si $Pr(D_0) < \alpha$ rechazamos la hipótesis nula, no existe normalidad.

A pesar de que la prueba arroje como resultado la no normalidad, si nuestros datos superan los 30 casos, se acepta como condición suficiente para el análisis.

Si el número de casos es inferior a 30 aplicamos la prueba de Shapiro-Wilk. Se realiza de la forma siguiente:

1. Formulación de las hipótesis
   $H_0$: La distribución es normal
   $H_A$: La distribución no es normal
2. Cálculo del valor del estadístico muestral
   Se realiza el cálculo del estadístico $W$ obtenido por simulaciones de Monte Carlo.
3. Determinación de la significación
   Se estima la probabilidad asociada al estadístico, con sus correspondientes grados de libertad.
4. Decisión sobre la significación del estadístico
   Tomando el valor de significación $\alpha=0,05$, la decisión se formaliza de la siguiente manera:
   Si $Pr(W_0) \geq \alpha$ aceptamos la hipótesis nula, existe normalidad.
   Si $Pr(W_0) < \alpha$ rechazamos la hipótesis nula, no existe normalidad.

Si realizamos las dos pruebas de normalidad presentadas con los datos del ejemplo de la valoración del producto de consumo de la Tabla III.8.10 se obtienen los resultados de la Tabla III.8.11.

Tabla III.8.11. Prueba de normalidad del ejemplo de Valoración del producto de consumo según la Clase social

| Prueba | Valoración del producto de consumo | Clase social | | |
|---|---|---|---|---|
| | | Baja | Media | Alta |
| Kolmogorov-Smirnov[a] | Estadístico | 0,201 | 0,220 | 0,209 |
| | Grados de libertad | 18 | 12 | 6 |
| | Sig. | 0,053 | 0,114 | 0,200* |
| Shapiro-Wilk | Estadístico | 0,886 | 0,920 | 0,907 |
| | Grados de libertad | 18 | 12 | 6 |
| | Sig. | 0,034 | 0,284 | 0,415 |

* Este es un límite inferior de la significación verdadera.
a. Corrección de la significación de Lilliefors

Como el número de casos en cada grupo es inferior a 30 debemos aplicar e interpretar la prueba de Shapiro-Wilk. En el caso de la clase baja no se cumple la normalidad, la significación es inferior a 0,05, mientras que en los grupos d clase media y alta la probabilidad asociada al estadístico, superior a 0,05, nos permite concluir la condición de normalidad. Veremos más adelante cómo estos dos resultados tendrán un tratamiento distinto al realizar el análisis de varianza.

Para determinar la homogeneidad o la igualdad de varianzas podemos aplicar la prueba de Levene como ya tuvimos ocasión de ver anteriormente. Como en el caso de la normalidad, el no cumplimiento no afecta de forma sensible el contraste de la *F*, si el tamaño de la muestra es suficientemente grande, sobre todo si son aproximadamente de igual tamaño. Cuando todos los grupos tienen el mismo número de casos el contraste es igualmente exacto. El efecto de las varianzas desiguales depende de la heterogeneidad entre el número de observaciones de cada grupo.

Como en el caso de la normalidad, ante el incumplimiento, se puede aplicar una transformación de los datos (logaritmo, función inversa,...) lo que permite solventar ambas condiciones.

La prueba de la hipótesis es la misma que presentamos en el apartado 4.1. La hipótesis nula de igualdad de varianzas (homoscedasticidad) se acepta cuando la probabilidad asociada al estadístico es superior o igual a 0,05.

## 5.2. El modelo ANOVA unifactorial

Para examinar el efecto de la variable independiente sobre la dependiente en un análisis de varianza[7] consideramos una población donde la media de la variable dependiente $Y$ es el valor $\mu$. Si se cumplen las condiciones que hemos señalado anteriormente lo que se pone a prueba es si existen o no diferencias significativas entre las medias de la variable dependiente estudiada para cada grupo de población (o nivel) determinado por la variable independiente (o factor), es decir, se establece como hipótesis nula de que las medias poblacionales son iguales, esto es, que no existen diferencias significativas entre ellas, y como alternativa que por lo menos una de ellas es diferente de cero:

$H_0$: Las medias poblacionales de las k submuestras son iguales, $\mu_1 = \mu_2 = ... = \mu_k = \mu$.
$H_A$: Las medias poblacionales de las k submuestras no son iguales.

La medida del efecto de cada valor de la variable factorial $X$ se define por la diferencia entre la media de cada grupo j ($\mu_j$) y la media poblacional total ($\mu$):

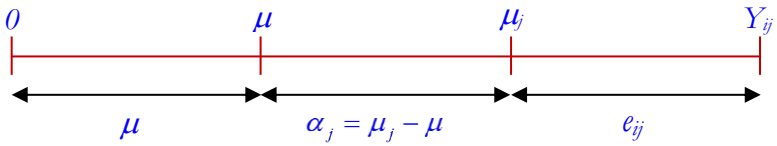
$$\alpha_j = \mu_j - \mu \qquad \textit{Ecuación 5}$$

De esta forma el modelo ANOVA con una variable independiente $X$ es:

$$Y_{ij} = \mu + \alpha_j + e_{ij} \qquad \textit{Ecuación 6}$$

La *Ecuación 6* del modelo del análisis de la varianza, a partir de los valores $Y_{ij}$ de la variable dependiente para el individuo $i$ del grupo $j$, contempla tres componentes que pueden representar según el Gráfico III.8.7.

Gráfico III.8.7. Representación gráfica del modelo ANOVA




1) $\mu$ es el valor de la media poblacional o "punto de partida" que determina la pauta del comportamiento de toda la población sin tener conocimiento del comportamiento de la variable independiente $X$, o de sus grupos ($X_1, X_2,..., X_k$). Si atribuimos a $Y$ este valor central como pauta de comportamiento de todos los individuos en la variable $Y$ se trata de "aceptar" un error por lo que de más o de menos difiera el valor de cada individuo en relación a este valor de la media. Se pretende que con el conocimiento de $X$ el error de $Y$ se reduzca. En este sentido decimos que $X$ determina o explica $Y$.
2) $\alpha_j$ es el efecto que se obtiene por la diferencia entre la media del grupo y la media global y nos acerca a una mayor precisión en el conocimiento de los valores de $Y$ pues $\mu_j$ es la media de cada valor de $X_j$. Es, por tanto, lo que el conocimiento de

[7] Cuando se dispone de una sola variable independiente al análisis se puede denominar como oneway o de una sola vía.

$X_j$ ayuda a precisar $Y_{ij}$, lo que se añade (o se retrae) a $\mu$ para acercarnos al valor de $Y_{ij}$.

3) $e_{ij}$ es la parte residual del modelo, la parte que no puede ser precisada por el conocimiento de $X_j$. Es el elemento estocástico o probabilístico del modelo. Se entiende que el conocimiento de $Y$ depende de otras variables, de las condiciones de medida de las variables, etc., siempre que haya un comportamiento aleatorio en relación a $Y$. De esta forma, el elemento residual es la diferencia entre el valor de $Y_{ij}$ y la parte explicada bien por el conocimiento de la media de $Y$, bien por la influencia de $X$ sobre $Y$: $e_{ij} = Y_{ij} - (\mu + \alpha_j)$.

Con esta nueva notación las hipótesis nula y alternativa anteriores se expresan de la siguiente forma:

$H_0$: Los efectos son nulos: $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$.
$H_A$: Los efectos no son iguales a cero.

## 5.3. Validación del modelo

La validación del modelo de dependencia, es decir, la aceptación de la influencia de la variable independiente sobre la dependiente implica rechazar la hipótesis nula y aceptar la alternativa, en caso contrario, se aceptaría hipótesis nula de ausencia de influencia. Para validar la hipótesis de influencia se deben estimar las medias poblacionales $\mu_j$ a partir de las medias muestrales $\bar{y}_j$ (media de $y$ para cada valor de $x$).

El problema que se plantea ahora es el de la significación, es decir, el de determinar hasta qué punto las estimaciones de las medias poblacionales a partir de las muestrales, que están afectadas por un error muestral y dado un nivel de confianza, son suficientemente diferentes entre sí. Para ello utilizaremos los conceptos de varianza intragrupos y varianza entregrupos. A continuación desarrollaremos estos conceptos introduciendo la noción de suma de cuadrados de las diferencias para la variable dependiente $y$ ($SCD_y$).

Para estimar los dos tipos de varianzas intra y entre se introduce el concepto de variación o de suma de cuadrados de las diferencias o desviaciones ($SCD$) en relación a la media, es decir, el numerador de la fórmula de la varianza[8]:

$$SCD = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \textit{Ecuación 7}$$

La suma de cuadrados de las diferencias se interpreta en términos espaciales (espacio euclidiano) como la suma de las distancias (cuadráticas) de los puntos (los individuos) al centro (la media), es decir, el grado de alejamiento o dispersión.

La variable independiente $x$ es la que configura los $k$ grupos, indexados con $j=1,...,k$. Las puntuaciones o los valores obtenidos de la variable $y$ se pueden distribuir entre los

[8] Como se vio en el capítulo de análisis descriptivo de una variable. Para recordar este concepto central y omnipresente de la estadística se puede consultar el apartado 3.3.2 del Capítulo III.3.

$k$ grupos. A partir de la distribución de los $n$ casos en tres grupos podemos expresar la suma de cuadrados con un doble sumatorio de la siguiente forma:

$$SCD = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y})^2 \qquad \textit{Ecuación 8}$$

La diferencia entre un valor individual y la media global, se puede descomponer en dos partes. Si a la expresión $(y_{ij} - \bar{y})^2$ le sumamos y restamos la media de cada grupo $j$, $\bar{y}_{+j}$, obtenemos:

$$(y_{ij} - \bar{y}) = (y_{ij} - \bar{y}_{+j} + \bar{y}_{+j} - \bar{y}) = (y_{ij} - \bar{y}_{+j}) + (\bar{y}_{+j} - \bar{y})$$

es decir, se puede expresar como la suma de a) las diferencias entre el valor y la media de su grupo, que es una medida de la variabilidad de cada muestra, más b) las diferencias entre la media del grupo y la media global, que es una medida de la variabilidad entre muestras.

A continuación, elevamos al cuadrado y obtenemos:

$$(y_{ij} - \bar{y})^2 = (y_{ij} - \bar{y}_{+j})^2 + 2(y_{ij} - \bar{y}_{+j})(\bar{y}_{+j} - \bar{y}) + (\bar{y}_{+j} - \bar{y})^2$$

Esta expresión para todos los individuos y grupos queda:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y}_{+j})^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y}_{+j})(\bar{y}_{+j} - \bar{y}) + \sum_{j=1}^{k}\sum_{i=1}^{n_k}(\bar{y}_{+j} - \bar{y})^2$$

Dado que la suma de desviaciones en relación a la media de cada grupo es cero, el término $2\sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y}_{+j})(\bar{y}_{+j} - \bar{y})$ es cero, y nos queda:

$$SCD_y = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y}_{+j})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_k}(\bar{y}_{+j} - \bar{y})^2 \qquad \textit{Ecuación 9}$$

$$SCD_T = SCD_I + SCD_E$$

Es decir, que la suma total de los cuadrados de las desviaciones de cada individuo en relación a la media global, $SCD_T$, es igual a:

- la suma de los cuadrados de las desviaciones entre cada individuo y la media de su grupo. Es la primera estimación de variabilidad que habíamos considerado anteriormente, llamada también suma de cuadrados intra, $SCD_I$ (llamada también variación intragrupo, variación no explicada o residual), más
- la suma de los cuadrados de las desviaciones entre cada media de grupo y la media global. Es la segunda estimación de variabilidad anterior, llamada también suma de cuadrados entre, $SCD_E$ (llamada también variación entre grupos o variación explicada).

Para ejemplificar estos conceptos supongamos que analizamos los ingresos de la población asalariada y queremos determinar la existencia de diferencias salariales entre hombres y mujeres. De cada submuestra o grupo podemos calcular la media de

ingresos, así como la media total de los ingresos para el conjunto de hombres y mujeres. A la hora calcular (estimar) la varianza podemos optar por dos vías:

1) Calcular la varianza como el promedio ponderado de las varianzas de cada grupo, de la misma forma que se calcula en la prueba de la diferencia entre dos medias cuando hay homoscedasticidad. Así calcularíamos la varianza de los ingresos de los hombres, de las mujeres, y finalmente calcularíamos la media ponderada entre ambas cantidades.
2) Alternativamente, podemos calcular la varianza de las medias de cada grupo en relación a la media global. En este caso supone considerar que en cada grupo, el de hombres y el de mujeres, hay un solo efectivo que representa al grupo a través de su media, después se evaluaría como difieren estas medias de la media total, considerando juntos hombres y mujeres.

La idea que hay detrás de estos cálculos es la de comparar el comportamiento de dos tipos de variabilidad: dentro de cada grupo y entre los grupos. El objetivo del ANOVA es dividir o establecer una partición del total de la $SCD_T$ en dos componentes y así mostrar la existencia de asociación y de determinación de la variable independiente sobre la variable dependiente. Así, cuanto mayor sea la variabilidad entre los grupos, más importantes serán las diferencias de las medias de cada grupo en relación a la media total y, por tanto, más homogéneos serán los grupos, es decir, menor será su variabilidad interna, y las diferencias en relación a su media, en la de cada grupo, no serán importantes. Esto será indicativo de que las medias son distintas. Inversamente, cuanto menor sea la variabilidad entre los grupos, menos importancia tiene la existencia de los grupos, tenderán a ser grupos heterogéneos, con una variabilidad interna alta, con medias que no difieren entre los distintos grupos, tienden a ser iguales.

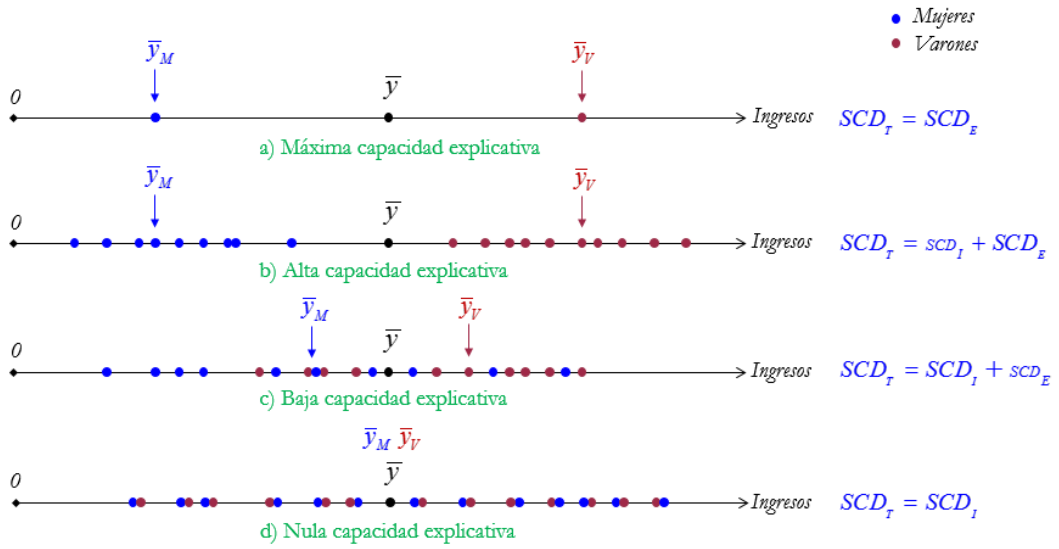
Ilustremos estas ideas en el caso del ejemplo que sugeríamos de la relación entre la variable de ingresos y el sexo. La distribución de los ingresos según estos dos grupos se puede representar de la forma que aparece en el Gráfico III.8.8 donde se contemplan cuatro situaciones.

En la situación a) presentamos un caso extremo, si todas las mujeres tienen el mismo sueldo y todos los hombres tienen el mismo sueldo, pero el de las mujeres inferior a la media y muy inferior al de los varones, reflejando una situación muy desigual, nos encontramos en una situación donde las diferencias internas de los varones y de las mujeres son nulas, constituyen grupos totalmente homogéneos, la variabilidad interna es cero. Por lo tanto, toda la variabilidad o toda la diferencia que da entre los salarios se debe a la diferencia entre el salario medio de los varones y el salario medio de las mujeres, la variabilidad entre los dos grupos (la variabilidad explicada) coincide con la variabilidad total, $SCD_T=SCD_E$, tan sólo necesitamos saber el sexo para predecir exactamente cuál será el sueldo medio que se tendrá. En esta situación la capacidad explicativa del modelo es máxima, del 100%.

En la situación b) se observa una cierta variabilidad interna, tanto en varones como en mujeres, unas mujeres cobran más que otras y unos varones cobran más que otros, pero siempre se da que las mujeres cobran por debajo de la media y menos que los varones, reflejando de nuevo una situación desigual pero menor que en el caso extremo anterior. Por lo tanto, son grupos internamente menos homogéneos pero siguen

teniendo un alto grado de similitud interna cada uno de los sexos. Como la variabilidad interna ha aumentado, de hecho ha aparecido, siendo la variabilidad total constante, ello implica que la variabilidad externa o entre los grupos disminuya, $SCD_T = SCD_I + SCD_E$, si bien seguimos teniendo una situación donde la variabilidad explicada es importante y alta la capacidad explicativa del modelo.

Gráfico III.8.8. Representación gráfica de la varianza intra y entre

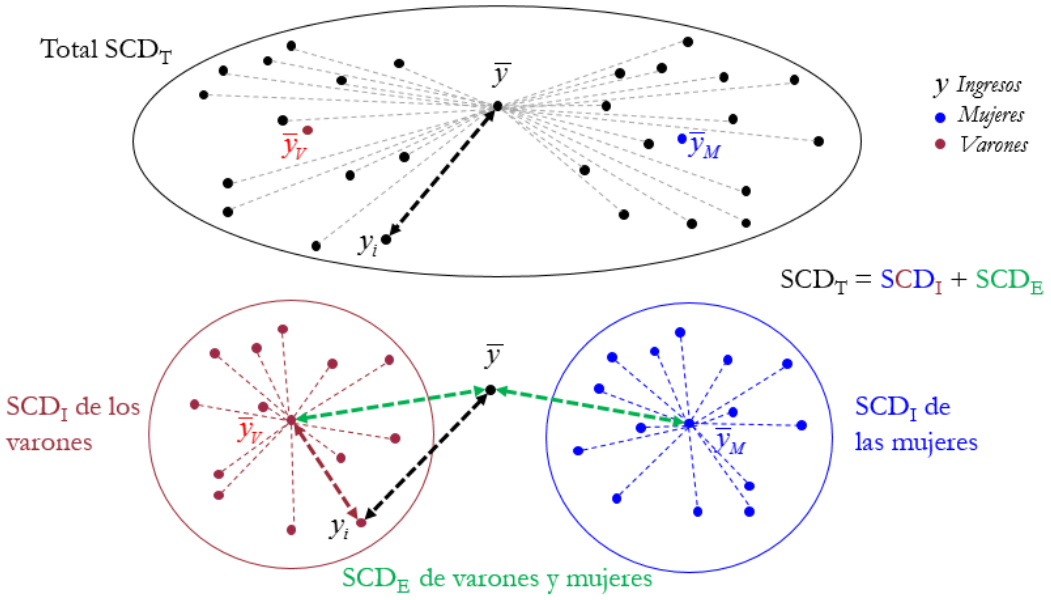



En la situación c) vemos que varones y mujeres se distribuyen a lo largo de la escala de ingresos salariales, dándose casos donde ahora las mujeres pueden estar por encima de los varones. No obstante, si bien la desigualdad de ingresos entre ellos y ellas ha disminuido, todavía nos encontramos en una situación donde, en promedio, las mujeres cobran por debajo de la media y menos que los varones. En consecuencia, la cierta variabilidad interna es mucho más importante que antes, reduciéndose las diferencias entre los grupos, $SCD_T = SCD_I + SCD_E$, las medias de varones y mujeres se han aproximado. Así pues, la capacidad explicativa del modelo se ha reducido de forma notable, el conocimiento del salario medio de varones y de mujeres es una información puede ser errónea en muchos casos al intentar estimar el valor de los ingresos de la población asalariada.

En la situación d), por último, disponemos de otra situación extrema, la contraria a la primera. En este caso se da la misma distribución de ingresos salariales para varones que para mujeres. En consecuencia, la media de varones y de mujeres es la misma, e igual a la media global, no existen diferencias entre los grupos, la variabilidad entre es nula y todo es variabilidad interna, $SCD_T = SCD_I$. Al no existir diferencias según el sexo nuestra capacidad explicativa es nula.

Otra forma de representar gráficamente estos conceptos se presenta en el Gráfico III.8.9 donde la variabilidad total, la distancia de cada punto $y_i$ a la media global, se descompone en la suma de dos distancias: la distancia al centro del grupo y la distancia del centro del grupo al centro global.

Gráfico III.8.9. Representación gráfica de la descomposición de la varianza




La estimación de la varianza, llamada también media cuadrática (*MC*), para cada suma de cuadrados se obtiene dividiendo ésta entre los correspondientes grados de libertad: $k-1$ para la $SCD_E$ y $n-k$ para la $SCD_I$, siendo los de la suma de cuadrados total, es decir, $n-1=(n-k)+(k-1)$. Así:

*Varianza o media cuadrática entre grupos*

$$VE = MC_E = \frac{SCD_E}{k-1} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_k}(\bar{y}_{+j} - \bar{y})^2}{k-1}$$

*Ecuación 10*

*Varianza o media cuadrática intra grupos*

$$VI = MC_I = \frac{SCD_I}{n-k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y}_{+j})^2}{n-k}$$

*Ecuación 11*

Cuando calculamos $MC_E$ y $MC_I$ estamos midiendo dos varianzas, una debida al tratamiento de los efectos (valores de la variable independiente) o explicada, $MC_E$, y otra debida a factores aleatorios y no explicada $MC_I$. La prueba estadística de contraste consiste en calcular un cociente que coloca la estimación de la variabilidad externa, entre grupos o explicada, $MC_E$, en el numerador, y la estimación de la varianza interna, dentro de los grupos o no explicada, $MC_I$, en el denominador. Así se obtiene el estadístico llamado F de Fisher-Snedecor:

$$F = \frac{Varianza\ \text{explicada}}{Varianza\ \text{no explicada}} = \frac{Varianza\ \text{entregrups}}{Varianza\ \text{intragrups}} = \frac{MC_E}{MC_I}$$

*Ecuación 12*

Esperamos que cuando $F = \frac{MC_E}{MC_I}$ sea 1, las varianzas entre e intra se igualan, y se cumpliría la hipótesis nula. Cuando el cociente F es superior a 1 se tiende a rechazar la hipótesis nula y se concluye la existencia de diferencias significativas entre las medias. Por tanto, si las medias son iguales o sus diferencias sólo se deben al azar, entonces ambos tipos de variabilidades tenderán a coincidir y el cociente tenderá a ser 1. Si por el contrario, las variabilidades difieren significativamente se debe a que nos encontramos con poblaciones con distinto comportamiento y no sólo se deben a fluctuaciones del azar.

La prueba de hipótesis se establece pues en los términos siguientes:

1. Formulación de las hipótesis
   $H_0$: Las $k$ medias poblacionales son iguales, $\mu_1 = \mu_2 = ... = \mu_k = \mu$.
   $H_A$: Las $k$ medias poblacionales no son iguales.

2. Cálculo del valor del estadístico muestral
   Se realiza un cálculo de varianzas intergrupos e intragrupos que de lugar al estadístico F de Fisher-Snedecor[9]:
   $$F = \frac{MC_E}{MC_I}$$

3. Determinación de la significación
   Se estima la probabilidad asociada al valor estadístico obtenido $F_o$ según el valor teórico de la distribución muestral $F$ [10].

4. Decisión sobre la significación del estadístico
   La probabilidad asociada se contrasta con el nivel de significación de $\alpha=0,05$ y con los grados de libertad $\nu_1=k-1$ y $\nu_2=n-k$, de la forma siguiente:
   Si $Pr(F_o) \geq \alpha$ aceptamos la hipótesis nula, las medias son iguales.
   Si $Pr(F_o) < \alpha$ rechazamos la hipótesis nula, las medias son diferentes.

Si obtenemos un valor de la $F$ igual a 1 o menor, entonces la varianza intergrupos es menor que la intragrupos, las diferencias dentro de cada grupo son mayores que entre ellos. En esta situación inmediatamente se deduce, sin necesidad de efectuar el

---

[9] Para calcular F manualmente se suele calcular primero la $VE$ y la varianza total $VT$, y por diferencias, se obtiene la $VI$. A efectos de cálculo se utilizan las fórmulas siguientes:

$$SCD_T = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij})^2 - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_k} y_{ij}\right)^2}{n}$$

$$SCD_E = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(\bar{y}_{+j} - \bar{y})^2 = \frac{\sum_{i=1}^{n_k}(y_{ij})^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_k} y_{ij}\right)^2}{n} = \sum_{j=1}^{k} n_j(\bar{y}_{+j} - \bar{y})^2$$

$$SCD_I = SCD_T - SCD_E$$

[10] Véase la tabla de valores críticos de la distribución F en el anexo.

contraste que las medias no son significativamente distintas, se acepta la hipótesis nula. Por el contrario, un valor mayor de la unidad indica la existencia de mayor variabilidad intergrupos, tienden a ser grupos homogéneos, y se trata de contrastar hasta qué punto es significativa esta mayor homogeneidad interna.
Retomamos el ejemplo anterior sobre la valoración de un producto de consumo. Nuestro objetivo es determinar si existen diferencias significativas entre la valoración que hacen la clase baja, media y alta. Por tanto, determinar si ambas variables están relacionadas y si la variabilidad de las valoraciones se puede explicar por el hecho de pertenecer a una categoría u otra. Los resultados que se obtienen con el software estadístico aparecen en la Tabla III.8.12.

Tabla III.8.12. Tabla ANOVA del ejemplo de Valoración del producto según la Clase social

| | Suma de cuadrados | Grados de libertad | Media cuadrática | F | Sig. |
|---|---|---|---|---|---|
| Inter-grupos | 82,139 | 2 | 41,069 | 24,166 | 0,000 |
| Intra-grupos | 56,083 | 33 | 1,699 | | |
| Total | 138,222 | 35 | | | |

Los resultados de la tabla Tabla III.8.12 se pueden reproducir con las fórmulas y los cálculos siguientes:

$$SCD_T = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij})^2 - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_k} y_{ij}\right)^2}{n} = 1120 - \frac{(188)^2}{36} = 138,222$$

$$SCD_E = \sum_{j=1}^{k} n_j(\bar{y}_{+j} - \bar{y})^2 = 18\times(6,5-5,22)^2 + 12\times(4,75-5,22)^2 + 6\times(2,33-5,22)^2$$

$$= 82,139$$

$$SCD_I = SCD_T - SCD_E = 138,22 - 82,14 = 56,083$$

Por tanto, las varianzas intragrupos y entre grupos, y el valor del estadístico F son:

$$VE = \frac{SCD_E}{k-1} = \frac{82,139}{3-1} = \frac{82,139}{2} = 41,069$$

$$VI = \frac{SCD_I}{n-k} = \frac{56,08}{36-3} = \frac{56,08}{33} = 1,699$$

$$F_0 = \frac{41,07}{1,699} = 24,166$$

A este valor $F_0$ se le asocia una probabilidad inferior al nivel de significación del 0,05. Se concluye pues que las medias son significativamente diferentes. Para llegar a esta conclusión podríamos haber consultado igualmente los valores de la tabla de distribución teórica de la $F$. El valor teórico de la F con 2 y 33 grados de libertad para un nivel de significación de 0,05 es $F_{2,33,0,05} = 3'28$, un valor inferior al valor observado de *24,166*, por tanto, las medias son diferentes[11].

[11] No hemos mirado las condiciones de aplicación de la prueba, estamos asumiendo los supuestos de normalidad y homogeneidad de varianzas.

Si se rechaza la hipótesis nula sabemos que existen diferencias entre las medias, es decir, que por lo menos una de ellas, sin determinar cuál, es diferente al resto. Para completar nuestro análisis deberemos establecer entre qué categorías se dan las diferencias así como la intensidad de la relación entre las variables.

## 5.4. Contraste entre grupos

Con el análisis de varianza hemos puesto a prueba la hipótesis nula según la cual hay igualdad de medias. En el caso en que la prueba estadística de la *F* nos lleve a rechazarla obtendremos como información que las medias no son todas iguales. La cuestión que se suscita inmediatamente es entre qué grupos se dan las diferencias, qué medias son diferentes, es decir, queremos responder a la pregunta de cómo se relacionan las variables, de cómo influye la variable independiente sobre la variable dependiente. Necesitamos en consecuencia un análisis adicional que nos ayude a comparar los diferentes grupos entre sí y determinar dónde se encuentran estas diferencias. Estas comparaciones se denominan contrastes.

En un análisis de varianza se pueden realizar contrastes a priori o a posteriori. Esta distinción obedece a razones de tipo metodológico. Un contraste a priori exige la presencia de una hipótesis previa definida en el modelo de análisis que está destinada a ser verificada con los datos observados. El contraste a posteriori se plantea a partir de la obtención de los datos y se busca efectuar una comparación sistemática entre todos los pares de grupos (o combinaciones de grupos).

Trataremos las pruebas de comparación a posteriori. Existen diversas pruebas alternativas que nos determinan entre qué grupos se dan las diferencias. En su aplicación se debe tener en cuenta una de las condiciones de aplicación del ANOVA: la homogeneidad de varianzas, si estamos ante varianzas iguales (homoscedasticidad) o diferentes (heteroscedasticidad) deberemos aplicar pruebas distintas. Entre ellas se encuentra las denominadas pruebas de Sheffé, Tukey, Bonferroni, Hotschberg, Hommel, Tamhane, Dunett, Duncan, entre otras.

La prueba de Scheffé es una de la más utilizadas cuando se dan condiciones de igualdad de varianza. La prueba está basada en la distribución F y permite la comparación binaria entre pares de medias y también la comparación múltiple. En las comparaciones de pares es conservadora, precisa grandes diferencias para obtener la significatividad. Se puede aplicar con muestras de tamaño desigual y es una prueba bastante robusta ante el incumplimiento del supuesto de homoscedasticidad.

Los resultados que se obtienen del ejemplo que seguimos de valoración se presentan en la Tabla III.8.13 a partir de los resultados generados con el software estadístico. La tabla contiene los valores de las medias de cada pareja de clases sociales que se comparan, la diferencia, así la como la significación y el intervalo de confianza. Cuando existen diferencias significativas al nivel 0,05 aparece un asterisco que las destaca. En este caso podemos observar que las diferencias de cada pareja comparada son estadísticamente significativas, es decir, que las distintas clases sociales valoran de forma diferente el producto de consumo y que esta valoración es mayor a medida que la clase social es inferior.

Tabla III.8.13. Contrastes de grupos del ejemplo de valoración del producto. Comparaciones múltiples según la prueba de Scheffé

| Clase social j | Clase social j' | Diferencia de medias j−j' | Error típico | Sig. | Intervalo de confianza al 95% | |
|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior |
| Baja | Media | | | | | |
| 6,5 | 4,75 | 1,75* | 0,486 | 0,004 | 0,50 | 3,00 |
| Baja | Alta | | | | | |
| 6,5 | 2,33 | 4,17* | 0,615 | 0,000 | 2,59 | 5,74 |
| Media | Alta | | | | | |
| 4,75 | 2,33 | 2,42* | 0,652 | 0,003 | 0,75 | 4,09 |

* Diferencias significativa entre las medias al nivel 0,05.

## 5.5. La fuerza de la relación

Una vez determinada la existencia de diferencias significativas y entre qué categorías, se trata de valorar la intensidad de la relación entre las variables, en qué medida el conocimiento del grupo de pertenencia determina el valor en la variable dependiente, es decir, el grado de dependencia entre las variables o su capacidad explicativa. Que hayamos encontrado una relación de dependencia entre las variables no quiere decir que el peso explicativo sea grande. La fuerza de la relación o el tamaño del efecto de la variable independiente se mide mediante el estadístico $\eta^2$ (eta cuadrado) que se define como:

$$\eta^2 = \frac{SCD_E}{SCD_T} = \frac{\sum_{j=1}^{k}(\bar{y}_{+j} - \bar{y})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_k}(y_{ij} - \bar{y})^2}$$

*Ecuación 13*

Es decir, la variabilidad explicada sobre la variabilidad total. El valor de $\eta^2$ que se obtiene se encuentra entre 0 y 1 ($0 \leq \eta^2 \leq 1$) y mide la parte explicada por la variable independiente. En el ámbito de ciencias sociales y cuando no disponemos más que una variable independiente este valor no es grande, difícilmente se superan los valores de 0,5 o 0,6, es decir, de poco más del 50% de variabilidad explicada. En el ejemplo que hemos utilizado se obtiene un valor del 59,4%:

$$\eta^2 = \frac{SCD_E}{SCD_T} = \frac{82,139}{138,222} = 0,594$$

# 6. El análisis de varianza multifactorial

El análisis de varianza multifactorial considera múltiples variables independientes cualitativas (factores) que explican la variabilidad de una variable dependiente cuantitativa. Cada variable independiente contribuye con un efecto principal individual, pero además se pueden considerar en el modelo las posibles interacciones entre los

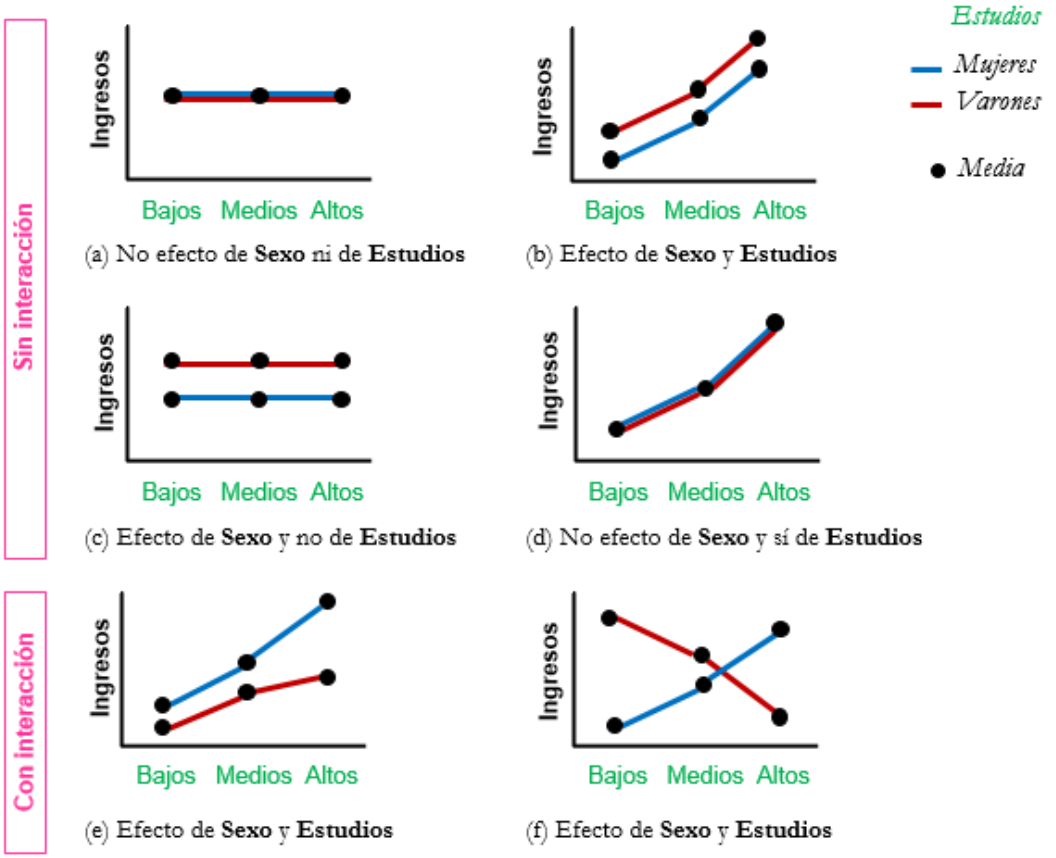
factores. Consideraremos modelos de muestras independientes, de factores fijos y sin la presencia de variables de control.

En el ANOVA multifactorial se consideran tantas poblaciones como casillas resultan de la combinación de todos los niveles (valores) de los factores considerados y sobre las que realizaremos las inferencias. En cada subpoblación se trata de verificar los supuestos de normalidad y de homoscedasticidad (igualdad de varianzas), a partir de observaciones que se han obtenido aleatoriamente de cada población de forma independiente.

Cuando tenemos más de un factor podemos obtener un modelo solo de efectos principales o bien donde también se la interacción entre los factores. El Gráfico III.8.10 ilustra las diversas situaciones que nos podemos encontrar con dos variables independientes. Se trata de los denominados gráficos de perfil, un gráfico de líneas con las medias de cada grupo que nos da una representación de utilidad para visualizar las diferencias entre las medias de las variables y derivar descriptivamente la existencia de patrones de asociación (efectos principales) así como de interacciones.

Se considera como ejemplo un análisis donde se busca explicar el nivel de ingresos según el sexo y el nivel de estudios.

Gráfico III.8.10. Gráficos de perfil de modelos ANOVA con y sin interacción

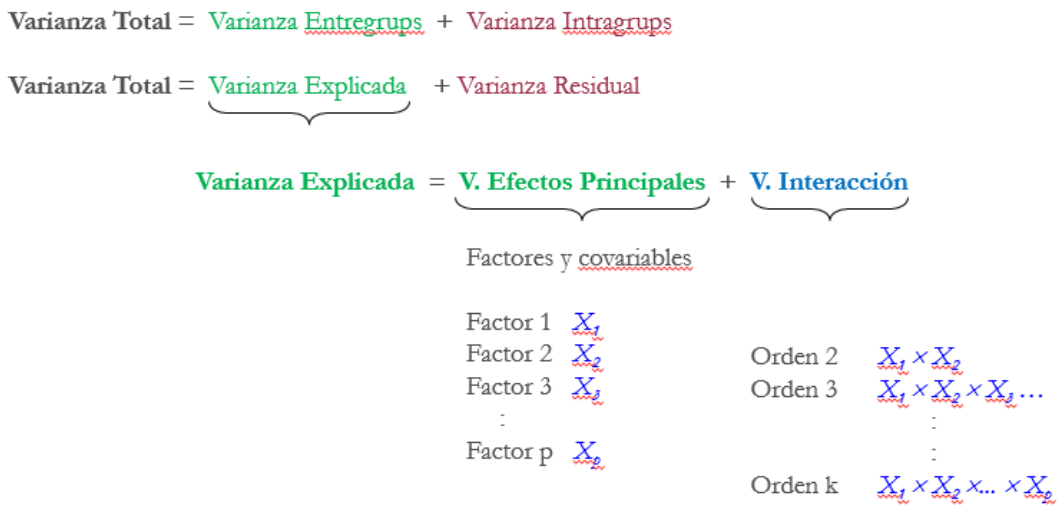


El gráfico muestra por tanto la existencia de medias distintas según los valores de cada variables (los puntos del gráfico) y líneas que marcan las tendencias de asociación. Cuando las líneas son paralelas estamos en situaciones de ausencia de asociación. Cuando las líneas se cruzan o cambian la pendiente estamos ante situaciones que pueden ser de interacción. La determinación de estos patrones de asociación será el resultado de las pruebas estadísticas del análisis de varianza multifactorial.

Para el caso unifactorial vimos como la significación del estadístico de la F de Fisher-Snedecor establecía la relación de dependencia al verificar las diferencias entre las medias permitiendo rechazar la hipótesis nula de igualdad de medias. En un análisis de varianza multifactorial existe una hipótesis nula para cada factor (las medias poblacionales definidas por los niveles del factor son iguales) y por cada posible combinación de factores (el efecto de la interacción es nulo, no se dan diferencias significativas entre las medias de cada combinación de niveles de los factores). Todos los contrastes de hipótesis se hacen en base al estadístico *F* en una tabla ANOVA que presenta tantas líneas como efecto se consideren en el modelo.

La descomposición de la varianza total entre varianza entregrupos (explicada) y varianza intragrupos (no explicada o residual) implica que la primera se descomponga en tantas parte como efectos intervengan en el modelo (Gráfico III.8.11).

Gráfico III.8.11. Esquema de descomposición de la varianza




Se denomina Modelo factorial completo al que incluye el efecto de la constante, los efectos principales y las interacciones. Los estadísticos F calculados evalúa para uno de los efectos la varianza específica sobre la varianza residual determinando su significación. Con dos variables independientes tendremos dos efectos principales y un efecto de interacción, además de la constante.

En la estimación de la suma de cuadrados se pueden emplear distintos métodos de cálculo. Para modelos equilibrados i no equilibrados sin casillas vacías, el método más utilizado para la suma de los cuadrados es el Tipo III, de los cuatro tipos posibles.

Llamado marginal u ortogonal, evalúa la contribución de cada factor más allá de todos los demás. Se utiliza como prueba conservadora para evaluar los efectos principales.

La suma de cuadrados Tipo I es secuencial y de descomposición jerárquica, se utiliza en modelos equilibrados y anidados, y donde cada término se corrige respecto a un anterior que le precede en el modelo evaluando primero los efectos de orden inferior. Por tanto el cálculo depende del orden de introducción de las variables o efectos y expresan los efectos añadidos sobre la suma de cuadrados. Esta opción se suele emplear menos.

La suma de cuadrados Tipo II es jerárquica se utiliza en modelos equilibrados, en aquellos que tan sólo hay efectos principales (y no interacciones), y en diseños anidados en el que cada efecto está anidado con el anterior; en estos casos se tienen en cuenta sólo los efectos pertinentes (aquel que no está contenido en el efecto evaluado).

Las sumas de cuadrados de Tipo IV son adecuados tanto para modelos equilibrados como no equilibrados con casillas vacías.

Una vez determinado el modelo de dependencia que se ajusta a los datos se trata de determinar entre qué categorías se dan las diferencias y también cómo se dan las interacciones si las hubiera.

Para dar cuenta de la capacidad explicativa de cada efecto contamos, además, con el estadístico eta cuadrado parcial que nos informa de la proporción de varianza explicada por cada efecto particular. El estadístico eta cuadrado parcial se obtiene a partir de la expresión siguiente:

$$\eta_e^2 = \frac{F_e \times gl_e}{F_e \times gl_e + gl_{error}} \qquad \textit{Ecuación 14}$$

Es decir, en relación a cada efecto $e$, es el resultado de dividir el valor del estadístico $F$ de este efecto multiplicado por los grados de libertad del efecto, y dividiendo por esa misma cantidad más los grados de libertad del error. Estos valores son estimaciones del grado en que cada factor o combinaciones de factores afectan a la variable dependiente, son las estimaciones del tamaño del efecto. Para el conjunto del modelo el valor del eta cuadrado coincide con el $R^2$.

El eta cuadrado parcial tiene una interpretación menos intuitiva que el general. En aquél el denominador no es la variación total de $Y$, sino la variación no explicada de $Y$ más la variación explicada solo por $X$. Así cualquier variación explicada por otra variable o efecto se excluye del denominador. Esto permite al analista comparar el efecto de la misma variable en dos estudios diferentes que contienen factores diferentes.

Para ilustrar el análisis de varianza multifactorial con un ejemplo presentamos a continuación los resultados de un análisis con los datos de la *Agència per a la Qualitat del Sistema Universitari de Catalunya* (AQU) de la Generalitat de Catalunya con los que se vienen realizando un seguimiento de los graduados universitarios desde el año 2001.

Con una periodicidad cuatrienal se realiza una encuesta[12] que indaga sobre la inserción laboral de los graduados/as universitarios a partir de una muestra de aproximadamente la mitad de los titulados/as[13]. Trataremos los datos del último estudio del año 2014, personas que egresaron en el año 2010 y fueron encuestadas cuatro años después, con la selección de algunas variables que se presentan en la matriz de datos de SPSS, AQU.sav, o de R, AQU.rda.

Consideramos como variable dependiente cuantitativa un índice de calidad ocupacional (ICO) construido según el criterio de Corominas et al. (2007) disponible directamente en la base de datos. El índice resume la calidad de la inserción laboral de los graduados/as universitarios combinando 4 indicadores: el tipo y duración del contrato de trabajo, el salario, la adecuación de los estudios realizados con el trabajo que realiza (*matching*) y la satisfacción en el trabajo en general. El índice varía entre 0 y 100, donde el valor más bajo refleja una menor calidad de los resultados de inserción laboral y el valor más alto representa la más alta calidad (Corominas et al., 2007: 127-136). Analizaremos el índice en función de dos factores o variables independientes: el área de conocimiento, a partir de la agrupación de las distintas titulaciones en 6 grupos: Humanidades, Sociales, Economía y Derecho, Experimentales, Salud y Técnica, y el sexo. Los datos de las medias se recogen en la Tabla III.8.14 y se representan en el Gráfico III.8.12.

Tabla III.8.14. Medias del Índice de calidad ocupacional según el Área de estudio y el Sexo

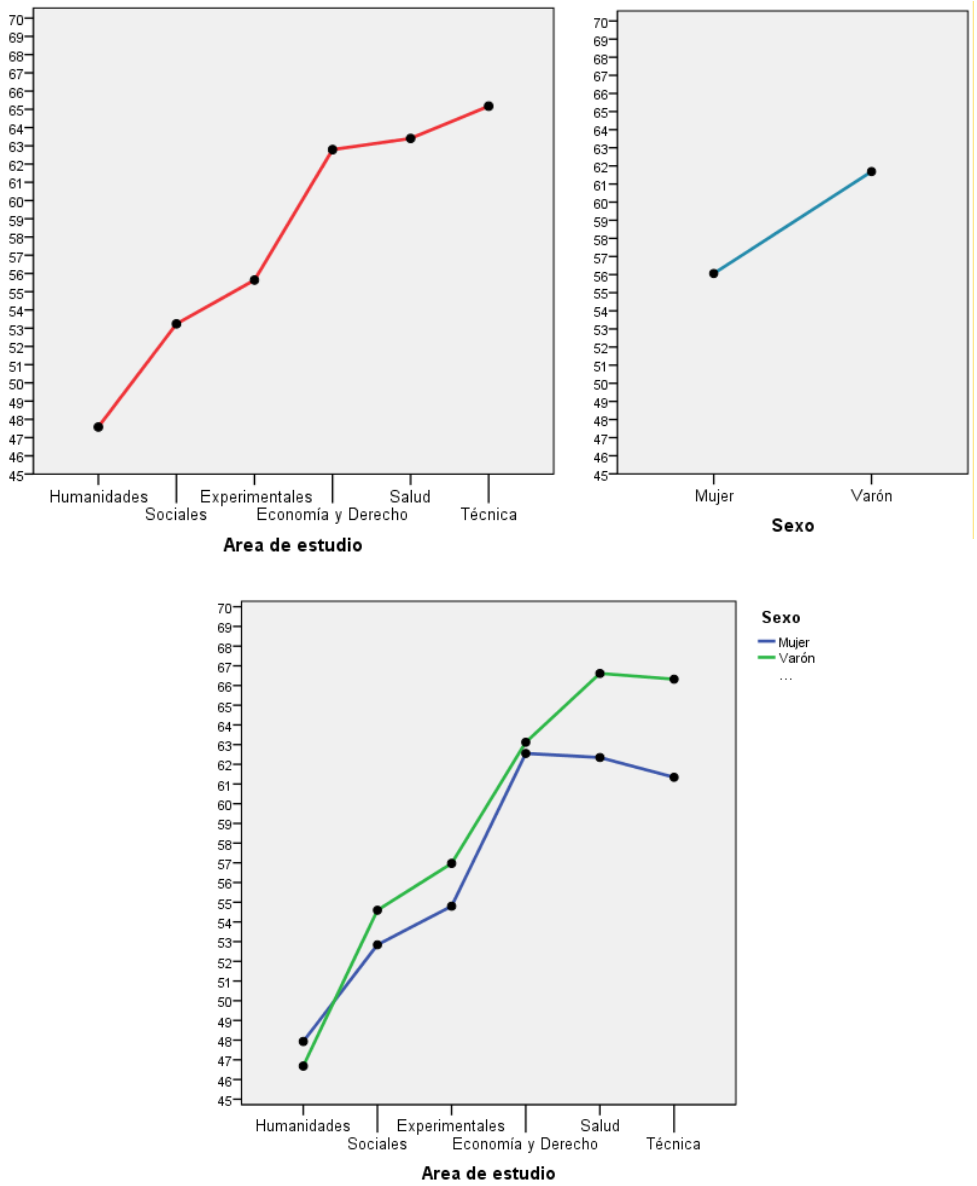
| Area de estudio | Media | | | Desviación típica | | | Casos | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sexo | | | Sexo | | | Sexo | | |
| | Mujer | Varón | Total | Mujer | Varón | Total | Mujer | Varón | Total |
| Humanidades | 47,9 | 46,7 | **47,6** | 21,5 | 21,8 | 21,6 | 649 | 255 | 904 |
| Sociales | 52,8 | 54,6 | **53,2** | 20,4 | 21,3 | 20,6 | 2.154 | 628 | 2.782 |
| Economía y Derecho | 54,8 | 57,0 | **55,6** | 22,1 | 22,3 | 22,2 | 388 | 244 | 632 |
| Experimentales | 62,6 | 63,1 | **62,8** | 19,8 | 20,8 | 20,2 | 886 | 633 | 1.519 |
| Salud | 62,3 | 66,6 | **63,4** | 17,3 | 17,4 | 17,4 | 705 | 232 | 937 |
| Técnica | 61,3 | 66,3 | **65,2** | 20,4 | 18,5 | 19,1 | 477 | 1.593 | 2.070 |
| Total | **56,1** | **61,7** | **58,3** | 20,8 | 20,8 | 21,0 | 5.259 | 3.585 | 8.844 |

Los datos muestran, desde un punto de vista descriptivo, cómo las titulaciones del área de Salud, de Economía y Derecho y Técnica alcanzan mayores valores en el índice de calidad ocupacional frente a las titulaciones del área de Experimentales, Sociales y Humanidades. Se observa asimismo que los varones tienen un valor superior a las mujeres.

[12] El cuestionario de 2014 se puede consultar en http://www.aqu.cat/doc/doc_24468900_1.pdf.

[13] Las características generales del estudio de AQU se pueden consultar en la página web http://www.aqu.cat/estudis/ilaboral_2014_es.html#.Vhpqsitzrfc . También se puede consultar AQU (2014), Planas y Fachelli (2010), Fachelli y Planas (2014), Fachelli y Montolio (2015).

Gráfico III.8.12. Gráficos de medias del Índice de calidad ocupacional según el Área de estudio y el Sexo




Las cuestiones que se suscitan a continuación son: ¿existen diferencias significativas entre las medias definidas por las categorías de cada una de las variables? ¿Existe un efecto de interacción entre ambas? Mediante un análisis de varianza comprobaremos estas cuestiones. La tabla ANOVA con las pruebas de los efectos (inter sujetos) es la siguiente es la Tabla III.8.15.

En conjunto tenemos una reducida capacidad explicativa del modelo, ya que el eta cuadrado o el R cuadrado nos indican que es del 9%. Si miramos el eta cuadrado parcial comprobamos que el efecto del sexo y de la interacción es muy reducido, la variabilidad en la calidad de la ocupación se explica básicamente por las diferencias entre las áreas de estudio. ¿Dónde se localizan esas diferencias? Debemos proceder a realizar las comparaciones entre los grupos para determinar entre qué categorías se dan esas diferencias.

Tabla III.8.15. Tabla ANOVA con las pruebas de los efectos del Índice de calidad ocupacional según el Área de estudio y el Sexo

| Fuente | Suma de cuadrados tipo III | Grados de libertad | Media cuadrática | F | Sig. | Eta cuadrado parcial |
|---|---|---|---|---|---|---|
| Modelo corregido | 347.513,398* | 11 | 31592,127 | 78,607 | 0,000 | 0,089 |
| Intersección | 19.110.871,254 | 1 | 19110871,254 | 47.551,485 | 0,000 | 0,843 |
| Area | 234.762,303 | 5 | 46952,461 | 116,827 | 0,000 | 0,062 |
| Sexo | 6.161,381 | 1 | 6161,381 | 15,331 | 0,000 | 0,002 |
| Area * Sexo | 6.842,236 | 5 | 1368,447 | 3,405 | 0,004 | 0,002 |
| Error | 3.549.567,661 | 8.832 | 401,899 | | | |
| Total | 33.999.458,162 | 8.844 | | | | |
| Total corregida | 3.897.081,059 | 8.843 | | | | |

* R cuadrado = 0,089 (R cuadrado corregido = 0,088)

Tabla III.8.16. Tabla de comparaciones múltiples del Índice de calidad ocupacional según el Área de estudio y el Sexo. Pruebas de Tukey

| | Hum | Soc | EyD | Exp | Sal | Tec |
|---|---|---|---|---|---|---|
| Hum | - | | | | | |
| Soc | *** | - | | | | |
| EyD | *** | *** | - | | | |
| Exp | *** | n.s. | | - | | |
| Sal | *** | *** | n.s. | *** | - | |
| Tec | *** | *** | ** | *** | n.s. | - |

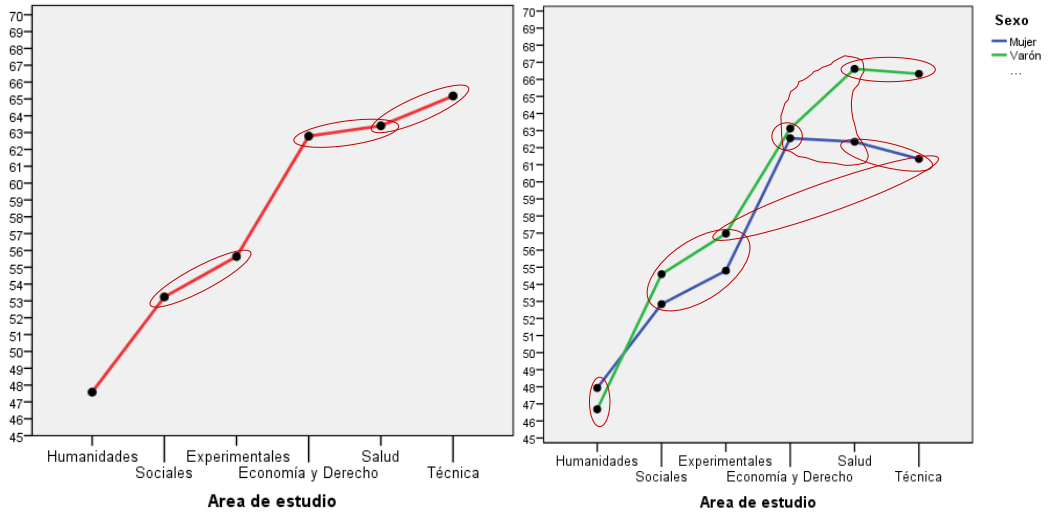
| | | Mujer | | | | | | Varón | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Hum | Soc | EyD | Exp | Sal | Tec | Hum | Soc | EyD | Exp | Sal | Tec |
| Mujer | Hum | - | | | | | | | | | | | |
| | Soc | *** | - | | | | | | | | | | |
| | EyD | *** | *** | - | | | | | | | | | |
| | Exp | *** | n.s. | | - | | | | | | | | |
| | Sal | *** | *** | n.s. | *** | - | | | | | | | |
| | Tec | *** | *** | n.s. | *** | n.s. | - | | | | | | |
| Varón | Hum | n.s. | *** | *** | *** | *** | *** | - | | | | | |
| | Soc | *** | n.s. | *** | n.s. | *** | *** | *** | - | | | | |
| | EyD | *** | *** | n.s. | *** | n.s. | n.s. | *** | *** | - | | | |
| | Exp | *** | n.s. | ** | n.s. | * | n.s. | *** | n.s. | ** | - | | |
| | Sal | *** | *** | n.s. | *** | n.s. | * | *** | *** | n.s. | *** | - | |
| | Tec | *** | *** | *** | *** | *** | *** | *** | *** | * | *** | n.s. | - |

***$p$<.001, **$p$<.01, *$p$<.05, n.s. no significativa

La Tabla III.8.16 resume las múltiples comparaciones que resultan del cruce de las distintas categorías. Se aplica método de la Diferencia Significativa Honesta de Tukey (HSD) que compara todos los posibles pares de medias, y se basa en una distribución de rango estudentizada, similar a la distribución de *t* de la prueba de la *t*. Se trata de una prueba conservadora cuando hay tamaños de muestra desiguales.

En primer lugar, concluimos, por la significación de la variable Sexo y por tener dos categorías, que entre varones y mujeres sus medias son diferentes, 56,1 y 61,7. Por áreas de conocimiento no se dan diferencias significativas en tres comparaciones: entre Experimentales y Sociales, Economía y Derecho y Salud, y entre Salud y Técnicas. Las comparaciones cruzadas permiten constatar que varones y mujeres tienen el mismo nivel de calidad ocupacional en cada área a excepción de las titulaciones técnicas donde las mujeres tienen índice ICO significativamente menor, 61,3 frente a 66,3. Estas diferencias son las que motivan la existencia de una muy débil interacción entre las variables. En el Gráfico III.8.13 se han marcado las diferencias que no son significativas.

Gráfico III.8.13. Comparaciones múltiples del Índice de calidad ocupacional según el Área de estudio y el Sexo




## 7. El análisis de varianza con SPSS

El análisis de varianza con SPSS se puede realizar a través de diversos procedimientos, contemplando en particular también los comandos complementarios de exploración o descripción de los datos con tablas y gráficos:

- El comando **MEANS**, en el menú: Analizar / Comparar medias / Medias, elabora tablas con la información de las medias y los estadísticos univariables de la variable dependiente cuantitativa para cada subgrupo definido por una o más variables independientes cualitativas. Además proporciona la tabla ANOVA y calcula el estadístico eta cuadrado.
- El comando **EXAMINE**, en el menú: Analizar / Estadísticos descriptivos / Explorar, genera igualmente estadísticos de resumen univariables y además proporciona diversas representaciones gráficas de los datos de gran utilidad, tanto

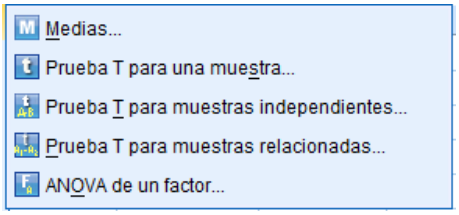
por el conjunto de los datos como para cada subgrupo de la variable cualitativa. Nos permite también comprobar algunas condiciones de nuestros datos como la normalidad y la homoscedasticidad, pero no realiza el análisis de varianza.

- El comando **T-TEST** efectúa el análisis de comparación de medias tanto con muestras independientes (en el menú Analizar / Comparar medias / Muestras independientes) como apareadas (en el menú Analizar / Comparar medias / Muestras relacionadas).
- El comando **ONEWAY**, en el menú: Analizar / Comparar medias / ANOVA de un factor, realiza el análisis de varianza de un factor para una variable dependiente cuantitativa respecto a una única variable de factor o variable independiente cualitativa.
- El comando **UNIANOVA**, en el menú: Analizar / Modelo lineal general / Univariante, considera una sola variable dependiente y realiza un análisis de varianza con uno o más factores o variables independientes, con opciones y tratamientos adicionales que no contempla el procedimiento anterior.
- El comando **GLM** realiza igualmente análisis multifactoriales con una dependiente y además permite un análisis multivariable de la varianza considerando diversas variables dependientes, en el menú: Analizar / Modelo lineal general / Multivariante, así como aplicar un modelo de medidas repetidas: Analizar / Modelo lineal general / Medidas repetidas.[14]
- El comando **VARCOMP**, Analizar / Modelo lineal general / Componentes de la varianza, realiza específicamente un análisis de componentes de la varianza.
- El comando **GENLIN**, Analizar / Modelo lineal generalizado, amplía el modelo lineal general considerando que la variable dependiente está relacionada linealmente con los factores y las covariables mediante una determinada función de enlace que permite reproducir diversas técnicas de análisis de datos.

En los apartados siguientes trataremos con diversos ejemplos los cinco primeros para obtener tanto los resultados descriptivos, complementados con algún gráfico, como el análisis de comparación de dos medias y los dos tipos de modelos de ANOVA, primero el análisis de varianza unifactorial y a continuación el multifactorial.

## 7.1. Análisis de la comparación de dos medias

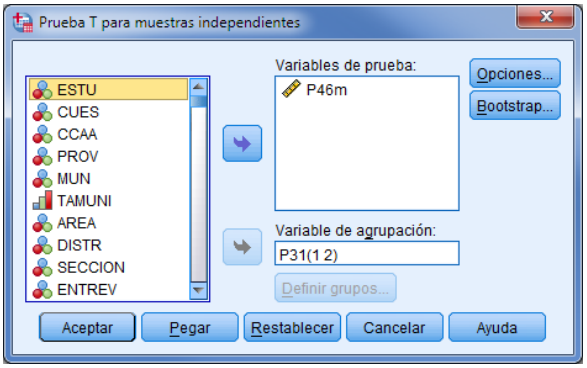
La prueba estadística de la comparación de dos medias supone aplicar la prueba de la t de student. En SPSS se realiza con el comando **T-TEST** el cual permite realizar diversos contrastes de hipótesis. Se accede a través del menú Analizar / Comparar medias, donde se puede optar por estas alternativas:



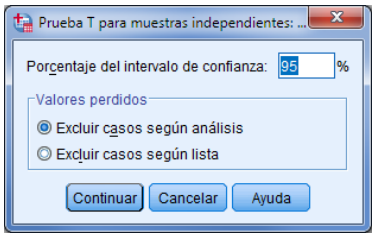

[14] El planteamiento de los modelos lineales general se extiende a otras técnicas y podemos utilizar el procedimiento **GLM** también para realizar un análisis de regresión lineal.

Las opciones de Prueba T para muestras independientes y Prueba T para muestras relacionadas ejecutan la comparación de dos medias para cada tipo de diseño.

Reproduciremos el ejemplo que vimos en la Tabla III.8.8 en el aparatado 4.1. donde se relacionan los ingresos individuales (P46m) y el sexo (P31) con los datos de la matriz CIS3041+.sav y el archivo de sintaxis AVA-Ingresos.sps. Se trata de un diseño de una comparación para muestras independientes que definen los dos valores de la variable sexo. A través del menú y desde la ventral principal del procedimiento:




Situaremos la variable P46m en el recuadro de Variables de prueba, y como Variable de agrupación la P31. En este último caso nos aparecerán dos interrogantes (? ?) donde debemos poner los valores de la variable que serán comparados a través del botón Definir grupos, en nuestro caso serán el 1 (de los varones) y el 2 (de las mujeres). Con esto es suficiente. En el botón Opciones se podría elegir el nivel de confianza, pero lo dejaremos con el valor por defecto del 95%.




Tras clicar sobre Continuar y Aceptar se obtienen estos resultados que comentamos en 4.1:

**Estadísticas de grupo**

| | P31 Sexo de la persona entrevistada | N | Media | Desviación estándar | Media de error estándar |
|---|---|---|---|---|---|
| P46m Ingresos individuales (marca de clase) | 1 Hombre | 923 | 986,46 | 852,507 | 28,061 |
| | 2 Mujer | 996 | 549,55 | 601,437 | 19,057 |

**Prueba de muestras independientes**

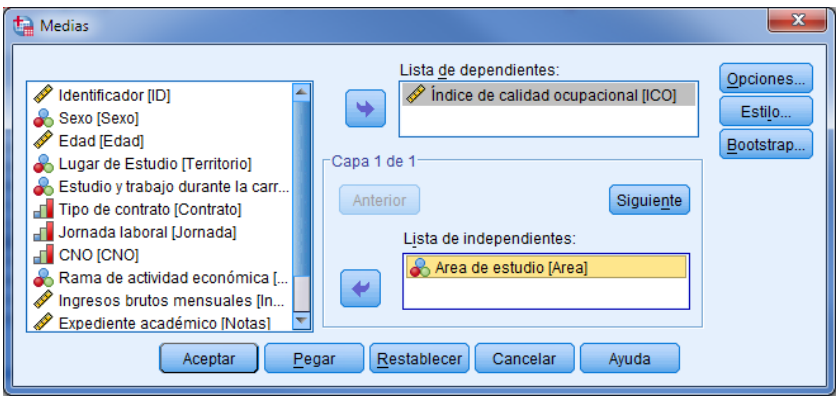
| | | Prueba de Levene de calidad de varianzas | | prueba t para la igualdad de medias | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% de intervalo de confianza de la diferencia | |
| | | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Diferencia de error estándar | Inferior | Superior |
| P46m Ingresos individuales (marca de clase) | Se asumen varianzas iguales | 28,3 | ,000 | 13,046 | 1917 | ,000 | 436,909 | 33,490 | 371,229 | 502,590 |
| | No se asumen varianzas iguales | | | 12,881 | 1644,489 | ,000 | 436,909 | 33,920 | 370,378 | 503,440 |

Es decir, que la diferencia de ingresos entre varones y mujeres es de 437€ y que esta diferencia es significativa (Sig=0,000), pues la probabilidad es inferior a 0,05.

## 7.2. Análisis unifactorial

Con los datos de la *Agència per a la Qualitat del Sistema Universitari de Catalunya* (AQU) de de la matriz de datos AQU.sav, en primer lugar, realizaremos un análisis de varianza unifactorial para explicar la calidad ocupacional en función del área de estudios. Como hipótesis planteamos que la calidad de la ocupación varía según el área de conocimiento, esperando obtener una media de calidad mayor en las carreras técnicas y de la salud que en las carreras de humanidades y sociales. El archivo de instrucciones que realiza los distintos análisis que se presentan a continuación es AVA-AQU.sps.

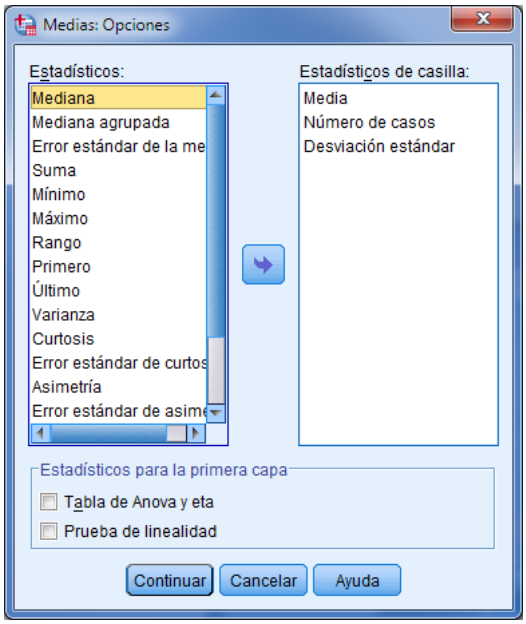
Nuestra variable dependiente cuantitativa es un índice de calidad ocupacional (ICO) que varía entre 0 y 100. La variable independiente del área de conocimiento (Area) tiene 6 grupos: Humanidades, Sociales, Economía y Derecho, Experimentales, Salud y Técnica. Iniciamos el análisis con el análisis descriptivo de la comparación de las medias del índice de calidad ocupacional según el área de conocimiento a través del procedimiento **MEANS** que por el menú se ejecuta en Analizar / Comparar medias / Medias. El procedimiento está destinado al cálculo de las medias de una variable dependiente cuantitativa en el interior de cada categoría o grupo definido por valores de una variable independiente cualitativa, o por la combinación de los valores de dos o más variables independientes. La información se presenta en forma de tabla con tantas casillas como grupos las haya y en el interior de las cuales, además de las medias, se pueden solicitar los diversos estadísticos univariantes. Este es el cuadro de diálogo del procedimiento:




ICO será la variable dependiente cuantitativa que queremos relacionar con la variable independiente Area, calculando la media del índice ocupacional en cada área de estudio para compararlas entre sí. Si hubiéramos tenido dos variables independientes para definir los grupos entonces habríamos tenido que usar la Capa 2 para especificar la segunda variable independiente, y sucesivas capas si tuviéramos más variables de grupo que se cruzan.

Para completar las especificaciones el botón de Opciones permite acceder al cuadro donde se detallan los estadísticos que se calcularán para cada grupo, así como la opción para pedir la tabla del análisis de varianza unifactorial, las medidas de asociación eta y

eta cuadrado para medir la varianza explicada, y los contrastes de linealidad del $R^2$ para medir la bondad del ajuste sobre si las categorías de la variable independiente están ordenadas[15]. Para realizar la comparación podemos optar por diversos estadísticos. Por defecto, se calculan la media, el número de casos y la desviación estándar o típica:




Podemos anadir a la lista que aparece en Estadísticos de casilla el mínimo, el máximo y la mediana. También podemos marcar los dos estadísticos denominados Tabla de Anova y eta y Prueba de linealidad. Como resultado de la ejecución obtenemos estas tablas que seguidamente comentaremos, las descriptivas:

**Resumen de procesamiento de casos**

| | Casos | | | | | |
|---|---|---|---|---|---|---|
| | Incluido | | Excluido | | Total | |
| | N | Porcentaje | N | Porcentaje | N | Porcentaje |
| ICO Índice de calidad ocupacional * Area Area de estudio | 8844 | 74,5% | 3022 | 25,5% | 11866 | 100,0% |

**Informe**

ICO Índice de calidad ocupacional

| Area Area de estudio | Media | N | Desviación estándar | Mínimo | Máximo | Mediana |
|---|---|---|---|---|---|---|
| 1 Humanidades | 47,5797 | 904 | 21,56933 | ,00 | 100,00 | 46,2963 |
| 2 Sociales | 53,2341 | 2782 | 20,58106 | ,00 | 100,00 | 55,5556 |
| 3 Economía y Derecho | 62,7892 | 1519 | 20,22911 | ,00 | 100,00 | 64,8148 |
| 4 Experimentales | 55,6376 | 632 | 22,16158 | ,00 | 100,00 | 59,2593 |
| 5 Salud | 63,4037 | 937 | 17,38771 | ,00 | 100,00 | 64,8148 |
| 6 Técnica | 65,1749 | 2070 | 19,06852 | ,00 | 100,00 | 72,2222 |
| Total | 58,3413 | 8844 | 20,99278 | ,00 | 100,00 | 60,1852 |

[15] Calcula la suma de cuadrados, los grados de libertad y la media cuadrática asociados con las componentes lineal y no lineal, así como la razón $F$, $R$ y $R^2$.

y las de las pruebas estadísticas:

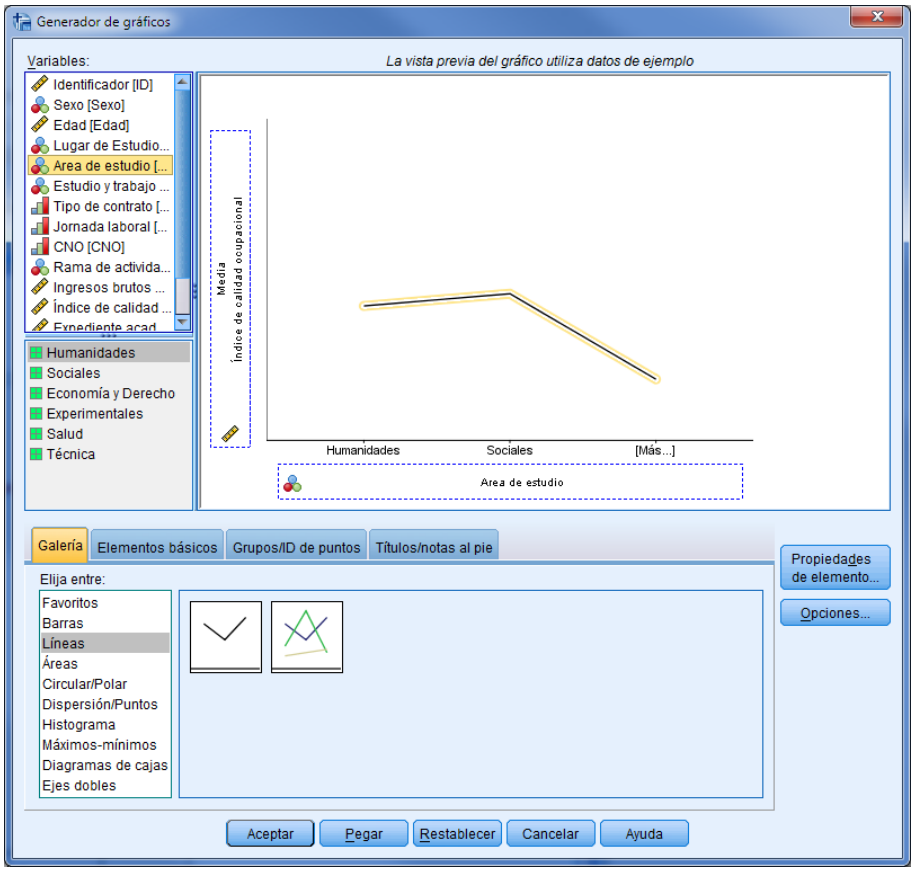
**Tabla de ANOVA**

| | | | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|---|---|---|---|---|---|---|---|
| ICO Índice de calidad ocupacional * Area Area de estudio | Entre grupos | (Combinado) | 332609,317 | 5 | 66521,863 | 164,939 | ,000 |
| | | Linealidad | 256007,454 | 1 | 256007,454 | 634,763 | ,000 |
| | | Desviación de la linealidad | 76601,862 | 4 | 19150,466 | 47,483 | ,000 |
| | Dentro de grupos | | 3564471,742 | 8838 | 403,312 | | |
| | Total | | 3897081,059 | 8843 | | | |

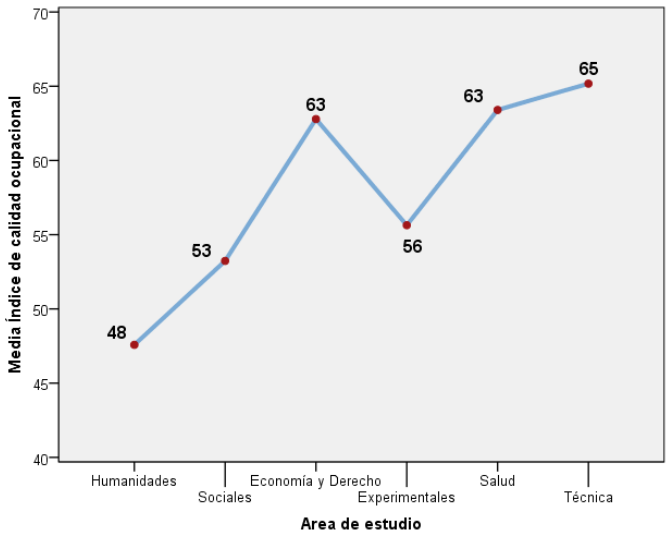
**Medidas de asociación**

| | R | R al cuadrado | Eta | Eta cuadrada |
|---|---|---|---|---|
| ICO Índice de calidad ocupacional * Area Area de estudio | ,256 | ,066 | ,292 | ,085 |

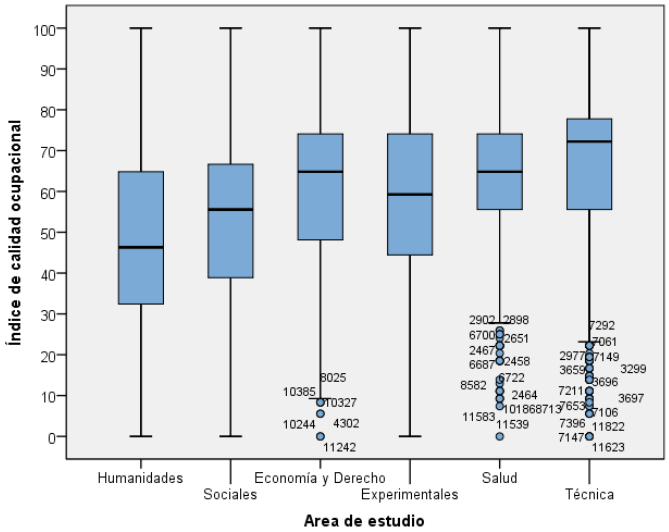
A continuación, para completar el ejercicio de análisis descriptivo, podemos generar un gráfico de medias a través del menú Gráficos / Generador de gráficos eligiendo un gráfico de líneas simple en la Galería y colocando sobre el eje Y la variable ICO y sobre el eje X la variable Area.




Una vez ejecutado el procedimiento se puede editar para cambiar su aspecto: color y grosor de la línea, mostrar los marcadores de líneas, etiquetar los datos, aumentar el tamaño de letra y cambiar la escala. El resultado puede ser el siguiente:




También podemos generar un diagrama de caja simple eligiendo esta opción de la Galería del generador de gráficos. Después de obtenerlo y editarlo adopta esta imagen:




Con la información generada podemos observar cómo Humanidades, Sociales y Experimentales alcanzan los niveles inferiores de calidad ocupacional mientras que Economía y Derecho junto a Salud y Técnica poseen valores de media y mediana superiores en el índice. La prueba estadística de la tabla ANOVA es significativa[16], la probabilidad de 0,000, inferior a 0,05, nos indica que entre las distintas áreas se da, por lo menos, que la media de una de ellas es diferente de las demás. No sabemos si se da entre más grupos ni entre cuáles, eso lo veremos más adelante.[17]

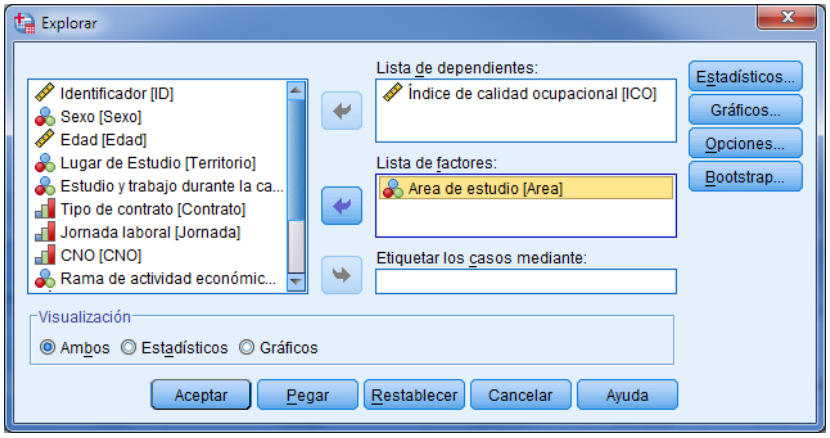
[16] Con heteroscedasticidad se recomienda introducir un ajuste con una prueba alternativa como veremos.

[17] La información de la tabla ANOVA se completa con las pruebas estadísticas de linealidad. Nos informa de la variación debida a una relación lineal entre las variables. Un nivel de significación inferior al 0,05 nos indica que existe una relación lineal. La desviación de la linealidad muestra de forma complementaria una variación debida a una relación no lineal entre las variables, que implicaría un nivel de significación inferior al 0,05. Por tanto, en este caso se da una relación lineal, se produce un aumento del índice de ocupación que permite ordenar a las diferentes áreas de conocimiento como se aprecia en el gráfico de medias. En el gráfico de medias anterior podemos situar en el eje al área de Experimentales antes del de Economía y Derecho y la tendencia creciente se evidenciaría claramente.

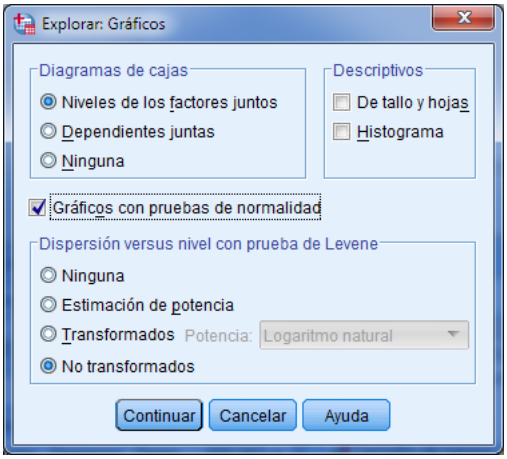
Las medidas de asociación R y R cuadrado se interpretan ante una situación de linealidad. El $R^2$, nos informa de la proporción de varianza explicada de la variable dependiente por el modelo lineal. En este caso un valor bajo del 0,066, es decir, del 6,6%. Los estadísticos eta y eta cuadrado no asumen una relación lineal entre las variables, se interpretan en el mismo sentido, dan la proporción de varianza explicada por la diferencia entre los grupos, y son preferibles cuando no se da linealidad. En este caso el valor es algo superior, de un 8,5%.

Para determinar si existen diferencias significativas y entre qué áreas procedemos a realizar un análisis de varianza más completo que el que nos ha proporcionado el procedimiento **MEANS**. Para ello debemos verificar primeramente los supuestos de normalidad y homoscedasticidad.

Este ejercicio de exploración sobre las condiciones que deben cumplir nuestros datos se puede realizar inicialmente con el procedimiento de análisis exploratorio, el comando **EXAMINE**, que se ejecuta a través del menú Analizar / Estadísticos descriptivos / Explorar. Colocaremos la variable dependiente cuantitativa ICO y la variable independiente o factor Area que configura los grupos de comparación:




Además de los estadísticos descriptivos y de los gráficos que aparecen por defecto (gráfico de tallo y hojas y diagrama de caja) en el apartado de Gráficos podemos elegir la generación de tablas y gráficos destinados a establecer las pruebas normalidad y homoscedasticidad, clicando sobre Gráficos con pruebas de normalidad, y sobre Estimación de potencia del apartado de la prueba de Levene.



De los resultados que se generan destacamos los dedicados a las dos pruebas citadas. En relación con la normalidad debemos resaltar en primer lugar que disponemos de una amplia muestra dentro de cada grupo que nos exime de la exigencia del cumplimiento de esta condición. Los resultados de la tabla de las pruebas de normalidad[18]:

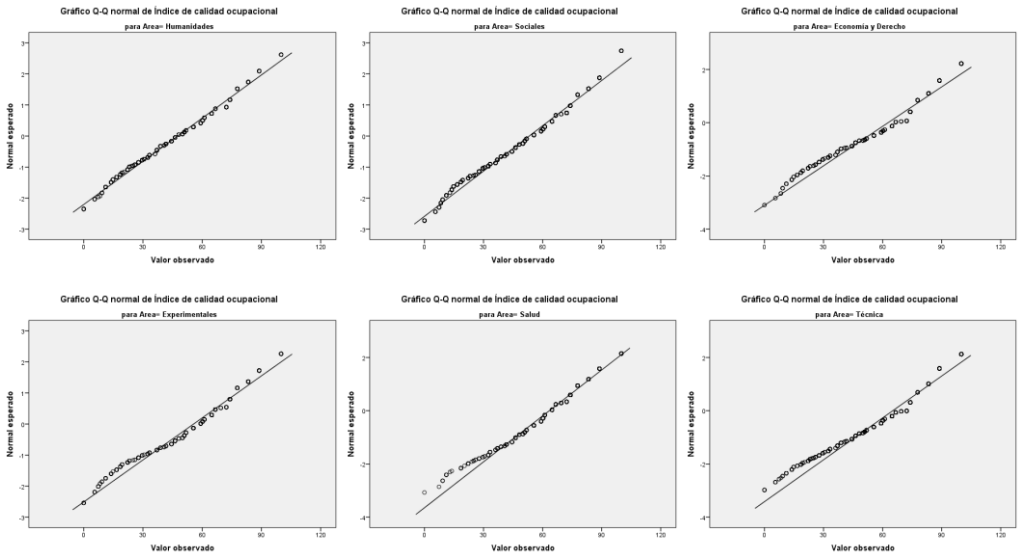
**Pruebas de normalidad**

| | | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | Area Area de estudio | Estadístico | gl | Sig. | Estadístico | gl | Sig. |
| ICO Índice de calidad ocupacional | 1 Humanidades | ,070 | 904 | ,000 | ,986 | 904 | ,000 |
| | 2 Sociales | ,085 | 2782 | ,000 | ,980 | 2782 | ,000 |
| | 3 Economía y Derecho | ,179 | 1519 | ,000 | ,944 | 1519 | ,000 |
| | 4 Experimentales | ,100 | 632 | ,000 | ,966 | 632 | ,000 |
| | 5 Salud | ,092 | 937 | ,000 | ,975 | 937 | ,000 |
| | 6 Técnica | ,179 | 2070 | ,000 | ,946 | 2070 | ,000 |

a. Corrección de significación de Lilliefors

muestran que en el interior de cada grupo (área de estudio) la variable ICO no sigue una distribución normal. Todos los valores de significación son inferiores a 0,05 por lo que se rechaza la hipótesis nula de normalidad. No obstante, como hemos comentado, esta sería una condición necesaria en el caso de muestras pequeñas, de 30 casos o menos. No es nuestra situación y podemos aplicar el análisis de varianza sin restricción.

[18] La prueba Kolmogorov-Smirnov en particular se realiza con un nivel de significación de Lilliefors para el caso en que la media y la varianza poblacionales se desconocen y necesitan ser estimadas. La prueba de Saphiro-Wilk se aplica cuando el tamaño muestral es inferior o igual a 50. La tabla se acompaña con resultados gráficos, los diagramas de normalidad: diagramas de probabilidad normal y de probabilidad sin tendencia. El gráfico Q-Q Normal representa los cuantiles de la distribución de una variable (valores de la variable que dividen los casos en un número de grupos de igual tamaño) y se compara cada valor observado con el valor tipificado que le correspondería en una distribución normal. Por tanto, nos indica en qué medida los datos se alejan de la distribución normal que representa la línea diagonal, se trata de una apreciación visual a veces difícil de establecer y que la tabla determina claramente.




También se genera el llamado gráfico Q-Q Normal sin tendencia que complementa a éstos y se basa en las diferencias entre los valores tipificados y los valores de la normal que corresponden a cada valor observado. Son pues el valor de las distancias verticales en el gráfico Q-Q normal anterior entre cada punto y la diagonal. Si estas diferencias se distribuyen aleatoriamente sin una pauta clara alrededor del eje de abscisas, del valor 0, se supone la hipótesis de normalidad.

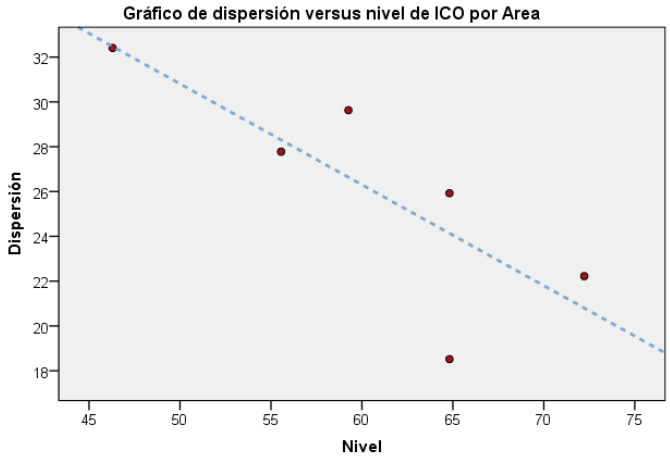
Para ver la prueba de igualdad de varianzas (homoscedasticidad) se aplica la prueba de Levene generándose esta tabla:

**Prueba de homogeneidad de varianza**

| | | Estadístico de Levene | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| ICO Índice de calidad ocupacional | Se basa en la media | 20,236 | 5 | 8838 | ,000 |
| | Se basa en la mediana | 18,572 | 5 | 8838 | ,000 |
| | Se basa en la mediana y con gl ajustado | 18,572 | 5 | 8692,400 | ,000 |
| | Se basa en la media recortada | 20,457 | 5 | 8838 | ,000 |

Las distintas pruebas llevan a la misma conclusión de significatividad, es decir, la probabilidad es inferior a 0,05 por lo que rechazamos la hipótesis nula de igualdad de varianzas y aceptamos que nuestros datos son heterocedásticos, es decir, la varianza difiere entre los distintos grupos, no es homogénea. Esta conclusión la deberemos tener en cuenta a la hora de aplicar el ANOVA unifactorial.

Al ejecutar el procedimiento elegimos la opción en Gráficos de No transformados para la prueba de Levene. Así obtenemos la tabla anterior y el llamado Gráfico de dispersión por nivel:

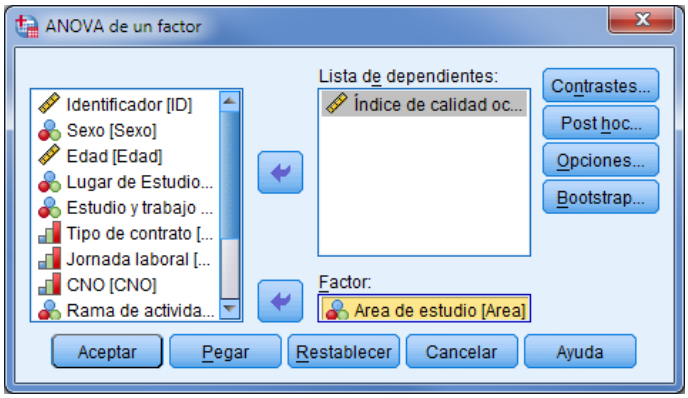



El gráfico de dispersión por nivel nos permite valorar la igualdad de varianzas entre los grupos. Si se observa que los puntos no se encuentran alineados horizontalmente y diera un patrón lineal de los puntos, cercano por ejemplo a la línea discontinua que se ha añadido, y el valor de la pendiente no fuera cercano a cero entonces no se puede asumir la igualdad de varianzas. Es así en este caso y nos indica que las varianzas no son homogéneas, se da heteroscedasticidad[19].

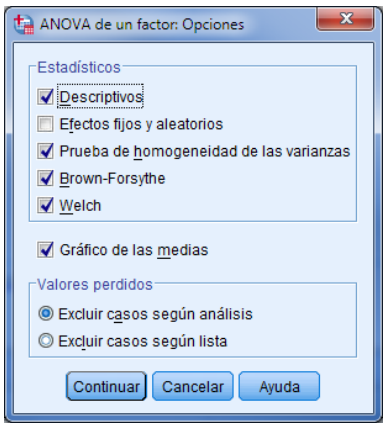
Analizados nuestros datos descriptivamente y comprobados los supuestos de normalidad y homogeneidad de varianzas, nos disponemos seguidamente a realizar un

[19] Cuando se constata la no homogeneidad es habitual aplicar algún tipo de transformación a los datos originales para conseguir homogenizarlas. Esta transformación se realiza en potencias. Una potencia para la transformación sugerida de 1 como en este caso no requiere ninguna transformación de los datos. Si fuera 0 sería conveniente aplicar una transformación logarítmica, si fuera ½ la raíz cuadrada, de ser 2 se elevaría al cuadrado, si fuera -½ la inversa de la raíz cuadrada y si fuera -1 la inversa.

análisis de varianza de un solo factor para determinar si existen diferencias entre las medias definidas por los grupos y detallar entre qué grupos. Para ello se puede optar por ejecutar el comando **ONEWAY**, que corresponde por el menú a Analizar / Comparar medias / ANOVA de un factor:




En el cuadro de diálogo principal de nuevo ubicamos la variable dependiente ICO y la variable independiente Area como Factor. El procedimiento presenta la tabla de análisis de varianza para cada variable dependiente considerada, pero con el botón de Opciones podemos también especificar la obtención de resultados adicionales: estadísticos descriptivos para cada grupo (número de casos, media, desviación típica, error típico de la media, mínimo, máximo e intervalo de confianza al 95% para la media), descriptivos adiconales referidos a los modelos de efectos fijos y aleatorios, además podemos pedir la prueba de Levene sobre la homogeneidad de varianzas, las pruebas de Brown-Forsythe y Welch (que también contrastan la igualdad de las medias y son preferibles al contraste *F* cuando no se puede mantener el supuesto de igualdad de varianzas), un gráfico de representación de las medias y precisar los tratamiento de los valores perdidos.

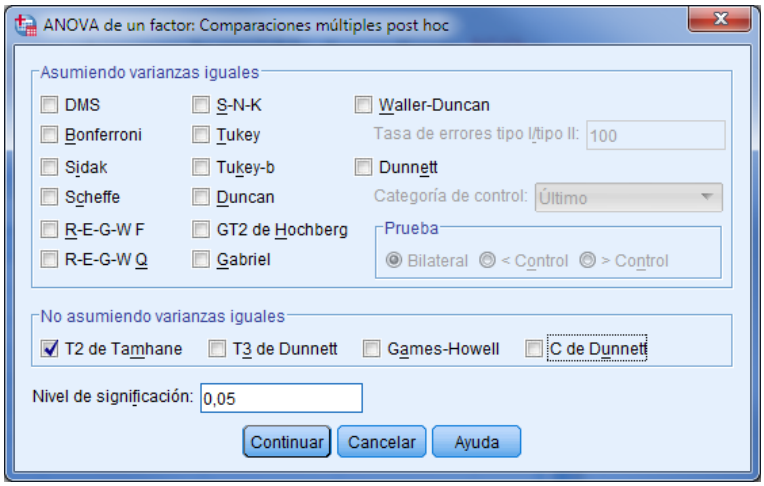



Como estamos ante una situación de heteroscedasticidad, el análisis deberá aplicarse con las pruebas equivalentes de Brown-Forsythe y Weclch que tienen en cuenta esta condición de los datos.

El análisis de varianza nos dice si hay diferencias significativas entre las medias de los grupos, pero no entre qué grupos en concreto. La especificación de los contrastes nos permitirá conocer este detalle. Los hay de dos tipos: los contrastes a priori (botón Contrastes) que implica un interés definido previamente por el analista de contrastar

determinados grupos, y los contrastes a posteriori (botón Post hoc) donde se hacen las comparaciones múltiples de todos los pares de medias posibles y que suele ser el tipo de prueba a aplicar analizando datos de encuesta[20].

Utilizaremos pues los contrastes post hoc en el que se comparan sistemáticamente todos los pares de medias a partir de la muestra global de casos. Hay varios métodos de comparación que se aplican distinguiendo dos situaciones según se de homocedasticidad o heteroscedasticidad. Si tuviéramos homogenidad de varianzas disponemos de una amplia variedad de alternativas. Podríamos elegir por ejemplo la prueba de Scheffé, una de las más utilizadas por su flexibilidad y robustez ya que se caracteriza por ser más exigente a la hora de encontrar diferencias significativas entre las medias, además es recomendado cuando los tamaños de los grupos son diferentes y es menos sensible a situaciones de alejamiento del supuesto de normalidad y de igualdad de varianzas poblacionales.




Ante una situación donde las varianzas no son iguales como es nuestro caso podemos optar por la prueba T2 de Tamhane, prueba conservadora de comparación por parejas basada en la prueba t para situaciones donde se infringe la homogeneidad de varianzas y también cuando el tamaño de los grupos es diferente.

Los resultados del ANOVA, o análisis de varianza unidireccional como lo denomina también el software SPSS, se presentan a continuación. En primer lugar aparece la

[20] No trataremos los contrastes a priori. Para especificarlos hay que seguir un procedimiento particular que presentamos brevemente. Todo contraste se basa en la especificación de una serie de coeficientes para cada categoría de la variable independiente, según el orden de codificación, y la suma de los cuales tiene que ser 0. Tomemos como ejemplo el contraste entre tres clases sociales, detallamos a continuación cinco pruebas de contraste diferentes:

| Contraste | Clase social | | |
|---|---|---|---|
| | 1 Baja | 2 Media | 3 Alta |
| 1 | -1 | ,5 | ,5 |
| 2 | ,5 | ,5 | -1 |
| 3 | 1 | -1 | 0 |
| 4 | 1 | 0 | -1 |
| 5 | 1 | -1 | 0 |

El primer contraste compara la categoría baja (−1) con las categorías media y alta conjuntamente (0,5 y 0,5). El segundo contraste es similar comparando las categorías baja y media, conjuntamente, con la categoría alta. En los otros tres contrastes comparamos sólo dos categorías entre sí. El contraste 3 compara la categoría baja (1) y la media (−1), la categoría alta no se considera y se codifica con el 0. El contraste 4 comparar la categoría baja (1) y la alta (−1), y el 5 la categoría media (1) y la alta (−1). En SPSS, en el cuadro de diálogo de contrastes se trata de especificar los diferentes coeficientes que se adjudican a los grupos (también llamados niveles) y que son los que determinan los contrastes a realizar, ya sea por comparaciones de niveles individuales o de combinaciones de niveles como hemos visto.

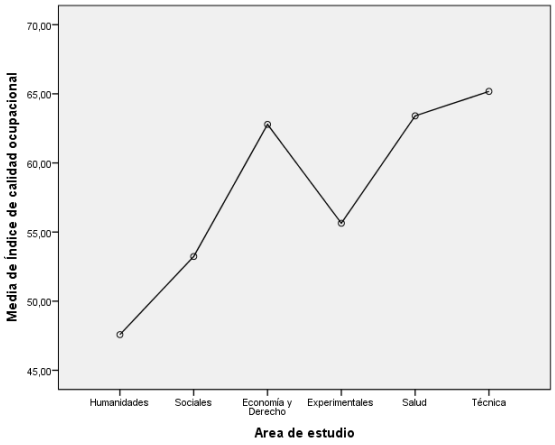
tabla de los estadísticos descriptivos que reproduce la información que vimos anteriormente, junto con el intervalo de confianza de cada media estimada en cada grupo:

**Descriptivos**

ICO Índice de calidad ocupacional

| | N | Media | Desviación estándar | Error estándar | 95% del intervalo de confianza para la media | | Mínimo | Máximo |
|---|---|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior | | |
| 1 Humanidades | 904 | 47,5797 | 21,56933 | ,71739 | 46,1718 | 48,9876 | ,00 | 100,00 |
| 2 Sociales | 2782 | 53,2341 | 20,58106 | ,39020 | 52,4690 | 53,9992 | ,00 | 100,00 |
| 3 Economía y Derecho | 1519 | 62,7892 | 20,22911 | ,51904 | 61,7711 | 63,8073 | ,00 | 100,00 |
| 4 Experimentales | 632 | 55,6376 | 22,16158 | ,88154 | 53,9065 | 57,3687 | ,00 | 100,00 |
| 5 Salud | 937 | 63,4037 | 17,38771 | ,56803 | 62,2889 | 64,5185 | ,00 | 100,00 |
| 6 Técnica | 2070 | 65,1749 | 19,06852 | ,41911 | 64,3530 | 65,9968 | ,00 | 100,00 |
| Total | 8844 | 58,3413 | 20,99278 | ,22323 | 57,9037 | 58,7789 | ,00 | 100,00 |

El análisis descriptivo se puede completar con la presentación del gráfico de medias:




La prueba de homogeneidad de varianza muestra un resultado ya conocido, que estamos en una situación de heteroscedasticidad pues Sig.=0,000<0,05.

**Prueba de homogeneidad de varianzas**

ICO Índice de calidad ocupacional

| Estadístico de Levene | df1 | df2 | Sig. |
|---|---|---|---|
| 20,236 | 5 | 8838 | ,000 |

La tabla ANOVA nos proporciona el contraste de la hipótesis sobre si las medias del índice ocupacional observadas en nuestra muestra son iguales entre cada área de conocimiento o al menos una de ellas es diferente. El estadístico *F* obtenido (164,939) y su significación (valor inferior a 0,05) que nos permite concluir que las medias son significativamente distintas.

**ANOVA**

ICO Índice de calidad ocupacional

| | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|---|---|---|---|---|---|
| Entre grupos | 332609,317 | 5 | 66521,863 | 164,939 | ,000 |
| Dentro de grupos | 3564471,742 | 8838 | 403,312 | | |
| Total | 3897081,059 | 8843 | | | |

A continuación, vemos las comparaciones por parejas entre los grupos, entre las distintas áreas de conocimiento. Aparecen a continuación dos tablas, la original de SPSS con la prueba de Tamhane (valores de diferencia de medias por parejas, significación e intervalo de confianza) y otra simplificada con la información seleccionada de las medias y la indicación de la significación mediante el asterisco (cuando la significación es inferior a 0,05 las diferencias de medias se acompañan de ese símbolo) y otra disposición de la información en filas y columnas después de pivotarla con el software SPSS.

**Comparaciones múltiples**

Variable dependiente: ICO Índice de calidad ocupacional

Tamhane

| (I) Area Area de estudio | (J) Area Area de estudio | Diferencia de medias (I-J) | Error estándar | Sig. | 95% de intervalo de confianza | |
|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior |
| 1 Humanidades | 2 Sociales | -5,65440* | ,81664 | ,000 | -8,0492 | -3,2596 |
| | 3 Economía y Derecho | -15,20955* | ,88546 | ,000 | -17,8054 | -12,6137 |
| | 4 Experimentales | -8,05791* | 1,13655 | ,000 | -11,3915 | -4,7243 |
| | 5 Salud | -15,82400* | ,91504 | ,000 | -18,5068 | -13,1412 |
| | 6 Técnica | -17,59521* | ,83084 | ,000 | -20,0315 | -15,1589 |
| 2 Sociales | 1 Humanidades | 5,65440* | ,81664 | ,000 | 3,2596 | 8,0492 |
| | 3 Economía y Derecho | -9,55515* | ,64935 | ,000 | -11,4578 | -7,6525 |
| | 4 Experimentales | -2,40352 | ,96404 | ,176 | -5,2336 | ,4266 |
| | 5 Salud | -10,16961* | ,68914 | ,000 | -12,1898 | -8,1494 |
| | 6 Técnica | -11,94081* | ,57264 | ,000 | -13,6182 | -10,2634 |
| 3 Economía y Derecho | 1 Humanidades | 15,20955* | ,88546 | ,000 | 12,6137 | 17,8054 |
| | 2 Sociales | 9,55515* | ,64935 | ,000 | 7,6525 | 11,4578 |
| | 4 Experimentales | 7,15164* | 1,02299 | ,000 | 4,1499 | 10,1533 |
| | 5 Salud | -,61445 | ,76945 | 1,000 | -2,8697 | 1,6408 |
| | 6 Técnica | -2,38566* | ,66712 | ,005 | -4,3403 | -,4310 |
| 4 Experimentales | 1 Humanidades | 8,05791* | 1,13655 | ,000 | 4,7243 | 11,3915 |
| | 2 Sociales | 2,40352 | ,96404 | ,176 | -,4266 | 5,2336 |
| | 3 Economía y Derecho | -7,15164* | 1,02299 | ,000 | -10,1533 | -4,1499 |
| | 5 Salud | -7,76609* | 1,04870 | ,000 | -10,8430 | -4,6892 |
| | 6 Técnica | -9,53730* | ,97610 | ,000 | -12,4025 | -6,6721 |
| 5 Salud | 1 Humanidades | 15,82400* | ,91504 | ,000 | 13,1412 | 18,5068 |
| | 2 Sociales | 10,16961* | ,68914 | ,000 | 8,1494 | 12,1898 |
| | 3 Economía y Derecho | ,61445 | ,76945 | 1,000 | -1,6408 | 2,8697 |
| | 4 Experimentales | 7,76609* | 1,04870 | ,000 | 4,6892 | 10,8430 |
| | 6 Técnica | -1,77121 | ,70592 | ,168 | -3,8405 | ,2981 |
| 6 Técnica | 1 Humanidades | 17,59521* | ,83084 | ,000 | 15,1589 | 20,0315 |
| | 2 Sociales | 11,94081* | ,57264 | ,000 | 10,2634 | 13,6182 |
| | 3 Economía y Derecho | 2,38566* | ,66712 | ,005 | ,4310 | 4,3403 |
| | 4 Experimentales | 9,53730* | ,97610 | ,000 | 6,6721 | 12,4025 |
| | 5 Salud | 1,77121 | ,70592 | ,168 | -,2981 | 3,8405 |

*. La diferencia de medias es significativa en el nivel 0.05.

**Comparaciones múltiples**

Variable dependiente: ICO Índice de calidad ocupacional

Tamhane

Diferencia de medias (I-J)

| | (J) Area Area de estudio | | | | | |
|---|---|---|---|---|---|---|
| (I) Area de estudio | 1 Humanidades | 2 Sociales | 3 Economía y Derecho | 4 Experimentales | 5 Salud | 6 Técnica |
| 1 Humanidades | | -5,7* | -15,2* | -8,1* | -15,8* | -17,6* |
| 2 Sociales | **5,7*** | | -9,6* | -2,4 | -10,2* | -11,9* |
| 3 Economía y Derecho | **15,2*** | **9,6*** | | 7,2* | -,6 | -2,4* |
| 4 Experimentales | **8,1*** | **2,4** | **-7,2*** | | -7,8* | -9,5* |
| 5 Salud | **15,8*** | **10,2*** | **0,6** | **7,8*** | | -1,8 |
| 6 Técnica | **17,6*** | **11,9*** | **2,4*** | **9,5*** | **1,8** | |

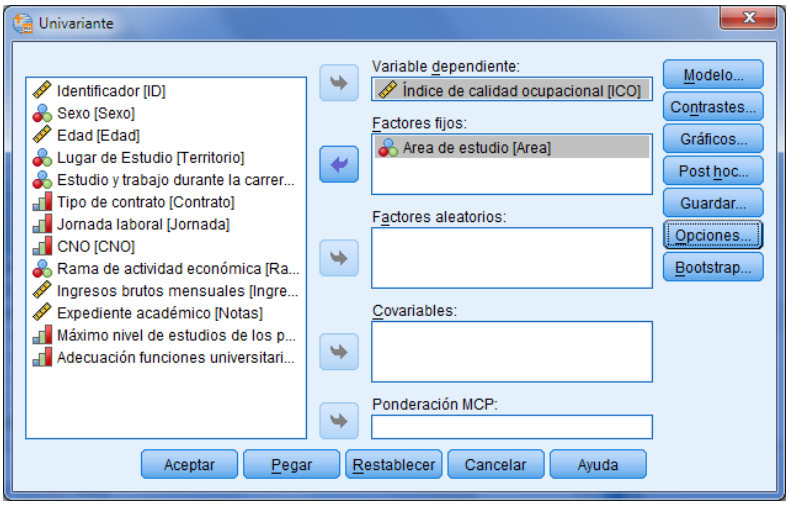
*. La diferencia de medias es significativa en el nivel 0,05.

En todos los casos excepto cuando se comparan las áreas de Experimentales y Sociales, Economía y Derecho con Salud, y Técnica con Salud, las diferencias son significativas. Podemos así validar nuestra hipótesis inicial según la cual se dan diferencias significativas por áreas de conocimiento.

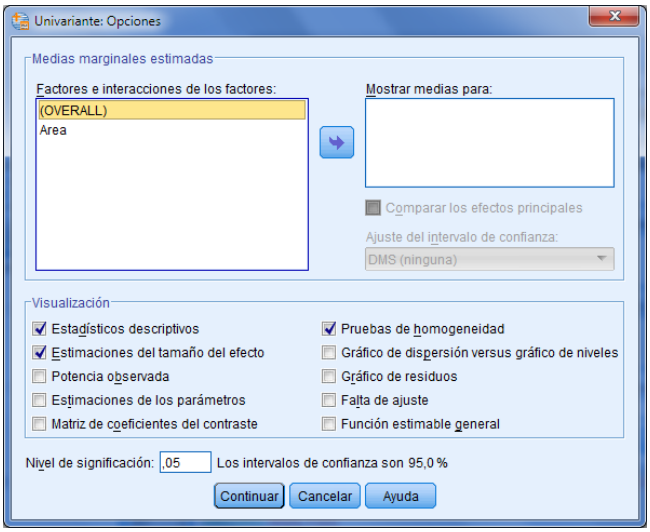
El siguiente paso consiste en cuantificar la magnitud de esas diferencias, es decir, hasta qué punto saber el área de conocimiento estudiada nos permite determinar el nivel de índice ocupacional. Nos preguntamos con estos resultados ¿qué capacidad explicativa alcanza el modelo de dependencia?, ¿en qué medida la variable independiente del área de conocimiento explica el índice ocupacional? Para determinar la intensidad de la relación calcularemos el estadístico eta cuadrado, $\eta^2$, equivalente al coeficiente de determinación $R^2$, como veremos también en el análisis de regresión, que nos mide qué parte de la variación de la variable dependiente ICO es atribuible a la variable Area.

Este resultado no se genera con el comando **ONEWAY**. Lo vimos como resultado del procedimiento **MEANS**, y lo veremos en el contexto del análisis multifactorial. El valor que se obtiene es de 0,085, es decir, que el 8,5% de la varianza total de la variable dependiente es explicada por haberse graduado en un área o en otro. Un valor relativamente bajo si tenemos en cuenta que podríamos llegar al 100%, pero un valor habitual en los estudios por encuesta donde escasas veces, con una variable, se alcanzan niveles del 60%. En todo caso, sin encontrar grandes diferencias y con un alto grado de variabilidad interna, es decir, que en todas las áreas de conocimiento encontramos personas con índices ocupacionales altos y bajos que motiva que el determinismo de la variable dependiente se reduzca. Con todo, se observa una tendencia a que la ocupación sea, en promedio, mejor si se cursan determinados estudios.

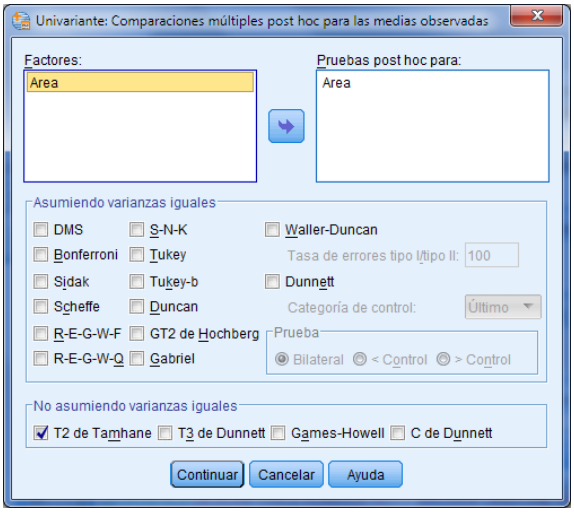
El análisis de varianza unifactorial también se puede realizar con el procedimiento **UNIANOVA** que veremos seguidamente aplicado al análisis multifactorial. Reproduciremos a continuación las opciones básicas. A través de Analizar / Modelo lineal general / Univariante, en primer lugar, colocamos la variable dependiente y la variable independiente o factor:



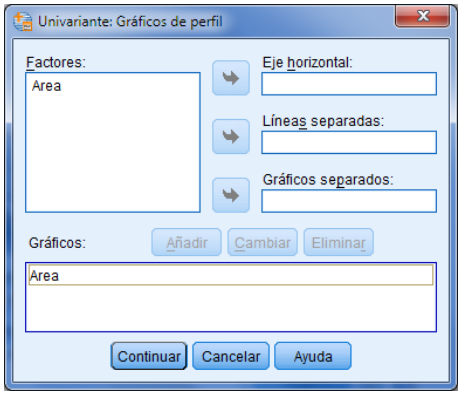
Podemos marcar en Opciones los Estadísticos descriptivos, Pruebas de homogeneidad y las Estimaciones del tamaño del efecto:




Para reproducir una prueba post hoc seleccionaremos la variable independiente y marcaremos la opción de T2 de Tamhane:




El gráfico de medias de cada grupo se puede obtener en el cuadro de diálogo correspondiente ubicando la variable independiente en la casilla de Eje horizontal y clicando Añadir:



Los resultados, son prácticamente los mismos a los obtenidos con **ONEWAY**, la única información presentada de forma diferente es la tabla ANOVA donde podemos ver que se la información del tamaño del efecto de la variable independiente, es decir, el valor de eta cuadrado:

**Pruebas de efectos inter-sujetos**

Variable dependiente: ICO Índice de calidad ocupacional

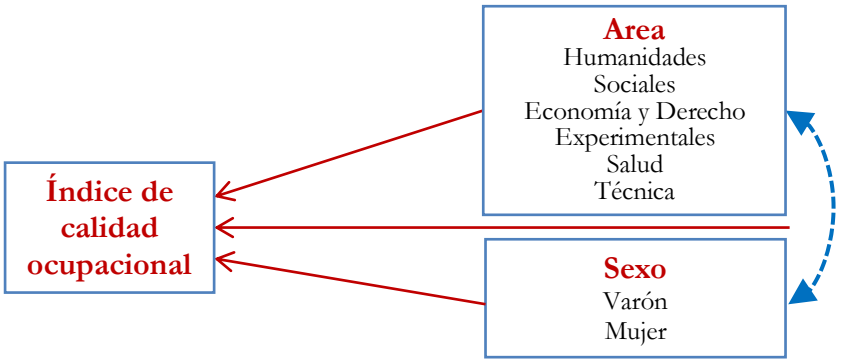
| Origen | Tipo III de suma de cuadrados | gl | Cuadrático promedio | F | Sig. | Eta parcial al cuadrado |
|---|---|---|---|---|---|---|
| Modelo corregido | 332609,317[a] | 5 | 66521,863 | 164,939 | ,000 | ,085 |
| Interceptación | 23014610,49 | 1 | 23014610,49 | 57064,031 | ,000 | ,866 |
| Area | 332609,317 | 5 | 66521,863 | 164,939 | ,000 | ,085 |
| Error | 3564471,742 | 8838 | 403,312 | | | |
| Total | 33999458,16 | 8844 | | | | |
| Total corregido | 3897081,059 | 8843 | | | | |

a. R al cuadrado = ,085 (R al cuadrado ajustada = ,085)

En el siguiente apartado nos detendremos en la interpretación de la misma.

## 7.3. Análisis multifactorial

Realizaremos un análisis de varianza multifactorial con el procedimiento **UNIANOVA**[21] y seguiremos el ejemplo de la relación entre el índice de calidad ocupacional (ICO) y las variables independientes Area y Sexo a partir del modelo que gráfica se representa a continuación.




Reproducimos a continuación los resultados que se obtienen de tablas y gráficos tras la ejecución por el menú de SPSS Analizar / Modelo lineal general / Univariante.[22]

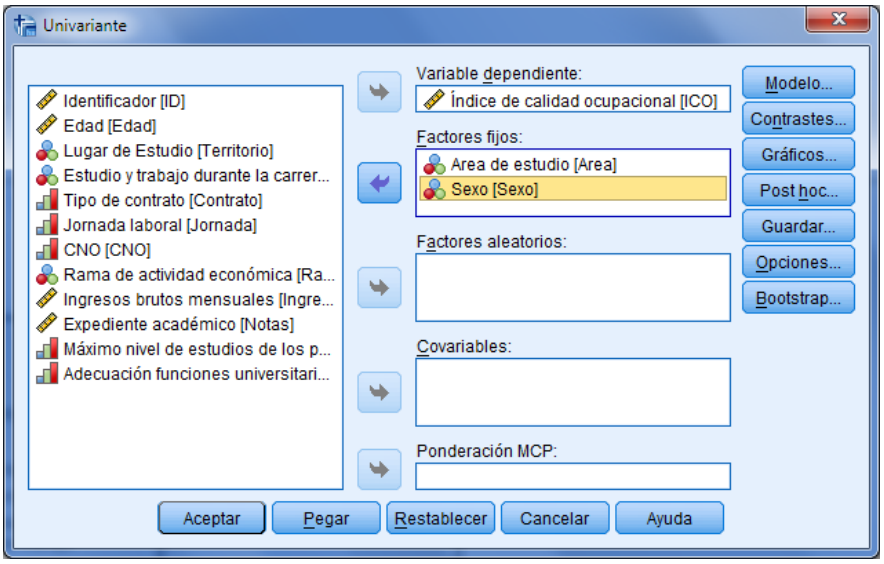
Este procedimiento trata modelos donde se considera una variable dependiente (o respuesta) que es medida con una escala cuantitativa, y con las que se debe dar el cumplimiento de varios supuestos paramétricos, junto con una o más variables independientes (factores) medidas como variables cualitativas o categóricas. Adicionalmente, se pueden considerar en el modelo una o más variables covariantes, es decir, variables predictoras cuantitativas que se relacionan con la variable dependiente y que se consideran con una función de control estadístico de la variable

[21] El análisis de varianza con múltiples variables independientes con SPSS se puede realizar también con el comenado **GLM** (*General Linear Model*), se generan los mismos resultados.
[22] El archivo de instrucciones que realiza los distintos análisis que se presentan es AVA-AQU.sps.

dependiente. Este planteamiento corresponde a la llamada análisis de covarianza (ANCOVA), y se trata de contrastar si los resultados de un ANOVA se ven alterados cuando se introduce la variable de control. El modelo puede incluir igualmente las posibles interacciones entre los factores, las interacciones entre las covariantes, y entre factores y covariantes.

Las variables independientes pueden considerarse como factores fijos o factores aleatorios. Si son factores fijos, se incluyen todos los niveles sobre los que se desea extraer conclusiones, el investigador/a fija los niveles (en un contexto experimental) o bien vienen dados por las características del factor (por ejemplo, si se considera la variable sexo, los niveles son hombre y mujer). Si son factores aleatorios, en este caso los niveles de un factor aleatorio son una muestra aleatoria de los posibles niveles, y luego se extrapolan los resultados al resto[23].

Considerando nuestras dos variables independientes, Area y Sexo, como factores fijos en el análisis, el modelo se especifica de esta en el cuadro de diálogo principal:

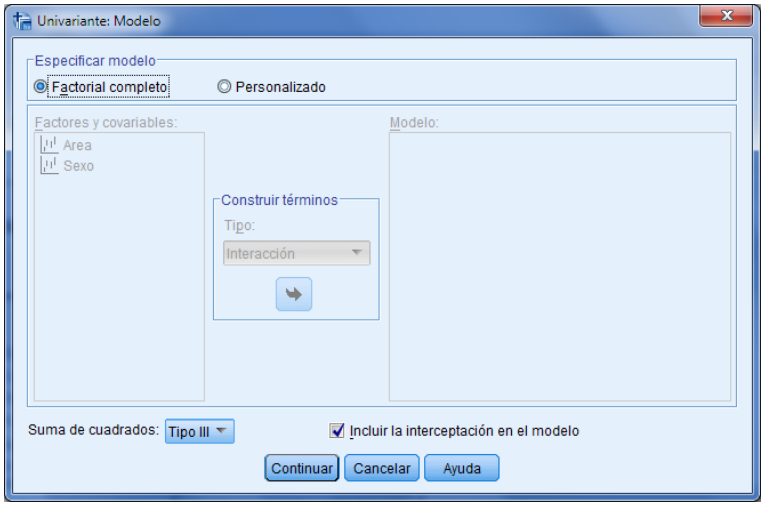



El procedimiento permite especificar el modelo factorial que será analizado a través del botón Modelo. El modelo factorial completo contiene los efectos principales del factor, todos los efectos principales de las covariables y todas las interacciones factor por factor. Esta es la opción por defecto. Alternativamente el modelo se puede personalizar especificando un subconjunto de todos estos efectos. En función de los términos incluidos el modelo recibe nombres diferentes, por ejemplo, de bloques aleatorios o jerárquicos (o anidados).

[23] Aquí nos limitaremos a tratar modelos de factores fijos en un ANOVA multifactorial con dos variables independientes y contrastes a posteriori (post hoc). Otros posibles modelos son posibles. Cuando se consideran factores no relacionados entre ellos el diseño se llama ortogonal; es no ortogonal cuando las variables independientes asociadas. Se pueden contrastar tanto los modelos equilibrados o balanceados (cuando se tiene igual número de casos por grupo o por celda, que se genera por combinación de valores de los factores) como los no equilibrados. También disponemos de los contrastes a priori de hipótesis. Por otra parte, se pueden considerar variables para realizar una ponderación MCP (mínimos cuadrados ponderados) cuando la varianza de la variable dependiente para cada casilla es diferente, es decir, se da heteroscedasticidad. De esta manera se ponderan las observaciones de forma diferente con el fin de compensar la distinta precisión de las medidas (las que tienen menos variabilidad serán más precisas y estas se consideran con mayor importancia).

La especificación del modelo incluye la concreción de la suma de cuadrados, es decir, el método para calcular las sumas de cuadrados. Para los modelos equilibrados y no equilibrados sin casillas vacías, el método más utilizado para la suma de cuadrados es el Tipo III, opción por defecto, de los cuatro tipos posibles. La suma Tipo I (descomposición jerárquica) se utiliza en modelos equilibrados y anidados, y donde cada término se corrige respecto a un anterior que le precede en el modelo evaluando primero los efectos de orden inferior. La suma Tipo II se utiliza en modelos equilibrados, en los que sólo hay efectos principales (y no interacciones), y en diseños anidados en el que cada efecto está anidado con el anterior; en estos casos se casos se tienen en cuenta sólo los efectos pertinentes (aquel que no está contenido en el efecto evaluado). Las sumas de cuadrados de Tipo IV son adecuados tanto para modelos equilibrados como no equilibrados con casillas vacías.
También la constante o interceptación se incluye normalmente en el modelo; si los datos pasan por el origen, se puede excluir.

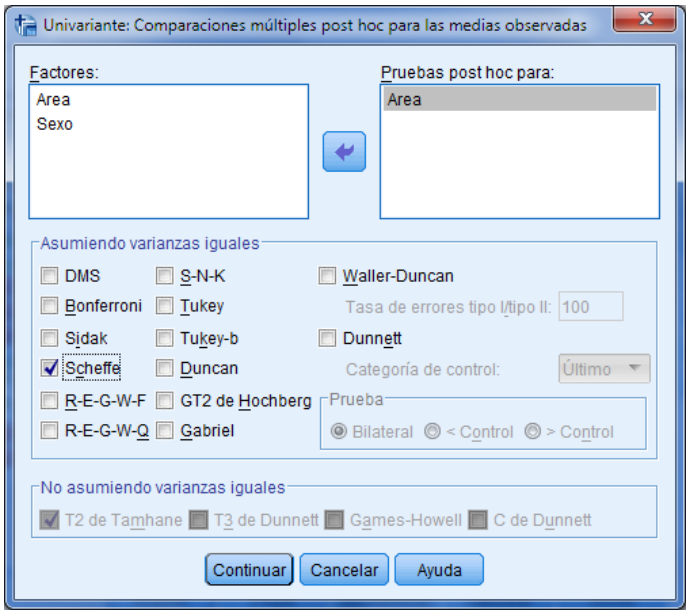
Consideraremos en este ejemplo todas las opciones por defecto: modelo completo con sumas de cuadrados Tipo III y con el término de intersección:




Cuando sabemos por el estadístico F del ANOVA que se dan diferencias significativas nos podemos plantear, a posteriori, pruebas para determinar donde se dan estas diferencias entre las medias mediante comparaciones múltiples llamadas post hoc. En el cuadro de diálogo correspondiente debemos elegir las variables independientes, los niveles de las que queremos comparar, y escoger un procedimiento de comparación.

Las opciones disponibles son las mismas que en el caso del mando **ONEWAY**. Todos ellos siguen la misma lógica interpretativa. En el caso que nos ocupa no tiene sentido comparar las categorías o niveles de la variable Sexo ya que sólo tiene dos valores y el estadístico F ya nos dice si se dan diferencias significativas. Pediremos pues las comparaciones para la variable Area. A la hora de elegir el procedimiento de comparación debemos tener en cuenta si se da una situación de igualdad o no de varianzas. Desde la versión 17 de SPSS las pruebas donde no se asumen varianzas iguales se consideran que no son válidas con dos o más factores, por eso aparece en la imagen siguiente con gris sombreado. Deberemos aplicar una estrategia alternativa. En este caso hemos escogido el método de Scheffé para las comparaciones por parejas de

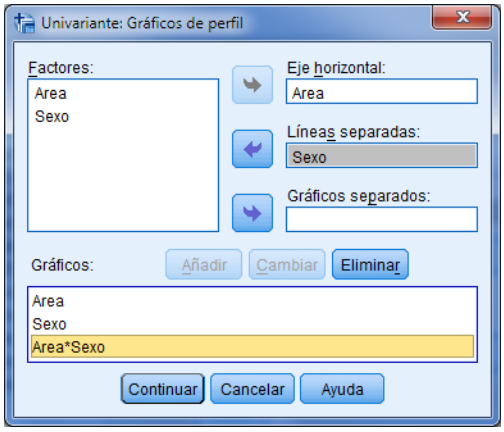
valores de cada variable individualmente. Veremos luego cómo realizar esta tarea de forma alternativa contemplando no solo los efectos de cada variable sino también la interacción.




Los resultados de las comparaciones múltiples nos permiten extraer conclusiones suficientes y correctas de los efectos principales. Pero no sucede lo mismo con las interacciones, por lo que es de gran utilidad inspeccionar gráficamente la interacción mediante un gráfico de líneas llamado también gráfico de perfil. Este gráfico, sobre la interacción entre dos factores, se representa sobre el eje de ordenadas la escala de la variable dependiente, y sobre el eje de abscisas los niveles del primer factor. Las líneas del gráfico representan los niveles del segundo factor. Si consideráramos tres variables interaccionando, se debería hacer un gráfico interacción doble para cada nivel de la tercera variable. Las interacciones a partir de este nivel no resultan fáciles de interpretar.

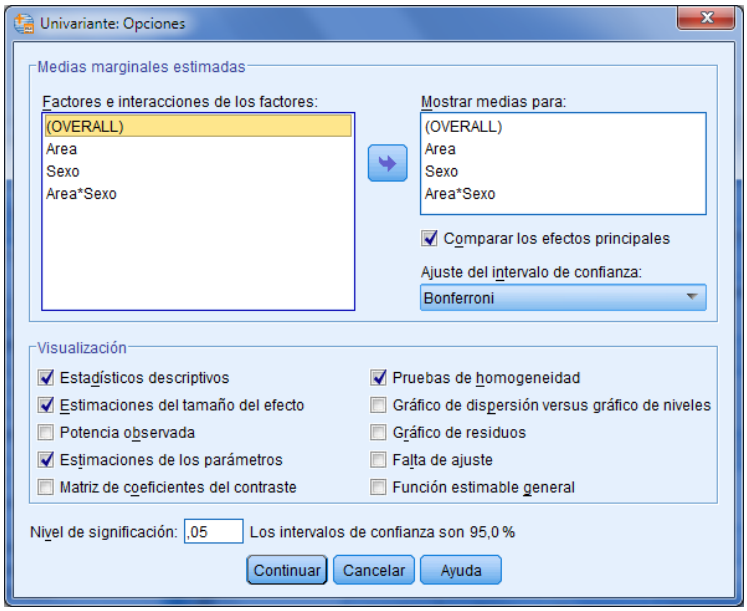
Para obtener un gráfico de perfil pulsamos sobre el botón Gráficos. El cuadro de diálogo permite obtenerlo con las medias estimadas bajo el modelo ANOVA[24], para cada variable y para combinaciones de dos y tres factores. Se trata de trasladar la primera variable al recuadro Eje horizontal solamente si el gráfico es para una sola variable y clicar sobre el botón Añadir. Si queremos un gráfico de perfil para observar la posible interacción añadiremos la segunda variable que interacciona al recuadro Líneas separadas. Si hubiera una tercera variable se colocaría en el recuadro Gráficos distintos. En este caso, pedimos tanto los gráficos individuales como con las dos variables independientes que aparece expresado como Area*Sexo (también se puede pedir el gráfico Sexo*Area intercambiando los factores):

[24] Si queremos el gráfico de medias con los datos observados y no con los estimados en el modelo ANOVA podemos generarlos a través el menú Gráficos. A diferencia del ejemplo anterior con una sola variable debemos incluir una segunda. Para ello debemos ir a la pestaña Grupos/ID fr puntos y marcamos la opción Variables de agrupación/Apilado. Aparecerá un recuadro en la paleta del gráfico donde colocaremos la segunda variable independiente.




Pero para interpretar y determinar correctamente la interacción, además de la ayuda de la representación gráfica, se pueden obtener resultados estadísticos que nos indiquen de forma concluyente qué medias difieren entre sí mediante comparaciones múltiples que no se pueden pedir a través del menú sino por sintaxis. Para ello nos ayudaremos de una especificación particular que se puede generar por el menú a través de Opciones, eligiendo Comparar los Efectos principales y modificando después la sintaxis que se genera si clicamos al final sobre Pegar en vez de Aceptar. Ahora lo detallamos.

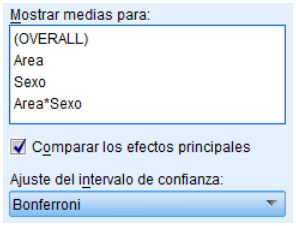
El cuadro de diálogo de Opciones que aparece a continuación incluye distintas alternativas que hemos elegido para nuestro ejercicio:




Facilita diferentes resultados de interés:

- Medias marginales estimadas: en este apartado podemos obtener, de cada factor o interacción, las medias estimadas (no las observadas) a partir de los parámetros del modelo. Activando además la opción de comparar los efectos principales podemos obtener todas las comparaciones dos a dos entre las medias correspondientes a los factores que tienen más de dos niveles. Las comparaciones implican realizar una

prueba de t-student de diferencias entre dos medias. Adicionalmente podríamos escoger una opción de ajuste del intervalo de confianza. Como comentábamos, para realizar esta comparación o contraste entre grupos a partir de la interacción de las variables deberemos incluir en el recuadro Mostrar medias para, la interacción Area*Sexo, además de los efectos principales Area y Sexo y el efecto global (OVERALL) si se consideran:




La sintaxis que se genera incluye estas líneas:

```
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(Area) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Sexo) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Area*Sexo)
```

Se trata de añadir a la última especificación lo siguiente:

```
/EMMEANS=TABLES(Area*Sexo) COMPARE (Sexo) ADJ (Bonferroni)
```

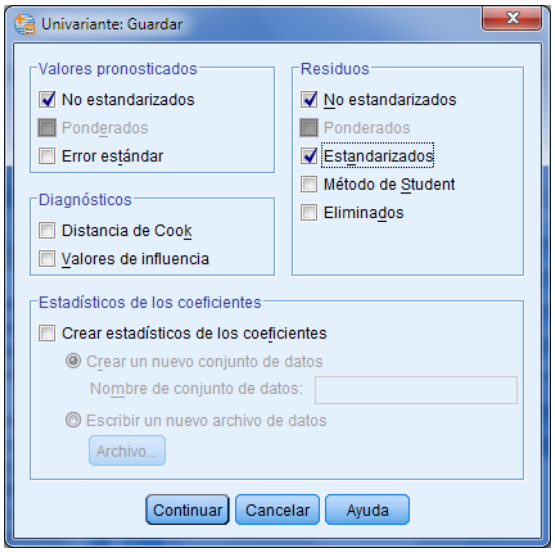
Con lo que dispondremos de una tabla de comparación entre cada Sexo dentro de cada nivel de Area.

– En Mostrar disponemos de diversas opciones:
  - Estadísticos descriptivos: media, desviación y número de casos de cada nivel y combinación de niveles.
  - Estimaciones del tamaño del efecto: estima el grado en que cada factor o combinaciones de factores afecta a la variable dependiente mediante la eta cuadrado parcial, interpretándose, por tanto, como la proporción de varianza explicada. Esta información se presenta en la tabla de resumen del ANOVA.
  - Potencia observada: información asociada al contraste de cada efecto que nos indica la capacidad del contraste para detectar una diferencia poblacional tan grande como la diferencia muestral observada, con un nivel de significación por defecto del 0,05.
  - Estimaciones de los parámetros: del modelo ANOVA especificado proporciona la tabla de estimaciones de los parámetros a partir de los cuales se obtienen las medias estimadas para cada nivel o combinación de niveles.
  - Matriz de coeficientes de contraste: matriz $L$ de coeficientes asociados a cada efecto.
  - Pruebas de homogeneidad: prueba de Levene de igualdad de varianzas para las combinaciones de factores.
  - Diagramas de dispersión por nivel: informan gráficamente sobre la igualdad de varianzas. Cuando las varianzas son iguales los puntos del gráfico se encuentran a la misma altura, alineados horizontalmente.
  - Gráfico de residuos. Los residuos se definen como la diferencia entre los valores observados y los valores pronosticados por el modelo ANOVA, y se suponen que son independientes entre sí y se distribuyen normalmente, en situación de homogeneidad de varianzas. Si los residuos son independientes el gráfico de

valores pronosticados y residuos tipificados no debe mostrar un patrón de comportamiento o de variación determinado (una línea recta, una curva, ...). Si las varianzas son homogéneas, la dispersión de los residuos tipificados debe ser similar a lo largo de todos los valores pronosticados. Si el modelo considerado se ajusta a los datos, la nube de puntos entre los valores observados y los pronosticados debe mostrar una clara relación lineal, más cuando mejor es el ajuste.

- Falta de ajuste: se puede pedir un contraste de falta de ajuste cuando se dispone de observaciones repetidas para una o más variables independientes. Si se rechaza el contraste necesario revisar el modelo.
- Función estimable general: se puede generar una tabla que muestra la forma general de las funciones estimables.
- Nivel de significación: habitualmente será del 0,05, por lo que los intervalos de confianza son del 95%.

Finalmente se pueden guardar como variables de la matriz los valores pronosticados y los residuos (tipificados y no) que se derivan del modelo, entre otras opciones, a través del cuadro de diálogo Guardar:



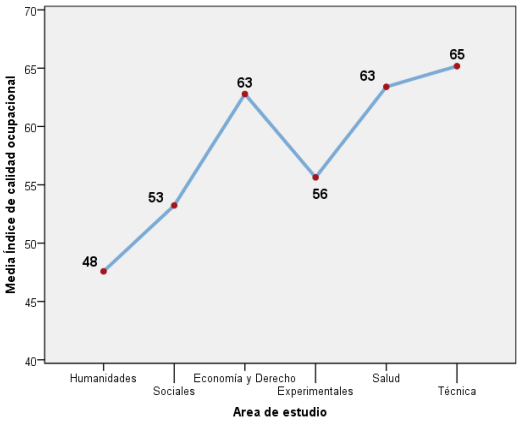

Veamos ahora los resultados generados con las distintas especificaciones que hemos elegido. En primer lugar, analizamos los resultados descriptivos leyendo la información de la tabla de medias y de los gráficos de perfil. La tabla original de estadísticos descriptivos se puede pivotar para facilitar la lectura de la información:

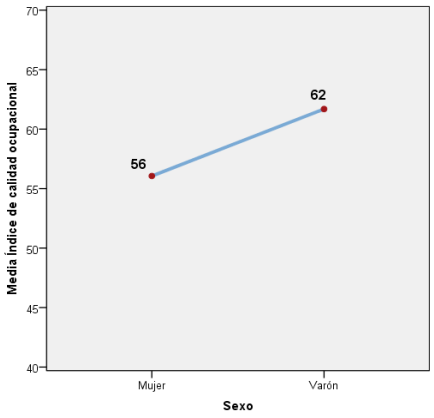
**Estadísticos descriptivos**

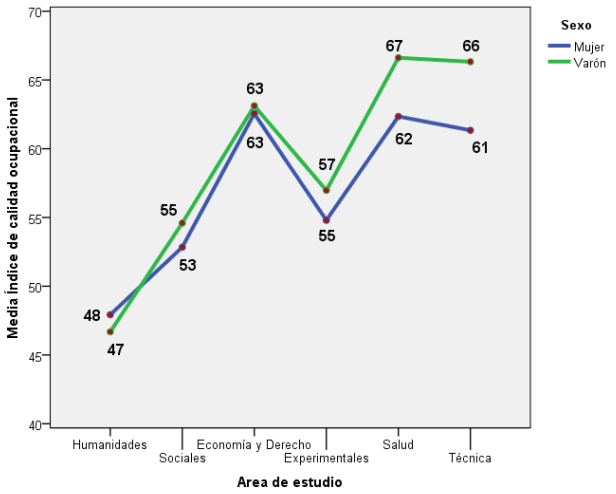
Variable dependiente: ICO Índice de calidad ocupacional

| Area Area de estudio | | Sexo Sexo | | |
|---|---|---|---|---|
| | | 1 Mujer | 2 Varón | Total |
| 1 Humanidades | Media | 47,9 | 46,7 | 47,6 |
| | Desviación estándar | 21,5 | 21,8 | 21,6 |
| | N | 649,0 | 255,0 | 904,0 |
| 2 Sociales | Media | 52,8 | 54,6 | 53,2 |
| | Desviación estándar | 20,4 | 21,3 | 20,6 |
| | N | 2154,0 | 628,0 | 2782,0 |
| 3 Economía y Derecho | Media | 62,6 | 63,1 | 62,8 |
| | Desviación estándar | 19,8 | 20,8 | 20,2 |
| | N | 886,0 | 633,0 | 1519,0 |
| 4 Experimentales | Media | 54,8 | 57,0 | 55,6 |
| | Desviación estándar | 22,1 | 22,3 | 22,2 |
| | N | 388,0 | 244,0 | 632,0 |
| 5 Salud | Media | 62,3 | 66,6 | 63,4 |
| | Desviación estándar | 17,3 | 17,4 | 17,4 |
| | N | 705,0 | 232,0 | 937,0 |
| 6 Técnica | Media | 61,3 | 66,3 | 65,2 |
| | Desviación estándar | 20,4 | 18,5 | 19,1 |
| | N | 477,0 | 1593,0 | 2070,0 |
| Total | Media | 56,1 | 61,7 | 58,3 |
| | Desviación estándar | 20,8 | 20,8 | 21,0 |
| | N | 5259,0 | 3585,0 | 8844,0 |

La tabla permite leer tres informaciones: los marginales de fila de la variable Area, los de columna de la variable Sexo, ambas en color rojo oscuro, y la interacción entre ambas en el interior de la tabla, en color azul. Esta misma información se dispone en forma de gráfico a continuación.



Las medias de la variable Area las vimos en el análisis unifactorial y nos muestran menores niveles para las carreras de Humanidades, Sociales y Experimentales en relación a las de Economía y Derecho, Salud y Técnica. La variable Sexo, por su parte, da lugar a un índice ocupacional algo mayor para los varones que para las mujeres, 62 frente a 56. Del cruce de ambas variables permite derivarse un patrón general similar, ambas líneas del gráfico marcan una disposición muy parecida, con dos matices visuales destacables: los varones siempre se sitúan por encima de las mujeres excepto en Humanidades, y el mayor índice de los mayores se agranda con las carreras de Salud y Técnica.

La tabla descriptiva anterior nos permite también ver que cada grupo tiene un número de efectivos diferentes (es un diseño no equilibrado) y un comportamiento bastante diverso de las desviaciones típicas, lo que nos sugiere la ausencia de una igualdad de varianzas.

Como el tamaño de la muestra es importante, mucho mayor que 30, en cada grupo definido por el cruce de cada área de conocimiento y sexo, no es preciso asumir el supuesto de normalidad. La heterogeneidad en el comportamiento de la dispersión que hemos comentado nos sugiere una situación de heteroscedasticidad que la prueba de Levene confirma al dar el estadístico *F* una significación inferior a 0,05:

**Prueba de igualdad de Levene de varianzas de error[a]**

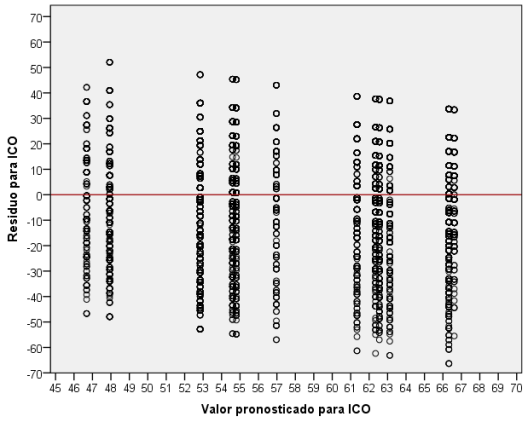
Variable dependiente: ICO Índice de calidad ocupacional

| F | df1 | df2 | Sig. |
|---|---|---|---|
| 20,236 | 5 | 8838 | ,000 |

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + Area

A partir de las variables guardadas del análisis: los residuos (variable RES_1, diferencia entre los valores observados y predichos) y los valores predichos o pronosticados (variable PRE_1)[25], podemos obtener un gráfico de dispersión de los residuos en relación a los valores pronosticados por el modelo (las medias de los grupos) ya que es un indicador de las buenas condiciones de aplicación en la medida en que se no observen distribuciones muy diferentes entre sí de los puntos en cada grupo. A través del menú Gráficos relacionando ambas variables obtenemos:




[25] Son los nombres que por defecto asigna el SPSS donde el número indica que se ha guardado una primera vez. Si se ejecutara una segunda vez, existiendo las anteriores, se denominarían de igual forma pero con un 2, y así sucesivamente cada vez que se ejecutara el procedimiento.

El gráfico muestra que en nuestro caso se dan esas buenas condiciones y la línea roja plana nos indica justamente este hecho[26].

Vistas las condiciones de aplicación y a la luz de los resultados descriptivos comentados se observan comportamientos diferenciados según el área de conocimiento, según el sexo, y también por efecto de la interacción (se entrecruzan las líneas y cambia la pendiente). La cuestión que se plantea es si estos comportamientos diferenciables observados son, desde un punto de vista estadístico, significativos, entre qué grupos de dan las diferencias y si la interacción es relevante. Para ello interpretaremos la tabla ANOVA de nuestro modelo, especificado como modelo factorial completo con los dos efectos principales y la interacción (factores intersujetos). En un análisis de varianza multifactorial existe una hipótesis nula para cada factor (las medias poblacionales definidas por los niveles del factor son iguales) y para cada posible combinación de factores (el efecto de la interacción es nulo, no se dan diferencias significativas entre las medias de cada combinación de niveles de los factores). Todos los contrastes de hipótesis se hacen en base al estadístico F que se presentan en la tabla ANOVA siguiente:

**Pruebas de efectos inter-sujetos**

Variable dependiente: ICO Índice de calidad ocupacional

| Origen | Tipo III de suma de cuadrados | gl | Cuadrático promedio | F | Sig. | Eta parcial al cuadrado |
|---|---|---|---|---|---|---|
| Modelo corregido | 347513,398[a] | 11 | 31592,127 | 78,607 | ,000 | ,089 |
| Interceptación | 19110871,25 | 1 | 19110871,25 | 47551,485 | ,000 | ,843 |
| Area | 234762,303 | 5 | 46952,461 | 116,827 | ,000 | ,062 |
| Sexo | 6161,381 | 1 | 6161,381 | 15,331 | ,000 | ,002 |
| Area * Sexo | 6842,236 | 5 | 1368,447 | 3,405 | ,004 | ,002 |
| Error | 3549567,661 | 8832 | 401,899 | | | |
| Total | 33999458,16 | 8844 | | | | |
| Total corregido | 3897081,059 | 8843 | | | | |

a. R al cuadrado = ,089 (R al cuadrado ajustada = ,088)

La columna Origen, también identificada como fuente de variación, nos proporciona la posible causa explicativa de la variabilidad del índice ocupacional. El modelo corregido hace referencia a todos los efectos del modelo considerados conjuntamente: el efecto de cada factor o efecto principal (Area y Sexo), el efecto de la interacción (Area * Sexo), sin el efecto de la constante o interceptación[27], una vez corregidos por la media. Como la probabilidad asociada al estadístico *F* es menor de 0,05, el modelo es significativo, el modelo explica de forma significativa la variación del salario.

La variación no explicada (residual o de error) se expresa en la media cuadrática o cuadrático promedio del error (es un estimador no sesgado de la varianza de las 12 poblaciones consideradas, la combinación de 6 áreas y 2 sexos). Esta cantidad es el denominador de cada estadístico *F* calculado en la tabla. La fila identificada con Total es la suma de cuadrados de la variable dependiente, mientras que Total corregido recoge la variación total corregida respecto de la media: la variación debida a cada efecto más la variación del error.

[26] La línea se obtiene editando el gráfico y eligiendo que se inserte la Línea de ajuste total, es decir, la recta de regresión.

[27] El efecto de la constante o interceptación prueba si la media global es cero y no es especialmente informativo, pero es un parámetro necesario para obtener las estimaciones de las medias de cada casilla.

**Pruebas de efectos inter-sujetos**

Variable dependiente: ICO Índice de calidad ocupacional

| Origen | Tipo III de suma de cuadrados | Eta cuadrado ×100 | gl | Cuadrático promedio | F | Sig. | Eta parcial al cuadrado |
|---|---|---|---|---|---|---|---|
| Modelo corregido | **347.513,4**[a] | | 11 | 31.592,1 | 78,6 | 0,000 | 0,089 |
| Interceptación | 19.110.871,2 | | 1 | 19.110.871,3 | 47.551,5 | 0,000 | 0,843 |
| Area | 234.762,3 | 6,2% | 5 | 46.952,5 | 116,8 | 0,000 | 0,062 |
| Sexo | 6.161,4 | 0,2% | 1 | 6.161,4 | 15,3 | 0,000 | 0,002 |
| Area * Sexo | 6.842,2 | 0,2% | 5 | 1.368,4 | 3,4 | 0,004 | 0,002 |
| Error | **3.549.567,7** | 93,5% | 8832 | 401,9 | | | |
| **SCTotal** | **3.797.333,6** | **100,0%** | | | | | |
| Total | 33.999.458,2 | | 8844 | | | | |
| Total corregido | **3.897.081,1** | | 8843 | | | | |

a. R al cuadrado = 0,089 (R al cuadrado ajustada = 0,088)

$$R^2 = \frac{SC\ Modelo\ corregido}{SC\ Total\ corregido} = \frac{347.513,4}{3.897.081,1}$$

$$\eta_E^2 = \frac{F_E \times gl_E}{F_E \times gl_E + gl_{error}} = \frac{116,8 \times 5}{116,8 \times 5 + 8832}$$

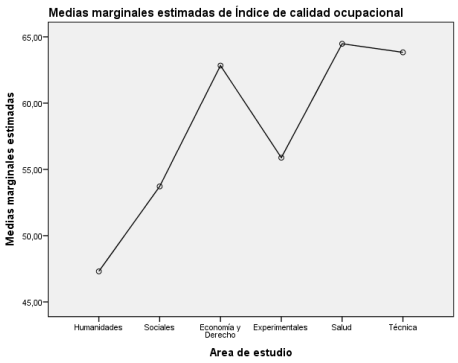
Si bien el modelo corregido es significativo, no obstante, la capacidad explicativa del modelo es reducida, el $R^2$ es del 8,9% (el resultado de dividir la suma de cuadrados del modelo corregido entre la suma de cuadrados total del modelo corregido). En consecuencia, el tamaño de los efectos parciales, es decir, la proporción de varianza explicada por cada efecto particular que nos da el eta cuadrado parcial, será bajo.

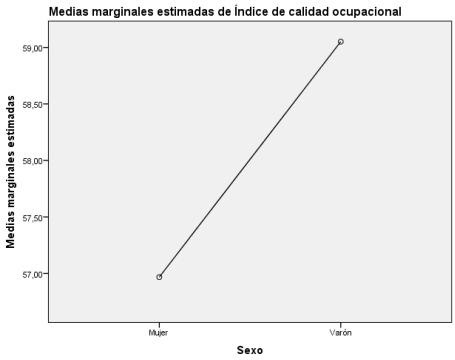
Así, podemos observar, en primer lugar, que todos los efectos son significativos, la probabilidad siempre es inferior a 0,05. En segundo término, que, de los tres efectos, tanto en el caso del efecto principal de la variable Sexo así como en el de la interacción Area*Sexo, la magnitud es muy débil, siendo el efecto principal de la variable Area el que determina las diferencias observadas en el índice de calidad ocupacional ICO.

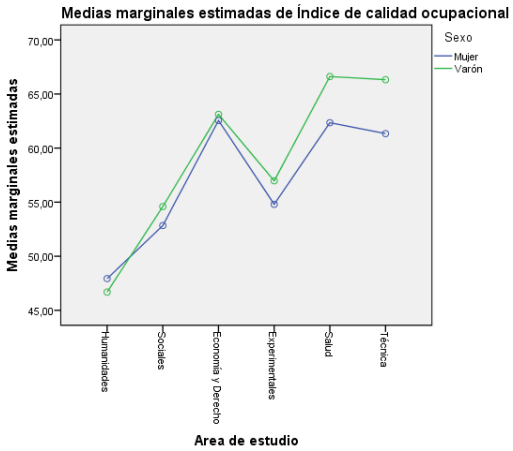
Sabemos pues que existen diferencias entre áreas de conocimiento, entre varones y mujes, y entre alguna combinación de área y sexo. La siguiente tarea es determinar entre qué categorías se dan estas diferencias. Para ello analizaremos las comparaciones múltiples que aparecen en el apartado de resultados que se identifica como Medias marginales estimadas, y que se refieren sucesivamente, a la variable Area, Sexo y a la interacción Area * Sexo. En cada caso aparece primero la tabla de medias estimadas bajo el modelo ANOVA, con su intervalo de confianza. Resumimos la información de las medias estimadas en esta tabla:

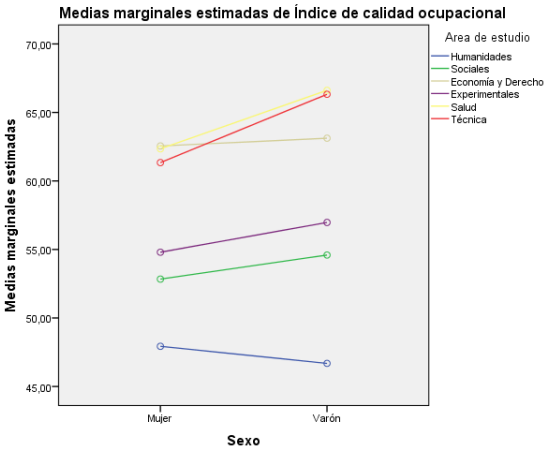
| | Sexo Sexo | | |
|---|---|---|---|
| Area Area de estudio | 1 Mujer | 2 Varón | Total |
| 1 Humanidades | 47,931 | 46,685 | **47,308** |
| 2 Sociales | 52,838 | 54,594 | **53,716** |
| 3 Economía y Derecho | 62,553 | 63,119 | **62,836** |
| 4 Experimentales | 54,799 | 56,971 | **55,885** |
| 5 Salud | 62,347 | 66,615 | **64,481** |
| 6 Técnica | 61,340 | 66,323 | **63,832** |
| Total | **56,968** | **56,100** | **58,010** |

Estas medias los valores que se representan en los gráficos de perfiles:




A continuación, se realiza en cada caso la comparación de medias por parejas estableciendo la significación a partir de la prueba de la diferencia de medias. Comentaremos las tres tablas que se generan[28].

En el primer caso, para la variable Area, se obtienen los mismos resultados que vimos en el análisis unifactorial con las pruebas de Tamhane, con la única excepción de que las personas que se licenciaron con carreras del área Técnica y de Economía y Derecho aquí no presentan diferencias significativas. Para resumir el conjunto de comparaciones se adjunta una segunda tabla de comparaciones con los resultados más sintéticos. En ella se observa, por tanto, que Humanidades difiere de todas las demás, Sociales y Experimentales no difieren entre sí y sí con las demás áreas, mientras que Economía y Derecho junto a Técnica y Salud tampoco muestran diferencias entre ellas y sí con las demás áreas.

[28] Las dos tablas comentadas se complementan con una tercera denominada Pruebas univariadas donde simplemente se reproduce la información de la tabla ANOVA que vimos anteriormente seleccionando la parte que es específica de cada caso.

**Comparaciones por parejas**

Variable dependiente: ICO Índice de calidad ocupacional

| (I) Area de estudio | (J) Area de estudio | Diferencia de medias (I-J) | Error estándar | Sig.[b] | 95% de intervalo de confianza para diferencia[b] | |
|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior |
| 1 Humanidades | 2 Sociales | -6,408* | ,869 | ,000 | -8,960 | -3,856 |
| | 3 Economía y Derecho | -15,528* | ,906 | ,000 | -18,189 | -12,868 |
| | 4 Experimentales | -8,577* | 1,104 | ,000 | -11,819 | -5,335 |
| | 5 Salud | -17,173* | 1,060 | ,000 | -20,286 | -14,060 |
| | 6 Técnica | -16,524* | ,907 | ,000 | -19,186 | -13,861 |
| 2 Sociales | 1 Humanidades | 6,408* | ,869 | ,000 | 3,856 | 8,960 |
| | 3 Economía y Derecho | -9,121* | ,692 | ,000 | -11,152 | -7,089 |
| | 4 Experimentales | -2,169 | ,937 | ,309 | -4,919 | ,581 |
| | 5 Salud | -10,765* | ,884 | ,000 | -13,362 | -8,168 |
| | 6 Técnica | -10,116* | ,693 | ,000 | -12,151 | -8,081 |
| 3 Economía y Derecho | 1 Humanidades | 15,528* | ,906 | ,000 | 12,868 | 18,189 |
| | 2 Sociales | 9,121* | ,692 | ,000 | 7,089 | 11,152 |
| | 4 Experimentales | 6,951* | ,971 | ,000 | 4,100 | 9,802 |
| | 5 Salud | -1,644 | ,921 | 1,000 | -4,348 | 1,059 |
| | 6 Técnica | -,995 | ,739 | 1,000 | -3,164 | 1,174 |
| 4 Experimentales | 1 Humanidades | 8,577* | 1,104 | ,000 | 5,335 | 11,819 |
| | 2 Sociales | 2,169 | ,937 | ,309 | -,581 | 4,919 |
| | 3 Economía y Derecho | -6,951* | ,971 | ,000 | -9,802 | -4,100 |
| | 5 Salud | -8,596* | 1,116 | ,000 | -11,874 | -5,318 |
| | 6 Técnica | -7,947* | ,972 | ,000 | -10,800 | -5,093 |
| 5 Salud | 1 Humanidades | 17,173* | 1,060 | ,000 | 14,060 | 20,286 |
| | 2 Sociales | 10,765* | ,884 | ,000 | 8,168 | 13,362 |
| | 3 Economía y Derecho | 1,644 | ,921 | 1,000 | -1,059 | 4,348 |
| | 4 Experimentales | 8,596* | 1,116 | ,000 | 5,318 | 11,874 |
| | 6 Técnica | ,649 | ,922 | 1,000 | -2,057 | 3,355 |
| 6 Técnica | 1 Humanidades | 16,524* | ,907 | ,000 | 13,861 | 19,186 |
| | 2 Sociales | 10,116* | ,693 | ,000 | 8,081 | 12,151 |
| | 3 Economía y Derecho | ,995 | ,739 | 1,000 | -1,174 | 3,164 |
| | 4 Experimentales | 7,947* | ,972 | ,000 | 5,093 | 10,800 |
| | 5 Salud | -,649 | ,922 | 1,000 | -3,355 | 2,057 |

Se basa en medias marginales estimadas

*. La diferencia de medias es significativa en el nivel ,05.

b. Ajuste para varias comparaciones: Bonferroni.

**Comparaciones por parejas**

Variable dependiente: ICO Índice de calidad ocupacional

Diferencia de medias (I-J)

| | (J) Area Area de estudio | | | | | |
|---|---|---|---|---|---|---|
| (I) Area de estudio | 1 Humanidades | 2 Sociales | 3 Economía y Derecho | 4 Experimentales | 5 Salud | 6 Técnica |
| 1 Humanidades | | -6,408* | -15,528* | -8,577* | -17,173* | -16,524* |
| 2 Sociales | 6,408* | | -9,121* | -2,169 | -10,765* | -10,116* |
| 3 Economía y Derecho | 15,528* | 9,121* | | 6,951* | -1,644 | -0,995 |
| 4 Experimentales | 8,577* | 2,169 | -6,951* | | -8,596* | -7,947* |
| 5 Salud | 17,173* | 10,765* | 1,644 | 8,596* | | 0,649 |
| 6 Técnica | 16,524* | 10,116* | 0,995 | 7,947* | -0,649 | |

*. La diferencia de medias es significativa en el nivel 0,05.

A esta misma conclusión se llega si se analizan las pruebas de comparación múltiples post hoc de Scheffé que aparecen más tarde en los resultados, cuando establece grupos homogéneos a partir de las comparaciones. En las columnas aparecen los subconjuntos de niveles del factor que son significativamente diferentes, si dos o más grupos o niveles de la variable factor no presentan diferencias significativas estas ubicarán en la misma columna. En este caso, como se puede observar en la tabla siguiente, se

obtienen tres subconjuntos de áreas de conocimiento: Humanidades (subconjunto 1), Sociales y Experimentales (subconjunto 2), y Economía y Derecho, Técnica y Salud (subconjunto 3).

**Subconjuntos homogéneos**

**ICO Índice de calidad ocupacional**

Scheffe[a,b,c]

| Area de estudio | N | Subconjunto 1 | Subconjunto 2 | Subconjunto 3 |
|---|---|---|---|---|
| 1 Humanidades | 904 | 47,5797 | | |
| 2 Sociales | 2782 | | 53,2341 | |
| 4 Experimentales | 632 | | 55,6376 | |
| 3 Economía y Derecho | 1519 | | | 62,7892 |
| 5 Salud | 937 | | | 63,4037 |
| 6 Técnica | 2070 | | | 65,1749 |
| Sig. | | 1,000 | ,145 | ,152 |

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = 401,899.

a. Utiliza el tamaño de la muestra de la media armónica = 1141,426.

b. Los tamaños de grupo no son iguales. Se utiliza la media armónica de los tamaños de grupo. Los niveles de error de tipo I no están garantizados.

c. Alfa = ,05.

El segundo efecto principal, el de la variable Sexo, no requiere de hecho de ninguna comparación como indicamos, pues solo tiene dos valores y ya sabemos que es un efecto significativo, donde se concluye que la calidad ocupacional es mayor entre los varones, aunque con diferencias no muy destacables. El resultado que se obtiene es el siguiente:

**Comparaciones por parejas**

Variable dependiente: ICO Índice de calidad ocupacional

| (I) Sexo | (J) Sexo | Diferencia de medias (I-J) | Error estándar | Sig.[b] | 95% de intervalo de confianza para diferencia[b] | |
|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior |
| 1 Mujer | 2 Varón | -2,083* | ,532 | ,000 | -3,126 | -1,040 |
| 2 Varón | 1 Mujer | 2,083* | ,532 | ,000 | 1,040 | 3,126 |

Se basa en medias marginales estimadas

*. La diferencia de medias es significativa en el nivel ,05.

b. Ajuste para varias comparaciones: Bonferroni.

Finalmente nos quedan las comparaciones cruzadas de ambas variables:

**Comparaciones por parejas**

Variable dependiente: ICO Índice de calidad ocupacional

| Area de estudio | (I) Sexo | (J) Sexo | Diferencia de medias (I-J) | Error estándar | Sig.[b] | 95% de intervalo de confianza para diferencia[b] | |
|---|---|---|---|---|---|---|---|
| | | | | | | Límite inferior | Límite superior |
| 1 Humanidades | 1 Mujer | 2 Varón | 1,246 | 1,482 | ,400 | -1,658 | 4,151 |
| | 2 Varón | 1 Mujer | -1,246 | 1,482 | ,400 | -4,151 | 1,658 |
| 2 Sociales | 1 Mujer | 2 Varón | -1,757 | ,909 | ,053 | -3,539 | ,025 |
| | 2 Varón | 1 Mujer | 1,757 | ,909 | ,053 | -,025 | 3,539 |
| 3 Economía y Derecho | 1 Mujer | 2 Varón | -,566 | 1,043 | ,587 | -2,611 | 1,479 |
| | 2 Varón | 1 Mujer | ,566 | 1,043 | ,587 | -1,479 | 2,611 |
| 4 Experimentales | 1 Mujer | 2 Varón | -2,172 | 1,638 | ,185 | -5,383 | 1,039 |
| | 2 Varón | 1 Mujer | 2,172 | 1,638 | ,185 | -1,039 | 5,383 |
| 5 Salud | 1 Mujer | 2 Varón | -4,268* | 1,517 | ,005 | -7,242 | -1,293 |
| | 2 Varón | 1 Mujer | 4,268* | 1,517 | ,005 | 1,293 | 7,242 |
| 6 Técnica | 1 Mujer | 2 Varón | -4,983* | 1,046 | ,000 | -7,034 | -2,932 |
| | 2 Varón | 1 Mujer | 4,983* | 1,046 | ,000 | 2,932 | 7,034 |

Se basa en medias marginales estimadas

*. La diferencia de medias es significativa en el nivel ,05.

b. Ajuste para varias comparaciones: Bonferroni.

Si comparamos las diferencias entre varones y mujeres constatamos que, en Humanidades, Sociales, Experimentales y Economía y Derecho, no se dan diferencias estadísticamente significativas y que la fuente de la interacción viene dada por las diferencias que se encuentran en las áreas Técnica y de Salud.

Por último, se adjunta la tabla de las estimaciones de los parámetros del modelo, con los cuales se obtienen las medias estimadas a partir de cada efecto presente en el modelo. Así, por ejemplo, teniendo en cuenta tabla de las medias estimadas que presentamos anteriormente, en el caso de la media estimada de las personas tituladas en Humanidades que son mujeres: 47,931, se obtiene por la suma de los valores de la intercepción o constante (66,323), del valor del efecto principal de Area (cuando [Area=1], del valor −19,638), del efecto principal de Sexo (cuando [Sexo=1], el valor −4,983), y del efecto interacción de administrativos que no son minoría ([Area=1] * [Sexo=1], el valor 6,229).

**Estimaciones de parámetro**

Variable dependiente: ICO Índice de calidad ocupacional

| Parámetro | B | Error estándar | t | Sig. | Intervalo de confianza al 95% | | Eta parcial al cuadrado |
|---|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior | |
| Interceptación | 66,323 | ,502 | 132,043 | ,000 | 65,339 | 67,308 | ,664 |
| [Area=1] | -19,638 | 1,352 | -14,524 | ,000 | -22,289 | -16,988 | ,023 |
| [Area=2] | -11,729 | ,945 | -12,417 | ,000 | -13,581 | -9,877 | ,017 |
| [Area=3] | -3,204 | ,942 | -3,401 | ,001 | -5,050 | -1,357 | ,001 |
| [Area=4] | -9,352 | 1,378 | -6,786 | ,000 | -12,054 | -6,651 | ,005 |
| [Area=5] | ,292 | 1,409 | ,207 | ,836 | -2,470 | 3,053 | ,000 |
| [Area=6] | 0[a] | . | . | . | . | . | . |
| [Sexo=1] | -4,983 | 1,046 | -4,762 | ,000 | -7,034 | -2,932 | ,003 |
| [Sexo=2] | 0[a] | . | . | . | . | . | . |
| [Area=1] * [Sexo=1] | 6,229 | 1,814 | 3,434 | ,001 | 2,674 | 9,785 | ,001 |
| [Area=1] * [Sexo=2] | 0[a] | . | . | . | . | . | . |
| [Area=2] * [Sexo=1] | 3,226 | 1,386 | 2,328 | ,020 | ,509 | 5,943 | ,001 |
| [Area=2] * [Sexo=2] | 0[a] | . | . | . | . | . | . |
| [Area=3] * [Sexo=1] | 4,417 | 1,478 | 2,989 | ,003 | 1,520 | 7,313 | ,001 |
| [Area=3] * [Sexo=2] | 0[a] | . | . | . | . | . | . |
| [Area=4] * [Sexo=1] | 2,811 | 1,944 | 1,446 | ,148 | -,999 | 6,621 | ,000 |
| [Area=4] * [Sexo=2] | 0[a] | . | . | . | . | . | . |
| [Area=5] * [Sexo=1] | ,715 | 1,843 | ,388 | ,698 | -2,898 | 4,328 | ,000 |
| [Area=5] * [Sexo=2] | 0[a] | . | . | . | . | . | . |
| [Area=6] * [Sexo=1] | 0[a] | . | . | . | . | . | . |
| [Area=6] * [Sexo=2] | 0[a] | . | . | . | . | . | . |

a. Este parámetro está establecido en cero porque es redundante.

Como la suma de los parámetros de cada efecto suman cero, la tabla no presenta las estimaciones de los parámetros redundantes. De cada parámetro se ofrece el contraste de si es significativamente diferente de cero.

Como ejercicio adicional se puede realizar un análisis de la misma variable dependiente del índice ocupacional considerando otras variables independientes, o bien tomar como variable a explicar los ingresos o las notas del expediente académico considerando también diversas variables independientes posibles.

# 8. El análisis de varianza con R

El análisis de varianza con R se puede realizar a través de diversos procedimientos: `aov`, `lm` y `glm` del paquete stats, `Anova` del paquete car, `lme` del paquete nlme, `lmer` del paquete lme4, o `ezANOVA` del paquete ez.

R utiliza el método de cálculo de la suma de cuadrados tipo I por defecto, mientras que SPSS utiliza el tipo III por lo que tendremos que especificar instrucciones particulares para obtener el habitual tipo III, donde se compara cada término con el modelo completo.

## 8.1. Análisis unifactorial

Con los datos de la *Agència per a la Qualitat del Sistema Universitari de Catalunya* (AQU) de de la matriz de datos AQU.rda, en primer lugar, realizaremos un análisis de varianza unifactorial para explicar la calidad ocupacional en función del área de estudios. Como hipótesis planteamos que la calidad de la ocupación varía según el área de conocimiento, esperando obtener una media de calidad mayor en las carreras técnicas y de la salud que en las carreras de humanidades y sociales.

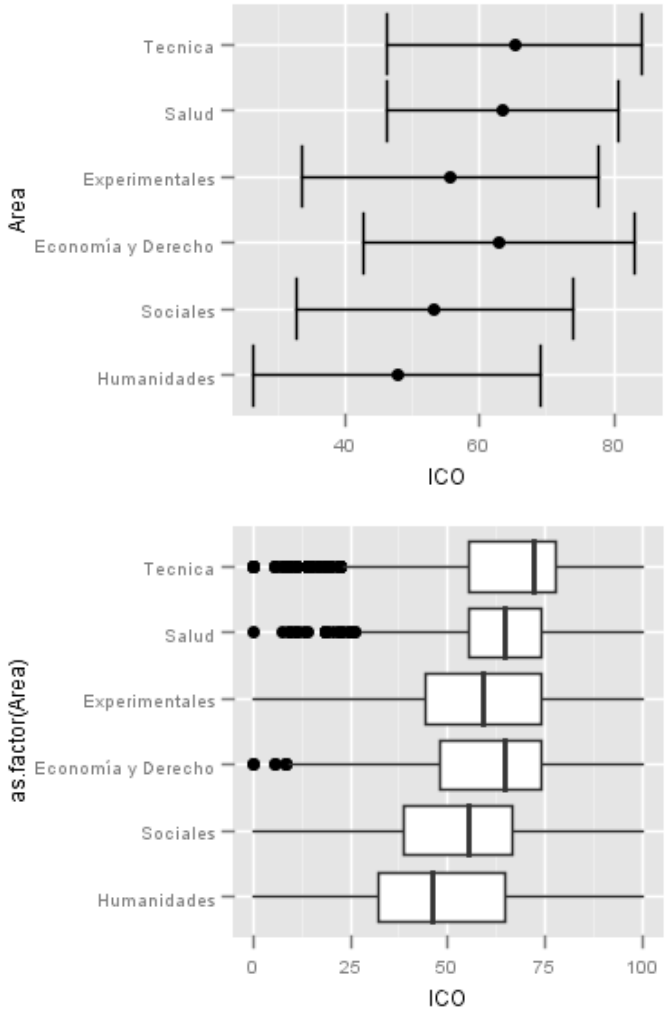
Nuestra variable dependiente cuantitativa es un índice de calidad ocupacional (ICO) que varía entre 0 y 100. La variable independiente del área de conocimiento (Area) tiene 6 grupos: Humanidades, Sociales, Economía y Derecho, Experimentales, Salud y Técnica. Iniciamos el análisis con el análisis descriptivo de la comparación de las medias del índice de calidad ocupacional según el área de conocimiento (menú de Deducer Descriptives):

**Descriptive Statistics**

**Variable: ICO**

| | Area | Mean | St. Deviation | Median | Minimum | Maximum | Valid N |
|---|---|---|---|---|---|---|---|
| 1 | Economía y Derecho | 62.79 | 20.23 | 64.81 | 0.00 | 100.00 | 1519 |
| 2 | Experimentales | 55.64 | 22.16 | 59.26 | 0.00 | 100.00 | 632 |
| 3 | Humanidades | 47.58 | 21.57 | 46.30 | 0.00 | 100.00 | 904 |
| 4 | Salud | 63.40 | 17.39 | 64.81 | 0.00 | 100.00 | 937 |
| 5 | Sociales | 53.23 | 20.58 | 55.56 | 0.00 | 100.00 | 2782 |
| 6 | Tecnica | 65.17 | 19.07 | 72.22 | 0.00 | 100.00 | 2070 |

Podemos pedir un gráfico de medias a través de Plot Builder (opción mean) y un diagrama de caja (opción group boxplot):




Podemos observar cómo Humanidades, Sociales y Experimentales alcanzan los niveles inferiores de calidad ocupacional mientras que Economía y Derecho junto a Salud y Técnica poseen valores medios superiores en el índice.

Para determinar si existen diferencias significativas y entre qué áreas procedemos a realizar un análisis de varianza. Disponemos de una amplia muestra dentro de cada grupo que nos exime de la exigencia del cumplimiento de la condición de normalidad.

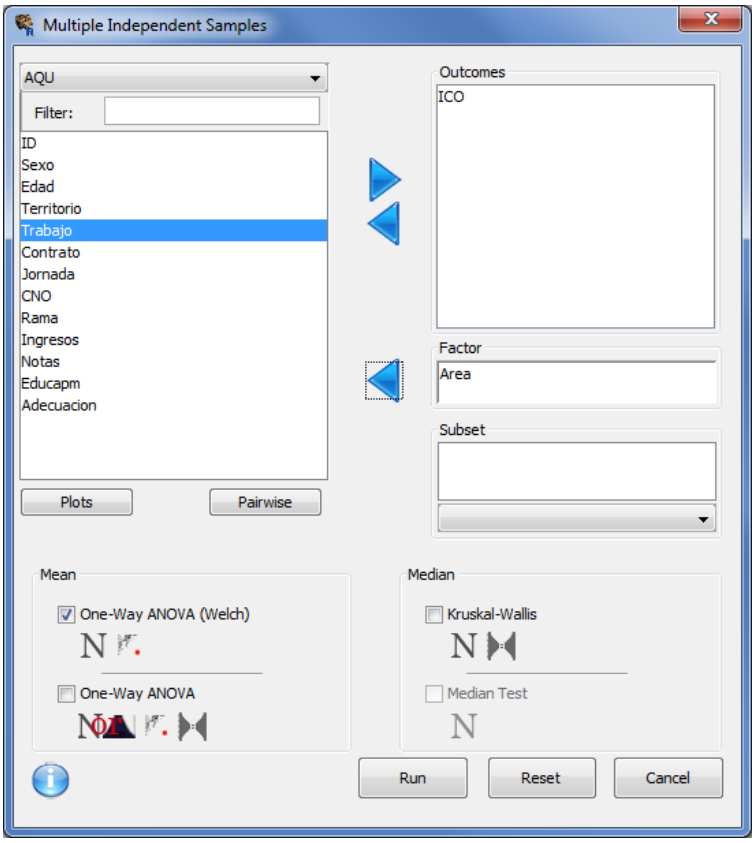
Si realizamos la prueba de homogeneidad de varianzas a través del test de Levene[29]:

```
> leveneTest(AQU$ICO,AQU$Area)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    5  18.572 < 2.2e-16 ***
      8838
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

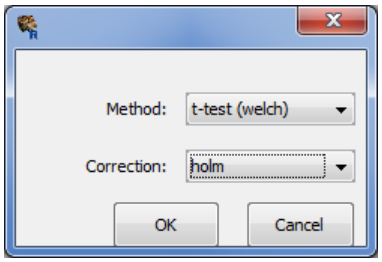
[29] Alternativamente se puede realizar a través el menú de Deducer: Extras / k-sample variance test.

Constatamos que la prueba es significativa (probabilidad muy inferior a 0,05), por lo que se da la hipótesis alternativa de heteroscedasticidad. Aplicamos un análisis de varianza a partir de esta conclusión.

Podemos realizar el ANOVA unifactorial a través de Deducer, por el menú Analysis / K-Sample Test:




La variable dependiente se sitúa en el recuadro de Outcomes y la independiente como Factor. Como estamos ante una situación de heteroscedasticidad un análisis unifactorial (oneway) con la prueba de Welch. Además solicitaremos las comparaciones de medias por parejas (pairwise) para determinar entre qué áreas se producen diferencias significativas. A través del botón Pairwise elegiremos el método t-test (Welch) y la corrección de Holm:




Una vez ejecutado se obtienen los resultados que se presentan a continuación.

En primer lugar se presenta el estadístico *F* (163,19) y su significación (valor inferior a 0,05) que nos permite concluir que las medias son significativamente distintas.

**K-Sample Test**

**Method: One-way analysis of means (not assuming equal variances)**

**Factor: Area ( Humanidades vs. Sociales vs. Economía y Derecho vs. Experimentales vs. Salud vs. Tecnica )**

| | *F* | *(num df,denom df)* | *p-value* |
|---|---|---|---|
| ICO | 163.19 | (5,2897.209) | <0.001 |

A continuación, vemos las comparaciones por parejas entre los grupos, entre las distintas áreas de conocimiento. En todos los casos excepto cuando se comparan las áreas de Economía y Derecho con Salud y Técnica con Salud, las diferencias son significativas. Podemos así validar nuestra hipótesis inicial y cuantificar la magnitud de las diferencias.

**Pairwise comparisons using t tests with pooled SD**

**data: AQUR$ICO and AQUR$Area**

| | *Humanidades* | *Sociales* | *Economía y Derecho* | *Experimentales* | *Salud* |
|---|---|---|---|---|---|
| Sociales | < 0.001 | - | - | - | - |
| Economía y Derecho | < 0.001 | < 0.001 | - | - | - |
| Experimentales | < 0.001 | 0.01986 | < 0.001 | - | - |
| Salud | < 0.001 | < 0.001 | 0.46141 | < 0.001 | - |
| Tecnica | < 0.001 | < 0.001 | 0.00176 | < 0.001 | 0.05024 |

*p*-value adjustment method: holm

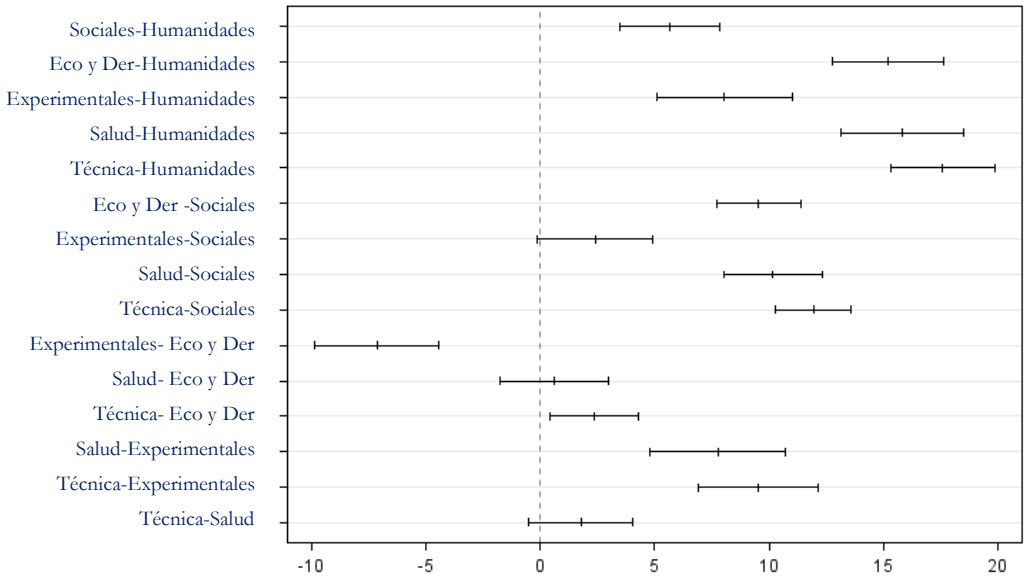
Un resultado como este ¿qué capacidad explicativa alcanza?, ¿en qué medida la variable independiente del área de conocimiento explica el índice ocupacional? Para determinar la intensidad de la relación calcularemos el estadístico eta cuadrado, $\eta^2$, equivalente al coeficiente de determinación $R^2$, que nos mide qué parte de la variación de ICO es atribuible a la variable Area.

Para obtener el estadístico eta cuadrado podemos instalar una librería específica, lsr, que contiene un conjunto de herramientas para el análisis estadístico[30]. Sobre un objeto de R que contenga un análisis de varianza se aplica la función `etaSquared` y se obtiene este valor que nos medirá el tamaño del efecto de la variable independiente. El objeto lo generaremos a través de otra función, `aov`, destinada al análisis de varianza y que se incluye en el paquete stats[31]. Se incluyen también los resultados las comparaciones múltiples con el método HSD de Tukey y las estimaciones de los parámetros donde obtenemos también una medida global de la capacidad explicativa del modelo con el coeficiente de determinación ($R^2$).

[30] Navarro, D. J. (2015). Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners, Version 0.5. School of Psychology, University of Adelaide, Adelaide, Australia. http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/.

[31] Alternativamente se puede ejecutar un modelo lineal en Deducer (menú Analysis / Linear Model) y obtener el estadístico equivalente $R^2$.

```
> library(lsr)
> install.packages("lsr")

> # Análisis de varianza
> AV=aov(ICO~Area, data=AQU)
> summary(AV)
             Df  Sum Sq Mean Sq F value Pr(>F)
Area          5  332609   66522   164.9 <2e-16 ***
Residuals  8838 3564472     403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
3022 observations deleted due to missingness
> # Capacidad explicativa
> etaSquared(AV)
        eta.sq eta.sq.part
Area 0.08534832  0.08534832
> # Comparaciones múltiples
> TukeyHSD(AV)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ICO ~ Area, data = AQU)

$Area
                                          diff        lwr       upr     p adj
Sociales-Humanidades                 5.6543972  3.4629504  7.845844 0.0000000
Economía y Derecho-Humanidades      15.2095506 12.8050236 17.614078 0.0000000
Experimentales-Humanidades           8.0579126  5.0898784 11.025947 0.0000000
Salud-Humanidades                   15.8240049 13.1553713 18.492638 0.0000000
Tecnica-Humanidades                 17.5952101 15.3132022 19.877218 0.0000000
Economía y Derecho-Sociales          9.5551534  7.7289744 11.381332 0.0000000
Experimentales-Sociales              2.4035154 -0.1188657  4.925897 0.0721752
Salud-Sociales                      10.1696077  8.0074827 12.331733 0.0000000
Tecnica-Sociales                    11.9408129 10.2792656 13.602360 0.0000000
Experimentales-Economía y Derecho -7.1516379 -9.8611983 -4.442078 0.0000000
Salud-Economía y Derecho             0.6144543 -1.7633798  2.992288 0.9774105
Tecnica-Economía y Derecho           2.3856596  0.4517386  4.319581 0.0058733
Salud-Experimentales                 7.7660922  4.8196414 10.712543 0.0000000
Tecnica-Experimentales               9.5372975  6.9358503 12.138745 0.0000000
Tecnica-Salud                        1.7712053 -0.4826592  4.025070 0.2195465

> plot(TukeyHSD(AV),las=1)
```



```
> # Capacidad explicativa y coeficientes
> etaSquared(AV)
        eta.sq eta.sq.part
Area 0.08534832  0.08534832
> summary.lm(AV)
Call:
aov(formula = Y ~ A)

Residuals:
    Min      1Q  Median      3Q     Max
-65.175 -12.478   2.321  12.605  52.420

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          47.5797     0.6679  71.234  < 2e-16 ***
ASociales             5.6544     0.7688   7.354 2.09e-13 ***
AEconomía y Derecho  15.2096     0.8436  18.029  < 2e-16 ***
AExperimentales       8.0579     1.0413   7.738 1.12e-14 ***
ASalud               15.8240     0.9363  16.901  < 2e-16 ***
ATecnica             17.5952     0.8006  21.977  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.08 on 8838 degrees of freedom
  (3022 observations deleted due to missingness)
Multiple R-squared: 0.08535,  Adjusted R-squared: 0.08483
F-statistic: 164.9 on 5 and 8838 DF,  p-value: < 2.2e-16
```

La medida de la intensidad de la relación arroja un valor de 0,085 que se interpreta como el porcentaje de varianza explicada en el modelo, es decir, un valor reducido del 8,5%. Ello quiere decir que las diferencias de las medias son relevantes entre la mayor parte de áreas de conocimiento pero que no se observan diferencias muy acentuadas, dentro de cada área se observan variabilidades internas importantes. Observamos no obstante que se dan diferencias entre la mayor parte de áreas como vimos anteriormente. El ajuste global del modelo arroja un $R^2$ de 0,085 que coincide con el valor del eta cuadrado.

Podemos reproducir el mismo análisis pero analizando las diferencias de medias de ICO según la variable Sexo. Evidenciaremos que la calidad de la ocupación las graduadas catalanas en 2010, cuatro años después, en 2014, alcanzan una calidad ocupacional inferior a la de los graduados varones: 56,06 frente a 61,69, diferencias que son estadísticamente significativas, siendo el valor del eta cuadrado de tan solo el 1,7%, diferencias de calidad por tanto significativas pero no muy relevantes. Constatamos asimismo que todas las estimaciones de los coeficientes son significativamente diferentes de la media global.

El lector/a puede preguntarse, lo proponemos de ejercicio, si el origen educativo de los padres (Educapm) tiene alguna influencia sobre el nivel de calidad ocupacional alcanzado por los hijos e hijas graduados. Otras hipótesis y análisis adicionales se pueden considerar con los datos de la matriz de datos de AQU.

## 8.2. Análisis multifactorial

El análisis de varianza multifactorial lo realizaremos a través de la función aov y seguiremos el ejemplo de la relación entre el índice de calidad ocupacional (ICO) y las variables Area y Sexo[32]. Reproducimos a continuación los resultados que se obtienen de tablas y gráficos.

Inicialmente asignamos los nombres Y, A y B a nuestras variables:

```
> # Asignación de nombres a las variables
> Y=AQU$ICO
> A=AQU$Area
> B=AQU$Sexo
```

Para obtener inicialmente una tabla descriptiva de las medias podemos optar por ejecutar a través de Deducer el procedimiento Analysis/Descriptives, colocando las dos factores en el recuadro de Stratify By. Se generará una tabla con los estadísticos elegidos para el cruce de categorías de ambas variables:

**Descriptive Statistics**

**Variable: ICO**

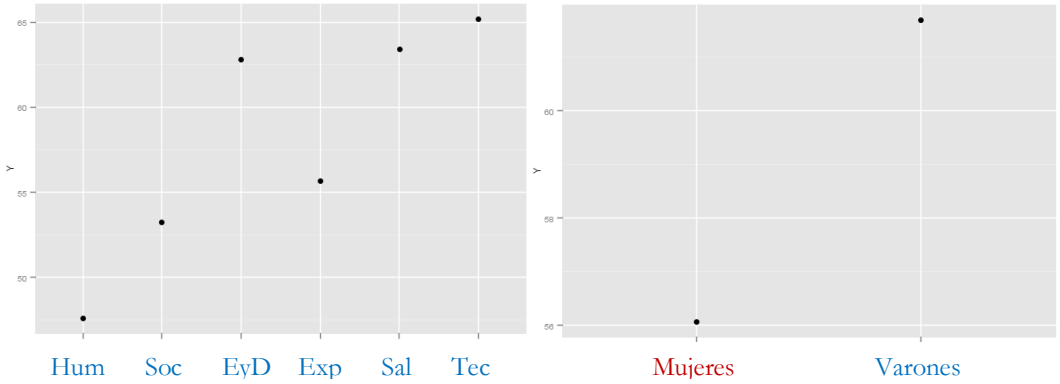
| | Area | Sexo | Mean | St. Deviation | Median | Minimum | Maximum | Valid N |
|---|---|---|---|---|---|---|---|---|
| 1 | Economía y Derecho | Mujer | 62.55 | 19.84 | 64.81 | 5.56 | 100.00 | 886 |
| 2 | Experimentales | Mujer | 54.80 | 22.08 | 55.56 | 0.00 | 100.00 | 388 |
| 3 | Humanidades | Mujer | 47.93 | 21.47 | 46.30 | 0.00 | 100.00 | 649 |
| 4 | Salud | Mujer | 62.35 | 17.26 | 64.81 | 0.00 | 100.00 | 705 |
| 5 | Sociales | Mujer | 52.84 | 20.35 | 55.56 | 0.00 | 100.00 | 2154 |
| 6 | Tecnica | Mujer | 61.34 | 20.40 | 64.81 | 0.00 | 100.00 | 477 |
| 7 | Economía y Derecho | Varon | 63.12 | 20.77 | 66.67 | 0.00 | 100.00 | 633 |
| 8 | Experimentales | Varon | 56.97 | 22.27 | 60.19 | 0.00 | 100.00 | 244 |
| 9 | Humanidades | Varon | 46.68 | 21.84 | 46.30 | 0.00 | 88.89 | 255 |
| 10 | Salud | Varon | 66.61 | 17.41 | 64.81 | 11.11 | 100.00 | 232 |
| 11 | Sociales | Varon | 54.59 | 21.31 | 55.56 | 0.00 | 100.00 | 628 |
| 12 | Tecnica | Varon | 66.32 | 18.50 | 74.07 | 0.00 | 100.00 | 1593 |

La información de la tabla graficar igualmente. A través de diversas instrucciones generamos gráficos de medias y diagramas de caja que nos muestran una representación descriptiva de las relaciones entre las variables.

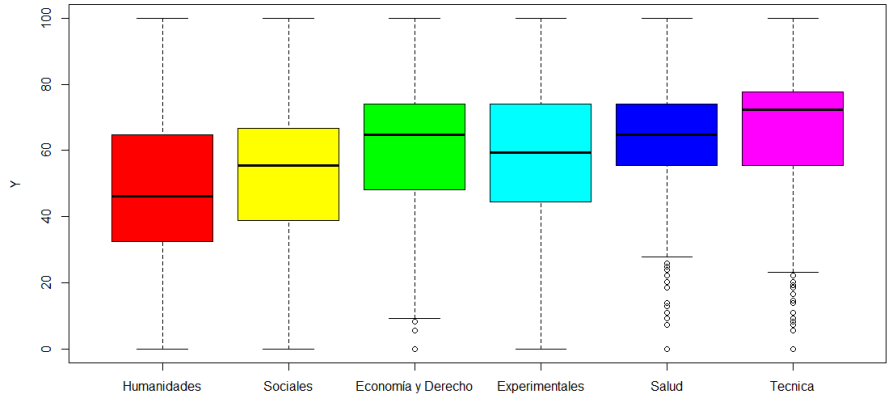
```
> # Gráficos
> qplot(A,Y, stat="summary", fun.y="mean")
> qplot(B,Y, stat="summary", fun.y="mean")
> plot(Y ~ A + B, col=rainbow(6))
> plot(A:B,Y, col=rainbow(2))
> boxplot(Y ~ A + B, col=rainbow(6))
> interaction.plot(A,B,Y)
```

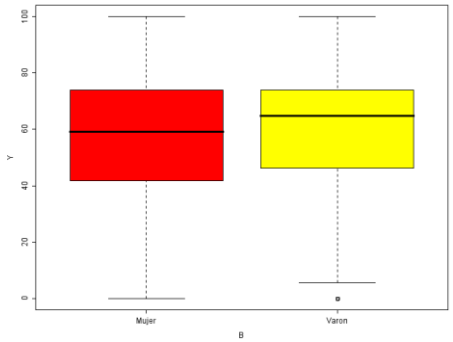
[32] El archivo de instrucciones que realiza los distintos análisis que se presentan es AV-AQU.R.

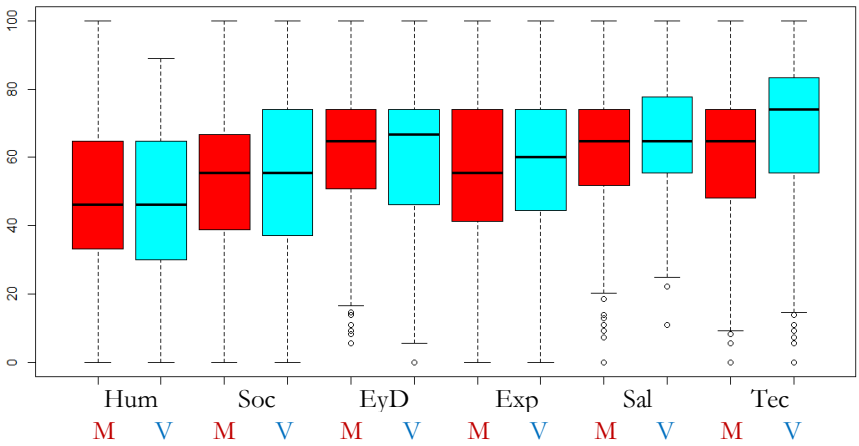
Gráficos de puntos con los valores de las medias:

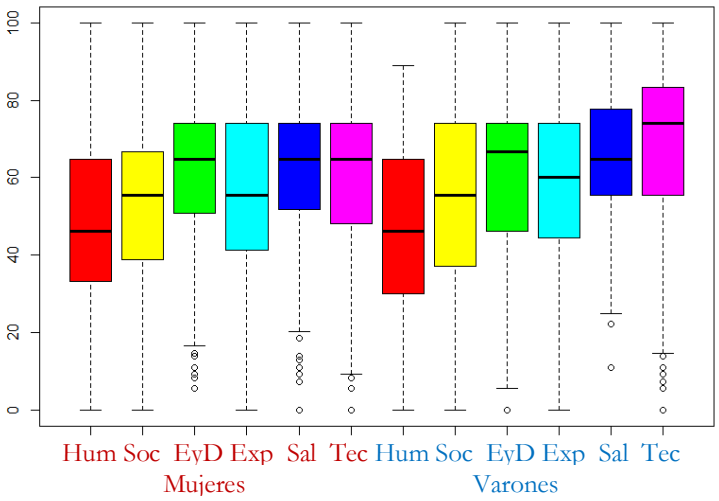



Diagramas de caja:





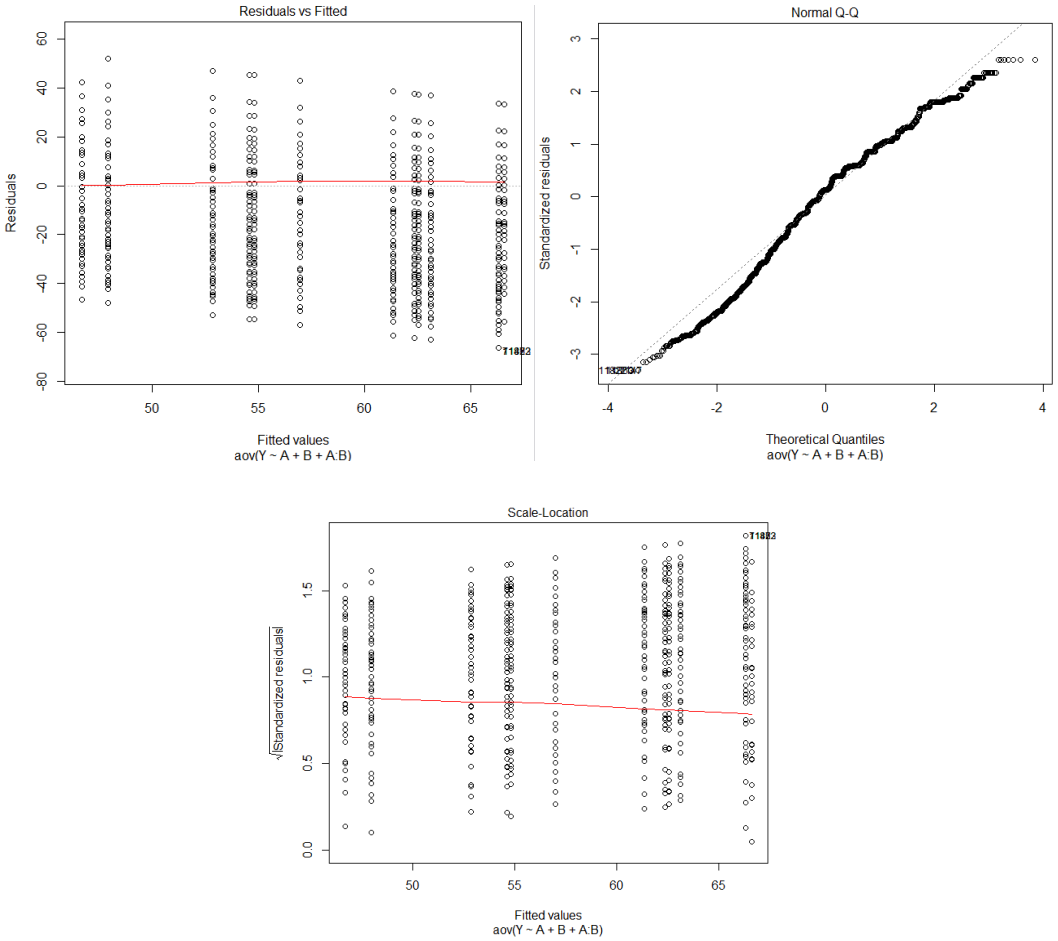

A continuación, se ejecuta el procedimiento de análisis de varianza donde se ha especificado el modelo completo con los dos efectos principales y la interacción. Adicionalmente se especifica con contrasts=list(A=contr.sum, B=contr.sum) una opción que nos permitirá obtener una estimación del modelo con un cálculo de la suma de cuadrados tipo III. Creamos así el objeto AV y solicitamos a continuación la información de las tablas de medias.

```
> # Anova
> AV=aov(Y ~ A + B + A:B, data=Empleados,
+ contrasts=list(A=contr.sum, B=contr.sum))
> model.tables(AV, "means") # Tabla de medias

Tables of means
Grand mean
58.34129

 A
    Humanidades Sociales Economía y Derecho Experimentales Salud Tecnica
          47.58    53.23              62.79          55.64  63.4   65.17
rep      904.00  2782.00            1519.00         632.00 937.0 2070.00

 B
      Mujer   Varon
      57.63   59.38
rep 5259.00 3585.00
```

```
 A:B
                    B
A                    Mujer  Varon
  Humanidades         47.9   46.7
  rep                649.0  255.0
  Sociales            52.8   54.6
  rep               2154.0  628.0
  Economía y Derecho  62.6   63.1
  rep                886.0  633.0
  Experimentales      54.8   57.0
  rep                388.0  244.0
  Salud               62.3   66.6
  rep                705.0  232.0
  Tecnica             61.3   66.3
  rep                477.0 1593.0
```

Una vez creado el objeto del análisis de varianza podemos ejecutar el comando plot para obtener diversos gráficos que nos informan del comportamiento de los datos en relación a algunos de los supuestos de un anova.

```
> plot(AV) # Gráficos de ajuste del modelo
```




El primer gráfico de los residuos en relación a los valores ajustados por el modelo (las medias de los grupos) es indicador de buenas condiciones de aplicación en la medida en que se no observan distribuciones muy diferentes entre sí de los puntos en cada grupo. En este caso tiende a ser así. El gráfico Normal Q-Q nos indica en qué medida los datos se alejan de la distribución normal que representa la línea diagonal. Vemos como el alejamiento es muy pequeño. El tercer gráfico es una variante del primero pero con una escala diferente del eje Y al considerar la raíz cuadrada de los residuos: como regla general los puntos por encima de 2 sobre el eje Y sugieren heterogeneidad de la varianza. En este caso no se observan y a pesar de que la prueba de Levene siguiente se muestre significativa.

```
> leveneTest(Y ~ A, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)
group    5  20.236 < 2.2e-16 ***
      8838
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Seguidamente se presenta la tabla ANOVA que muestra la significación de los efectos. La primera se corresponde con un cálculo de la suma de cuadrado Tipo I, la que se estima por defecto en R, la segunda solicita un cálculo Tipo III que es la se corresponde a nuestros datos.

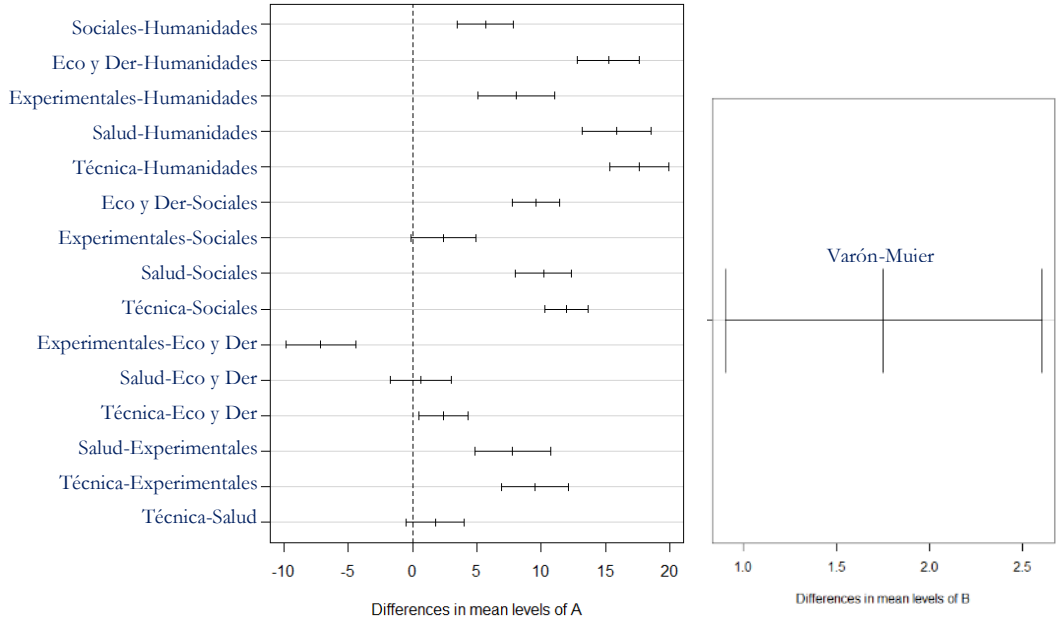
```
> summary(AV)  # SC Tipo I
             Df  Sum Sq Mean Sq F value  Pr(>F)
A             5  332609   66522 165.519 < 2e-16 ***
B             1    8062    8062  20.059 7.6e-06 ***
A:B           5    6842    1368   3.405 0.00448 **
Residuals  8832 3549568     402
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
3022 observations deleted due to missingness
> Anova(AV, type="III")  # SC Tipo III
Anova Table (Type III tests)

Response: Y
             Sum Sq   Df   F value    Pr(>F)
(Intercept) 19110871    1 47551.485 < 2.2e-16 ***
A             234762    5   116.827 < 2.2e-16 ***
B               6161    1    15.331 9.092e-05 ***
A:B             6842    5     3.405  0.004478 **
Residuals    3549568 8832
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

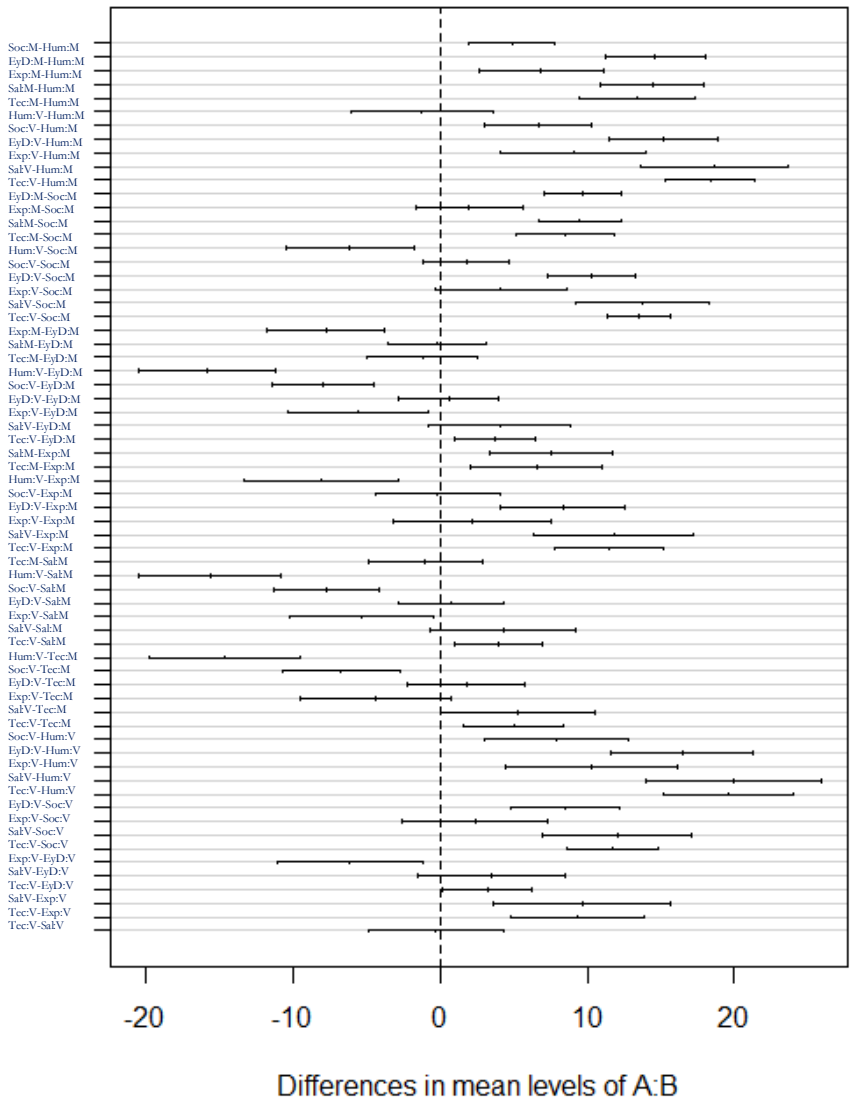
Comprobamos la significación de todos ellos y concluimos la existencia de diferencias en las medias entre alguna categoría de las dos variables independiente y también comparando parejas de valores de ambos factores. La siguiente tarea por tanto es determinar entre qué categorías se dan estas diferencias. Para ello ejecutaremos el procedimiento de comparaciones múltiples a posteriori (post-hoc) denominado Diferencia Significativa Honesta (HSD) de Tukey. Los resultados de las pruebas se presentan a continuación junto con una representación gráfica de las mismas. La interpretación la realizamos anteriormente y resumimos la significación de cada pareja de comparaciones en la Tabla III.8.16.

```
> # Comparación de medias: Tukey's Honest Significance Test
> TukeyHSD(AV)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ A + B + A:B, data = Empleados, contrasts = list(A =
contr.sum, B = contr.sum))

$A
                                       diff        lwr       upr     p adj
Sociales-Humanidades                5.6543972  3.4667937  7.842001 0.0000000
Economía y Derecho-Humanidades     15.2095506 12.8092406 17.609861 0.0000000
Experimentales-Humanidades          8.0579126  5.0950837 11.020742 0.0000000
Salud-Humanidades                  15.8240049 13.1600515 18.487958 0.0000000
Tecnica-Humanidades                17.5952101 15.3172043 19.873216 0.0000000
Economía y Derecho-Sociales         9.5551534  7.7321772 11.378130 0.0000000
Experimentales-Sociales             2.4035154 -0.1144420  4.921473 0.0712680
Salud-Sociales                     10.1696077  8.0112746 12.327941 0.0000000
Tecnica-Sociales                   11.9408129 10.2821796 13.599446 0.0000000
Experimentales-Economía y Derecho -7.1516379 -9.8564463 -4.446830 0.0000000
Salud-Economía y Derecho           0.6144543 -1.7592096  2.988118 0.9772346
Tecnica-Economía y Derecho         2.3856596  0.4551303  4.316189 0.0057439
Salud-Experimentales               7.7660922  4.8248089 10.707376 0.0000000
Tecnica-Experimentales             9.5372975  6.9404127 12.134182 0.0000000
Tecnica-Salud                      1.7712053 -0.4787064  4.021117 0.2178081
```

```
$B
                diff       lwr      upr    p adj
Varon-Mujer 1.751829 0.9007022 2.602955 5.51e-05

$`A:B`
                                                          diff         lwr         upr     p adj
Sociales:Mujer-Humanidades:Mujer                     4.9062385   1.97180226   7.8406747 0.0000031
Economía y Derecho:Mujer-Humanidades:Mujer          14.6220079  11.23611431  18.0079015 0.0000000
Experimentales:Mujer-Humanidades:Mujer               6.8677742   2.66235308  11.0731954 0.0000063
Salud:Mujer-Humanidades:Mujer                       14.4157021  10.85077417  17.9806300 0.0000000
Tecnica:Mujer-Humanidades:Mujer                     13.4088759   9.45661152  17.3611402 0.0000000
Humanidades:Varon-Humanidades:Mujer                 -1.2464682  -6.08986562   3.5969292 0.9995439
Sociales:Varon-Humanidades:Mujer                     6.6629536   2.99476546  10.3311418 0.0000002
Economía y Derecho:Varon-Humanidades:Mujer          15.1881878  11.52736980  18.8490058 0.0000000
Experimentales:Varon-Humanidades:Mujer               9.0397176   4.11856542  13.9608698 0.0000001
Salud:Varon-Humanidades:Mujer                       18.6834926  13.67069763  23.6962875 0.0000000
Tecnica:Varon-Humanidades:Mujer                     18.3918596  15.34013108  21.4435881 0.0000000
Economía y Derecho:Mujer-Sociales:Mujer              9.7157694   7.10026530  12.3312735 0.0000000
Experimentales:Mujer-Sociales:Mujer                  1.9615357  -1.65262438   5.5756958 0.8322786
Salud:Mujer-Sociales:Mujer                           9.5094636   6.66599787  12.3529293 0.0000000
Tecnica:Mujer-Sociales:Mujer                         8.5026374   5.18646793  11.8188068 0.0000000
Humanidades:Varon-Sociales:Mujer                    -6.1527067 -10.49264734  -1.8127661 0.0002280
Sociales:Varon-Sociales:Mujer                        1.7567151  -1.21518505   4.7286153 0.7389524
Economía y Derecho:Varon-Sociales:Mujer             10.2819493   7.31915085  13.2447477 0.0000000
Experimentales:Varon-Sociales:Mujer                  4.1334791  -0.29306866   8.5600268 0.0941074
Salud:Varon-Sociales:Mujer                          13.7772541   9.24904257  18.3054656 0.0000000
Tecnica:Varon-Sociales:Mujer                        13.4856211  11.32006501  15.6511772 0.0000000
Experimentales:Mujer-Economía y Derecho:Mujer       -7.7542337 -11.74365880  -3.7648086 0.0000000
Salud:Mujer-Economía y Derecho:Mujer                -0.2063058  -3.51366985   3.1010582 1.0000000
Tecnica:Mujer-Economía y Derecho:Mujer              -1.2131321  -4.93473644   2.5084723 0.9960110
Humanidades:Varon-Economía y Derecho:Mujer         -15.8684761 -20.52556150 -11.2113908 0.0000000
Sociales:Varon-Economía y Derecho:Mujer             -7.9590543 -11.37746769  -4.5406409 0.0000000
Economía y Derecho:Varon-Economía y Derecho:Mujer   0.5661799  -2.84432362   3.9766834 0.9999944
Experimentales:Varon-Economía y Derecho:Mujer       -5.5822903 -10.32018906  -0.8443916 0.0066173
Salud:Varon-Economía y Derecho:Mujer                 4.0614847  -0.77153291   8.8945022 0.2035150
Tecnica:Varon-Economía y Derecho:Mujer               3.7698517   1.02340076   6.5163026 0.0004531
Salud:Mujer-Experimentales:Mujer                     7.5479279   3.40547107  11.6903846 0.0000002
Tecnica:Mujer-Experimentales:Mujer                   6.5411016   2.06096429  11.0212390 0.0001172
Humanidades:Varon-Experimentales:Mujer              -8.1142425 -13.39720899  -2.8312759 0.0000337
Sociales:Varon-Experimentales:Mujer                 -0.2048206  -4.43646822   4.0268270 1.0000000
Economía y Derecho:Varon-Experimentales:Mujer        8.3204136   4.09515315  12.5456740 0.0000000
Experimentales:Varon-Experimentales:Mujer            2.1719434  -3.18239842   7.5262851 0.9758946
Salud:Varon-Experimentales:Mujer                    11.8157184   6.37702822  17.2544085 0.0000000
Tecnica:Varon-Experimentales:Mujer                  11.5240854   7.81406055  15.2341102 0.0000000
Tecnica:Mujer-Salud:Mujer                           -1.0068262  -4.89202566   2.8783732 0.9995127
Humanidades:Varon-Salud:Mujer                      -15.6621703 -20.45099892 -10.8733417 0.0000000
Sociales:Varon-Salud:Mujer                          -7.7527485 -11.34857738  -4.1569195 0.0000000
Economía y Derecho:Varon-Salud:Mujer                 0.7724857  -2.81582441   4.3607958 0.9999203
Experimentales:Varon-Salud:Mujer                    -5.3759845 -10.24343967  -0.5085293 0.0161252
Salud:Varon-Salud:Mujer                              4.2677905  -0.69229966   9.2278807 0.1746140
Tecnica:Varon-Salud:Mujer                            3.9761575   1.01179786   6.9405172 0.0007198
Humanidades:Varon-Tecnica:Mujer                    -14.6553441 -19.73909879  -9.5715894 0.0000000
Sociales:Varon-Tecnica:Mujer                        -6.7459222 -10.72608154  -2.7657629 0.0000020
Economía y Derecho:Varon-Tecnica:Mujer               1.7793119  -2.19405591   5.7526798 0.9501752
Experimentales:Varon-Tecnica:Mujer                  -4.3691583  -9.52704566   0.7887291 0.1934878
Salud:Varon-Tecnica:Mujer                            5.2746167   0.02922095  10.5200125 0.0472628
Tecnica:Varon-Tecnica:Mujer                          4.9829837   1.56258748   8.4033800 0.0001233
Sociales:Varon-Humanidades:Varon                     7.9094218   3.04323516  12.7756085 0.0000072
Economía y Derecho:Varon-Humanidades:Varon           16.4346560  11.57402264  21.2952894 0.0000000
Experimentales:Varon-Humanidades:Varon               10.2861858   4.41746303  16.1549086 0.0000007
Salud:Varon-Humanidades:Varon                        19.9299608  13.98418233  25.8757393 0.0000000
Tecnica:Varon-Humanidades:Varon                      19.6383278  15.21823558  24.0584200 0.0000000
Economía y Derecho:Varon-Sociales:Varon              8.5252342   4.83431789  12.2161504 0.0000000
Experimentales:Varon-Sociales:Varon                  2.3767640  -2.56681908   7.3203470 0.9193920
Salud:Varon-Sociales:Varon                          12.0205390   6.98572140  17.0553565 0.0000000
Tecnica:Varon-Sociales:Varon                        11.7289060   8.64113630  14.8166756 0.0000000
Experimentales:Varon-Economía y Derecho:Varon       -6.1484702 -11.08658696  -1.2103534 0.0027734
Salud:Varon-Economía y Derecho:Varon                 3.4953048  -1.53414564   8.5247552 0.4970248
Tecnica:Varon-Economía y Derecho:Varon               3.2036718   0.12466135   6.2826823 0.0328611
Salud:Varon-Experimentales:Varon                     9.6437750   3.63448867  15.6530613 0.0000103
Tecnica:Varon-Experimentales:Varon                   9.3521420   4.84698320  13.8573008 0.0000000
Tecnica:Varon-Salud:Varon                           -0.2916330  -4.89672043   4.3134544 1.0000000
```

```
plot(TukeyHSD(AV),las=1)
```


Sociales-Humanidades
Eco y Der-Humanidades
Experimentales-Humanidades
Salud-Humanidades
Técnica-Humanidades
Eco y Der-Sociales
Experimentales-Sociales
Salud-Sociales
Técnica-Sociales
Experimentales-Eco y Der
Salud-Eco y Der
Técnica-Eco y Der
Salud-Experimentales
Técnica-Experimentales
Técnica-Salud
-10
-5
0
5
10
15
20
Differences in mean levels of A
Varón-Muier
1.0
1.5
2.0
2.5
Differences in mean levels of B


Soc:M-Hum:M
EyD:M-Hum:M
Exp:M-Hum:M
SaE:M-Hum:M
Tec:M-Hum:M
Hum:V-Hum:M
Soc:V-Hum:M
EyD:V-Hum:M
Exp:V-Hum:M
SaE:V-Hum:M
Tec:V-Hum:M
EyD:M-Soc:M
Exp:M-Soc:M
SaE:M-Soc:M
Tec:M-Soc:M
Hum:V-Soc:M
Soc:V-Soc:M
EyD:V-Soc:M
Exp:V-Soc:M
SaE:V-Soc:M
Tec:V-Soc:M
Exp:M-EyD:M
SaE:M-EyD:M
Tec:M-EyD:M
Hum:V-EyD:M
Soc:V-EyD:M
EyD:V-EyD:M
Exp:V-EyD:M
SaE:V-EyD:M
Tec:V-EyD:M
SaE:M-Exp:M
Tec:M-Exp:M
Hum:V-Exp:M
Soc:V-Exp:M
EyD:V-Exp:M
Exp:V-Exp:M
SaE:V-Exp:M
Tec:V-Exp:M
Tec:M-SaE:M
Hum:V-SaE:M
Soc:V-SaE:M
EyD:V-SaE:M
Exp:V-SaE:M
SaE:V-SaE:M
Tec:V-SaE:M
Hum:V-Tec:M
Soc:V-Tec:M
EyD:V-Tec:M
Exp:V-Tec:M
SaE:V-Tec:M
Tec:V-Tec:M
Soc:V-Hum:V
EyD:V-Hum:V
Exp:V-Hum:V
SaE:V-Hum:V
Tec:V-Hum:V
EyD:V-Soc:V
Exp:V-Soc:V
SaE:V-Soc:V
Tec:V-Soc:V
Exp:V-EyD:V
SaE:V-EyD:V
Tec:V-EyD:V
SaE:V-Exp:V
Tec:V-Exp:V
Tec:V-SaE:V
-20
-10
0
10
20
Differences in mean levels of A:B

Finalmente determinamos la capacidad explicativa del modelo y la importancia diferenciada de cada efecto. Podemos concluir, como comentamos anteriormente, la relativa baja capacidad explicativa del modelo y la debilidad del efecto de la variable Sexo y de la interacción Area:Sexo, siendo el efecto principal de la variable Area el que determina las diferencias observadas en el índice de calidad ocupacional ICO.

```
> # Capacidad explicativa y coeficientes
> etaSquared(AV, type=3)
        eta.sq eta.sq.part
A   0.060240549 0.062035368
B   0.001581025 0.001732804
A:B 0.001755734 0.001923917
> summary.lm(AV)

Call:
aov(formula = Y ~ A + B + A:B, data = Empleados, contrasts = list(A = contr.sum,
    B = contr.sum))

Residuals:
    Min      1Q  Median      3Q     Max
-66.323 -11.627   2.593  12.734  52.069

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.00965    0.26602 218.063  < 2e-16 ***
A1          -10.70160    0.66080 -16.195  < 2e-16 ***
A2           -4.29377    0.45665  -9.403  < 2e-16 ***
A3            4.82674    0.50218   9.612  < 2e-16 ***
A4           -2.12462    0.71967  -2.952  0.00316 **
A5            6.47124    0.67417   9.599  < 2e-16 ***
B1           -1.04160    0.26602  -3.915 9.09e-05 ***
A1:B1         1.66483    0.66080   2.519  0.01177 *
A2:B1         0.16324    0.45665   0.357  0.72075
A3:B1         0.75851    0.50218   1.510  0.13097
A4:B1        -0.04438    0.71967  -0.062  0.95083
A5:B1        -1.09230    0.67417  -1.620  0.10522
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.05 on 8832 degrees of freedom
  (3022 observations deleted due to missingness)
Multiple R-squared: 0.08917,  Adjusted R-squared: 0.08804
F-statistic: 78.61 on 11 and 8832 DF,  p-value: < 2.2e-16
```

Como ejercicio adicional se puede realizar un análisis de la misma variable dependiente del índice ocupacional considerando otras variables independientes, o bien tomar como variable a explicar los ingresos o las notas del expediente académico considerando también diversas variables independientes posibles.

## 9. Bibliografía


Ajenjo, M.; García, J. (2014). La transmissió intergeneracional de rols de gènere a la llar. *Revista Catalana de Sociologia*, 29, desembre, 35-47. DOI: 10.2436/20.3005.01.55. http://revistes.iec.cat/index.php/RCS/article/view/56199/pdf_587

Agresti, A.; Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Upper Saddle River, New Jersey: Pearson Prentice Hall.

AQU (2014). *Universitat i treball a Catalunya 2014. Estudi de la inserció laboral de la població titulada de les universitats catalanes*. Barcelona: Agència per a la Qualitat del Sistema Universitari de Catalunya.

Arnau, J. (1990). *Psicología experimental. Un enfoque metodológico*. 2a. ed. México: Trillas.

Barker, R. H.; Barker, B. M. (1984). *Multivariate Analysis of Variance (MANOVA): a practical guide to its use in scientific decision making*. Alabama: The University of Alabama Press.

Bisquerra, R. (1987). *Introducción a la estadística aplicada a la investigación educativa. Un enfoque informático con los paquetes BMDP y SPSSX*. Barcelona: Promociones y Publicaciones Universitarias.

Bisquerra, R. (1989). *Introducción conceptual al análisis multivariable*. Barcelona: Promociones y Publicaciones Universitarias.

Blalock, H. M. Jr. (1978). *Estadística Social*. México: Fondo de Cultura Económica. Wa. Edición.

Bray, J. H.; Maxwell, S. E. (1985). *Multivariate Analysis of Variance*. Beverly Hills: Sage Publications.

Bryman, A.; Cramer, D: (1990). *Quantitative Data Analysis for Social Scientist*. London: Routledge.

Catena, A.; Ramos, M. M.; Trujillo, H. M. (2003). *Análisis multivariado. Un manual para investigadores*. Madrid: Biblioteca Nueva.

Corominas, E.; Villar, E; Saurina, C.; Fàbregas, M. (2007). El mercat laboral qualificat i la qualitat de l'ocupació. En *Educació superior i treball a Catalunya. Anàlisi dels factors d'inserció laboral*. Barcelona: Agència per a la Qualitat del Sistema Universitari de Catalunya.

Crawley, M. J. (2005). *Statistics. An Introduction using R*. West Sussex: John Wiley & Sons.

Dalgaard, P. (2008). *Introductory Statistics with R*. New York: Springer.

Domènech, J. M. (1982). *Bioestadística. Métodos estadísticos para investigadores*. Barcelona: Herder.

Domenech, J. M.; Riba, M. D. (1981). *Una síntesis de los métodos estadísticos bivariantes*. Barcelona: Herder.

Everitt, B. S.; Hothorn, T. (2006). Analysis of Variance: Weight Gain, Foster Feeding in Rats, Water Hardness and Male Egyptian Skulls. En B. S. Everitt y T. Hothorn, *A Handbook of Statistical Analyses Using R*, 55-72.

Fachelli, S.; Montolio D. (2015). Valuation of the training received in university regarding the utility for work by Catalan graduates. *Multidisciplinary journal for education, social and technological sciences*, 2, 2, 14-37.

Fachelli, S; Planas, J. (2014). Inserción profesional y movilidad intergeneracional de los universitarios: de la expansión a la crisis. *Revista Española de Sociología*, 21, 68-98.

Faraway, J. J. (2015). *Linear Models with R*. 2nd edition. Boca Raton, FL: Taylor and Francis.

Gamst, G.; Meyers, L. S.; Guarino, A. J. (2008). *Analysis of Variance Designs. A Conceptual and Computational Approach with SPSS and SAS*. New York: Cambridge University Press.

García Ferrando, Manuel (1987) *Socioestadística. Introducción a la estadística en sociología*. 2a edició amp. Madrid: Alianza. Alianza Universidad Textos, 96.

Glass, G. V.; Stanley, J. C. (1970). *Métodos Estadísticos Aplicados a las Ciencias Sociales*. Méxixo: Prentice-Hall.

Hair, J. F. et al. (2011). *Multivariate Data Analysis*. Upper Saddle River, New jersey: Pearson Prentice Hall.

Hand, D. J.; Taylor, C. C. (1987) .*Multivariate Analysis of Variance and Repeated Measures: a practical approach for behavioural scientist*. London: Chapman and Hall.

Ho, R. (2006). *Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS*. Boca Raton: Chapman and Hall/CRC.

Hoaglin, D. C.; Mosteller, F.; Tukey, J. W. (1991). *Fundamentals of Exploratory Analysis of Variance*. New York: John Wiley & Sons.

Iversen, G. R.; Norpoth, H. (1987). *Analysis of Variance*. Beverly Hills: Sage Publications.

Lebart, L.; Morineau, A.; Fénelon, J.-P. (1985). *Tratamiento estadístico de datos. Métodos y Programas*. Barcelona: Marcombo.

Lévy, J.-P.; Varela, J. (2003). *Análisis multivariable para las ciencias sociales*. Madrid: Pearson-Prentice Hall.

MacFarland, T. W. (2012). *Two-Way Analysis of Variance: Statistical Tests and Graphics Using R*. New York: Springer.

Maxwell, S. E.; Delaney, H. D. (2004). *Designing experiments and analyzing data. A Model Comparison Perspective*. 2nd Edition. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Moncho, J. (2015). *Estadística aplicada a las ciencias de la salud*. Barcelona: Elsevier.

Navarro, D. J. (2015). *Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners, Version 0.5*. School of Psychology, University of Adelaide, Adelaide, Australia. http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/

Page, M. C.; Braver, S. L.; MacKinnon, D. P. (2003). *Levine's Guide to SPSS for Analysis of Variance*. 2nd Edition. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Pardo, A.; Ruíz, M. A. (2001a). Análisis de varianza factorial. El procedimiento Modelo lineal general: Univariante. En A. Pardo y M. A. Ruíz, *SPSS 10.0. Guía para el análisis de datos*. Madrid: Hispanoportuguesa SPSS.

Pardo, A.; Ruíz, M. A. (2001b). Análisis de varianza con medidas repetidas: el procedimiento MLG: medidas repetidas. En A. Pardo y M. A. Ruíz, *SPSS 10.0. Guía para el análisis de datos*. Madrid: Hispanoportuguesa SPSS.

Pardo, A. et al. (2007). La interacción entre factores en el análisis de varianza: errores de interpretación. Psicothema, 17, 2, 343-349.

Pérez, C. (2004). T*écnicas de análisis multivariante de datos. Aplicaciones con SPSS*. Madrid: Pearson Prentice Hall.

Planas, J.; Fachelli, S. (2010*). Les universitats catalanes, factor d'equitat i de mobilitat professional. Una anàlisi sobre les relacions entre l'estatus familiar, el bagatge acadèmic i la inserció professional l'any 2008 dels titulats l'any 2004 a les universitats catalanes*. Barcelona: Agència per a la Qualitat del Sistema Universitari de Catalunya.

Riba, M. D. (1990). *Modelo lineal de análisis de la varianza*. Barcelona: Herder.

Ruiz-Maya, L. (1977). *Métodos estadísticos de investigación. Introducción al análisis de la varianza*. Madrid: Instituto Nacional de Estadística.

Sánchez Carrión, J. J. (1995). *Manual de análisis de datos*. Madrid: Alianza. Alianza Universidad Textos, 150.

Tejedor, F. J. (1984). *Análisis de varianza aplicada a la investigación en pedagogía y psicología*. Madrid: Anaya.

Tejedor, F. J. (1999). *Análisis de varianza: introducción conceptual y diseños básicos*. Madrid: La Muralla.

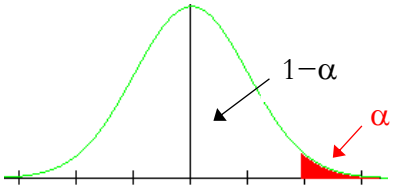
Webster, A. L. (2000). *Estadística Aplicada a los negocios y la economía*. Bogotá: Irwin McGraw-Hill

Ximénez, M. C.; San Martín, R. (2000). *Análisis de varianza con medidas repetidas*. Madrid: La Muralla.

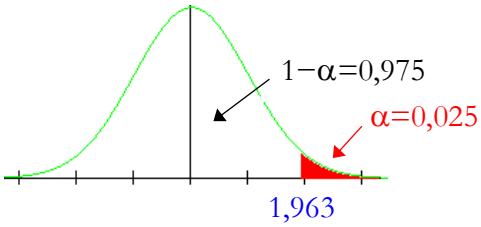
## Anexo I. Tabla de distribución teórica de la t de Student

| Grados de libertad ν | Probabilidades α (asociadas al valor crítico en un contraste a una cola) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0,10 | 0,075 | 0,05 | 0,025 | 0,01 | 0,005 | 0,001 |
| 1 | 3,078 | 4,165 | 6,314 | **12,706** | 31,821 | 63,657 | 318,309 |
| 2 | 1,886 | 2,282 | 2,920 | **4,303** | 6,965 | 9,925 | 22,327 |
| 3 | 1,638 | 1,924 | 2,353 | **3,182** | 4,541 | 5,841 | 10,215 |
| 4 | 1,533 | 1,778 | 2,132 | **2,776** | 3,747 | 4,604 | 7,173 |
| 5 | 1,476 | 1,699 | 2,015 | **2,571** | 3,365 | 4,032 | 5,893 |
| 6 | 1,440 | 1,650 | 1,943 | **2,447** | 3,143 | 3,707 | 5,208 |
| 7 | 1,415 | 1,617 | 1,895 | **2,365** | 2,998 | 3,499 | 4,785 |
| 8 | 1,397 | 1,592 | 1,860 | **2,306** | 2,896 | 3,355 | 4,501 |
| 9 | 1,383 | 1,574 | 1,833 | **2,262** | 2,821 | 3,250 | 4,297 |
| 10 | 1,372 | 1,559 | 1,812 | **2,228** | 2,764 | 3,169 | 4,144 |
| 11 | 1,363 | 1,548 | 1,796 | **2,201** | 2,718 | 3,106 | 4,025 |
| 12 | 1,356 | 1,538 | 1,782 | **2,179** | 2,681 | 3,055 | 3,930 |
| 13 | 1,350 | 1,530 | 1,771 | **2,160** | 2,650 | 3,012 | 3,852 |
| 14 | 1,345 | 1,523 | 1,761 | **2,145** | 2,624 | 2,977 | 3,787 |
| 15 | 1,341 | 1,517 | 1,753 | **2,131** | 2,602 | 2,947 | 3,733 |
| 16 | 1,337 | 1,512 | 1,746 | **2,120** | 2,583 | 2,921 | 3,686 |
| 17 | 1,333 | 1,508 | 1,740 | **2,110** | 2,567 | 2,898 | 3,646 |
| 18 | 1,330 | 1,504 | 1,734 | **2,101** | 2,552 | 2,878 | 3,610 |
| 19 | 1,328 | 1,500 | 1,729 | **2,093** | 2,539 | 2,861 | 3,579 |
| 20 | 1,325 | 1,497 | 1,725 | **2,086** | 2,528 | 2,845 | 3,552 |
| 21 | 1,323 | 1,494 | 1,721 | **2,080** | 2,518 | 2,831 | 3,527 |
| 22 | 1,321 | 1,492 | 1,717 | **2,074** | 2,508 | 2,819 | 3,505 |
| 23 | 1,319 | 1,489 | 1,714 | **2,069** | 2,500 | 2,807 | 3,485 |
| 24 | 1,318 | 1,487 | 1,711 | **2,064** | 2,492 | 2,797 | 3,467 |
| 25 | 1,316 | 1,485 | 1,708 | **2,060** | 2,485 | 2,787 | 3,450 |
| 26 | 1,315 | 1,483 | 1,706 | **2,056** | 2,479 | 2,779 | 3,435 |
| 27 | 1,314 | 1,482 | 1,703 | **2,052** | 2,473 | 2,771 | 3,421 |
| 28 | 1,313 | 1,480 | 1,701 | **2,048** | 2,467 | 2,763 | 3,408 |
| 29 | 1,311 | 1,479 | 1,699 | **2,045** | 2,462 | 2,756 | 3,396 |
| 30 | 1,310 | 1,477 | 1,697 | **2,042** | 2,457 | 2,750 | 3,385 |
| 32 | 1,309 | 1,475 | 1,694 | **2,037** | 2,449 | 2,738 | 3,365 |
| 34 | 1,307 | 1,473 | 1,691 | **2,032** | 2,441 | 2,728 | 3,348 |
| 36 | 1,306 | 1,471 | 1,688 | **2,028** | 2,434 | 2,719 | 3,333 |
| 38 | 1,304 | 1,469 | 1,686 | **2,024** | 2,429 | 2,712 | 3,319 |
| 40 | 1,303 | 1,468 | 1,684 | **2,021** | 2,423 | 2,704 | 3,307 |
| 42 | 1,302 | 1,466 | 1,682 | **2,018** | 2,418 | 2,698 | 3,296 |
| 44 | 1,301 | 1,465 | 1,680 | **2,015** | 2,414 | 2,692 | 3,286 |
| 46 | 1,300 | 1,464 | 1,679 | **2,013** | 2,410 | 2,687 | 3,277 |
| 48 | 1,299 | 1,463 | 1,677 | **2,011** | 2,407 | 2,682 | 3,269 |
| 50 | 1,299 | 1,462 | 1,676 | **2,009** | 2,403 | 2,678 | 3,261 |
| 60 | 1,296 | 1,458 | 1,671 | **2,000** | 2,390 | 2,660 | 3,232 |
| 70 | 1,294 | 1,456 | 1,667 | **1,994** | 2,381 | 2,648 | 3,211 |
| 80 | 1,292 | 1,453 | 1,664 | **1,990** | 2,374 | 2,639 | 3,195 |
| 90 | 1,291 | 1,452 | 1,662 | **1,987** | 2,368 | 2,632 | 3,183 |
| 100 | 1,290 | 1,451 | 1,660 | **1,984** | 2,364 | 2,626 | 3,174 |
| 110 | 1,289 | 1,450 | 1,659 | **1,982** | 2,361 | 2,621 | 3,166 |
| 120 | 1,289 | 1,449 | 1,658 | **1,980** | 2,358 | 2,617 | 3,160 |
| 150 | 1,287 | 1,447 | 1,655 | **1,976** | 2,351 | 2,609 | 3,145 |
| 200 | 1,286 | 1,445 | 1,653 | **1,972** | 2,345 | 2,601 | 3,131 |
| 250 | 1,285 | 1,444 | 1,651 | **1,969** | 2,341 | 2,596 | 3,123 |
| 500 | 1,283 | 1,442 | 1,648 | **1,965** | 2,334 | 2,586 | 3,107 |
| 750 | 1,283 | 1,441 | 1,647 | **1,963** | 2,331 | 2,582 | 3,101 |
| 1000 | 1,282 | 1,441 | 1,646 | **1,962** | 2,330 | 2,581 | 3,098 |
| 2000 | 1,282 | 1,440 | 1,646 | **1,961** | 2,328 | 2,578 | 3,094 |
| 3000 | 1,282 | 1,440 | 1,645 | **1,961** | 2,328 | 2,577 | 3,093 |
| 4000 | 1,282 | 1,440 | 1,645 | **1,961** | 2,327 | 2,577 | 3,092 |
| 5000 | 1,282 | 1,440 | 1,645 | **1,960** | 2,327 | 2,577 | 3,092 |
| 10000 | 1,282 | 1,440 | 1,645 | **1,960** | 2,327 | 2,576 | 3,091 |
| ∞ | 1,282 | 1,440 | 1,645 | **1,960** | 2,326 | 2,576 | 3,090 |

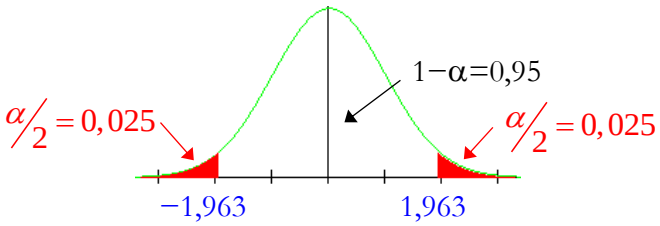
Cada valor de la tabla hace referencia al valor *t* que deja un área (probabilidad) a la derecha del mismo, igual al valor de la columna:




Por ejemplo, el valor 1,963 se corresponde con la línea ν=750 grados de libertad y la columna de probabilidad o nivel de significación α=0,025.




Este valor es el que corresponde a un contraste a dos colas considerando una significación del 0,05, dejando a cada lado el 0,025:



## Anexo II. Tabla de distribución teórica de la F

**Probabilidad α = 0,05**

| *Grados de libertad del denominador* | *Grados de libertad del numerador* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **15** | **20** |
| **1** | 161,4 | 199,5 | 215,7 | 224,6 | 230,2 | 234,0 | 236,8 | 238,9 | 240,5 | 241,9 | 246,0 | 248,0 |
| **2** | 18,51 | 19,00 | 19,16 | 19,25 | 19,30 | 19,33 | 19,35 | 19,37 | 19,39 | 19,40 | 19,43 | 19,45 |
| **3** | 10,13 | 9,55 | 9,28 | 9,12 | 9,01 | 8,94 | 8,89 | 8,85 | 8,81 | 8,79 | 8,70 | 8,66 |
| **4** | 7,71 | 6,94 | 6,59 | 6,39 | 6,26 | 6,16 | 6,09 | 6,04 | 6,00 | 5,96 | 5,86 | 5,80 |
| **5** | 6,61 | 5,79 | 5,41 | 5,19 | 5,05 | 4,95 | 4,88 | 4,82 | 4,77 | 4,74 | 4,62 | 4,56 |
| **6** | 5,99 | 5,14 | 4,76 | 4,53 | 4,39 | 4,28 | 4,21 | 4,15 | 4,10 | 4,06 | 3,94 | 3,87 |
| **7** | 5,59 | 4,74 | 4,35 | 4,12 | 3,97 | 3,87 | 3,79 | 3,73 | 3,68 | 3,64 | 3,51 | 3,44 |
| **8** | 5,32 | 4,46 | 4,07 | 3,84 | 3,69 | 3,58 | 3,50 | 3,44 | 3,39 | 3,35 | 3,22 | 3,15 |
| **9** | 5,12 | 4,26 | 3,86 | 3,63 | 3,48 | 3,37 | 3,29 | 3,23 | 3,18 | 3,14 | 3,01 | 2,94 |
| **10** | 4,97 | 4,10 | 3,71 | 3,48 | 3,33 | 3,22 | 3,14 | 3,07 | 3,02 | 2,98 | 2,85 | 2,77 |
| **11** | 4,84 | 3,98 | 3,59 | 3,36 | 3,20 | 3,10 | 3,01 | 2,95 | 2,90 | 2,85 | 2,72 | 2,65 |
| **12** | 4,75 | 3,89 | 3,49 | 3,26 | 3,11 | 3,00 | 2,91 | 2,85 | 2,80 | 2,75 | 2,62 | 2,54 |
| **13** | 4,67 | 3,81 | 3,41 | 3,18 | 3,03 | 2,92 | 2,83 | 2,77 | 2,71 | 2,67 | 2,53 | 2,46 |
| **14** | 4,60 | 3,74 | 3,34 | 3,11 | 2,96 | 2,85 | 2,76 | 2,70 | 2,65 | 2,60 | 2,46 | 2,39 |
| **15** | 4,54 | 3,68 | 3,29 | 3,06 | 2,90 | 2,79 | 2,71 | 2,64 | 2,59 | 2,54 | 2,40 | 2,33 |
| **16** | 4,49 | 3,63 | 3,24 | 3,01 | 2,85 | 2,74 | 2,66 | 2,59 | 2,54 | 2,49 | 2,35 | 2,28 |
| **17** | 4,45 | 3,59 | 3,20 | 2,97 | 2,81 | 2,70 | 2,61 | 2,55 | 2,49 | 2,45 | 2,31 | 2,23 |
| **18** | 4,41 | 3,56 | 3,16 | 2,93 | 2,77 | 2,66 | 2,58 | 2,51 | 2,46 | 2,41 | 2,27 | 2,19 |
| **19** | 4,38 | 3,52 | 3,13 | 2,90 | 2,74 | 2,63 | 2,54 | 2,48 | 2,42 | 2,38 | 2,23 | 2,16 |
| **20** | 4,35 | 3,49 | 3,10 | 2,87 | 2,71 | 2,60 | 2,51 | 2,45 | 2,39 | 2,35 | 2,20 | 2,12 |
| **22** | 4,30 | 3,44 | 3,05 | 2,82 | 2,66 | 2,55 | 2,46 | 2,40 | 2,34 | 2,30 | 2,15 | 2,07 |
| **24** | 4,26 | 3,40 | 3,01 | 2,78 | 2,62 | 2,51 | 2,42 | 2,36 | 2,30 | 2,26 | 2,11 | 2,03 |
| **26** | 4,23 | 3,37 | 2,98 | 2,74 | 2,59 | 2,47 | 2,39 | 2,32 | 2,27 | 2,22 | 2,07 | 1,99 |
| **28** | 4,20 | 3,34 | 2,95 | 2,71 | 2,56 | 2,45 | 2,36 | 2,29 | 2,24 | 2,19 | 2,04 | 1,96 |
| **30** | 4,17 | 3,32 | 2,92 | 2,69 | 2,53 | 2,42 | 2,33 | 2,27 | 2,21 | 2,17 | 2,02 | 1,93 |
| **40** | 4,09 | 3,23 | 2,84 | 2,61 | 2,45 | 2,34 | 2,25 | 2,18 | 2,12 | 2,08 | 1,92 | 1,84 |
| **50** | 4,03 | 3,18 | 2,79 | 2,56 | 2,40 | 2,29 | 2,20 | 2,13 | 2,07 | 2,03 | 1,87 | 1,78 |
| **60** | 4,00 | 3,15 | 2,76 | 2,53 | 2,37 | 2,25 | 2,17 | 2,10 | 2,04 | 1,99 | 1,84 | 1,75 |
| **120** | 3,92 | 3,07 | 2,68 | 2,45 | 2,29 | 2,18 | 2,09 | 2,02 | 1,96 | 1,91 | 1,75 | 1,66 |
| **∞** | 3,84 | 3,00 | 2,60 | 2,37 | 2,21 | 2,10 | 2,01 | 1,94 | 1,88 | 1,83 | 1,67 | 1,57 |