# Sobrevivir a la Estadística en 40 páginas y con 7 ejercicios

**Dr. Alberto Nájera López**

Facultad de Medicina de Albacete
Universidad de Castilla – La Mancha



Albacete, 14 de abril de 2014

# Contenido de esta guía

## Motivación

Ya he olvidado el número de cursos de Estadística a los que he asistido y de los que no me acuerdo de nada. Bien porque el enfoque era demasiado teórico y la aplicación práctica no se veía por ningún lado, o porque eran tal vez demasiado prácticos o, mejor dicho, los casos prácticos estaban diseñados de tal manera, que no se podían aplicar a mis datos. He de añadir que, en general, nunca he terminado los cursos pues me parecían una pérdida de tiempo.

Y es un problema generalizado: no conozco a nadie que diga que le encanta la Estadística, más bien todo lo contrario (hablo de personas normales). Y es un problema pues la Estadística es una herramienta básica en investigación, necesitamos saber Estadística, debemos saber Estadística, al menos lo básico, no nos queda más remedio.

Por este motivo y debido a las presiones de muchos alumnos y compañeros que desean disponer de un manual lo más breve posible con las recetas básicas y ejemplos claros. He creído conveniente sacrificar rigor académico en favor de explicar las cosas de forma clara, aunque en algunos casos puedan no ser del todo correctas, y elaborar esta breve guía de forma muy práctica que espero que te sea útil. Espero cumplir mis y, claro como no también tus, expectativas. Para ello, espero que este documento sirva de receta básica para saber qué hacer y cómo interpretar los resultados, sin necesidad de unos conocimientos de Estadística.

Más información en: http://blog.uclm.es/albertonajera/curso-de-supervivencia-en-estadistica/

## Objetivos de esta breve guía

Pretendo proporcionar la receta básica para realizar las operaciones estadísticas más sencillas e indispensables en Estadística Descriptiva e Inferencial mediante Excel (cuando sea posible) y con SPSS (cuando no quede más remedio), así como técnicas de contraste de hipótesis. ¿Por qué esta distinción? SPSS es una herramienta sumamente potente, pero es una herramienta más a añadir a nuestra lista de aplicaciones informáticas que no dominamos y que algún día pretendemos, sin conseguirlo, aprender. En cambio Excel, en general, es una herramienta que todos en algún momento hemos usado; conocemos la interfaz pues es similar a la de Word, nos es más familiar y, en muchas ocasiones, hay cálculos estadísticos que podremos realizar sin demasiados problemas. Descarto el paquete estadístico libre y gratuito "R" por considerarlo demasiado complicado y que requiere unos mínimos conocimientos de programación[1].

Así intentaré proporcionar unos mínimos conocimientos de Estadística teórica sin entrar en detalles, dejando de lado en algunos casos parte del rigor matemático, y realizando diversos ejemplos en Excel y SPSS para ayudar a asentar los conceptos y las técnicas. Para ello te propongo los siguientes objetivos:

1. Histograma de frecuencias y Estadística Descriptiva.
2. Distribución normal y predicciones. Test Kolmogorov-Smirnov/Shapiro-Wilk.

[1] http://www.r-project.org/

3. Estadística inferencial: Estimar la media de una población a partir de una muestra. Determinar el tamaño de una muestra.
4. Estadística inferencial: Prueba z. Test t. Wilcoxon. ANOVA. U de Mann-Whitney. Test de Kruskal-Wallis. Chi-cuadrado.
5. Cálculo de correlaciones y Regresión.

Puedes descargar el archivo de Excel para realizar los ejercicios desde: http://goo.gl/UEd2pg

## Introducción

En el siguiente diagrama se resumen los contenidos básicos de Estadística que debería dominar cualquier persona que tenga que analizar un conjunto de datos.

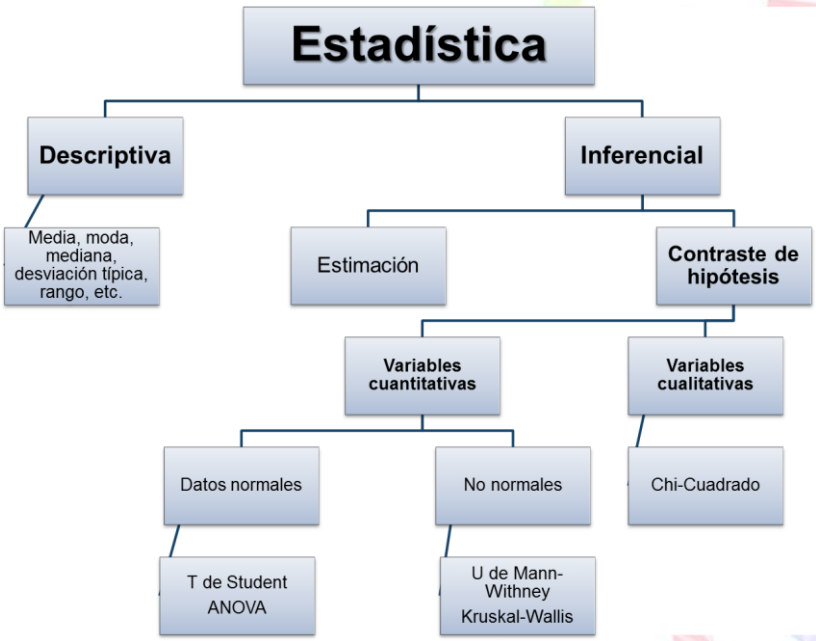



**Figura 1. Esquema que resume las diferentes partes de la Estadística y los principales test estadísticos.**

Comenzaremos por la Estadística descriptiva: realización de un histograma de frecuencias para después realizar estimaciones de valores de una población a partir de mi muestra, evaluando el error cometido y su relación con el tamaño de la muestra, para después realizar estudios comparativos, comparar muestras para determinar posibles diferencias y controlar el error o el grado de confianza del análisis dependiendo del tipo de datos que tengamos: normales o no, variables cualitativas o no. Por último veremos cómo realizar estudios de correlación y regresiones.

## Histograma de frecuencias

El desarrollo de la Ciencia se basa en la prueba y el error, en la reproducción sistemática de resultados. Esta metodología frecuentemente genera un gran conjunto de datos que debemos resumir, que debemos mostrar de forma sencilla. Si tengo un estudio con 20 millones de datos, no se me ocurriría entregar o intentar publicar esos datos en crudo (que sería lo ideal para no perder información). Pero debemos sacrificar parte de esa información "tan detallada" y

proporcionar un resumen. Esto lo conseguiremos mediante la Estadística descriptiva. Y en este sentido el primer paso es presentar los datos de forma muy visual, para lo cual lo normal es realizar una representación gráfica de los mismos mediante un histograma de frecuencias que nos mostrará los datos que he registrado y cuántas veces se repite cada uno de ellos.

Si tenemos que comunicar los datos de pulsaciones por minuto de 50 alumnos, lo mejor sería entregarlos tal cual, si acaso ordenados de menor a mayor y así estaríamos dando toda la información. Pero poder sacar algún tipo de información útil es otra cosa. A la vista de los datos mostrados en la Fig.2 puedo comprobar que hay un valor mínimo (62) un valor máximo (96) pero poco más.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 89 | 68 | 92 | 74 | 76 | 65 | 77 | 83 | 75 | 87 |
| 85 | 64 | 79 | 77 | 96 | 80 | 70 | 85 | 80 | 80 |
| 82 | 81 | 86 | 71 | 90 | 87 | 71 | 72 | 62 | 78 |
| 77 | 90 | 83 | 81 | 73 | 80 | 78 | 81 | 81 | 75 |
| 82 | 88 | 79 | 79 | 94 | 82 | 66 | 78 | 74 | 72 |

**Figura 2. Datos de partida. Pulsaciones por minuto de 50 personas.**

Si realizo un histograma de frecuencias (Fig. 3) puedo obtener más información, como por ejemplo que el 50% de los datos se encuentra entre los valores de 75 y 84 pulsaciones por minuto o que los valores más repetidos son el 80 y el 81. Si tengo muchos datos diferentes, mucha variabilidad, puedo agruparlos en intervalos (Fig. 3 abajo).

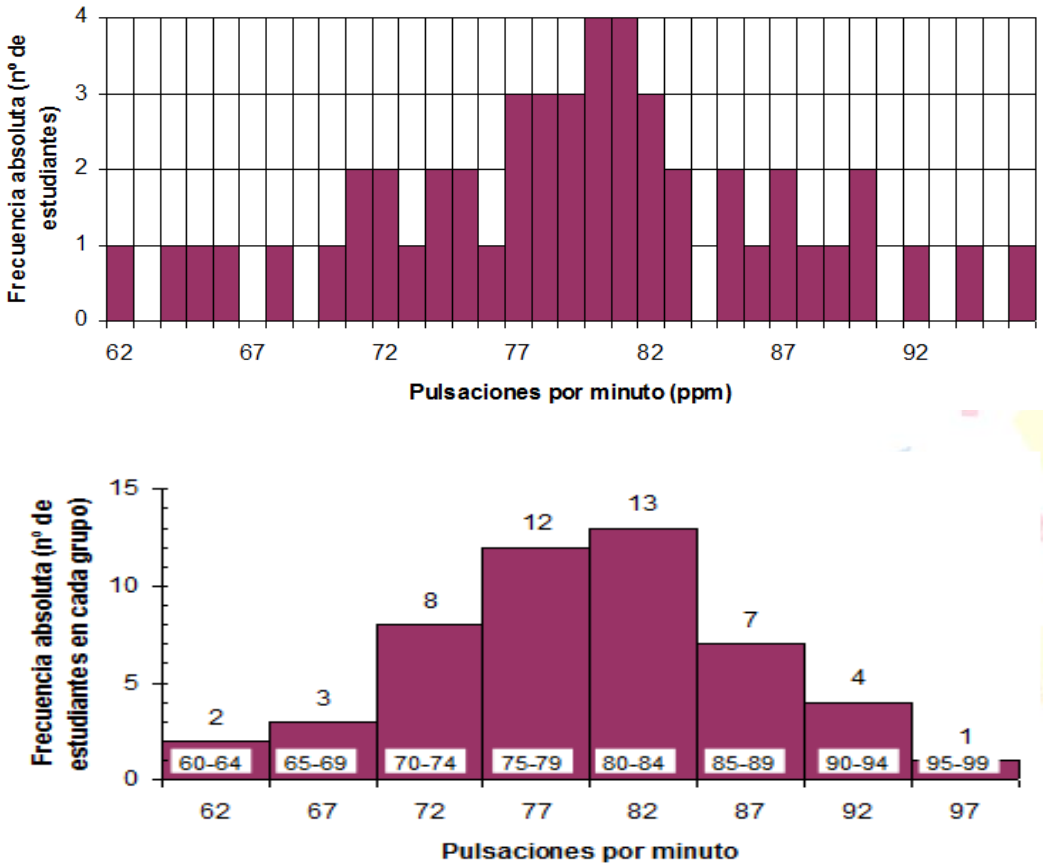



**Figura 3. Histograma de frecuencias (arriba) o con datos agrupados (abajo).**

Veamos primero cómo realizar este histograma en Excel. Para ello debemos tener activada la herramienta de "análisis de datos". Para activarla debemos ir a "Archivo" (o el botón de Office arriba a la izquierda dependiendo de la versión), "Opciones", "Complementos", abajo ir a "Administrar complementos de Excel" y darle a "Ir". Activar las dos "Herramientas de Análisis" y darle a "Aceptar" (Fig. 4).

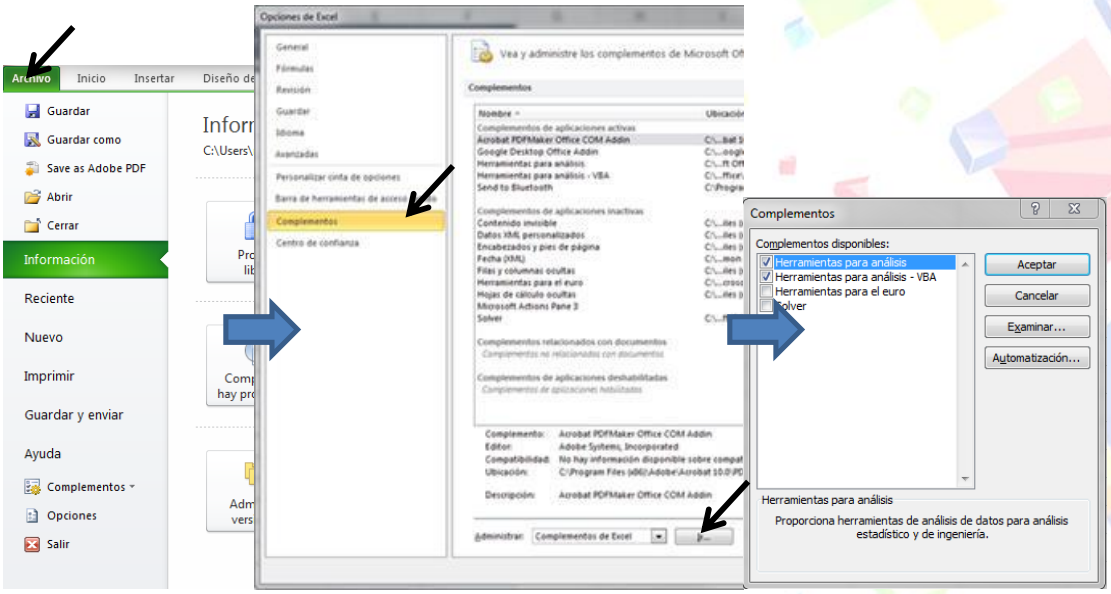



**Figura 4. Cómo activar las herramientas de análisis estadístico en Excel.**

Una vez tenemos estas herramientas activadas, encontraremos un nuevo botón en la barra de herramientas, en la pestaña "Datos", en la parte superior derecha que dirá "Análisis de Datos".

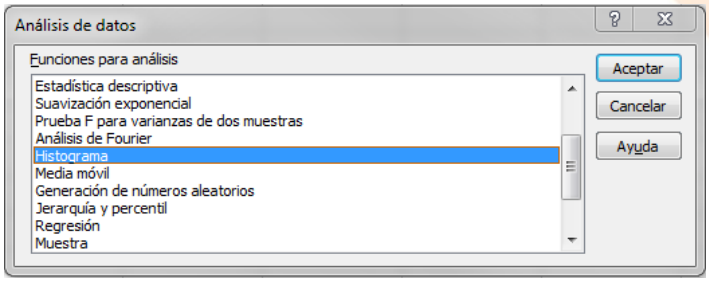



**Figura 5. Opciones de "Análisis de datos" en Excel.**

A través de ese botón, accederemos a un menú (Fig. 5) desde donde podremos seleccionar la opción "Histograma" (echa un vistazo al resto de pruebas que hay, trataremos algunas de ellas más adelante) y nos aparecerá el cuadro de diálogo de la Fig. 6. Indicaremos el "Rango de entrada" que será la fila de datos que queremos representar. Si elegimos también la celda donde está el rótulo o nombre de la serie, debemos marcar la opción "Rótulos". También podremos indicar el "Rango de clases" para que Excel agrupe mis datos de partida en unos intervalos determinados. Si no indico nada, Excel tomará la serie de intervalos que él estime. En las "Opciones de salida" podemos indicar dónde queremos que se muestren los resultados. Si elijo "Rango de salida" sólo tendré que indicar una celda a partir de la cual se generará el gráfico. Además puedo pedirle a Excel que genere un gráfico marcando "Crear gráfico" o que me haga un resumen del porcentaje acumulado.

En el archivo de ejercicios de Excel[2], en la pestaña "Ej. 1 y 2 – Est. Descriptiva" encontrarás dos tablas de datos meteorológicos de Albacete (arriba) y de Ávila (abajo). Crearemos el gráfico a partir de los 24 datos de, por ejemplo, presión registrada en Ávila durante unas 12 horas (Fig. 7). Por qué no intentas hacer lo mismo con los 50 datos de pulsaciones por minuto de la Fig. 2.

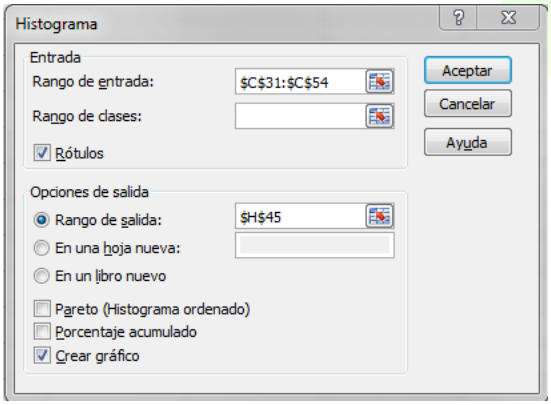



**Figura 6. Opciones para realizar un histograma de frecuencias en Excel.**

| *Clase* | *Frecuencia* |
|---|---|
| 922,5 | 1 |
| 923,9 | 7 |
| 925,3 | 11 |
| 926,7 | 3 |
| y mayor... | 2 |

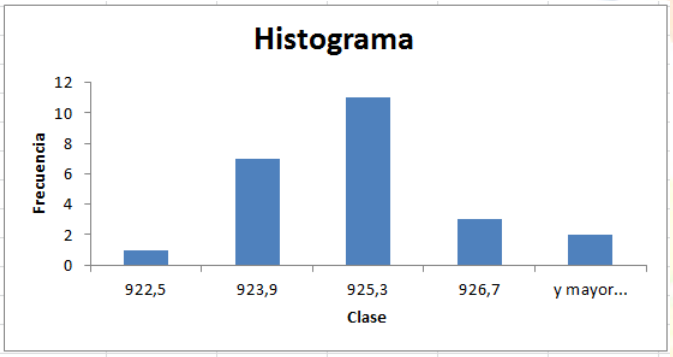



**Figura 7. Histograma de frecuencias realizado con Excel.**

Veamos ahora cómo se realiza esto mismo en SPSS (no se proporciona archivo, deberás ir haciéndolo tú mismo). Pasamos nuestra columna de datos a SPSS (copiar y pegar) y vamos al menú "Analizar", "Estadísticos descriptivos", "Frecuencias" y accederemos al menú de la Fig. 8.

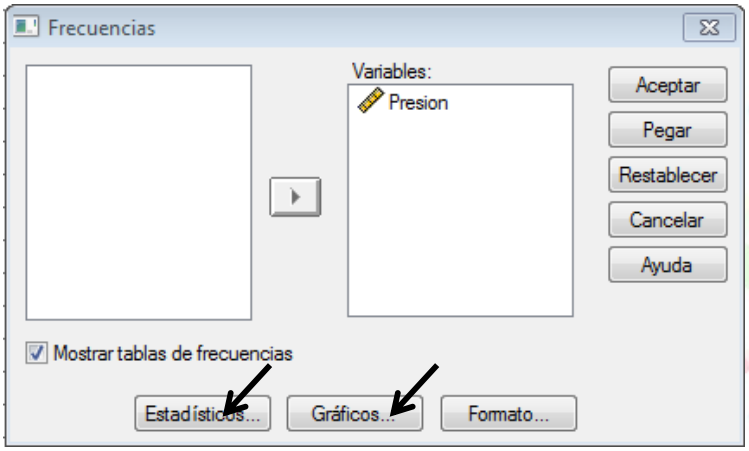



**Figura 8. Menú de opciones para elaborar un histograma de frecuencias en SPSS.**

[2] Te recuerdo que puedes descargarlo desde: http://goo.gl/UEd2pg

Selecciono la variable en el listado de la izquierda y la paso a la derecha. En esta ventana también podré, como veremos más adelante, pedirle a SPSS que nos calcule algunos estadísticos descriptivos a través del botón "Estadísticos…". En el botón "Gráficos…", cuyas opciones se muestran en la Fig. 9, podremos indicar el tipo de gráfico que deseamos.

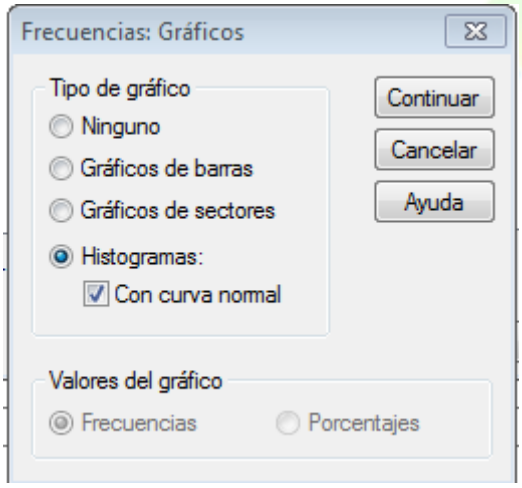



**Figura 9. Opciones de gráficos de frecuencias en SPSS.**

Marcaremos la opción "Histograma" y podremos también pedirle que incluya la curva normal (veremos qué es más adelante). Tras darle a aceptar, obtendremos la tabla de frecuencias de los datos y el gráfico que se muestra en la Fig. 10. Las diferencias entre el gráfico de Excel (Fig. 7) y el de SPSS ( ig. 7) se deben a los diferentes intervalos seleccionados.

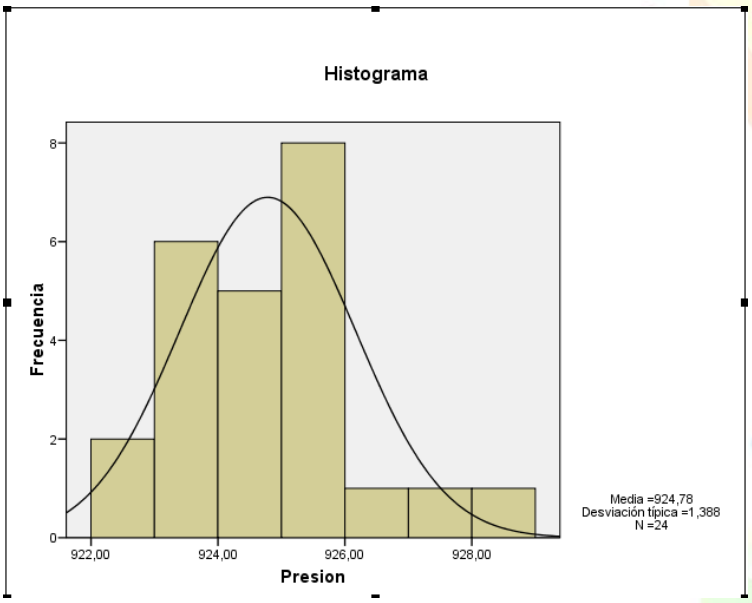



**Figura 10. Gráfico de frecuencias elaborado con SPSS que muestra la curva normal.**

## Estadística Descriptiva

Mediante el uso de estadísticos descriptivos como la media o la desviación típica que supongo familiares para todos, podremos resumir nuestros datos de manera que serán más fácilmente interpretables. Como se ha indicado y se puede comprobar en la Fig. 3, la mitad de los datos se

agrupan entre 75 y 84, muestran una tendencia central, no muestran una gran dispersión. El conjunto de estadísticos descriptivos como la media, la moda, la mediana, el máximo, el mínimo, el rango, la desviación típica, el rango, asimetría, curtosis, es fundamental para caracterizar nuestro conjunto de datos.

Tanto con Excel como con SPSS, calcular la Estadística descriptiva es sumamente sencillo. En Excel debemos ir a la pestaña "Datos", como en el caso anterior, "Análisis de datos" y elegir "Estadística descriptiva"; se mostrará el cuadro de diálogo de la Fig. 11.

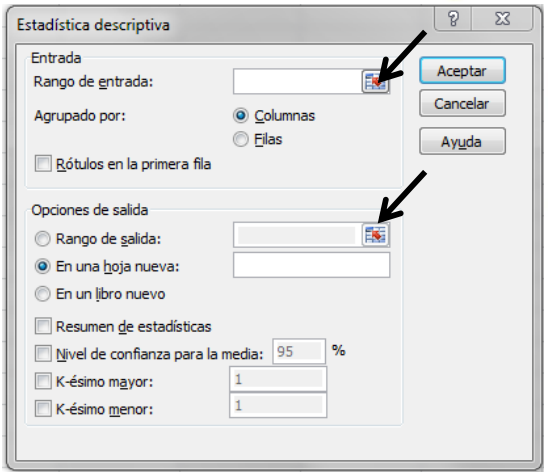



**Figura 11. Cuadro de diálogo con las opciones para realizar la Estadística descriptiva en Excel.**

Las opciones son similares a las disponibles para la inserción de un histograma. Indicaremos el "Rango de entrada", si elegimos los "rótulos en la primera fila" habrá que indicarlo, seleccionaremos también el "Rango de salida" y marcaremos el "Resumen de Estadísticas". Obtenemos el resultado que se muestra en la Fig. 12.

| *Tem (ºC)* | | *Racha vien. (km/h)* | | *Precipit (mm)* | | *Presión (hPa)* | | *Humedad (%)* | |
|---|---|---|---|---|---|---|---|---|---|
| Media | 6,2125 | Media | 31,875 | Media | 0,2208333 | Media | 924,78333 | Media | 87,875 |
| Error típico | 0,2083678 | Error típico | 0,8412217 | Error típico | 0,0892763 | Error típico | 0,2833333 | Error típico | 1,15205 |
| Mediana | 6,6 | Mediana | 32 | Mediana | 0 | Mediana | 924,85 | Mediana | 85,5 |
| Moda | 6,9 | Moda | 32 | Moda | 0 | Moda | 925,2 | Moda | 84 |
| Desviación es | 1,0207893 | Desviación es | 4,121128 | Desviación es | 0,4373628 | Desviación es | 1,3880442 | Desviación es | 5,6438693 |
| Varianza de l | 1,0420109 | Varianza de l | 16,983696 | Varianza de l | 0,1912862 | Varianza de l | 1,9266667 | Varianza de l | 31,853261 |
| Curtosis | 2,8184667 | Curtosis | -0,530984 | Curtosis | 2,5307629 | Curtosis | 0,5256345 | Curtosis | -0,91111 |
| Coeficiente d | -1,677078 | Coeficiente d | 0,4510045 | Coeficiente d | 1,955826 | Coeficiente d | 0,6322774 | Coeficiente d | 0,3993861 |
| Rango | 4,1 | Rango | 15 | Rango | 1,4 | Rango | 5,6 | Rango | 21 |
| Mínimo | 3,1 | Mínimo | 26 | Mínimo | 0 | Mínimo | 922,5 | Mínimo | 78 |
| Máximo | 7,2 | Máximo | 41 | Máximo | 1,4 | Máximo | 928,1 | Máximo | 99 |
| Suma | 149,1 | Suma | 765 | Suma | 5,3 | Suma | 22194,8 | Suma | 2109 |
| Cuenta | 24 | Cuenta | 24 | Cuenta | 24 | Cuenta | 24 | Cuenta | 24 |

**Figura 12. Estadística descriptiva tal cual la muestra Excel.**

Un "fallo" molesto de Excel es la forma que tiene de ordenar los resultados. En la Fig. 13 se muestra el resultado tras ordenar las etiquetas, los datos, etc. Lo que he hecho ha sido reordenar los encabezados, borrar los que se repetían, ordenador los resultados un poco… buscando un aspecto más adecuado.

| | Tem (ºC) | Racha vien. (km/h) | Precipit (mm) | Presión (hPa) | Humedad (%) |
|---|---|---|---|---|---|
| Media | 6,2125 | 31,875 | 0,220833333 | 924,7833333 | 87,875 |
| Error típico | 0,2083678 | 0,841221722 | 0,089276311 | 0,283333333 | 1,15205 |
| Mediana | 6,6 | 32 | 0 | 924,85 | 85,5 |
| Moda | 6,9 | 32 | 0 | 925,2 | 84 |
| Desviación es | 1,0207893 | 4,121127959 | 0,437362815 | 1,388044188 | 5,643869317 |
| Varianza de l | 1,0420109 | 16,98369565 | 0,191286232 | 1,926666667 | 31,85326087 |
| Curtosis | 2,8184667 | -0,530984428 | 2,530762857 | 0,525634458 | -0,911110046 |
| Coeficiente d | -1,677078 | 0,451004504 | 1,955825981 | 0,632277426 | 0,399386149 |
| Rango | 4,1 | 15 | 1,4 | 5,6 | 21 |
| Mínimo | 3,1 | 26 | 0 | 922,5 | 78 |
| Máximo | 7,2 | 41 | 1,4 | 928,1 | 99 |
| Suma | 149,1 | 765 | 5,3 | 22194,8 | 2109 |
| Cuenta | 24 | 24 | 24 | 24 | 24 |

**Figura 13. Estadística descriptiva realizada con Excel tras retocar el aspecto.**

No creo necesario explicar la interpretación de todos estos estadísticos descriptivos. Si acaso, es posible que la curtosis o el coeficiente de asimetría no sean demasiado famosos, pero veremos su importancia y utilidad en el siguiente apartado. Estos resultados no incluyen cuartiles ni percentiles o el rango intercuartílico que nos indicarán los valores que permitirían clasificar nuestro conjunto de datos en grupos. Por ejemplo el primer cuartil será el valor de nuestra variable que deja por debajo de él el 25% de los datos y que coincidirá con el percentil 25. De similar manera el segundo cuartil coincidirá con la mediana y dejará un 50% de las observaciones por encima y el otro 50% por debajo. Otra medida de dispersión útil es el rango intercuartílico que nos indicará la anchura del 50% de los datos que quedan en el centro de la muestra (alrededor de la media o, lo que es lo mismo, entre el cuartil 1 y 3 o igualmente entre los percentiles 25 y 75. Podremos calcular los cuartiles en Excel utilizando la función "**=CUARTIL(matriz de datos; cuartil deseado (0 a 4))".** Calcularemos el primer y el tercer cuartil y después restaremos ambos valores, así conocemos el rango del 50% de los datos y nos servirá para determinar las condiciones para poder considerar y desestimar los "puntos atípicos". Estos puntos serán aquellos que quedan fuera del intervalo [Q1-1,5·RI , Q3+1,5·RI]. Seguro que alguna vez has eliminado datos que se salían de "lo normal" con un criterio objetivo "porque sí" o "porque afean el resultado". Con este criterio, podrás utilizar algo realmente objetivo y eliminar sin problema los datos que queden fuera de ese intervalo.

Mediante la función "**=PERCENTIL(matriz de datos; percentil deseado entre 0 y 1)"** podremos calcular los diferentes percentiles que deseemos. El percentil deseado se indicará entre 0 y 1 de manera que el percentil 75 se calculará indicando como percentil deseado 0,75.

Los estadísticos Asimetría y Curtosis nos informan sobre la forma de los datos (si son simétricos y si siguen una forma acampanada más o menos alta, respectivamente). Si estos estadísticos se encuentran en torno a ±0,5, la distribución de frecuencias de nuestros datos podría ser aproximadamente "normal", que más adelante explicaremos qué quiere decir y cómo realizar un test más apropiado y objetivo.

En nuestro ejemplo, ¿qué variables presentan una dispersión mayor? No podemos saberlo comparando las desviaciones típicas pues cada variable utiliza unidades diferentes. Un

estadígrafo útil para comparar dispersiones de variables con diferentes unidades es el Coeficiente de Variación, que calcularemos como el cociente entre la desviación típica y la media expresado en tanto por ciento: 100 · S/M. Por ejemplo si hacemos esto para la temperatura y la racha de viento que tienen desviaciones típicas de 1,46 y 4,12 respectivamente, obtenemos unos coeficientes de variación de 24,5% y 12,9%, por tanto a pesar de que la racha de viento tiene una desviación típica mayor, está menos dispersa que la temperatura.

En SPSS tenemos dos maneras de calcular los estadísticos descriptivos. Ambos métodos están en "Analizar" → "Estadísticos descriptivos" → "Descriptivos" o "Frecuencias". Vamos a calcularlos primeramente mediante la segunda opción "Frecuencias", que nos permitió también realizar e histograma de la Fig. 10. En el cuadro de diálogo (Fig. 8) podremos indicar a través del botón "Estadísticos…" cuáles, de todos los disponibles, queremos que SPSS nos calcule. Mediante la primera opción, "Descriptivos", accederemos a las opciones que se muestran en la Fig. 14 (izquierda) desde donde podremos elegir las series que queremos analizar. A través del botón "Opciones" podremos acceder al cuadro de diálogo que se muestra en la Fig. 14 (derecha) y elegir los estadísticos descriptivos que queremos calcular.

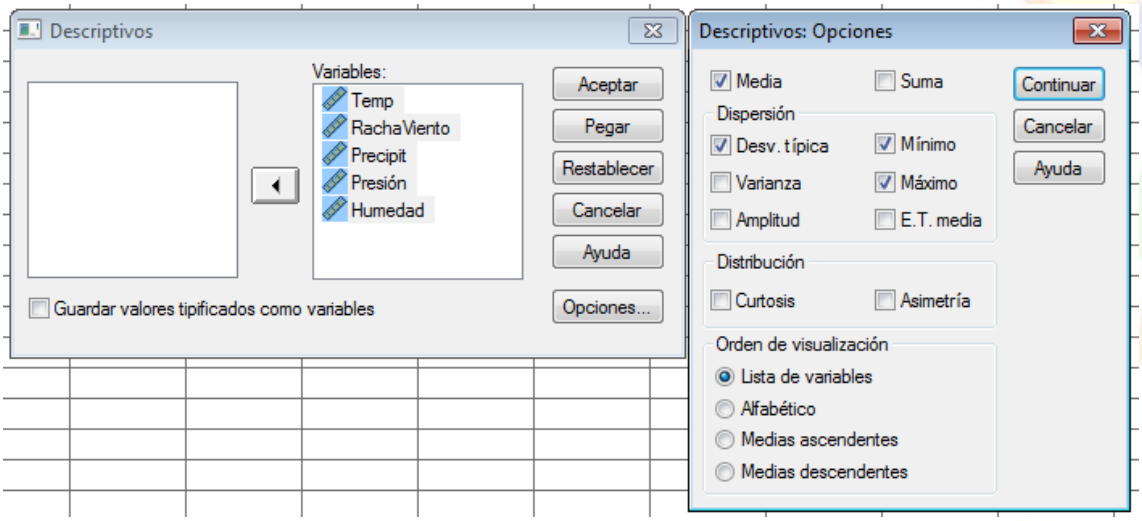



**Figura 14. Ventanas de opciones en SPSS para calcular la estadística descriptiva.**

Una vez aceptadas ambas ventanas, SPSS ofrece los resultados mostrados en la Fig. 15.

**Estadísticos descriptivos**

| | N | Mínimo | Máximo | Media | Desv. típ. | Asimetría | | Curtosis | |
|---|---|---|---|---|---|---|---|---|---|
| | Estadístico | Estadístico | Estadístico | Estadístico | Estadístico | Estadístico | Error típico | Estadístico | Error típico |
| Temp | 24 | 3,10 | 7,20 | 6,2125 | 1,02079 | -1,677 | ,472 | 2,818 | ,918 |
| RachaViento | 24 | 26,00 | 41,00 | 31,8750 | 4,12113 | ,451 | ,472 | -,531 | ,918 |
| Precipit | 24 | ,00 | 1,40 | ,2208 | ,43736 | 1,956 | ,472 | 2,531 | ,918 |
| Presión | 24 | 922,50 | 928,10 | 924,7833 | 1,38804 | ,632 | ,472 | ,526 | ,918 |
| Humedad | 24 | 78,00 | 99,00 | 87,8750 | 5,64387 | ,399 | ,472 | -,911 | ,918 |
| N válido (según lista) | 24 | | | | | | | | |

**Figura 15. Estadística descriptiva calculada mediante SPSS.**

Será importante caracterizar convenientemente nuestra muestra mediante los estadísticos descriptivos. Es una vía rápida de analizar nuestros datos en busca de posibles anomalías.

Otra opción interesante de SPSS es la elaboración de "Tablas de Contingencia" de manera que si tenemos los datos en crudo, esto es, cada caso con sus diferentes variables y valores, SPSS nos hará una tabla resumen de proporciones (Mira el apartado de la Prueba $\chi^2$).

## La Distribución Normal

Analizando los histogramas de frecuencias, éstos presentan una determinada forma; los datos se reparten por el gráfico de una manera particular. Bien, pues supongamos que realizamos un experimento con 50 ratones a los que sometemos a una determinada dieta. Queremos saber cuántos meses viven de media y si puedo extraer alguna información sobre toda la población. Para ello registramos el tiempo que vive cada uno de ellos. Pues imagina que al hacer esta representación, comprobamos que nuestro histograma de frecuencias se parece mucho al de la Fig. 16. Además, realizamos la estadística descriptivas y comprobamos que la media y la desviación típica y de nuestra muestra son 40 y 6,3 meses, respectivamente.

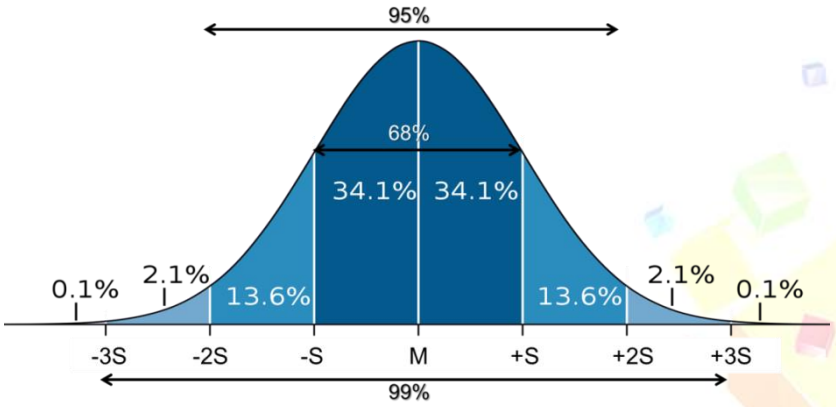



**Figura 16. La Distribución Normal.**

Esta forma particular de campana tiene unas características y propiedades muy interesantes y se denomina "Distribución Normal" o "curva Gaussiana" (ahora veremos cómo comprobar si nuestros datos se adaptan a esta forma no sólo cualitativamente sino también cuantitativamente). Entre otras propiedades, esta distribución cumple que el 68% de las observaciones se distribuyen en torno a la media más/menos una desviación típica, que es una propiedad sumamente importante. Además sabemos que el 95% de los datos están en torno a la media más/menos 1,96 desviaciones típicas y que el 99% están en torno a la media más/menos 2,58 desviaciones típicas[3].

Con esto en la cabeza, ¿sería probable que un ratón viviera menos de 32 meses? ¿Y más de 55 meses? Pues sin hacer cuentas, podemos ver que si la vida media es de 40 meses y la desviación típica es de 6,4, el 68% de los datos estarán en el intervalo 33,6 y 46,4. Por lo que podemos suponer que existe una probabilidad de 0,68 (recuerda que las probabilidades, al contrario de lo que estamos acostumbrados, se expresan en tanto por uno y no en porcentaje) de que al criar un ratón alimentado con nuestra dieta (controlando otros factores), éste viva

[3] Los valores de 1,96 y 2,54 provienen de los valores de la variable z, variable tipificada, que proporcionan las áreas de 0,95 y 0,99 en torno a la media de la distribución normal tipificada de media 0 y desviación típica 1.

entre 33,6 y 46,4 meses. Por debajo del valor 33,6 y por encima de 46,4, puesto que la curva es simétrica, tendremos el 32% de los datos, 16% por debajo de 33,6 y 16% por encima de 46,4, por lo que vivir menos de 32 meses tendrá una probabilidad inferior al 0,16 y el 16% de nuestros ratones vivirán más de 46,4 meses. El 95% de los ratones vivirán entre 27,5 y 52,5 meses (que será el valor medio más/menos 1,96 desviaciones típicas), por lo que sólo el 2,5% (recuerda que la curva es simétrica, tenemos otro 2,5% por debajo de 27,5) vivirán más de 52,5 meses. De manera similar, tener un ratón que viva más de 55 meses será bastante improbable (menor a 0,025).

Excel permite calcular directamente el área (y por tanto directamente la probabilidad) de la curva normal para un valor medio y una desviación típica dadas. Para ello usaremos la función **"=DISTR.NORM.N(x;media;desviación;verdadero)**" de manera que "x" es el valor para el cual queremos calcular el área de la curva desde -∞ hasta ese valor dado. La función siempre proporciona el área acumulada. Debemos recordar/saber que el área total vale 1 (desde -∞ a +∞ es 1), y por tanto desde -∞ al valor de la media, el área es 0,5. Necesitaremos indicar la media y la desviación típica de nuestro ejemplo y poner "verdadero" para que nos proporcione el área acumulada. Haciendo esto, obtenemos un área hasta 32 meses de 0,11 y para 55 meses de 0,99. Por tanto la probabilidad de que un ratón viva menos de 32 meses será de 0,11 y de que viva más de 55 meses será de 0,01 (recuerda que la curva es simétrica y ahora calculamos los valores superiores a 55 meses, por lo que la probabilidad será el área total menos la probabilidad acumulada hasta el valor de 55 meses, esto es 1-0,99).

El teorema central del límite establece que "*en condiciones muy generales, si Sn es la suma de **n variables aleatorias independientes**, entonces la función de distribución de Sn «se aproxima bien» a una distribución normal, distribución gaussiana, curva de Gauss o campana de Gauss*"[4]. Así, si nuestra distribución fuera normal podríamos calcular cuál es la probabilidad de que ocurra un determinado suceso ¿pero cómo sé de forma precisa si mis datos son normales o no? Como se indicó podemos aproximar una idea analizando los valores de asimetría y curtosis, pues la curva normal es simétrica y mesocúrtica, pero veremos un par de métodos mejores. Sigue leyendo.

## Test de normalidad Kolmogorov-Smirnov y Shapiro-Wilk

Como hemos indicado, mediante los valores de asimetría y curtosis podemos realizar una primera estimación de si nuestros datos se ajustan o no a una distribución normal o curva de Gauss: ambos deben estar en torno a ±0,5. Si así fuera, podremos aprovechar las propiedades de la curva normal y realizar predicciones, cálculos de probabilidades, etc.

Veamos cómo comprobar estadísticamente si nuestros datos realmente se distribuyen o no de manera normal. Para ello recurrimos al test Kolmogorov-Smirnov (KS) o al de Shapiro-Wilk (SW) que nos indicarán si se cumple o no esta hipótesis de normalidad y que en SPSS se puede hacer de dos formas diferentes, no así en Excel. **El test KS es adecuado para muestras de más de 50 datos, en caso contrario, lo recomendable es utilizar el test SW**.

[4] http://es.wikipedia.org/wiki/Teorema_central_del_límite

Los test estadísticos nos informan de si una hipótesis planteada es válida o no. Seguro que has oído hablar de las dichosas hipótesis nula (H0) y alternativa (H1). En el caso de estos dos test KS y SW, la hipótesis nula es que la muestra se distribuye normalmente. El test en SPSS proporcionará un valor de significación, "Sig." o p-valor para un intervalo de confianza que por defecto será del 95% (p=0,05). Pues el criterio general siempre será en SPSS que **se aceptará la hipótesis nula si el valor de "Sig." es mayor que el valor de p que fijamos**.

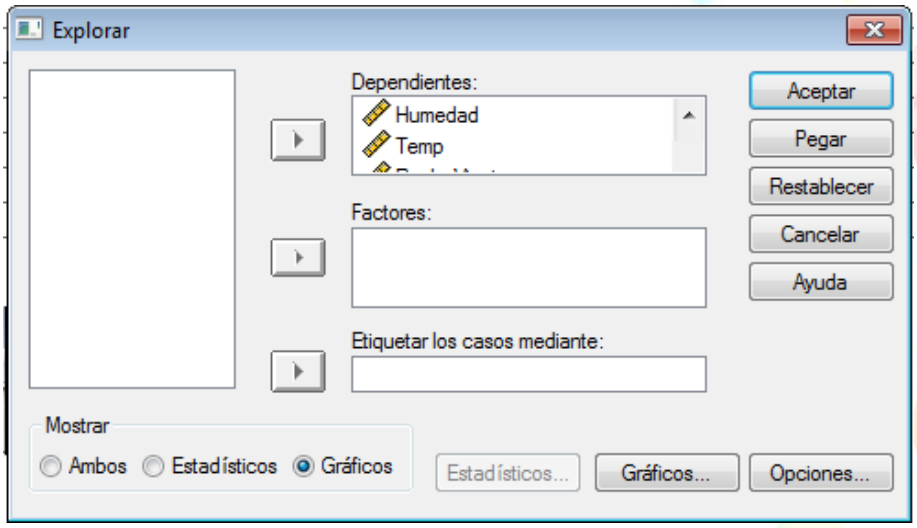



**Figura 17. Cuadro de Diálogo "Explorar" en SPSS para la realización de test de normalidad.**

Para realizar el test KS deberemos ir a "Analizar" → "Estadísticos descriptivos" → "Explorar…" (Fig. 17). Debemos ir al botón "Gráficos" y seleccionar "Gráficos con prueba de normalidad" y, tras añadir la/s variable/s a estudiar, daremos a "Aceptar". Obtenemos, entre otras cosas, la tabla que se muestra en la Fig. 18.

**Pruebas de normalidad**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Estadístico | gl | Sig. | Estadístico | gl | Sig. |
| Humedad | ,212 | 24 | ,007 | ,924 | 24 | ,071 |
| Temp | ,194 | 24 | ,020 | ,804 | 24 | ,000 |
| RachaViento | ,155 | 24 | ,143 | ,949 | 24 | ,262 |
| Precipit | ,402 | 24 | ,000 | ,571 | 24 | ,000 |
| Presión | ,147 | 24 | ,198 | ,954 | 24 | ,337 |

a. Corrección de la significación de Lilliefors

**Figura 18. Pruebas de normalidad Kolmogorov-Smirnov y Shapiro-Wilk en SPSS.**

En el ejemplo que estamos utilizando, tenemos 24 datos por lo que debemos atender a los resultados del test SW. En la columna "Sig." nos aparece el valor para confirmar o no la normalidad de los datos para un intervalo de confianza del 95% (que veremos más adelante exactamente qué quiere decir). El valor de "Sig." debe ser mayor a 0,05 para quedarnos con la hipótesis nula (normalidad), en caso contrario, deberemos quedarnos con la hipótesis alternativa. Esta condición la cumple la variable Humedad, Racha Viento y Presión, por tanto estas tres variables sí son normales. En el caso de la Temp y la Precipit el valor de "Sig." es inferior a 0,05 y por tanto ambas variables no son normales. Si miramos la columna de la "Sig."

de la prueba KS, comprobaremos que hay diferencias con respecto a la prueba SW, debido a que el tamaño de la muestra es inferior a 50.

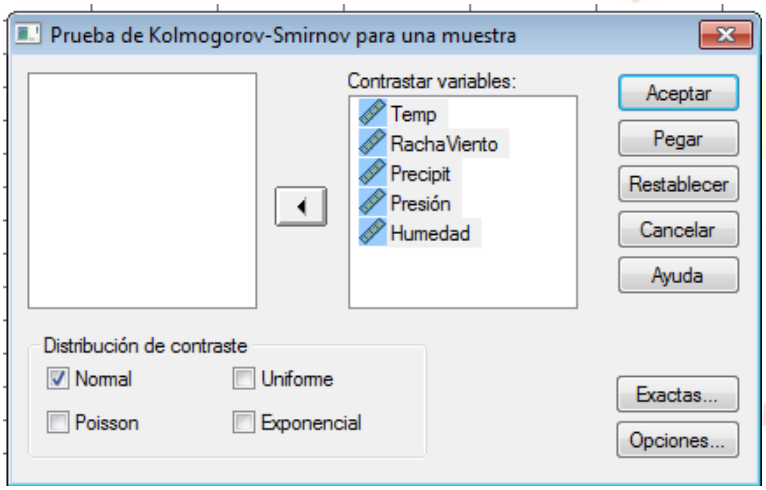



**Figura 19. Opciones para realizar una prueba Kolmogorov-Smirnov en SPSS.**

La otra forma de realizar esta comprobación es ir a "Analizar" → "Pruebas no paramétricas" → "K-S para una muestra". Así accederemos a las opciones que se muestran en la Fig. 19. Indicamos las series y le damos a Aceptar. El resultado de la prueba se muestra en la Fig. 20. De nuevo debemos atender a los valores de "Sig." aquellos que sean mayores a la p fijada (0,05) cumplirán la hipótesis nula (Ho) y por tanto serán normales (p=0,05) mientras que aquellos que sean menores no cumplirán H0 y por tanto debemos asumir la hipótesis alternativa (H1), esto es, no son normales.

**Prueba de Kolmogorov-Smirnov para una muestra**

| | | Humedad | Temp | RachaViento | Precipit | Presión |
|---|---|---|---|---|---|---|
| N | | 24 | 24 | 24 | 24 | 24 |
| Parámetros normales[a,b] | Media | 87,8750 | 6,2125 | 31,8750 | ,2208 | 924,7833 |
| | Desviación típica | 5,64387 | 1,02079 | 4,12113 | ,43736 | 1,38804 |
| Diferencias más extremas | Absoluta | ,212 | ,194 | ,155 | ,402 | ,147 |
| | Positiva | ,212 | ,179 | ,155 | ,402 | ,147 |
| | Negativa | -,111 | -,194 | -,102 | -,307 | -,062 |
| Z de Kolmogorov-Smirnov | | 1,039 | ,952 | ,757 | 1,967 | ,718 |
| Sig. asintót. (bilateral) | | ,230 | ,326 | ,615 | ,001 | ,681 |

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

**Figura 20. Resultado de la prueba KS a través de "Pruebas no paramétricas" en SPSS.**

De nuevo, las diferencias que observamos, probablemente se deban a tener una muestra inferior a 50 datos.

## Estimación del tamaño de la muestra

Hagamos un ejercicio mental. Imagina que medimos una determinada variable de todos los miembros de una población que se distribuye de manera normal y determinamos que su media vale $\mu$ y su desviación típica $\sigma$. Ahora tomamos una muestra y medimos esa variable

obteniendo una media, $M_1$, y una desviación típica, $S_1$. Si repetimos el proceso n veces, podremos construir un conjunto de n valores de M y S, de manera que unas veces M será mayor que $\mu$ y otras veces será menor; también podremos encontrar muestras en las que $M=\mu$. Imagina que tomamos todas las medias de todas esas muestras y calculamos su media (la media de las medias). Si el conjunto es suficientemente grande, esa media coincidirá con la media de la población $\mu$. Pues bien, cada una de las muestras será normal y formará parte de una nueva distribución de medias que denominaremos Distribución de las Medias Muestrales (DMM), también normal y cuya media será igual a la media de la población, $\mu$. Además se puede demostrar que su desviación típica será $\sigma/\sqrt{N}$ (que se denomina error típico, SE), donde N es el tamaño de mi muestra. Por tanto, cuando yo tomo una muestra de la población y calculo la media M, sabemos que pertenece a esa teórica DMM($\mu$, $\sigma/\sqrt{N}$) y, puesto que esa DMM es normal, tendremos que ese valor M tendrá una probabilidad de 0,68 de estar en torno a la media de la población ($\mu$) más/menos un error típico ($\sigma/\sqrt{N}$). De la misma manera podremos calcular los intervalos (de confianza) del 95% y del 99%, de manera que sabemos que la media de mi muestra tendrá un valor entre: $\mu - 1\cdot SE \leq M \leq \mu + 1\cdot SE$ que es lo mismo que $M - 1\cdot SE \leq \mu \leq M + 1\cdot SE$, por lo que podemos estimar el valor de la media de la población ($\mu$) a partir de la media de nuestra muestra (M) con un error definido por el error típico (SE). Además, con un 0,95 de probabilidades (recuerda que la probabilidad se expresa en tanto por uno) sabemos que $M - 1{,}96\cdot SE \leq \mu \leq M + 1{,}96\cdot SE$ y que con una probabilidad de 0,99tendremos que $M - 2{,}54\cdot SE \leq \mu \leq M + 2{,}54\cdot SE$.

Pero ¿de dónde sacamos $\sigma$? Pues se puede demostrar que para muestra superiores a 30, podemos aproximar $\sigma$ con nuestra S, así $SE=S/\sqrt{N}$.

Recordando el caso de los ratones (M=40 meses, S=4,6 meses y N=50 ratones), podríamos estimar el valor medio de la población, de manera que:

- Para un 95% → $M - 1{,}96\cdot SE \leq \mu \leq M + 1{,}96\cdot SE$ → $\mu \in [38{,}26, 41{,}74]$
- Para un 99% → $M - 2{,}54\cdot SE \leq \mu \leq M + 2{,}54\cdot SE$ → $\mu \in [37{,}74, 42{,}26]$

Si queremos reducir el error de predicción, o bien reducimos el "intervalo de confianza" o bien incrementamos el tamaño de la muestra. El problema es que la dependencia de SE con N es con el inverso de una raíz cuadrada, de manera que incrementar N mucho, supone reducir SE poco. Si incrementamos N hasta los 100 ratones, obtenemos que:

- Para un 95% → $M - 1{,}96\cdot SE \leq \mu \leq M + 1{,}96\cdot SE$ → $\mu \in [39{,}10, 40{,}90]$
- Para un 99% → $M - 2{,}54\cdot SE \leq \mu \leq M + 2{,}54\cdot SE$ → $\mu \in [38{,}83, 41{,}17]$

Duplicar el tamaño de la muestra mejora la determinación del valor de $\mu$, pero ¿a qué coste y con qué beneficio? Pues esa es la pregunta que debe contestar el investigador. Puedes comprobar qué precisión se consigue incrementando la muestra a 1000 ratones… ¿merecería la pena?

Si quieres, puedes avanzar hasta el "test t para una sola muestra" que se introduce más adelante y que permite comprobar la hipótesis de que la media de la población tenga un valor determinado, por ejemplo saber si el valor M de nuestra muestra podría ser la media de la población $\mu$, en esto consiste la Inferencia Estadística.

## Comparación de muestras

Lo que veremos a continuación será la manera de determinar si una diferencia que nosotros observamos entre dos muestras, por ejemplo al comparar los valores de sus medias, se debe al muestreo (al azar) o es estadísticamente significativa y con qué intervalo de confianza o nivel de significación, p.

Supongamos que medimos la presión sanguínea a 50 varones y a 50 mujeres, todos ellos de 25 años de edad y obtenemos:

- La media de la presión sanguínea de los varones es 120 mmHg (S=11,3 mmHg)
- La media de la presión sanguínea de las mujeres es de 110 mmHg (S=11,3 mmHg)

Observando las medias, claramente los hombres tienen de media 10 mmHg de presión sanguínea superior a la de las mujeres. Esto lo pilla un periodista y ya tiene noticia, pero ¿es esto real o se debe al muestreo? ¿Puede ser que por casualidad he cogido a las mujeres con la presión más baja? Veamos cómo podemos hacer un primer análisis sencillo. De la misma manera que hacíamos para el caso de los ratones, podemos estimar la media de la población a partir de ambas muestras, obteniendo:

| Muestra (50 datos) | M (mm Hg) | S (mm Hg) | SE=S/√N (mm Hg) | μ (mm Hg) Certeza del 99% |
|---|---|---|---|---|
| Varones | 120 | 11.3 | 1.6 | [116, 124] |
| Mujeres | 110 | 11.3 | 1.6 | [106, 114] |

Si suponemos que ambas muestras provienen de la misma población (hipótesis nula en nuestro caso), al estimar los intervalos donde debería estar la media de la población, éstos deberían coincidir. No obstante, con un intervalo de confianza del 99% (recuerda que en este caso la media de la población estará en torno a la media de la muestra más/menos 2,58 errores típicos) obtenemos que los dos intervalos estimados donde se encontraría la media de la población no coinciden, ni se tocan.

Por tanto, podemos suponer que es improbable que esas dos muestras provengan de la misma población y debemos rechazar H0, o lo que es lo mismo, es muy improbable que esa diferencia que yo veo se deba al muestreo, es estadísticamente significativa, y por tanto que ambas muestras sean iguales. Dicho de otro modo, la presión media entre hombres y mujeres de 25 años difiere en 10 mm de Hg (p=0,99).

## Test estadísticos de contraste de hipótesis

Lo que hemos hecho en el apartado anterior es comparar muestras de una manera más o menos rudimentaria. A continuación introduciremos los principales test de comparación de hipótesis en Estadística. Debes tener claro cuándo deberás utilizar cada uno de ellos; como norma general, se puede resumir el cuadro de la Fig. 21.

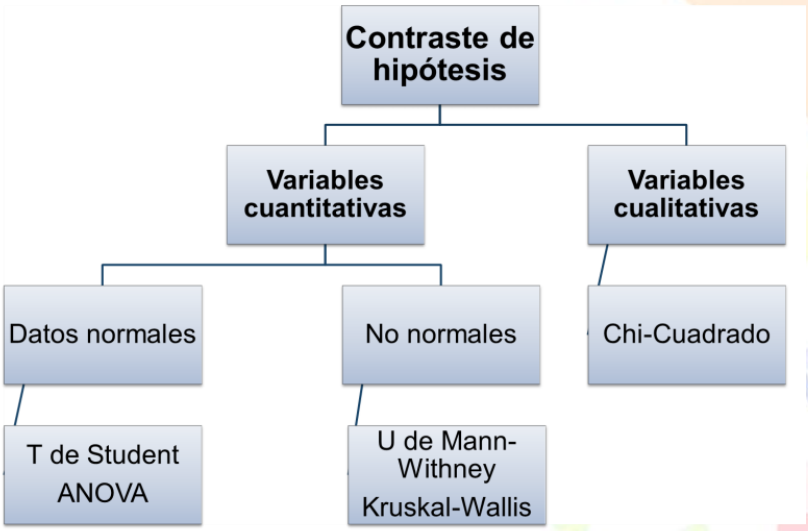



**Figura 21. Esquema test de contraste de hipótesis más frecuentes.**

## Test z

Antes de introducir los test más famosos **para variables cuantitativas normales**, comenzaremos con un caso particular, que sólo puede aplicarse cuando N>30 pero que nos ayudará a entender lo que estamos haciendo. El método del apartado anterior es aproximado, veamos un primer test para el contraste de hipótesis. Realicemos un nuevo ejercicio mental. Imagina que de la población del ejercicio mental anterior tomamos parejas de dos muestras (de medias y desviaciones típicas conocidas). La diferencia entre sus medias variará con cada pareja de muestras. Habrá casos en los que la media de la primera sea mayor que la segunda, otras veces en las que ocurra lo contrario. Incluso habrá casos en los que ambas medias coincidan y por tanto esa diferencia sea cero. Pues bien, si tomamos un conjunto muy grande de parejas de muestras y realizamos las sucesivas diferencias, podremos calcular la media de ese conjunto de diferencias de las medias de las dos muestras. Si ambas muestras provienen de la misma población (nuestra hipótesis nula), la media de la nueva Distribución Muestral de las Diferencias (DMD) constituida con ese conjunto de N diferencias de otras tantas parejas de muestras, será cero. Además se puede demostrar que su error típico ($SE_{dif}$) se puede calcular a partir de los errores típicos de ambas muestras mediante:

$$SE_{dif} = \sqrt{SE_A^2 + SE_B^2} = \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}}$$

Al igual que antes, también se cumple que para N>30 podemos estimar la desviación típica de la población, $\sigma$, por la de cada muestra. Y como es una distribución normal, tendremos que la diferencia observada entre nuestras dos muestras estará en el intervalo [$-SE_{dif}$, $SE_{dif}$] con un intervalo de confianza del 68% puesto que en este caso la media de las diferencias o valor medio de la DMD es 0. Del mismo modo, tal y como hemos venido haciendo hasta ahora, la diferencia observada estará en el intervalo [$-1{,}96 \cdot SE_{dif}$, $1{,}96 \cdot SE_{dif}$] con un intervalo de confianza del 95% y en el intervalo [$-2{,}54 \cdot SE_{dif}$, $2{,}54 \cdot SE_{dif}$] al 99%. En definitiva, si yo tomo dos muestras que supongo provienen de la misma población, con un 99% de intervalo de confianza, la diferencia entre sus medias estará en el intervalo [$-2{,}54 \cdot SE_{dif}$, $2{,}54 \cdot SE_{dif}$] y por tanto la

diferencia que observamos se debe al muestreo. Sin embargo, si la diferencia entre ambas medias estuviera fuera de este intervalo, podremos concluir que es muy improbable (99%) que al tomar dos muestras de una misma población, la diferencia de sus medias sea tan grande y por tanto ambas muestras deben provenir de poblaciones diferentes, esto es, las medias realmente son diferentes.

En este punto **podemos introducir el concepto de nivel de significación** (ya era hora), p, o p-valor. Que será complementario al intervalo de confianza pero expresado en tanto por uno, de manera que al intervalo del 95% le corresponde un p=0,05 y al del 99% un p=0,01. Cuanto menor sea p, mejor será el resultado obtenido puesto que mayor será el intervalo de confianza. En Medicina se suele trabajar al 95% (p=0,05) y hablaremos de resultado "significativo". No está de más comprobar nuestra hipótesis al 99%, o lo que es lo mismo p=0,01 y comprobar si además es "muy significativo".

En el ejemplo de la presión sanguínea de hombres y mujeres, podemos hacer cálculos. Lo que queremos estimar es cuán probable es obtener una diferencia de 10 mmHg entre dos muestras, que de partida, suponemos que provienen de la misma población. Lo primero que debemos calcular es el valor de $SE_{dif}$ = 2,3 mmHg. Podríamos ver qué área queda debajo de una curva de media 0 y desviación típica $SE_{dif}$ para un valor de 10 (en Excel con la función "**=DITRIB.NORM.N**". Obtenemos un valor de 0,999993. Recuerda que como la curva es simétrica y queremos ver diferencias de presiones de ±10 mmHg (en este caso lo que queremos comprobar es si la diferencia en sí es probable, una muestra mayor que la otra), la probabilidad de que esto ocurra es inferior a 0,000001, por tanto es extremadamente improbable que al tomar dos muestras de la misma población, se obtenga una diferencia entre sus medias tan grande. Si calculamos los intervalos:

- Para 95% → [-1,96·2,3 , 1,96·2,3] = [-4,51, 4,51] → p<0,05
- Para 99% → [-2,54·2,3 , 2,54·2,3] = [-5,84, 5,84] → p<0,01

Esto quiere decir que el 95% de las diferencias están entre los valores -4,51 y 4,51, ampliándose al intervalo de entre -5,84 y 5,84 para el 99% de los casos. Por tanto una diferencia de 10 mmHg es sumamente improbable suponiendo que ambas muestras provienen de la misma población. Concluimos, por tanto, que la diferencia observada es real, no se debe al muestro, ambas muestras provienen de poblaciones diferentes, rechazamos H0 y nos debemos quedar con la hipótesis alternativa (H1) con una p<0,01, en definitiva, los hombres de 25 años tienen una presión sanguínea 10 mmHg mayor que las mujeres (p=0,01). Recuerda que esto es sólo aplicable a muestras de variables cuantitativas continuas de más de 30 datos. El caso general es la prueba t, así que si quieres, puedes saltarte lo siguiente y avanzar al apartado correspondiente.

Veamos cómo realizar una prueba z en Excel. Para ello vamos a utilizar los datos de la segunda hoja "Ej.3 – Prueba z y t"[5]. Encontrarás dos conjuntos de datos del tiempo que han durado unos tubos transtimpánicos[6] (medido en meses). Tenemos dos tipos de tubos (T y D) y queremos comprobar cuál dura más. La duración media de los tubos T es de 23,1 y la de los

[5] Te recuerdo que puedes descargarlo desde: http://goo.gl/UEd2pg
[6] http://es.wikipedia.org/wiki/Tubo_de_timpanostom%C3%ADa

tubos D es de 8,3 meses. A simple vista parecen muy diferentes, pero debemos comprobarlo estadísticamente.

Como este test sólo se puede aplicar a muestras normales, antes debemos realizar un test de normalidad de Shapiro-Wilk pues las muestras son menores a 50 datos. Para ello copio y pego los datos en SPSS y realizo el test SW, obteniendo el resultado que se muestra en la Fig. 22.

**Pruebas de normalidad**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Estadístico | gl | Sig. | Estadístico | gl | Sig. |
| T | ,096 | 33 | ,200* | ,937 | 33 | ,054 |
| D | ,096 | 33 | ,200* | ,941 | 33 | ,075 |

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

**Figura 22. Resultado de la prueba de normalidad para los datos de duración de los tubos transtimpánicos.**

Como siempre, nuestra hipótesis nula es normalidad, como "Sig." es mayor a 0,05 comprobamos que en ambos casos son normales y por tanto podemos continuar.

Veamos cómo realizar, entonces, el test z en Excel. Debemos ir a la pestaña "Datos"→ "Análisis de datos" y seleccionaremos el test "Prueba z para medias de dos muestras", accedemos al cuadro de diálogo de la Fig. 23. Indicaremos el rango de cada una de las dos variables (no es preciso que sean del mismo tamaño), indicaremos que la "Diferencia hipotética entre las medias es "0" puesto que suponemos que realmente son iguales, e indicaremos los valores de la varianza de cada caso que habremos calculado previamente.

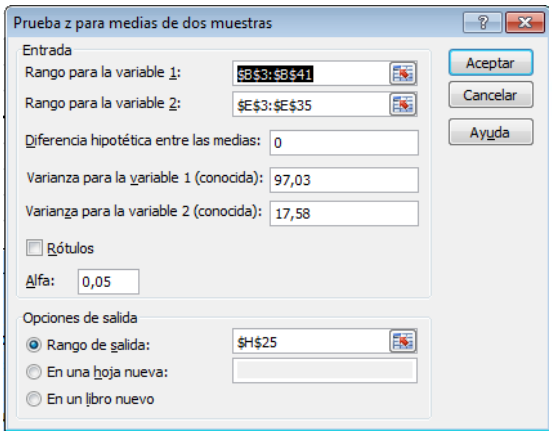



**Figura 23. Cuadro de diálogo para calcular una prueba z en Excel.**

Tras indicar un valor para "Alfa" que será nuestro nivel de significación p=0,05 y una celda "Rango de salida" donde mostrar los resultados, Excel devuelve los datos que se recogen en la Fig. 24. En las columnas inferiores se nos indica el resultado para "una cola" y para "dos colas".

Pensad en la gráfica de la distribución normal y las colas será la parte de la gráfica en cada uno de sus extremos. Un test de "una cola" marca una diferencia y un orden: que A es mayor que B o viceversa (entonces habría que calcular 1-P(Z<=z), mientras que el test de "dos colas" nos da

el valor para la diferencia en valor absoluto, da igual que A sea mayor que B o que B sea mayor que A, nos da igual, queremos ver la diferencia. Es responsabilidad del investigador decidir cuándo usar una u otra.

| | *Variable 1* | *Variable 2* |
|---|---|---|
| Media | 23,0717949 | 8,27474747 |
| Varianza (conocida) | 97,03 | 17,58 |
| Observaciones | 39 | 33 |
| Diferencia hipotética de | 0 | |
| z | 8,51379116 | |
| P(Z<=z) una cola | 0 | |
| Valor crítico de z (una co | 1,64485363 | |
| P(Z<=z) dos cola | 0 | |
| Valor crítico de z (dos co | 1,95996398 | |

**Figura 24. Resultado de la prueba z en Excel[7].**

Recuerda que **la hipótesis nula es que ambas muestras son iguales**, provienen de la misma población y que las diferencias se deben al muestreo, lo cual sería cierto si el p-valor es mayor a 0,05, lo cual no ocurre pues tenemos en ambos casos, para "una cola" y para "dos colas", un valor de 0. La interpretación en este caso es que debemos rechazar H0 y aceptar la hipótesis alternativa H1, esto es, que ambas muestras son diferentes, que la diferencia encontrada es real y estadísticamente significativa con un intervalo de confianza del 95% o un nivel de significación p=0,05. Por tanto los tubos T duran más que los tubos D (p=0,05).

## Test t

¿Pero qué pasa si N<30? Pues la distribución no es completamente normal, pero se parece mucho. Utilizamos una nueva distribución que se denomina Distribución t de Student. No entraremos en detalles sobre qué se está haciendo, qué cálculos se están haciendo, sólo indicaremos cómo realizar el análisis en Excel y en SPSS y cómo debemos interpretarlos, pero en esencia es lo mismo que hemos explicado con la construcción de la DMD y la aplicación del test z.

Vamos a repetir el caso anterior, comparar los datos de duración de los tubos T y D, pero ahora utilizaremos una prueba t. Aunque la muestra sea mayor que 30 y podríamos aplicar el test z, lo normal es que puesto que la prueba t es general se aplique en todas las condiciones, de ahí que sea más popular que la prueba z. En Excel iremos a "Datos" → "Análisis de datos". Tenemos tres posibilidades de prueba t. La primera es para muestras emparejadas que aplicaremos a una muestra en la que analizamos un estado inicial y otro posterior, ambas muestras están relacionadas (por ejemplo un grupo de células antes y después de un determinado experimento o tras someterlas a unas determinadas condiciones). En este caso ambas muestras en realidad son la misma, pero en tiempos diferentes.

[7] En la versión 2010 y en las anteriores, Excel no pone bien la etiqueta de la penúltima fila de la tabla, indicando dos veces "Valor crítico de z (dos colas). En la imagen lo he corregido.

Las otras dos opciones de prueba t son para varianzas iguales o diferentes que tomaremos dependiendo de las varianzas de nuestras muestras. En la Fig. 25 se muestra el cuadro de diálogo para la realización de esta última prueba en Excel.

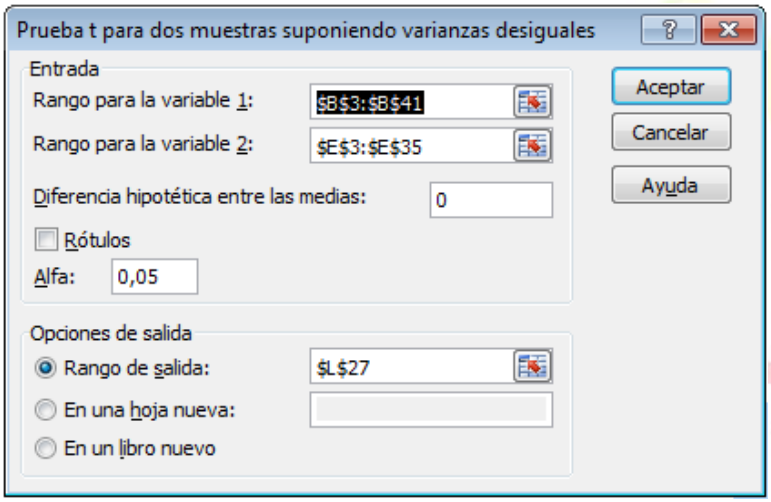



**Figura 25. Cuadro de diálogo para la realización de una prueba t en Excel.**

Los resultados de esta prueba se muestran en la Fig. 26. Podemos interpretarlos de dos maneras. La primera es atendiendo al p-valor, de manera que si éste es superior a nuestro alfa, entonces aceptaremos H0. En nuestro caso, tanto para una cola como para dos colas, ese valor es inferior a 0,05, por lo que debemos rechazar H0 y aceptar H1, esto es, que las diferencias observadas son reales, ambas muestras proceden de poblaciones diferentes, como era de esperar pues ya lo sabíamos.

La otra forma de interpretar este resultado es comprobar el valor de t y el valor de t crítico (que nos proporciona Excel y que condiciona el cumplimiento de nuestra hipótesis). Si t fuera menor que t crítico (estaría dentro del intervalo de confianza, similar al que calculábamos en la prueba z), entonces aceptaríamos H0, ambas muestras serían iguales. En nuestro caso t es mayor que t crítico y por tanto no podemos aceptar H0, una vez más, como no podía ser de otra forma, debemos concluir que ambas muestras son diferentes (p=0,05). Si obtuviéramos un valor de t negativo, esto sería porque la segunda muestra tiene una media superior a la primera, nos quedamos con el valor absoluto.

| | *Variable 1* | *Variable 2* |
|---|---|---|
| Media | 23,0717949 | 8,27474747 |
| Varianza | 97,0379847 | 17,5831965 |
| Observaciones | 39 | 33 |
| Diferencia hipotética de las media | 0 | |
| Grados de libertad | 53 | |
| Estadístico t | 8,51336616 | |
| P(T<=t) una cola | 8,5434E-12 | |
| Valor crítico de t (una cola) | 1,67411624 | |
| P(T<=t) dos colas | 1,7087E-11 | |
| Valor crítico de t (dos colas) | 2,005746 | |

**Figura 26. Resultados de la prueba t en Excel.**

Para realizar este test en SPSS, antes debemos ordenar los datos en una única serie y con una nueva variable que indique de qué grupo se trata. En el caso de los tubos, pondremos todos los datos en una columna "Tubo" y generaremos otra columna ("TipoTubo") en la que "1" se corresponderá con los tubos T y "2" con los tubos D. Iremos al menú "Analizar", "Comparar medias" y elegiremos la prueba que queremos realizar (prueba t para muestras independientes); el cuadro de diálogo es el que se muestra en la Fig. 27.

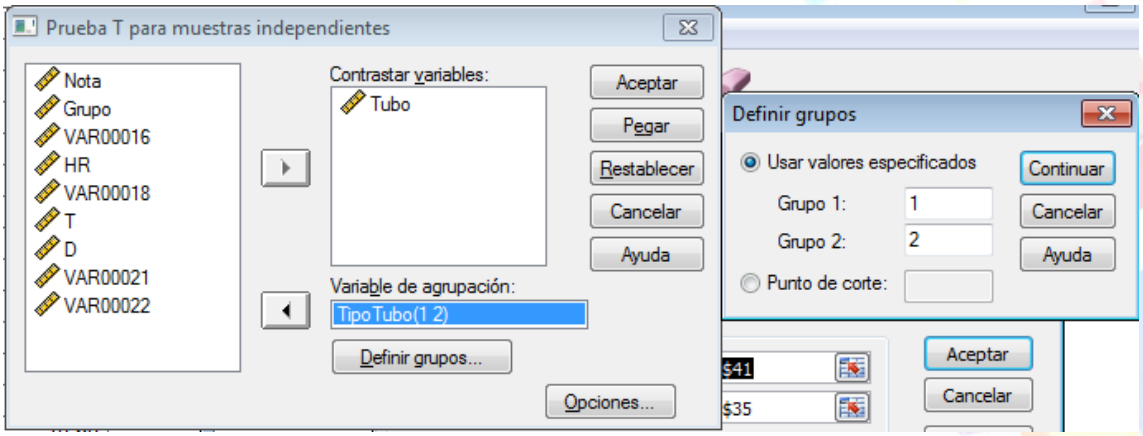



**Figura 27. Cuadros de diálogo para la realización del test t en SPSS (izquierda) y definición de grupos (derecha).**

Indicaremos la variable que utilizaremos (Tubo) y la variable de agrupación (TipoTubo). En la Fig. 27 derecha indicaremos los dos grupos que hemos definido (1-T, 2-D). En el botón "Opciones" podremos indicar el intervalo de confianza.

El resultado se muestra en la Fig. 28. En las dos primeras columnas se muestra la "Prueba de Levenne para igualdad de varianzas" miraremos el valor de "Sig." si éste es mayor que 0,05, tendremos varianzas iguales; en caso contrario, como es el caso, las varianzas serán diferentes. Por tanto debemos mirar a la última fila del análisis donde se indica "No se han asumido varianzas iguales". La hipótesis de partida en estos casos, como venimos haciendo hasta ahora, es que ambas muestras son iguales, provienen de la misma población, y se cumplirá si el valor de "Sig." es mayor al nivel de significación fijado (0,05 en nuestro caso). Como "Sig." es igual a 0,00 < 0,05, no podemos suponer que son iguales, por tanto son diferentes, como ya habíamos calculado con Excel.

**Estadísticos de grupo**

| | TipoTubo | N | Media | Desviación típ. | Error típ. de la media |
|---|---|---|---|---|---|
| Tubo | T | 39 | 23,0718 | 9,83926 | 1,57554 |
| | D | 33 | 8,2697 | 4,19363 | ,73002 |

**Prueba de muestras independientes**

| | | Prueba de Levene para la igualdad de varianzas | | Prueba T para la igualdad de medias | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Intervalo de confianza para la diferencia | |
| | | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | Inferior | Superior |
| Tubo | Se han asumido varianzas iguales | 18,150 | ,000 | 8,040 | 70 | ,000 | 14,80210 | 1,84116 | 11,13001 | 18,47418 |
| | No se han asumido varianzas iguales | | | 8,524 | 53,158 | ,000 | 14,80210 | 1,73645 | 11,31946 | 18,28473 |

**Figura 28. Resultado de la prueba t en SPSS.**

## Test t para una sola muestra

Vimos anteriormente cómo podíamos calcular, a partir de nuestra muestra, un intervalo en el que se encontraría la media de mi población con un cierto intervalo de confianza (el ejemplo de las presiones sanguíneas). La prueba t de Student para una sola muestra, permite contrastar la hipótesis nula de que la media de mi muestra es la media de la población (o se parece mucho). Esta prueba se puede hacer en SPSS a través de "Analizar", "Compara medias", "Prueba t para una muestra…". Elegiríamos la variable y el "Valor de prueba" que nosotros queramos comprobar (el valor de la media de nuestra media). Una vez más si "Sig." es mayor a la p que fijemos, entonces nuestra hipótesis de que el valor que indico es el de la población será cierta. En caso contrario, deberé asumir la hipótesis alternativa, o lo que es lo mismo, que ese valor que yo indico no es el valor de la media poblacional.

## Test F o ANOVA

¿Y qué pasa si tengo más de dos muestras? Imaginemos que tenemos tres muestras lo cual es bastante frecuente, por ejemplo cuando tenemos un grupo de estudio, otro placebo y otro control. Tendríamos que realizar pruebas t dos a dos, en total tres pruebas t. A lo mejor no es un problema realizar esas tres pruebas t (ya hemos visto que es sumamente sencillo tanto en Excel como en SPSS). Así podríamos identificar cuáles son iguales y cuáles son diferentes. Pero ¿y si tenemos 4 muestras? El número de pruebas t crece hasta seis.

La solución es aplicar un test ANOVA que nos indicará si alguna de las muestras es diferente. A priori no nos dirá cuál o cuáles son diferentes, sino que al menos una de ellas es diferente.

Para este apartado utilizaremos los datos de la pestaña "Ej.4 – ANOVA" de la hoja de Excel proporcionada[8]. Son datos de calificaciones de 5 grupos de alumnos y queremos saber si hay algún grupo mejor o peor que otro o, por el contrario (hipótesis nula) que todos los grupos son iguales. El método docente, el material y las clases prácticas han sido iguales, sólo hemos cambiado de profesor…

Para realizar este test en Excel, iremos a la pestaña "Datos" → "Análisis de datos" y elegiremos la prueba "Análisis de varianza de un factor". El cuadro de diálogo se muestra en la Fig. 29.

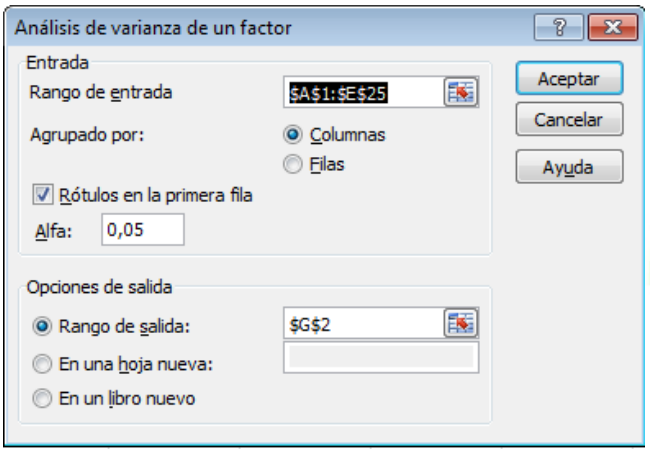



**Figura 29. Cuadro de diálogo para realizar una prueba ANOVA en Excel.**

[8] Te recuerdo que puedes descargarlo desde: http://goo.gl/UEd2pg

Indicamos en el "Rango de entrada" la matriz de datos, incluyendo los rótulos y por eso marcamos "Rótulos en la primera fila". Indicamos el valor de alfa (nuestro p-valor) y el "Rango de Salida". Obtenemos el resultado que se muestra en la Fig. 30. Como siempre, buscamos el valor de p, en este caso en la columna probabilidad y vemos que es 1,56E-05[9] menor que 0,05, por tanto rechazamos la hipótesis nula que asume que todas las muestras son iguales, y por tanto al menos una de ellas es diferente. Para determinar cuál, deberíamos hacer las 10 pruebas t de tomar las diferentes parejas posibles. Otra manera de analizar el resultado del test, al igual que hacíamos para la prueba t en Excel, consiste en comparar el valor de F calculado y el valor de F crítico, si nuestro F es menor que el F crítico, entonces aceptamos H0. En conclusión, en nuestro caso, el grupo 1 tiene una nota media de 7,4, el grupo 2 de 6,7, el grupo 3 de 6,3, el grupo 4 de 6,3 y el grupo 5 de 4,1 hay un grupo que es diferente. A simple vista parece que el grupo 5 tiene una nota media inferior al resto, aunque podría haber más de un grupo diferente, incluso podrían ser todos diferentes. En Excel deberíamos hacer las pruebas t, en SPSS no hace falta, veamos cómo.

| *Origen de las variaciones* | *Suma de cuadrados* | *Grados de libertad* | *Promedio de los cuadrados* | *F* | *Probabilidad* | *Valor crítico para F* |
|---|---|---|---|---|---|---|
| Entre grupos | 133,5132 | 4 | 33,37830003 | 7,738316 | 1,56317E-05 | 2,45421339 |
| Dentro de los grup | 474,471858 | 110 | 4,313380524 | | | |
| | | | | | | |
| Total | 607,985058 | 114 | | | | |

**Figura 30. Resultado de la prueba ANOVA en Excel.**

Antes de nada, tendremos que poner todos los datos en una única variable (Nota) y generar una nueva variable de agrupación (Grupo) que codificaremos del 1 al 5. Como en otras ocasiones vamos a "Analizar", "Comparar medias" y elegimos "ANOVA de un factor…", llegando al cuadro de diálogo que se muestra en la Fig. 31.

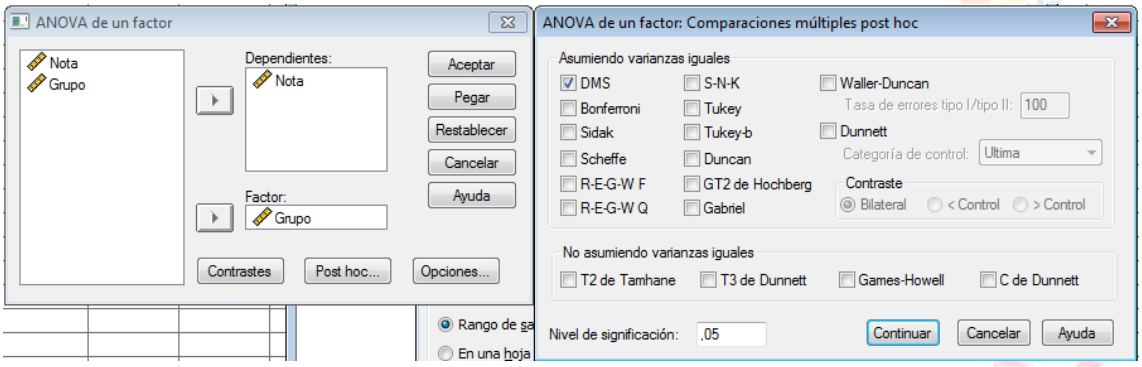



**Figura 31. Cuadro de diálogo para realizar una prueba ANOVA en SPSS.**

En la Fig. 31 (derecha) se muestran las opciones a las que accederemos a través del botón "Post hoc…" a través del cual podremos indicar si queremos que SPSS realice test dos a dos a posteriori. Esta es una ventaja con respecto a Excel y es que si marcamos alguna de estos post-test, podremos saber cuáles de las muestras, en caso de serlo, son diferentes. Marcaremos el primero, DMS.

[9] Notación científica que equivale a $1{,}56 \cdot 10^{-5} = 0{,}0000156$.

| | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|---|---|---|---|---|---|
| Inter-grupos | 133,908 | 4 | 33,477 | 7,754 | ,000 |
| Intra-grupos | 474,930 | 110 | 4,318 | | |
| Total | 608,838 | 114 | | | |

**Figura 32. Resultado de la prueba ANOVA en SPSS.**

En la Fig. 32 se muestra el resultado de la prueba ANOVA en SPSS. Como "Sig." es cero, es menor que 0,05, rechazamos de nuevo H0, esto es, los grupos son diferentes, en realidad sólo sabemos que al menos uno de ellos es diferente.

| (I) Grupo | (J) Grupo | Diferencia de medias (I-J) | Error típico | Sig. | Intervalo de confianza al 95% | |
|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior |
| G1 | G2 | ,67841 | ,61331 | ,271 | -,5370 | 1,8938 |
| | G3 | 1,08750 | ,59983 | ,073 | -,1012 | 2,2762 |
| | G4 | 1,13533 | ,60631 | ,064 | -,0662 | 2,3369 |
| | G5 | 3,25568* | ,61331 | ,000 | 2,0402 | 4,4711 |
| G2 | G1 | -,67841 | ,61331 | ,271 | -1,8938 | ,5370 |
| | G3 | ,40909 | ,61331 | ,506 | -,8063 | 1,6245 |
| | G4 | ,45692 | ,61965 | ,462 | -,7711 | 1,6849 |
| | G5 | 2,57727* | ,62650 | ,000 | 1,3357 | 3,8189 |
| G3 | G1 | -1,08750 | ,59983 | ,073 | -2,2762 | ,1012 |
| | G2 | -,40909 | ,61331 | ,506 | -1,6245 | ,8063 |
| | G4 | ,04783 | ,60631 | ,937 | -1,1537 | 1,2494 |
| | G5 | 2,16818* | ,61331 | ,001 | ,9527 | 3,3836 |
| G4 | G1 | -1,13533 | ,60631 | ,064 | -2,3369 | ,0662 |
| | G2 | -,45692 | ,61965 | ,462 | -1,6849 | ,7711 |
| | G3 | -,04783 | ,60631 | ,937 | -1,2494 | 1,1537 |
| | G5 | 2,12036* | ,61965 | ,001 | ,8923 | 3,3484 |
| G5 | G1 | -3,25568* | ,61331 | ,000 | -4,4711 | -2,0402 |
| | G2 | -2,57727* | ,62650 | ,000 | -3,8189 | -1,3357 |
| | G3 | -2,16818* | ,61331 | ,001 | -3,3836 | -,9527 |
| | G4 | -2,12036* | ,61965 | ,001 | -3,3484 | -,8923 |

*. La diferencia de medias es significativa al nivel .05.

**Figura 33. Resultado del post-test DMS tras la ANOVA en SPSS.**

En la Fig. 33 se muestra el resultado del post-test DMS para este caso. Comprobamos que los valores de "Sig." para todas las parejas de grupos son superiores a 0,05, salvo para el caso de los emparejamientos con el grupo 5. Por tanto comprobamos que el grupo 5 es diferente al resto, mientras que los grupos 1 al 4 son iguales. Por tanto podemos concluir que el grupo 5 tiene peores calificaciones que el resto de grupos (p=0,05).

Si haces la prueba ANOVA tomando sólo los grupos 1 a 4, podrás comprobar la validez de la hipótesis nula y, por tanto, los cuatro grupos no presentan diferencias significativas.

## ANOVA de dos (o más factores)

En el apartado anterior vimos en qué consiste el análisis de la varianza con un factor, pero nuestro estudio podría investigar el efecto de dos, tres, cuatro, o incluso más factores. En tal caso, se debería aplicar otro test estadístico: el análisis de la varianza con dos (o más) factores.

Por ejemplo podríamos calcular la variabilidad de las observaciones dentro de cada grupo (intra-grupo) así como entre los grupos (inter-grupo), de manera que podríamos preguntarnos si las medias observadas son diferentes entre hombres y mujeres, fumadores y no fumadores, introvertidos y extrovertidos, etc. Además, el análisis de la varianza de dos, tres o más factores permitirá investigar la existencia de posibles efectos de interacción entre los factores analizados.

En la pestaña "Ej.4b - ANOVA de dos factores" del archivo de ejemplos, hemos insertado una nueva columna antes de la columna "A", desplazando las existentes hacia la derecha. En esta nueva columna hemos insertado una nueva variable de agrupación, por ejemplo el sexo: hombre y mujer, de manera que las primeras 12 muestras son calificaciones de hombres y las siguientes 12 muestras serán de mujeres.

Para realizar este test en Excel, iremos a la pestaña "Datos" → "Análisis de datos" y elegiremos la prueba "Análisis de varianza de dos factores con varias muestras por grupo". El cuadro de diálogo se muestra en la Fig. 34.

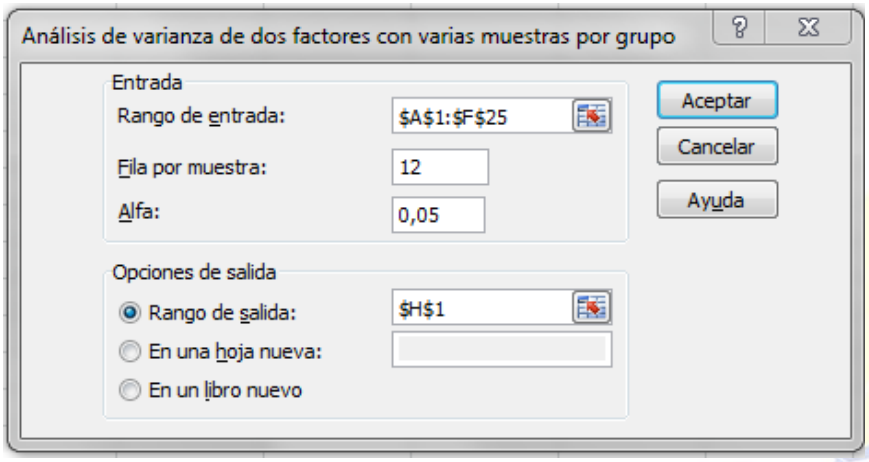



**Figura 34. Cuadro de diálogo de Excel para el cálculo de una ANOVA de dos o más factores.**

La diferencia con respecto al caso anterior es que debemos indicar la fila por muestra, en nuestro caso "12" filas por muestra de hombre y de mujeres. El resultado se muestra en la Fig. 35 y se interpretará atendiendo a la fila "Columna" que nos indicará la posible igualdad de las muestras con respecto al grupo y la fila "Interacción" que nos informa de la posible interacción de los grupos con el sexo de los individuos. Tanto en el primer caso (Muestra) como en el segundo (Columnas) obtenemos una "probabilidad" inferior a 0,05 y por tanto existe al menos una muestra diferente atendiendo estos factores. En cuanto a una posible interacción o dependencia entre el grupo y el sexo, observamos que p es mayor que 0,05 y por tanto no podemos rechazar la hipótesis nula de igualdad de las varianzas.

| ANÁLISIS DE VARIANZA | | | | | | |
|---|---|---|---|---|---|---|
| *Origen de las variaciones* | *Suma de cuadrados* | *Grados de libertad* | *Promedio de los cuadrados* | *F* | *Probabilidad* | *Valor crítico para F* |
| Muestra | 166,2312411 | 1 | 166,2312411 | 59,7649401 | 5,43306E-12 | 3,927393633 |
| Columnas | 133,6207015 | 4 | 33,40517538 | 12,0101269 | 3,91531E-08 | 2,45421339 |
| Interacción | 14,85768059 | 4 | 3,714420148 | 1,3354415 | 0,261347917 | 2,45421339 |
| Dentro del grupo | 305,9559083 | 110 | 2,781417348 | | | |
| | | | | | | |
| Total | 620,6655315 | 119 | | | | |

**Figura 35: Resultado de la prueba ANOVA de dos o más factores en Excel.**

En SPSS se realiza de manera similar a lo realizado anteriormente. Necesitaremos poner las calificaciones en una única columna. Insertaremos una nueva columna/variable, además de la de grupo, en la que codificaremos el sexo (por ejemplo 1 para hombres y 2 para mujeres). Así cada caso tendrá dos factores: el grupo y el sexo (pero podría haber más).

Para realizar el ANOVA de dos o más factores debemos ir a "Analizar", "Modelo lineal general" y elegimos "Univariante…", llegando al cuadro de diálogo que se muestra en la Fig. 36.

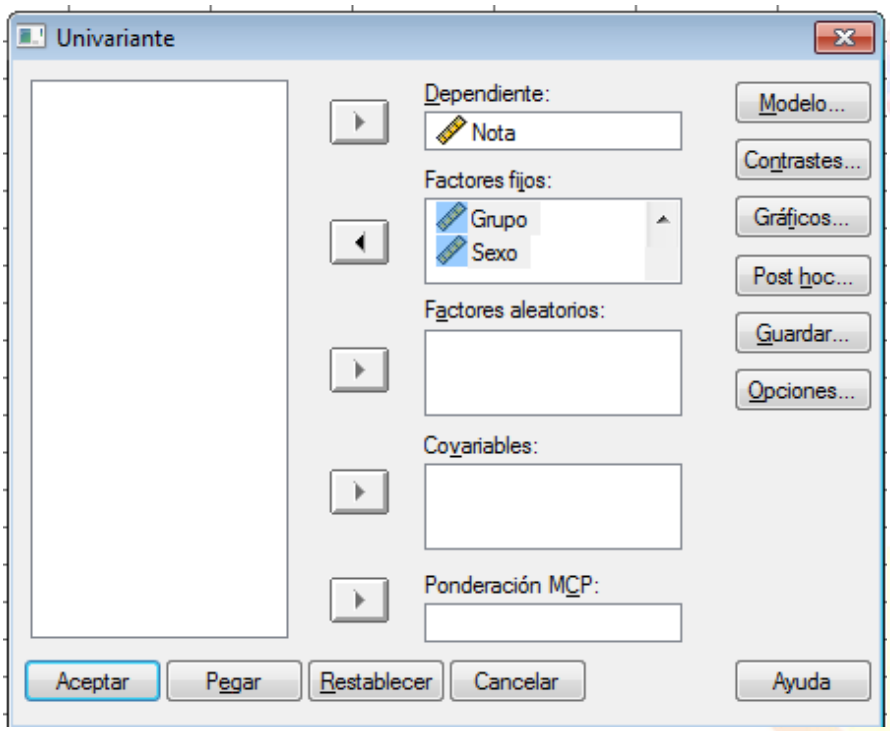



**Figura 36: Cuadro de diálogo para elegir las diferentes opciones del test ANOVA con dos o más factores en SPSS.**

El resultado de esta prueba se muestra en la Fig. 37. En SPSS no es preciso que las muestras tengan el mismo tamaño. SPSS presenta el resultado para "Grupo", "Sexo" y la interacción "Grupo*Sexo". Como ya sabíamos, hay diferencias entre los grupos teniendo en cuenta los dos factores pero en cambio no existe una interacción entre grupo y sexo.

**Pruebas de los efectos inter-sujetos**

Variable dependiente: Nota

| Fuente | Suma de cuadrados tipo III | gl | Media cuadrática | F | Significación |
|---|---|---|---|---|---|
| Modelo corregido | 315,149[a] | 9 | 35,017 | 12,590 | ,000 |
| Intersección | 4614,040 | 1 | 4614,040 | 1658,963 | ,000 |
| Grupo | 133,887 | 4 | 33,472 | 12,035 | ,000 |
| Sexo | 166,381 | 1 | 166,381 | 59,822 | ,000 |
| Grupo * Sexo | 14,881 | 4 | 3,720 | 1,338 | ,261 |
| Error | 305,941 | 110 | 2,781 | | |
| Total | 5235,130 | 120 | | | |
| Total corregida | 621,090 | 119 | | | |

a. R cuadrado = ,507 (R cuadrado corregida = ,467)

**Figura 37: Resultado del test ANOVA con dos o más factores en SPSS.**

## Test U de Mann-Whitney

La siguiente pregunta que debemos hacernos es ¿y qué pasa si los datos no son normales? ¿Existe algún test estadístico para poder comprobar este tipo de hipótesis con datos que no se ajustan a una distribución normal? Pues por suerte la respuesta es sí y se aplicarán exactamente igual que hasta ahora, pero sólo mediante SPSS. El equivalente al test t de Student para dos muestras es la U de Mann-Whitney (también de Wilcoxon) y el test de Kruskal-Wallis será el equivalente a un análisis ANOVA para más de dos muestras no normales.

En el archivo de Excel, en la hoja "Ej.5 U MW y KW" encontrarás tres series de datos que se corresponden con las temperaturas de Albacete, Ávila y Madrid. Lo primero que deberás hacer es comprobar la normalidad. Aunque los datos se proporcionan en Excel, utilizaremos SPSS para los siguientes cálculos.

Como ocurría en los casos anteriores, al pasar los datos a SPSS los pondremos todos en una misma columna en una misma variable (T), creando una nueva variable (Ciudad) con tres valores 1, 2 y 3 que definan los datos de cada ciudad. A continuación iremos a "Analizar", "Pruebas no paramétricas", y finalmente a "2 muestras independientes". El cuadro de diálogo se muestra en la Fig. 38.

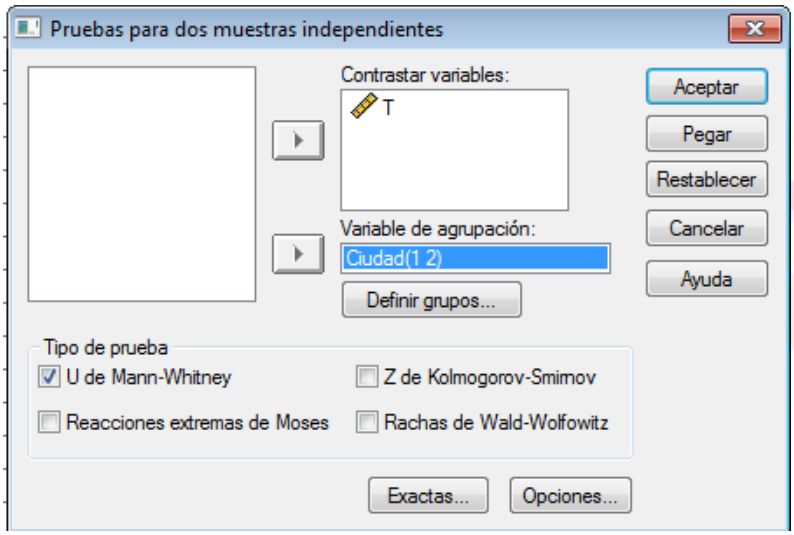



**Figura 38. Cuadro de diálogo para la realización de una prueba U de Mann-Whitney en SPSS.**

Como se trata de una prueba de dos muestras, tomaremos los grupos 1 y 2 correspondientes a Albacete y Ávila. Haremos lo mismo para Albacete y Madrid y por último Ávila y Madrid. Obtenemos los resultados que se muestran en la Fig. 39.

**Estadísticos de contraste[a]**

| | T |
|---|---|
| U de Mann-Whitney | 16,500 |
| W de Wilcoxon | 316,500 |
| Z | -5,605 |
| Sig. asintót. (bilateral) | ,000 |

a. Variable de agrupación: Ciudad

**Estadísticos de contraste[a]**

| | T |
|---|---|
| U de Mann-Whitney | 205,000 |
| W de Wilcoxon | 505,000 |
| Z | -1,713 |
| Sig. asintót. (bilateral) | ,087 |

a. Variable de agrupación: Ciudad

**Estadísticos de contraste[a]**

| | T |
|---|---|
| U de Mann-Whitney | 275,500 |
| W de Wilcoxon | 575,500 |
| Z | -,258 |
| Sig. asintót. (bilateral) | ,797 |

a. Variable de agrupación: Ciudad

**Figura 39. Resultados de la prueba U de Mann-Whitney para AB-AV (izquierda), AB-M (centro) y AV-M. (AV-M).**

La hipótesis nula es que no existen diferencias significativas entre las muestras y esto ocurrirá cuando la "Sig." sea mayor que 0,05. Podemos comprobar que esto se cumple entre AB y M (0,087) y entre AV y M (0,797), pero no se cumple para AB y AV (AB-AV). Por tanto no observamos diferencias significativas en las parejas AB-M y AV-M pero sí que existen diferencias estadísticamente significativas entre los datos AB-AV (p=0,05).

Si vuelves a abrir el menú de "Pruebas no paramétricas" encontrarás la misma prueba para muestras relacionadas, que aplicaremos cuando las dos muestras estén relacionadas o sean la misma muestra a dos tiempos diferentes.

## Test de Kruskal-Wallis

Como se ha indicado, la prueba de Kruskal-Wallis es equivalente al test ANOVA pero para muestras no normales. Lo que hemos hecho en el apartado anterior, comparar las tres muestras dos a dos, lo podríamos haber hecho directamente con este test, que nos dirá si hay alguna muestra diferente.

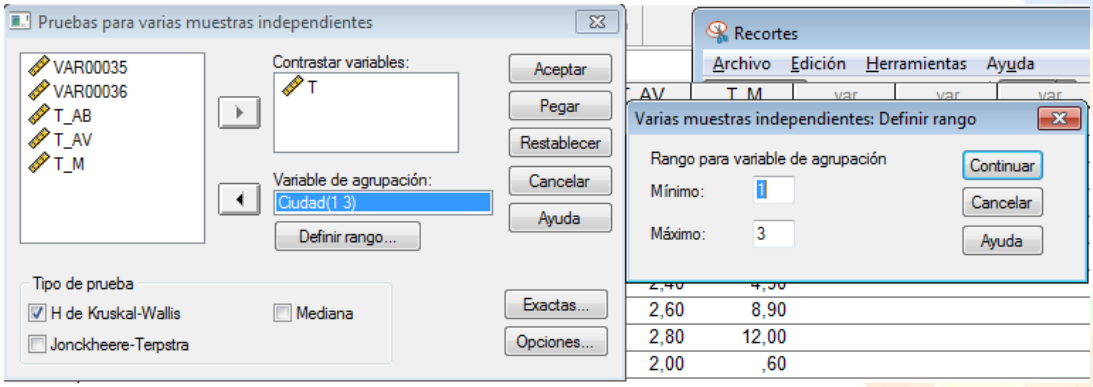



**Figura 40. Cuadro de diálogo para la realización del test de Kruskal-Wallis en SPSS.**

Para ello iremos a "Analizar", "Pruebas no paramétricas", y finalmente a "K muestras independientes" y accederemos al cuadro de diálogo de la Fig. 40.

| | T |
|---|---|
| Chi-cuadrado | 19,221 |
| gl | 2 |
| Sig. asintót. | ,000 |

**Figura 41. Resultados de la prueba de Kruskall-Wallis para tres muestras en SPSS.**

Ahora en vez de dos muestras tendremos tres, por lo que en la "Variable de agrupación" definiremos el rango (Fig. 36 derecha) entre 1 y 3, que son los tres valores de las 3 ciudades. Los resultados se muestran en la Fig. 41.

Como era de esperar, pues ya lo sabíamos, el valor de la "Sig." es menor a 0,05, por tanto debemos rechazar la hipótesis nula de igualdad de las muestras: al menos una es diferente. Ahora realizaríamos un test U de Mann-Whitney dos a dos para determinar cuáles son significativamente diferentes.

## Test $\chi^2$

Hasta ahora hemos visto pruebas de contraste para variables cuantitativas pero, ¿qué ocurre cuando tenemos variables cualitativas como el sexo, si se fuma o no o el color de pelo? Debemos recurrir a la prueba $\chi^2$ de Pearson.

Imagina que durante los últimos años se ha ido registrando el número de astronautas que presentaron trastornos tras una misión en la Estación Espacial Internacional dependiendo del tiempo que permanecieron en el espacio. Los datos registrados se muestran en la siguiente tabla:

| | Datos Observados | | |
|---|---|---|---|
| | **Astronautas con trastornos** | **Astronautas sin trastornos** | **Total** |
| **Menos de 1 mes** | 12 | 17 | 29 |
| **1-3 meses** | 11 | 14 | 25 |
| **3-6 meses** | 7 | 13 | 20 |
| **Más de 6 meses** | 11 | 15 | 26 |
| **Total** | 41 | 59 | 100 |

Queremos saber si existe alguna relación entre el tiempo que permanecieron en el espacio y haber sufrido algún tipo de trastorno.

El contraste mediante el test $\chi^2$ se realiza al comparar nuestros datos con los resultados esperados debido a las proporciones totales, me explico: supondremos que ambas muestras provienen de la misma población, por tanto si junto ambas muestras, la nueva muestras se parecerá más a la población que las muestras por separado, esto es, las proporciones de la nueva muestra conjunta, se parecerán más a la de la población. Por tanto calculo las proporciones que debería tener en cada una de mis dos muestras suponiendo que deberían conservar las proporciones de la muestra completa. En el archivo de Excel, en la pestaña "Ej.6 – Chi-2" encontraras la tabla de valores observados y la tabla de valores esperados que hemos calculado a partir de las proporciones totales. ¿Cómo? Pues fácil: En los valores observados tenemos 12 astronautas con trastornos que estuvieron menos de un mes y 17 que no tuvieron trastornos. Si no hubiera diferencias entre tener trastorno o no tenerlo, ambos grupos pertenecerían a la misma población y por tanto la proporción esperada de ambos casos sería de 29 casos (la suma) del total de 100. Por tanto esa proporción nos dará la regla para calcular el valor esperado, pues hemos tenido 41 astronautas con trastornos. Por tanto se hace una regla de tres sencilla de manera que si de 100 astronautas, 29 estuvieron menos de un mes, de 41 (que son los que tienen trastornos en total) deberíamos tener X. Así rellenamos la tabla de manera que:

| | Datos Esperados | | |
|---|---|---|---|
| | **Astronautas con trastornos** | **Astronautas sin trastornos** | **Total** |
| **Menos de 1 mes** | 11,89 | 17,11 | 29 |
| **1-3 meses** | 10,25 | 14,75 | 25 |
| **3-6 meses** | 8,2 | 11,8 | 20 |
| **Más de 6 meses** | 10,66 | 15,34 | 26 |
| **Total** | 41 | 59 | 100 |

La forma de calcular esta prueba en Excel es bastante manual, pero es fácil. En Excel existe la función "**=PRUEBA.CHI(rango observado; rango esperado)**" que devuelve la probabilidad.

Seleccionaremos el rango observado o esperado en cada caso y obtendremos un valor de probabilidad de 0,938 que como es superior a 0,05 nos indica que no podemos rechazar H0, por tanto las diferencias se deben al muestreo. Si queremos calcular y contrastar los valores de $\chi^2$ y de $\chi^2$ crítico y ver si el primero es mayor o menor que el segundo, debemos utilizar la función "**=PRUEBA.CHI.INV(probabilidad; grados de libertad)**" que calcula, a partir de una probabilidad dada para unos grados de libertad determinados, el valor del estadístico $\chi^2$. Así conoceremos el valor de $\chi^2$ para poder compararlo con el valor crítico teórico que nos marca el límite, el intervalo, para el cual se cumple la hipótesis de que ambas muestras pertenecen a la misma población, esto es, que las diferencias observadas se deben al muestreo, o que el hecho de estar en el espacio más o menos tiempo, no afectaría a la probabilidad de padecer un trastorno.

Por último tendremos que calcular el valor crítico de $\chi^2$ para el intervalo de confianza dado ("probabilidad" en la función de Excel = 0,05) con "**=PRUEBA.CHI.INV(probabilidad; grados de libertad)**" y ver si el valor del problema es menor que el valor crítico, en este caso, no habría diferencias.

Los grados de libertad se calculan como el número de filas menos 1 por el número de columnas menos 1, en nuestro caso (4-1)*(2-1) = 3.

Haciendo estos cálculos (que tienes en el archivo de Excel) obtenemos que Chi-2 vale 0,410 mientras que Chi-crítico vale 7,81, por lo que no existen diferencias estadísticamente significativas (p=0,05).

| | Nombre | Tipo | Anchura | Decimales | Etiqueta | Valores | Perdidos | Columnas | Alineación | Medida |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Id_Astrona | Numérico | 4 | 0 | Id del Astronauta | Ninguno | Ninguno | 8 | Derecha | Escala |
| 2 | t_en_EEI | Numérico | 1 | 0 | Tiempo en la EEI | {1, Menos de un mes}... | Ninguno | 8 | Derecha | Escala |
| 3 | Trastorno | Numérico | 1 | 0 | Trastorno (sí o no) | {0, No}... | Ninguno | 8 | Derecha | Escala |

**Figura 42. Vista de variables en SPSS.**

Ten en cuenta que en este ejemplo hemos partido de la tabla de contingencia y hemos calculado los valores de $\chi^2$ crítico y el de nuestro caso. Pero SPSS permite, partiendo de los datos en crudo, que sería cada caso con sus características, generar la tabla de contingencia y determinar la "Sig." de la prueba $\chi^2$. Para ello necesitamos los datos, cada caso por separado en una fila de SPSS. En la misma hoja de Excel (Ej.6 – Chi-2) encontrarás tres columnas con esos datos desglosados. Tenemos cada astronauta identificado por un número de 3 dígitos, el "tiempo de permanencia en la EEI" codificado (1, 2, 3, 4) y en la columna "trastorno" codificado con un 0 si no ha tenido trastorno y con un 1 si lo ha tenido.

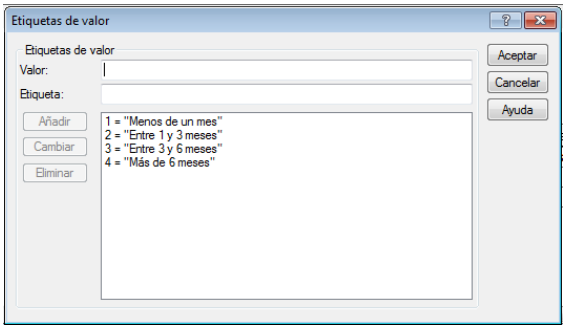



**Figura 43. Etiquetas de valor en SPSS.**

Debemos copiar y pegar estas tres columnas en SPSS y en la pestaña "Vista de Variables" que está en la parte inferior de la ventana principal del programa, podremos codificar convenientemente cada una de ellas como se muestra en la Fig. 43. En la columna "Valores" (Fig. 43) podremos codificar cada valor con su etiqueta correspondiente para que luego los resultados nos salgan más bonitos.

Una vez tenemos los datos preparados, podremos hacer el análisis. Para ello debemos ir a "Analizar" → "Estadísticos descriptivos" y "Tablas de contingencia" (Fig. 44 izquierda). Indicaremos qué variable figurará en las filas y cuál en las columnas, además debemos pulsar sobre el botón "Estadísticos" y marcar la opción de "Chi-cuadrado" (Fig. 44 derecha).

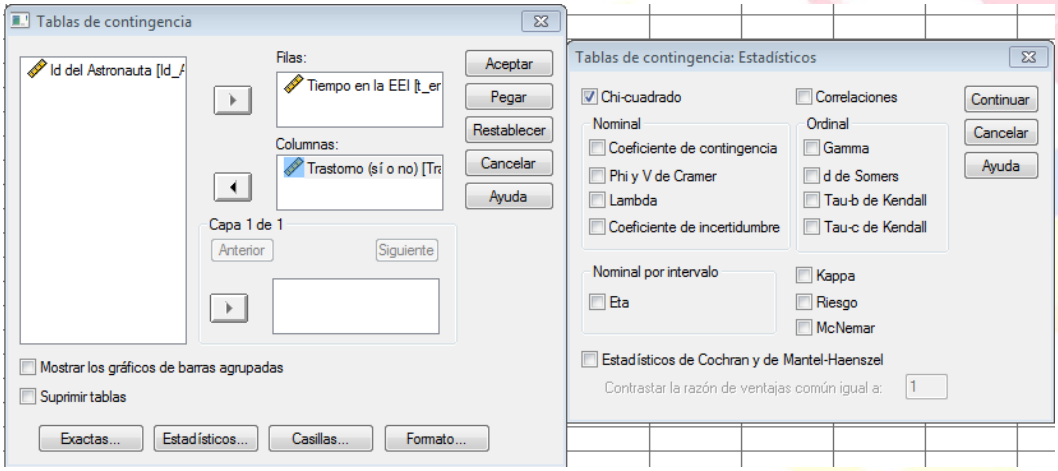



**Figura 44. Opciones de tablas de contingencia en SPSS.**

Al pulsar "Aceptar", SPSS devuelve dos tablas que se muestran en la Fig. 45. La primera tabla (izquierda) es la tabla de contingencia, resumen de datos (que como indicábamos en el apartado de descriptivos, puede ser útil para resumir los datos de un problema). La segunda tabla (derecha) muestra la "Sig." de prueba Chi cuadrado de Pearson que vale 0,938 que al ser mayor que 0,05 (nuestra p) significa que, como ya sabíamos, no existen diferencias entre los diferentes conjuntos de datos.

**Tabla de contingencia Tiempo en la EEI * Trastorno (sí o no)**

Recuento

| | | Trastorno (sí o no) | | |
|---|---|---|---|---|
| | | No | Sí | Total |
| Tiempo en la EEI | Menos de un mes | 12 | 17 | 29 |
| | Entre 1 y 3 meses | 11 | 14 | 25 |
| | Entre 3 y 6 meses | 7 | 13 | 20 |
| | Más de 6 meses | 11 | 15 | 26 |
| Total | | 41 | 59 | 100 |

**Pruebas de chi-cuadrado**

| | Valor | gl | Sig. asintótica (bilateral) |
|---|---|---|---|
| Chi-cuadrado de Pearson | ,411[a] | 3 | ,938 |
| Razón de verosimilitudes | ,415 | 3 | ,937 |
| Asociación lineal por lineal | ,012 | 1 | ,912 |
| N de casos válidos | 100 | | |

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 8,20.

**Figura 45. Resultado de la prueba chi-cuadrado en SPSS.**

## Correlación y Regresión

Hasta ahora hemos visto cómo buscar relaciones entre muestras, pero no podemos hacer predicciones sobre una característica particular de un individuo de una población al medir otra.

Antes, necesitaremos estimar el grado de "correlación" que tienen esas dos variables entre sí, esto es, caracterizar el grado de relación entre dos variables medidas para cada individuo de nuestra muestra. Para hacer las predicciones, una vez estimado el grado de correlación, necesitaremos realizar una "regresión" que nos informará sobre la naturaleza de la relación entre ambas variables, esto es la forma en la que dependen una de otra, y así poder hacer predicciones.

En el curso 2011-12 pedimos a los alumnos de primero de medicina de la Facultad de Albacete que midieran su estatura, la envergadura de sus brazos, longitud de sus piernas hasta la ingle y la altura de su ombligo. Los datos se encuentran en la hoja de Excel, en la pestaña "Ej.7 – Correl y Regres"[10]. Queremos determinar el grado y el tipo de relación existente entre estas variables.

Una forma sencilla de hacer este estudio en Excel es representar cada una de las variables frente a la estatura, gráfico de dispersión X/Y donde la variable en el eje X será siempre la estatura (Fig. 46).

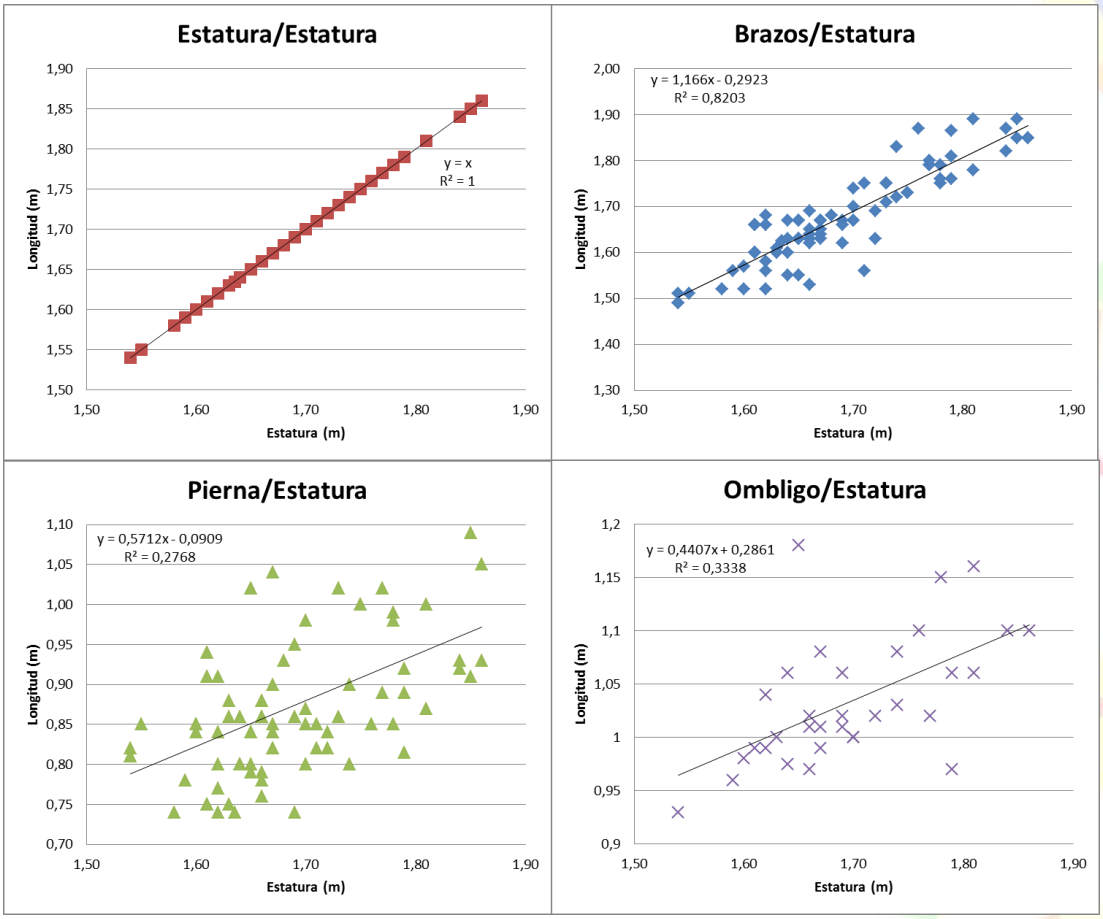



**Figura 46. Gráficos de dispersión de las diferentes variables frente a la estatura con la recta de ajuste y $R^2$.**

Una vez tenemos las gráficas, hacemos *click* con el botón derecho del ratón sobre los datos representados en cada una de las gráficas y elegiremos "Agregar línea de tendencia…". En el cuadro de diálogo que se muestra en la Fig. 47, marcaremos el tipo de tendencia, en este caso

[10] Te recuerdo que puedes descargarlo desde: http://goo.gl/UEd2pg

"lineal" e indicaremos que Excel presente la ecuación en el gráfico y el valor de R cuadrado (las dos opciones del final). En la Fig. 46 ya se muestran las líneas de tendencia y las ecuaciones de las diferentes rectas que mejor se ajustan a los datos.

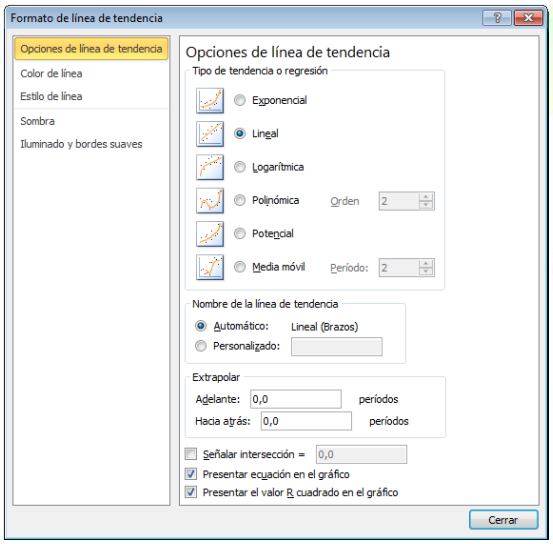
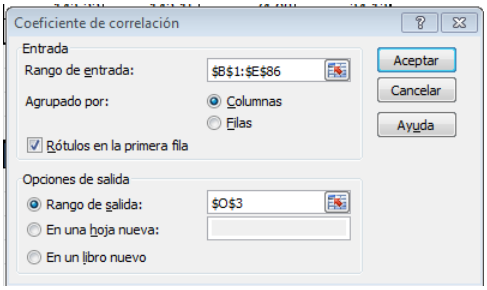



**Figura 47. Opciones para agregar la línea de tendencia en Excel.**

Si miramos a los gráficos, podemos ver que hay variables más relacionadas que otras. Por ejemplo la estatura y la envergadura muestran una dependencia lineal más fuerte que la longitud de las piernas y la estatura.

Veamos cómo podemos cuantificar el grado de correlación de una forma más precisa y objetiva. En los gráficos aparece el valor de $R^2$ que nos informa objetivamente sobre esta correlación, podemos interpretarlo como la proporción (que podremos transformar a porcentaje) de la estatura que podemos explicar con las otras variables. En el caso de la envergadura, este valor es del 82%.

La forma precisa para determinar esa relación es mediante el coeficiente de correlación de Pearson. En Excel podemos usar la fórmula "**=COEF.DE.CORREL(matriz1, matriz2)**" donde "matriz1" serán los datos de estatura y la "matriz2" serán los de las diferentes variables respectivamente. Obtenemos los siguientes valores: 0,906, 0,526, 0,578 que podrás comprobar que coinciden con los valores de la raíz de R cuadrado en cada caso. Pues bien, R es el coeficiente de correlación de Pearson en cada caso y cuanto más cercano sea este valor a 1, más fuerte será el grado de relación. Elevándolo al cuadrado tendremos la proporción de "explicación".




| | *Estatura* | *Brazos* | *Pierna* | *Ombligo* |
|---|---|---|---|---|
| Estatura | 1 | | | |
| Brazos | 0,90572268 | 1 | | |
| Pierna | 0,52608813 | 0,5469092 | 1 | |
| Ombligo | 0,57772705 | 0,59760024 | 0,72640215 | 1 |

**Figura 48. Cuadro de diálogo para el cálculo del coeficiente de correlación (izquierda) y correlaciones (derecha).**

Si quisiéramos estudiar la posible relación entre el resto de variables entre sí, deberíamos ir calculando el coeficiente de correlación en cada caso. Para eso hay una forma más sencilla de calcular el coeficiente de correlación. Iremos al menú "Datos" → "Análisis de Datos" → "Coeficiente de correlación". En la Fig. 48 (izquierda) se muestra el cuadro de diálogo a través del cual podremos indicar la matriz de datos (recuerda que si seleccionas los rótulos, deberás indicarlo), así como el rango de salida. El resultado se muestra en la Fig. 48 (derecha) y como podrás comprobar, se muestran todas las relaciones. Podemos ver que la altura del ombligo presenta una relación más fuerte con la longitud de la pierna que con la propia estatura, lo cual es lógico ¿no?

Antes de continuar, es importante insistir en que el hecho de que dos variables presenten una correlación alta, no quiere decir que están relacionadas, que exista una relación de causalidad. Podríamos estudiar el grado de correlación entre la cantidad de lluvia que se recoge en marzo y las ventas de un determinado color de tinte de cabello. Es posible que casualmente encuentre una correlación, pero eso no significa que exista causalidad.

Con las ecuaciones que aparecen en cada gráfico, podremos hacer predicciones y determinar la envergadura o la altura del ombligo de un alumno sólo con medir su estatura, lógicamente con un error que será menor cuanto mayor sea la muestra.

Para terminar, SSPS también permite realizar cálculos de correlación y regresión. Para ello debemos llevarnos los datos a SPSS e ir al menú "Analizar" → "Correlaciones" → "Correlaciones bivariadas". En la Fig. 49 se muestra el cuadro de diálogo a través del cual podremos indicar las diferentes opciones.

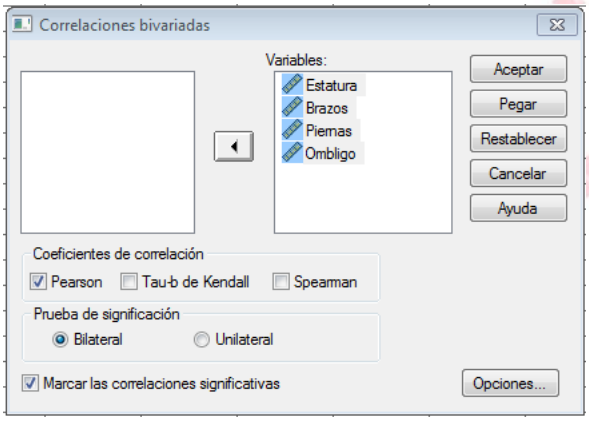



**Figura 49. Cuadro de diálogo para el cálculo de correlaciones en SPSS.**

Los resultados de esta prueba se muestran en la Fig. 50. Los valores de SPSS difieren un poco con respecto a los proporcionados por Excel, pero indica el grado de significación de aquellos que presentan una significación de p=0,01.

Una ventaja de SPSS frente a Excel a la hora de realizar una regresión es que tiene muchas más opciones. Permite hacer muchos tipos de regresiones diferentes (además de la lineal) a través de "Analizar" → "Regresión". En SPSS Podremos elegir "regresión lineal" cuyo cuadro de diálogo de opciones se muestra en la Fig. 51. Además permite calcular regresiones múltiples, esto es, calcular una variable no sólo con una variable independiente sino con varias, además

podemos determinar el grado de mejora de la predicción a la hora de tener en cuenta las diferentes variables en el modelo de predicción.

**Correlaciones**

| | | Estatura | Brazos | Piernas | Ombligo |
|---|---|---|---|---|---|
| Estatura | Correlación de Pearson | 1 | ,972** | ,864** | ,493** |
| | Sig. (bilateral) | | ,000 | ,000 | ,000 |
| | N | 85 | 85 | 85 | 85 |
| Brazos | Correlación de Pearson | ,972** | 1 | ,905** | ,526** |
| | Sig. (bilateral) | ,000 | | ,000 | ,000 |
| | N | 85 | 85 | 85 | 85 |
| Piernas | Correlación de Pearson | ,864** | ,905** | 1 | ,547** |
| | Sig. (bilateral) | ,000 | ,000 | | ,000 |
| | N | 85 | 85 | 85 | 85 |
| Ombligo | Correlación de Pearson | ,493** | ,526** | ,547** | 1 |
| | Sig. (bilateral) | ,000 | ,000 | ,000 | |
| | N | 85 | 85 | 85 | 85 |

**. La correlación es significativa al nivel 0,01 (bilateral).

**Figura 50. Coeficientes de correlación calculados con SPSS.**

En la tabla de correlación (Fig. 50) vemos que la estatura tiene una correlación alta con la envergadura y la longitud de las piernas, aunque algo menor con la altura del ombligo. Podemos realizar un modelo con las dos primeras, pero podemos decirle a SPSS que tenga en cuenta la altura del ombligo pero solo si al incluir en el modelo, la predicción mejora un determinado valor (Fig. 51 derecha).

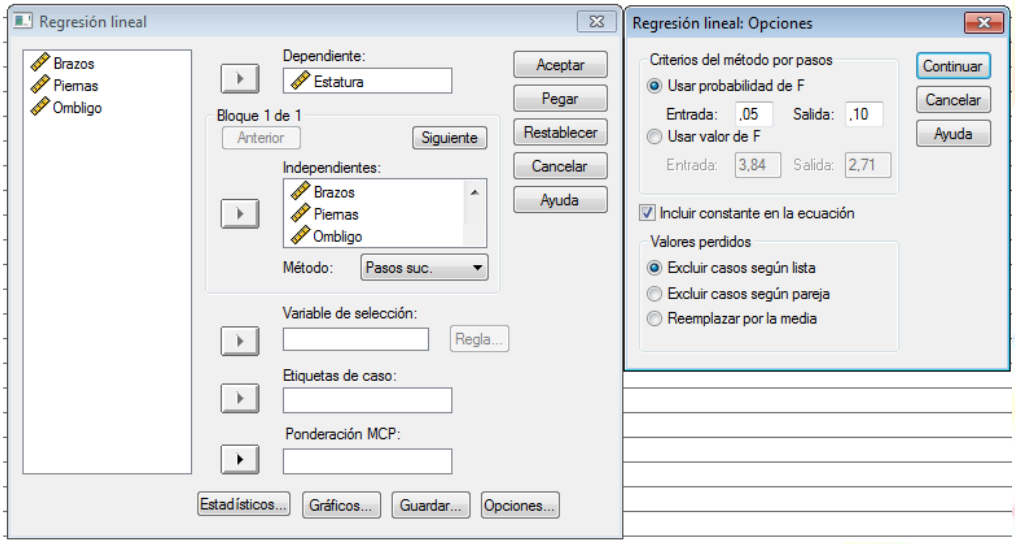



**Figura 51. Cuadro de diálogo para el modelo de regresión lineal en SPSS (izquierda) y opciones (derecha).**

En "Método" hemos elegido "Pasos suc." así SPSS irá añadiendo variables si mejora el resultado. El resultado se muestra en la Fig. 52.

**Coeficientes[a]**

| Modelo | | Coeficientes no estandarizados | | Coeficientes estandarizados | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Error típ. | Beta | | |
| 1 | (Constante) | -468,960 | 13,631 | | -34,405 | ,000 |
| | Brazos | 303,844 | 8,081 | ,972 | 37,600 | ,000 |

a. Variable dependiente: Estatura

**Variables excluidas[b]**

| Modelo | | Beta dentro | t | Sig. | Correlación parcial | Estadísticos de colinealidad |
|---|---|---|---|---|---|---|
| | | | | | | Tolerancia |
| 1 | Piernas | -,090[a] | -1,492 | ,140 | -,163 | ,180 |
| | Ombligo | -,025[a] | -,833 | ,408 | -,092 | ,723 |

a. Variables predictoras en el modelo: (Constante), Brazos

b. Variable dependiente: Estatura

**Figura 52. Resultados de la regresión en SPSS (método "Pasos suc.").**

SPSS proporciona los coeficientes de la regresión e indica que ha excluido las variables "Piernas" y "Ombligo" porque no cumplen los criterios de tolerancia. Si queremos el modelo completo con las tres variables, en "Método" deberemos elegir "Introducir". En la Fig. 53 se muestra los resultados de la predicción. A veces es preferible hacer una predicción con una única variable, que mejorar la predicción con tres variables, pues el esfuerzo de determinar el valor de tres variables, podría no compensar la mejora de la predicción.

**Resumen del modelo[b]**

| Modelo | R | R cuadrado | R cuadrado corregida | Error típ. de la estimación |
|---|---|---|---|---|
| 1 | ,973[a] | ,946 | ,944 | 5,82877 |

**Coeficientes[a]**

| Modelo | | Coeficientes no estandarizados | | Coeficientes estandarizados | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Error típ. | Beta | | |
| 1 | (Constante) | -475,367 | 14,262 | | -33,331 | ,000 |
| | Brazos | 330,303 | 19,052 | 1,057 | 17,337 | ,000 |
| | Piernas | -20,238 | 15,010 | -,083 | -1,348 | ,181 |
| | Ombligo | -4,965 | 8,897 | -,017 | -,558 | ,578 |

a. Variable dependiente: Estatura

**Figura 53. Resultados de la regresión en SPSS (método "introducir").**

Como podrás comprobar, en SPSS existen innumerables opciones a la hora de realizar cálculos de regresión que escapan los propósitos de este texto.

## Recetas

En este apartado se reúnen las diferentes recetas, los pasos a seguir, para poder hacer, tanto en Excel como en SPSS, los diferentes análisis básicos que se han descrito a lo largo del texto. Para muchos, probablemente éste sea el apartado más útil, ya veremos[11]… En cada caso se indica la página en la que encontrarás más detalles.

**Histograma de frecuencias**

- Excel: Datos→Análisis de datos→Histograma. Pág. 5.
- SPSS: Analizar→Estadísticos descriptivos→Frecuencias. Botón gráficos. Pág. 6.

**Estadística Descriptiva**

- Excel: Datos→Análisis de datos→Estadística descriptiva. Pág. 8.
- SPSS: Analizar→Estadísticos descriptivos→Frecuencias (botón estadísticos) o Descriptivos. También útil "Tablas de contingencia para agrupar los datos. Pág. 9.

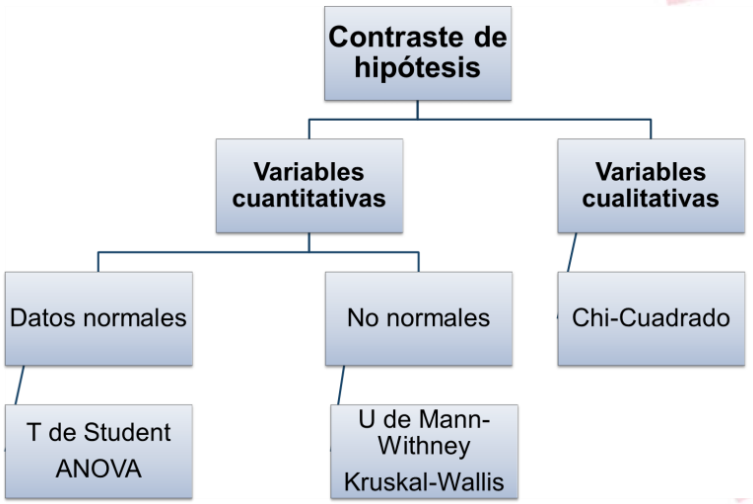
**Test de Normalidad**

- No diponible en Excel.
- SPSS: Kolmogorov-Smirnov (si muestras mayores a 50 datos) y Shapiro-Wilk: Analizar→Estadísticos descriptivos→Explorar, gráficos, gráficos con pruebas de normalidad. La hipótesis es que los datos son normales, por lo que se comprobará si la "Sig." es mayor que 0,05. Otra opción, sólo para WS: Analizar→Pruebas no paramétricas, K-S para una muestra. Págs. 11 y 12.

**Estimar la media de una población**

- Para un 95% → $M - 1{,}96 \cdot SE \leq \mu \leq M + 1{,}96 \cdot SE$
- Para un 99% → $M - 2{,}54 \cdot SE \leq \mu \leq M + 2{,}54 \cdot SE$
- Siendo $SE = S/\sqrt{N}$ válido para N>30. Pág. 14.

El siguiente esquema resumen los test para la comparación de muestras dependiendo del tipo de variables.




[11] En Excel necesitarás tener activadas las "Herramientas de Análisis". Para ello debes ir al botón de Oficce o menú Archivo (dependiendo de la versión, arriba a la izquierda) → "Opciones" → "Complementos" → "Administrar complementos de Excel" → "Ir". Activar las dos "Herramientas de Análisis" y darle a "Aceptar". Está explicado en la página 4.

## Variables cuantitativas – Datos normales

**Prueba z (N>30)**

- Excel: Datos→Análisis de datos→Prueba z para medias de dos muestras. Pra contrastar igualdad de medias hay que indicar "Diferencia hipotética entre las medias igual a 0. Pág. 19. Hipótesis nula es igualdad entre muestras, por tanto se cumplirá si el valor de P es mayor a 0,05. También si el valor de z es menor que el valor de z crítico.

**Prueba t (Para todo N)**

- Excel: Datos→Análisis de datos→Prueba t. Hay tres opciones de prueba t, dependiendo del valor de la varianza o si se trata de muestras emparejadas. De nuevo la hipótesis nula es igualdad entre muestras que se cumplirá si P es mayor que 0,05 o si el valor de nuestra t es menor que el t crítico. Pág. 21.
- SPSS: Analizar→Comparar medias→prueba t para muestras independientes. Hay qu eindicar la variable de agrupación que es la que diferencia unos casos de otros y en "opciones" podremos indicar el intervalo de confianza. Devuelve la prueba de Levenne para arianzas iguales que se cumple si la "Sig." es mayor que 0,05. Esto nos dirá qué columna debemos mirar. Como siempre, la hipótesis es igualdad de muestras, que se cumple si la "Sig." es mayor que 0,05. Pág. 22.
- Para una sola muestra (SPSS) compara mi media con la media de la población (supuesta). El valor de prueba es la media de nuestra muestra. Si la "Sig." es mayor que p, entonces nuesta media es una buena estimación de la población. Pág. 23.

**ANOVA**

- Excel: Datos→Análisis de datos→Análisis de la Varianza de un factor. Pág. 24. O bien miramos al valor de la probabilidad y si es mayor que nuestra p se cumplirá la hiipótesis de igualdad entre las muestras. También si F es menor que F crítico, las muetras serán iguales.
- SPSS: Analizar→Comparar medias→ANOVA de un factor. Pág. 25. Indicaremos el factor que determina los diferentes grupos y en "Post Hoc" podremos indicar un post-test que nos ayude a identificar las muestras diferentes. Si el valor de "Sig." es mayor que 0,05, las muestras son iguales.

## Variables cuantitativas – Datos no normales

**U de Mann-Whitney (dos muestras)**

- SPSS: Analizar→Pruebas no paramétricas→2 muestras independientes. Pág. 28. La variable de agrupación indicará los grupos a comprar. Como siempre, hipótesis nula que no existen diferencias que se verifica si "Sig." es mayor que 0,05.

**Kruskal-Wallis (más de dos muestras)**

- SPSS: Analizar→Pruebas no paramétricas→K muestras independicentes. Pág. 29. Si "Sig." es mayor que 0,05, entonces son iguales.

## Variables cualitativas

**Chi-Cuadrado**

- Excel: Se hace casi a mano. Pág. 31. Calculamos "**=PRUEBA.CHI(rango observado; rango esperado)"** con las tablas de valores medidos y la de valores esperados que calculamos nosotros. Con "**=PRUEBA.CHI.INV(probabilidad; grados de libertad)**" calculamos el valor de la chi -2de la probabilidad que calculamos en el paso anterior; los grados de libertad son (columnas-1)x(filas-1). Finalmente con "**=PRUEBA.CHI.INV(probabilidad; grados de libertad)**" calculamos el valor de chi-crítico para una probabilidad de 0,05. Si nuestro valor de chi-2 es menor que el de chi-crítico, entonces no hay diferencias estadísticamente significativas.
- SPSS: Hay que disponer de los datos en crudo, no de la tabla de contingencia. Pág. 32. Analizar→Estadísticos descriptivos→Estadísticos descriptivos o Tablas de contingencia. En Estadísticos indicaremos "chi-.cuadrado". Como siempre, si el valor de "Sig." es mayor que 0,05 entonces no existen diferencias.

## Correlación y Regresión

**Correlación**

- Excel: Representamos las dos variables de estudio, dispresión una frente a la otra Pág. 33. Botón derecho sobre los puntos y "Agregar línea de tendencia" marcando las opciones de morar la ecuación y el valor de R cuadrado. Así tenemos calculada también la recta de regresión con sus coeficientes. R cuadrado nos dice el porcentaje de la relación explicada. Su raiz es el coeficiente de correlación de Pearson que también podemos calcular con "**=COEF.DE.CORREL(matriz1, matriz2)**". Cuanto más próximo a 1, más correlación. También se puede hacer en Datos→Análisis de Datos→Coeficiente de correlación, y podremos indicar varias variables a la vez y nos ofrece todas las combinaciones.
- SPSS: Analizar→Correlaciones→Correlaciones bivariadas. Pág. 34. Devuelve el coeficiente de correlación y la significación.

**Regresión**

- Excel: ver punto anterior. Se hace mediante la representación gráfica y agregando la línea de tendencia.
- SPSS: Analizar→Regresión→Elegir la regresión que deseemos. Pág. 36.

## Vídeos

Todos los ejemplos, o casi todos, los tienes resueltos en vídeo en los siguientes enlaces:

1. Estadística descriptiva en Excel: http://youtu.be/ce1Z9lqafTU
2. Estadística descriptiva en SPSS: http://youtu.be/52ztEG4vbWU
3. Test de normalidad en SPSS: http://youtu.be/_7lBiSQz3sk
4. Test t de Student en Excel: http://youtu.be/xgnAUintXWw
5. Test t de Student en SPSS: http://youtu.be/xgnAUintXWw
6. ANOVA de un factor en Excel: http://youtu.be/WIOOBna4B5g
7. ANOVA de un factor en SPSS: http://youtu.be/3EXYi4CK0ss
8. ANOVA de dos o más factores en EXCEL: http://youtu.be/xYq2NMfX4dk
9. ANOVA de dos o más factores en SPSS: http://youtu.be/NShNBWbsYVI
10. U de Mann-Whitney con SPSS: http://youtu.be/t4YVEcPgmfw
11. Kruskal-Wallis con SPSS: http://youtu.be/GwRucDOZCCQ
12. Test Chi-2 en Excel: http://youtu.be/kA-sQa5VHBc
13. Test Chi-2 en SPSS: http://youtu.be/OMNfGehb02Q

Todos los vídeos en este canal de YouTube: http://goo.gl/tjAkkL