Minería de Datos, 2008-2009
Repaso de Estadística Descriptiva con Excel

Introducción a Excel como herramienta estadística

1. Como es sabido, Excel es una hoja de cálculo. Como tal, no es un paquete estadístico especializado y, en principio, los cálculos estadísticos con Excel son un poco más laboriosos que con uno de esos paquetes. Sin embargo, Excel es útil pues posee funciones estadísticas, gráficas y asistentes que facilitan la realización de los cálculos y gráficos necesarios.

2. **Ayuda**. Excel dispone de una ayuda (?) en la que se puede consultar el propósito y sintáxis de sus funciones y, en particular, de sus funciones estadísticas y gráficas. Es conveniente acudir a esta ayuda al menos la primera vez que se va a utilizar una función.

3. **Funciones estadísticas.** Entrar en la ayuda y buscar las *Funciones estadísticas*. Si no se encuentra en la lista, utilizar el campo *Buscar* de la ayuda para localizar y repasar el contenido de las funciones de estadística descriptiva: media aritmética, media geométrica, media armónica, mediana, cuartiles, percentiles, varianza, desviación típica, cuasivarianza, cuasidesviación típica, coeficiente de asimetría y coeficiente de curtosis, covarianza, coeficiente de correlación lineal de Pearson y análisis de regresión (estimación lineal).

4. **Complemento Herramientas para análisis**. Entrar de nuevo en la ayuda y en el campo *buscar* introducir la palabra "estadística". Leer las entradas "Realización de un análisis estadístico" y "Herramientas de análisis estadístico".

    - En el menú Herramientas/Complementos seleccionar "Herramientas para análisis" si no está seleccionado.

Manejo de Datos

1. **Variables y Casos**. En cualquier trabajo estadístico por ordenador, es preciso distinguir entre "variables" y "casos". Las variables se disponen en **columnas** y los casos en **filas**. Por tanto, una columna es el conjunto de valores que toma una sóla variable en la muestra de datos. Por otra parte, una fila de datos representa un caso, es decir, el conjunto de los valores que las diferentes variables definidas, toman en uno de los individuos que es objeto de estudio.

2. Abrir el Fichero de datos *ordenadores.xls* que se facilita junto con este boletín y guardarlo con el nombre *ordenadores1.xls* en la carpeta de trabajo.

   El fichero contiene datos de 20 ordenadores: tipo (portátil o sobremesa), origen (Nacional, Extranejero o Mixto), marca de su procesador, frecuencia de reloj, memoria RAM, tiempo que tarda en ejecutar una tarea, el tiempo que tarda en ejecutar una segunda tarea, y valoración media asignada por un conjunto de usuarios.

3. **Ventanas**. En el menú Ventana/Nueva ventana crear una nueva ventana.

- Organizar las dos ventanas existentes mediante el menú Ventana/Organizar/Horizontal seleccionando "Ventanas del libro activo".
- En la ventana inferior, situarse en la Hoja2 y denominarla "Descriptiva".
  Para una mejor organización del trabajo, es conveniente que los resultados se obtengan en esta hoja en dirección descendente. Pueden realizarse títulos y anotaciones a los mismos a medida que se van obteniendo.

4. En la hoja Datos (ventana superior), añadir un nuevo ordenador con los siguientes valores:

tipo	origen	proc	clock	ram	tiempo1	tiempo2	valoración
portátil	Nacional	0	1.0	64	12.5	16.4	3.2

5. **Ordenar**. Mediante Datos/Ordenar ordenar los ordenadores según la frecuencia de reloj de forma ascendente (indicar que "el rango de datos tiene fila de encabezamiento" y comprobar qué sucede si no es así).

6. **Filtros**. En la ventana Datos, seleccionar menú Datos/Filtro/Autofiltro. Posteriormente, en el selector de la variable *tipo*, seleccionar los ordenadores de sobremesa.

   (a) Para los ordenadores de sobremesa, calcular en la hoja Descriptiva la media artimética, varianza, desviación típica, cuasivarianza y cuasidesviación típica de la variable *ram*. Obsérvese la diferencia entre la varianza y la cuasivarianza (y entre la desviación y la cuasidesviación). Deshacer finalmente el filtro.

   (b) Repetir los cálculos anteriores ordenando previamente los ordenadores según su tipo (portátil/sobremesa). Comparar los resultados y decidir cuáles son los correctos y por qué.

7. **Subtotales**. Vamos a comparar algunas características según la marca de procesador. Antes, ordenar los datos según tipo de procesador.

   (a) Mediante Datos/Subtotales obtener la media de todas las variables cuantitativas dentro de cada marca de procesador. Pinchar en los signos menos del margen y ver el resultado. Para cada variable, comparar la media total con la media de cada grupo.

   (b) Repetir la acción anterior, pero ahora obteniendo la desviación típica (quitar selección "Reemplazar subtotales actuales"). Comparar las desviaciones totales de las variables *tiempo1* y *tiempo2* con las obtenidas dentro de cada grupo.

   (c) Eliminar los subtotales creados mediante el botón *Quitar todos* del menú Datos/Subtotales.

   (d) Obtener el número (la "cuenta") de ordenadores de cada marca. Después de ver el resultado, eliminar el subtotal creado.

8. Eliminar todos los subtotales creados y obtener la media, desviación típica y coeficiente de variacion de Pearson para las variables $clock$, ram y $tiempo1$. Comparar el grado de dispersión de las tres variables.

9. Crear una nueva variable denominada *calificación* cuyos valores sean resultado de discretizar la variable *valoración* de la siguiente forma:

$valoración$	< 5	$[5, 7)$	≥ 7
$calificación$	1	2	3

Estadística Descriptiva I

Distribuciones Unidimensionales

1. Realizar un análisis estadístico descriptivo de la variable *ram*.

   (a) Histograma. Queremos construir un histograma de frecuencias acumuladas y no acumuladas para la variable *ram* con los siguientes intervalos: $(-\infty, 32]$, $(32, 128]$, $(128, 512]$, $(512, 1024]$, $(1024, \infty)$.

      1. En la hoja Descriptiva, escribir el título "Intervalo" y debajo los valores 32, 128, 512 y 1024.
      2. Mediante Herramientas-Análisis de Datos-Histograma, especificar: el rango de celdas donde está la variable *ram* (incluyendo el rótulo), el rango de celdas donde se definieron las clases (los intervalos) en el punto anterior (incluyendo el título), seleccionar "Rótulos", elegir la celda de la hoja Descriptiva donde se ubicarán los resultados y elegir Crear gráfico.
      3. Repetir el punto anterior, pero ahora seleccionando además la opción *Porcentaje acumulado*.
      4. Con las frecuencias no acumuladas obtenidas anteriormente, realizar un gráfico circular (tarta) para la variable *ram*.

   (b) Mediante Herramientas-Análisis de Datos-Estadística Descriptiva, determinar el rango de celdas de la variable *ram* (incluyendo el rótulo), seleccionar "rótulos en la primera fila", señalar una celda en la hoja Descriptiva donde se impriman los resultados y elegir Resumen de Estadísticas.

      1. Comparar las medidas de posición (media, mediana y moda).
      2. Analizar la simetría y curtosis de la distribución de la variable *ram*.

Distribuciones Bidimensionales. Informe de tablas y gráficos dinámicos

1. Veamos la distribución conjunta de las variables *tipo* y *proc*. (Los procedimientos indicados son válidos para Excel 2003; con otras versiones investigar el método).

   (a) Mediante Datos/Informe para tablas y gráficos dinámicos realizar las siguientes acciones (pulsar siguiente después de cada punto): 1) dejar las opciones como están, 2) seleccionar como rango toda la tabla de datos, 3) situar el informe en una celda con espacio suficiente a su derecha y abajo, 4) arrastrar la variable *tipo* como variable columna, arrastrar la variable *proc* como variable fila y arrastrar la variable *tipo* al centro de la tabla.

   (b) Para ver los datos en frecuencias relativas (%) de cada tipo de procesador, pinchar con el botón derecho dentro de la tabla y seleccionar *Configuración de campo* (o bien, hacer doble click en el campo *Cuenta de tipo*) y en *Opciones*, seleccionar *Mostrar datos como:* % de la fila.

2. A partir de la tabla dinámica anterior, realizar un histograma en 3D para la distribución conjunta $(tipo, proc)$ de la siguiente forma:

   (a) Obtener un gráfico dinámico (diagrama de barras compuesto) pinchando con el botón derecho dentro de la tabla y seleccionando *Gráfico dinámico*.

   Identificar gráficamente la distribución marginal de $proc$ y, manipulando las leyendas, las distribuciones condicionadas: $tipo|proc = 0$ y $proc|tipo = 1$.

(b) Obtener un histograma en 3D pulsando en el icono gráficos, seleccionar el gráfico de columnas en 3D y seleccionar las opciones deseadas en el asistente de gráficos hasta finalizar.

- Una vez realizado, pinchar con el botón derecho en el fondo blanco del gráfico y seleccionar *vista en 3D* para recolocar el gráfico a gusto. Interpretar el histograma a la vista de la tabla de frecuencias original.
- Identificar en el gráfico las distribuciones condicionadas del punto anterior.

3. Comparemos algunas características de las variable *tiempo1* y $tiempo2$ condicionadas a las variables *tipo* y *proc*.

(a) Con el botón derecho pinchar dentro de la tabla y seleccionar *Ocultar*. (O bien, pinchar en el campo *Cuenta de proc* de la esquina superior izquierda de la tabla y arrastrarlo fuera de la tabla).

(b) Arrastrar *tiempo1* al centro de la tabla (pinchar en la tabla si no se ven las variables). Pinchar con el botón derecho en el contenido de la tabla y seleccionar *Configuración de campo* y *promedio*. Comparar las medias de *tiempo1* para ambos tipo de procesador según que los ordenadores sean portátiles o sobremesa.

(c) Hacer una copia de la tabla y pegarla un poco más abajo. En esta segunda tabla, pinchar con el botón derecho en el contenido de la tabla y seleccionar *Ocultar*. (O seleccionar el campo *Promedio de tiempo1* de la esquina superior izquierda y arrastrar fuera de la tabla).

(d) Arrastrar $tiempo2$ al centro de la tabla y seleccionar *Configuración de campo* y *promedio*.

(e) Comparar los resultados de $tiempo1$ y $tiempo2$ de las dos tablas.

(f) Repetir la acción anterior para las varianzas, el máximo y el mínimo, en lugar del promedio.

4. Construyendo una tabla dinámica, responder a las siguientes preguntas sobre las variables $proc$ y ram:

(a) Identificar las siguientes distribuciones:

1. **Conjunta**: la distribución de $(proc, ram)$.
2. **Marginales**: la distribución de $proc$ y la distribución de ram (para cada una, escribir en un papel la variable de que se trate, sus valores y sus frecuencias).
3. Las siguientes distribuciones **condicionadas**: la de $ram|proc = 0$ y la de $proc|ram = 256$. (para cada una, escribir en un papel la variable de que se trate, sus valores y sus frecuencias).

(b) ¿Cuál es la frecuencia absoluta y relativa de $(proc = 1, ram = 128)$?.

(c) ¿Cuál es la frecuencia absoluta y relativa de $proc = 0$?. ¿Y de $ram = 32$?.

(d) ¿Cuál es la frecuencia relativa de $proc = 1$ condicionado a $ram = 64$?. ¿Y la de $ram = 256$ condicionada a $proc = 0$?.

5. A partir de la tabla dinámica anterior, realizar un histograma en 3D para la distribución conjunta $(proc, ram)$ e identificar en el gráfico las respuestas anteriores.

6. Por último, se nos comunica que el ordenador sobremesa de tipo 2 con 2.8Ghz, tiene realmente 2048Gb de ram en lugar de los 1024 que se registraron por error. Modificar el dato y pichando en la tabla con el botón derecho seleccionar *Actualizar datos*. Observad el cambio en la tabla y en el gráfico dinámicos.

7. **Ejercicio**: Identificar y graficar la distribución conjunta de la variable $(tipo, calificación)$. Finalmente, identificar las distribuciones marginales y condicionadas.

8. Distribución conjunta de tres variables.

   (a) Construir una tabla dinámica con las variables $tipo$ en filas y $proc$ en columnas.

   (b) Arrastrar por ejemplo $proc$ como contenido de la tabla y obtener las frecuencias (operación *Cuenta*).

   (c) Arrastrar $calificación$ a la cabecera de las columnas, debajo de $proc$.

   (d) Obtener las frecuencias relativas (mostrar *Cuenta* como porcentaje del total) y determinar:

      1. La distribución de: $calificación|(proc = 1, tipo = sobremesa)$ y de $calificación|(proc = 0, tipo = sobremesa)$.
      2. A juzgar por la muestra disponible, qué marca de procesador es más valorado en portátiles. Justificar.

   (e) Obtener con la tabla el *Promedio* de *clock* en lugar de la *Cuenta* de *proc* e interpretar los resultados. (Notar que no tiene sentido mostrar el promedio en porcentaje).

9. **Ejercicio:** Identificar y graficar la distribución conjunta de $(tipo, proc, origen)$. Obtener la valoración media de los ordenadores en cada una de las categorías definidas por la variable tridimensional $(tipo, proc, origen)$ graficándola mediante un gráfico dinámico de diagrama de barras tridimensional.

Estadística Descriptiva II

Asociación entre Variables Cuantitativas. Regresión

1. Mediante Herramientas/Análisis de datos seleccionar *Covarianza* y posteriormente *Coeficiente de correlación* para obtener la matriz de covarianzas y la matriz de correlaciones con el fin de analizar la relación lineal entre las variables $clock$, ram, $tiempo1$, $tiempo2$ y $valoración$.

2. Realizar un gráfico radial (Excel 2007) con las variables $tiempo1$, $tiempo2$ y $valoración$. Interpretar.

3. Analicemos la relación entre las variables *clock* y *valoración.*

   (a) Realizar un gráfico de dispersión entre ambas variables, tomando $clock$ como variable x. Para ello, es mejor seleccionar antes las dos variables (columnas) incluyendo su rótulo y pinchar después en el icono de gráficos.

   (b) Una vez realizado, pinchar con el botón derecho en los puntos y seleccionar *Agregar línea de tendencia/Lineal* y en *Opciones* seleccionar mostrar la ecuación y el coeficiente R^2 en el gráfico.

   (c) Interpretar la ecuación y el coeficiente de determinación R^2. ¿Cómo aumenta la valoración por cada Mb de RAM adicional?. ¿Qué podemos decir de la bondad del ajuste?.

   (d) Modificar las diferentes opciones estéticas del gráfico según preferencias (título, nombre de las variables en ambos ejes, etc).

(e) Evaluar las diferencias en la ecuación de regresión y en el R^2 cuando se impone que el valor de intersección con el eje Y sea cero.

(f) Añadiendo nuevas líneas de tendencia con distintos colores, realizar sucesivamente ajustes de tipo polinómico de grado 2, 4 y 6 evaluando el coeficiente R^2. Reflexionar sobre cuál será la ecuación que mejor refleja la verdadera relación entre $clock$ y $valoración$.

4. Realizamos un análisis de regresión más detallado entre las variables $clock$ y $valoración$. En Herramientas-Análisis de Datos-Regresión, 1) especificar el rango de celdas de la variable y ($valoración$) y de la variable x ($clock$), 2) señalar Rótulos para que el programa sepa que en la primera fila están los nombres de las variables, 3) señalar una celda de la hoja Descriptiva donde aparecerán los resultados, 4) señalar *Gráfico de residuales* y *Curva de regresión ajustada*.

5. Comparemos las regresiones de $clock$ frente a $valoración$ para cada tipo de ordenador (portátil o sobremesa).

   (a) Ordenar los datos según tipo.

   (b) En el gráfico obtenido en el ejercicio 3b borrar todas las curvas ajustadas, sus ecuaciones y R^2, salvo lo relativo al ajuste lineal con constante.

   (c) Pinchando con el botón derecho en el marco del gráfico, seleccionar *Datos de origen/Serie* y agregar una serie denominada *portátiles* y otra serie denominada *sobremesa* especificando el rango donde se encuentran los valores de X y de Y para cada una.

   (d) Para cada una de las dos nuevas nubes de puntos, pinchar con el botón derecho y seleccionar *Agregar línea de tendencia* y en *Opciones*, mostrar la ecuación y R^2 para cada una. Comparar los resultados.

6. Realizar un análisis de regresión para los tres juegos de datos siguientes, interpretando y comparando los resultados:

x	-1	-0.25	0	0.5	1
y	1	0.25	0	0.25	1

;

x	-1	-0.25	0	0.5	1
y	-2	-1	0	1	2

;

x	0	0.5	1	1.5	2
y	0	0.25	1	2.25	4

7. Crear una nueva variable $tipon$ con dos valores (portátil = 1 y sobremesa = 0) y realizar un gráfico de dispersión con línea de tendencia lineal entre la variable $tiempo1$ como regresora y $tipon$ como respuesta. Utilizando el modelo obtenido, proponer un método para predecir el tipo de un nuevo ordenador cuando nos facilitan su valor en $tiempo1$ justificando la propuesta.

0.0.1 Asociación entre Variable Cuantitativa y Variable Cualitativa

1. Estudiemos la posible relación entre el grado de "españolidad" de los ordenadores y su valoración. Recodificar la variable $origen$ en una nueva variable $origen2$ con la codificación (0: extranjero, 1 : mixto y 2 : español)

   (a) Mediante un gráfico de dispersión realizar una regresión entre origen recodificado y la valoración, estudiando la bondad del ajuste.

   (b) Realizar la regresión por grupos (p. ej. utilizando tablas dinámicas), obtener el coeficiente η^2 y compararlo con el R^2 obtenido anteriormente.

Asociación entre Variables Cualitativas

1. La siguiente tabla de contigencia recoge los resultados de una prueba de resistencia realizada a 305 dispositivos de 3 clases diferentes A, B y C:

Resultado\Dispositivo	A	B	C
Averiados	100	25	40
No averiados	70	40	30

   Determinar si, según esa prueba, los tres tipos de dispositivo pueden considerarse igual de resistentes. Utilizar la función *prueba.chi* que calcula el coeficiente χ^2 (consultar esta función en la ayuda). Calcular el coeficiente de Contingencia de Pearson o el V de Cramer.