ESTADÍSTICA APLICADA
2^0 curso de la diplomatura en Nutrición Humana y Dietética
Curso 2006-2007

NOTAS SOBRE LA UTILIZACIÓN DEL *EXCEL* EN CÁLCULOS ESTADÍSTICOS

1. Introducción

El Excel es una hoja de cálculo y base de datos (asociada al entorno Windows) que presenta una gran versatilidad y un gran número de aplicaciones.

Cuando se abre el programa, la pantalla principal aparece organizada en una serie de columnas divididas en celdas. En cada una de éstas se puede poner un número (que será lo más habitual en nuestro caso) o información alfanumérica. Típicamente cada columna se contempla como una "variable" que corresponde al conjunto de valores de una magnitud que se ha medido en un experimento. Por ejemplo, la cantidad diaria de calorías que consumen las personas seleccionadas como muestra en un estudio dietético.

En la fila superior de la pantalla principal aparece un menú con diferentes opciones. Puede observarse que algunas de ellas (Archivo, Edición, Ver, Formato...) son comunes a muchos programas basados en Windows.

En particular, las opciones Insertar y Herramientas son muy útiles para el análisis estadístico.

2. Cómo cargar los datos de un fichero

Tomemos por ejemplo el fichero de datos `Mercurio.txt`, correspondiente al problema 2 de la relación de ejercicios propuestos.

Se sigue la siguiente ruta:

Archivo $\hookrightarrow$ Abrir $\hookrightarrow$ Se selecciona el fichero.

A continuación aparece un *Asistente para importar texto*, con tres cuadros de diálogo. Seleccionar las opciones

Delimitados $\hookrightarrow$ Siguiente $\hookrightarrow$ Espacio $\hookrightarrow$ Siguiente $\hookrightarrow$ Finalizar.

También se puede utilizar la ruta alternativa que comienza con

Datos $\hookrightarrow$ Obtener datos externos.

3. Cómo calcular la media, varianza, etc.

El procedimiento básico es "insertar funciones".

Por ejemplo para calcular la media de los niveles de contaminación por mercurio en los 171 peces del problema 2, se sitúa el cursor en una celda que no esté ocupada por ningún dato y se sigue la ruta

Insertar $\hookrightarrow$ Función $\hookrightarrow$ Estadísticas $\hookrightarrow$ Promedio $\hookrightarrow$ Aceptar.

En el recuadro que aparece a continuación, donde pone "Número 1", poner F1:F171 (suponiendo que F fuera el nombre de la columna donde están los datos cuya media nos interesa calcular).

A continuación hacer clic en Aceptar. La media de los 171 números que ocupan los primeros lugares de la columna F aparecerá entonces en la casilla donde se hubiese situado el cursor al comienzo del procedimiento.

De manera análoga se calcularía la varianza y otros estadísticos de interés.

4. Cómo realizar análisis estadísticos más completos

Además del anterior procedimiento (basado en la opción de "insertar funciones"), el Excel ofrece otra opción más avanzada y completa que, básicamente, constituye un pequeño paquete de software estadístico. Tiene un alcance más limitado que otros programas comerciales (como el SPSS o el SAS) pero permite realizar fácilmente la mayoría de los análisis estadísticos elementales más frecuentes en la práctica.

Esta opción se activa a través de la ruta

Herramientas ↪ Complementos ↪ Herramientas para análisis ↪ Aceptar.

Esta activación hay que realizarla sólo una vez. Es posible que ya esté hecha en los ordenadores del laboratorio. Después de activada la opción se puede acceder a un menú estadístico a través de la ruta

Herramientas ↪ Análisis de datos

A continuación se selecciona (en el cuadro que aparece en pantalla) la opción deseada.

Por ejemplo, para realizar una análisis descriptivo bastante completo (calculando, al mismo tiempo, la media, la desviación típica, la varianza, la mediana, etc.) de 20 datos situados en la columna A, se activa la ruta

Herramientas ↪ Análisis de datos Estadística descriptiva ↪ Aceptar

En el cuadro de diálogo que aparece se indica en "Rango de entrada", cuáles son los datos que se quieren analizar. En este caso se pondría A1:A20 en "Rango de entrada". En el mismo cuadro de diálogo hay que seleccionar "Resumen de estadísticas" y se indica también "En una hoja nueva" si se desea que los resultados aparezcan en una nueva "hoja" o pantalla nueva del programa Excel. Pinchar en Aceptar. Inmediatamente aparece en pantalla un rectángulo sombreado que incluye todos los resultados del análisis.

5. Gráficos con Excel

Se pueden obtener fácilmente varios tipos de gráficos a través de la ruta Insertar ↪ Gráfico.

Por ejemplo, para obtener el diagrama de dispersión del par de variables (longitud, peso) en el fichero `Mercurio.txt`, procederemos así:

- Seleccionar las dos columnas que contienen las variables cuyo diagrama de dispersión interesa representar. Esto se consigue haciendo clic sobre la casilla que contiene los respectivos nombres (manteniendo la tecla Ctrl pulsada, si no se trata de columnas consecutivas).

- Seguir la ruta Insertar ↪ Gráfico ↪ Dispersión.

- En el menú con muestras de gráficos que aparece a la derecha se elige la opción deseada. Por ejemplo "Dispersión. Compara pares de valores" o bien "Dispersión con puntos de datos conectados por líneas". Elijamos, por ejemplo, la primera de estas opciones y hagamos clic en Siguiente.

- En el cuadro de diálogo que aparece a continuación puede verse el gráfico en pequeño, junto con el rango de datos para el que se ha obtenido. Dicho rango puede cambiarse ahora, en caso necesario.

- Al pulsar Siguiente se accede a otro cuadro de diálogo en el que pueden elegirse diferentes opciones acerca de la rotulación y el formato de los ejes.

- Haciendo nuevamente clic en Siguiente se obtiene un último cuadro de diálogo en el que se puede elegir si se desea que el gráfico se guarde en la misma "hoja" del programa Excel en la que figuran los datos (el gráfico aparecería como un recuadro superpuesto que se puede "cortar", "pegar" y "mover"), o bien se incluiría en una nueva "hoja", a la que sólo se accedería seleccionando la correspondiente pestaña en la parte inferior izquierda de la pantalla. Una vez realizada la selección hacer clic en Finalizar.

- A continuación se puede editar el gráfico para introducir en él diversas modificaciones. Esto puede conseguirse haciendo doble clic sobre la parte del gráfico que se desea modificar. Por ejemplo, para cambiar el color del fondo, se hace doble clic sobre una zona del fondo en la que no haya otros elementos (como puntos del gráfico o líneas de división) y se accede a un cuadro de diálogo que da la opción de elegir el color deseado.

- También se pueden incorporar diferentes cambios y opciones suplementarias siguiendo las rutas

  Gráfico $\hookrightarrow$ Opciones de gráfico y Gráfico $\hookrightarrow$ Agregar línea de tendencia.

  Con esta última opción se puede, en particular, visualizar la recta de ajuste por mínimos cuadrados.

Hay alguna otra opción gráfica a la que se puede acceder a través de la ruta
Herramientas $\hookrightarrow$ Análisis de datos
(para ello situar el cursor en una celda libre de la hoja de trabajo, fuera del área de cualquier gráfico que se hubiera creado previamente). Por ejemplo, se puede obtener un histograma a través de Herramientas $\hookrightarrow$ Análisis de datos $\hookrightarrow$ Histograma $\hookrightarrow$ Aceptar. En el recuadro que aparece hay que seleccionar la opción "Crear gráfico" e indicar (en "Rango de entrada") los lugares ocupados por los datos para los que se desea obtener el histograma (por ejemplo, F2:F150). En "Rango de clases" se pueden indicar los valores extremos de los intervalos que definen las clases del histograma. Si no se pone nada en este recuadro el programa asigna automáticamente un rango de clases. A continuación se hace clic en Aceptar y aparece el gráfico. Puede ocurrir que el histograma aparezca en una escala distorsionada, con las barras muy cortas. Esto se puede modificar arrastrando con el ratón la línea divisoria superior o inferior del recuadro en el que aparece el gráfico, cambiando a voluntad la escala.

6. Cómo calcular probabilidades de la distribución normal usando Excel

Supongamos, por ejemplo, que se desea calcular la probabilidad $P\,(5.7 < X \leq 6.6)$ de que una v.a. $X \sim N(5, 2)$ tome un valor comprendido entre 5.7 y 6.6.

Recordemos que $P\,(5.7 < X \leq 6.6) = P\,(X \leq 6.6) - P(X \leq 5.7)$.

Para calcular esto con el Excel se procede de la siguiente forma:

1. Se sitúa el cursor en una celda libre (sin datos) y se activa la ruta

   Insertar $\hookrightarrow$ Función $\hookrightarrow$ Estadísticas $\hookrightarrow$ DIST.NORM.

2. En el cuadro de diálogo que aparece poner el valor 6.6 en la casilla indicada con "X", y 5 y 2, respectivamente, en las casillas "Media" y "Desv. estándar". En la

casilla "Acum" poner VERDADERO. Esto último indica que se desea calcular la función de distribución "acumulada" (que es la misma que se encuentra en las tablas usuales).

3. En la celda seleccionada aparecerá la cantidad 0.788144666 que corresponde a la probabilidad $P(X \leq 6.6)$.

4. A continuación se va a la línea de comandos, donde debe aparecer la expresión =DISTR.NORM(6,6;5;2;VERDADERO) y, detrás de esta expresión, se pone un signo menos y se repite la operación anterior para insertar de nuevo la misma función, cambiando el valor 6.6 por 5.7. En este caso puede resultar más cómodo usar la opción Usadas recientemente, en lugar de Estadísticas, ya que de este modo aparecerá en primer lugar la función DIST.NORM y no será necesario buscarla en la lista de funciones estadísticas.

5. Como resultado final del procedimiento anterior, debe aparecer en la celda seleccionada el número 0.151314076 que corresponde a $P(5.7 < X \leq 6.6)$.

Obsérvese que no es necesario "estandarizar", es decir, pasar a la distribución $N(0,1)$ (lo hace el programa automáticamente) ya que la función DIST.NORM permite seleccionar la media y la desviación típica deseadas. Existe también otra función, llamada DIST.NORM.ESTAND que corresponde a la distribución "estandarizada" $N(0,1)$. Por supuesto, cuando se utiliza esta función ya no es necesario especificar la media ni la desviación típica.

7. Intervalos de confianza

El intervalo de confianza para la media μ de una normal $N(\mu, \sigma)$ (donde σ no se supone conocida) a partir de una muestra $x_1, \ldots, x_n$ tiene la siguiente expresión

$$\bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}. \tag{1}$$

- Un intervalo de confianza de esta forma puede obtenerse con el EXCEL mediante la opción

  Herramientas $\hookrightarrow$ Análisis de datos $\hookrightarrow$ Estadística descriptiva

- En el cuadro de diálogo que aparece poner en Rango de entrada la situación de los datos (por ejemplo A1:A12). Puede seleccionarse el rango también poniendo el cursor en el recuadro de "Rango" y marcando los datos en la pantalla con el ratón (manteniendo el botón izquierdo pulsado).

- A continuación seleccionar las casillas

  Resumen de estadísticas y Nivel de confianza para la media.

  Junto a esta última aparece un recuadro en el que se indica como opción por defecto 95%. Esto significa que el intervalo que obtendrá el programa (después de pinchar en Aceptar) tendrá un nivel de 0.95.

- Este intervalo aparece indicado en la línea final del cuadro de resultados: concretamente, el número que aparece junto a la (impropia) indicación

  Nivel de confianza para la media (95%)

  es $t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}$ (si se ha elegido 95% en el cuadro de diálogo, se tendrá $\alpha = 0.05$). Por tanto, el intervalo completo (**??**) se obtendría a partir de este valor, y del valor de la media muestral que figura en la primera línea del cuadro de resultados.

- Si se desea, por ejemplo, obtener un intervalo de nivel de confianza 0.99 habría que poner 99% en el recuadro que hay junto a la casilla Nivel de confianza para la media.

8. Contraste de la t de Student para la comparación de dos medias, en el caso de muestras independientes, suponiendo que las varianzas de las dos poblaciones son iguales

A) Contraste bilateral

Supongamos en primer lugar que se desea contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$.

- En este caso, para realizar el test se sigue la ruta

  Herramientas $\hookrightarrow$ Análisis de datos $\hookrightarrow$

  Prueba t para dos muestras suponiendo varianzas iguales

- En el cuadro de diálogo que aparece se indican los lugares en que están situadas ambas muestras, por ejemplo A1:A10 y B1:B10, en las casillas "Rango para la variable 1" y "Rango para la variable 2".

- En el recuadro Alfa indicar el valor deseado del nivel de significación (por ejemplo 0.05).

- En principio, no poner nada en el recuadro Diferencia hipotética entre las variables

- En el cuadro de salida se incluye diversa información sobre ambas muestras (medias, varianzas,...) y, en los lugares séptimo y undécimo respectivamente, aparece Estadístico t y Valor crítico de t (dos colas). Si el primero de estos valores es mayor que el segundo, debe rechazarse (al nivel de significación elegido) la hipótesis $H_0 : \mu_1 = \mu_2$. En caso contrario, debe aceptarse.

B) Contraste unilateral

Supongamos ahora que se desea ver si hay suficiente evidencia estadística (al nivel de significación elegido) a favor de la hipótesis "unilateral" $H_1 : \mu_1 < \mu_2$ o, lo que es equivalente,se quiere contrastar la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a la alternativa $H_1 : \mu_1 < \mu_2$.

- Si se ha obtenido $\bar{x} \geq \bar{y}$ es obvio que no se tiene ninguna evidencia estadística a favor de $H_1 : \mu_1 < \mu_2$ y, por tanto, debe aceptarse $H_0 : \mu_1 \geq \mu_2$.

- Si, por el contrario, se tiene $\bar{x} < \bar{y}$, debe aceptarse $H_1 : \mu_1 < \mu_2$, al nivel prefijado, si el valor que aparece junto a Estadístico t es mayor que el correspondiente a Valor crítico de t (una cola).

C) p-valores

Por último, hay que hacer notar que el cuadro de resultados proporciona también los **p-valores** correspondientes a los contrastes bilateral y unilateral. Estos p-valores aparecen, respectivamente, junto a los recuadros indicados con P(T<=t) dos colas y P(T<=t) una cola. Recordemos que el p-valor es el **mínimo valor del nivel de significación para el cual se rechazaría una hipótesis nula con unos datos determinados**. Puede decirse que el p-valor mide el grado de evidencia estadística que se ha obtenido a favor de la hipótesis alternativa (o en contra de la hipótesis nula): cuando el p-valor es "muy pequeño" (digamos, menor que 0.01) se considera que se ha obtenido una fuerte evidencia estadística a favor de H_1 (o en contra de H_0). Cuando el p-valor está entre 0.01 y 0.05 se considera que se ha obtenido una moderada evidencia estadística a favor de H_1. Si el p-valor es "grande" se considera que no se ha obtenido suficiente evidencia estadística a favor de H_1. En resumen, si se conoce el p-valor (y el EXCEL lo da en su cuadro de resultados) se conoce el resultado del test para cualquier nivel de significación prefijado: si se hubiese tomado un nivel de significación menor que el p-valor, hubiéramos aceptado H_0 (y rechazado H_1) a ese nivel; si el nivel prefijado hubiese sido mayor que el p-valor, la decisión hubiera sido contraria. Por tanto, el p-valor es la información más completa y útil en la resolución de un problema de contraste de hipótesis.

9. Contraste de la t de Student para la comparación de dos medias, en el caso de muestras independientes, cuando las varianzas de las dos poblaciones no se pueden suponer iguales

En algunos casos prácticos no es razonable suponer que ambas poblaciones tienen la misma varianza. En estas situaciones pueden contrastarse las hipótesis bilaterales o unilaterales de forma totalmente análoga a la indicada en el apartado anterior pero reemplazando el test exacto de la t de Student por un test aproximado al que se accede mediante la ruta

Herramientas $\hookrightarrow$ Análisis de datos $\hookrightarrow$
Prueba t para dos muestras suponiendo varianzas desiguales.

10. Contraste de la t de Student para la comparación de dos medias en el caso de muestras emparejadas

Se accede a este test mediante la ruta

Herramientas $\hookrightarrow$ Análisis de datos $\hookrightarrow$
Prueba t para medias de dos muestras emparejadas.

11. Crear y copiar fórmulas con EXCEL

Supongamos, por ejemplo, que tenemos 20 valores (positivos y mayores que 1) en la columna A de una hoja Excel. Supongamos que deseamos expresar estos valores en escala logarítmica (es decir, trabajar con los valores log(A1), log(A2),...) y a continuación trabajar con la serie de las diferencias log(A2)-log(A1), log(A3)-log(A2),...). El Excel

proporciona un procedimiento muy sencillo y cómodo para hacer esto, creando la fórmula deseada en una sola celda y "arrastrándola" luego a todo el rango de aplicación. Por ejemplo, vamos a situar la nueva serie en la columna B. Situemos el cursor en la celda B2 (empezamos en la segunda fila porque no tenemos ningún elemento que restarle a log(A1)) y escribamos en la línea de comandos (en la parte superior de la pantalla):

=log(A2)-log(A1)

a continuación pulsemos Intro (log indica aquí logaritmo decimal). En la celda B2 aparecerá el valor numérico de log(A2)-log(A1). Ahora, **situemos el cursor en la esquina inferior derecha de la celda B2** (el puntero se transforma en una pequeña cruz) y **arrastremos (con el botón izquierdo del ratón pulsado) hacia abajo hasta la columna B144**. En la sucesivas celdas B3, B4, se ha "actualizado" el valor de la fórmula log(A2)-log(A1), de manera que en B3 se ha calculado log(A3)-log(A2), en B4 log(A4)-log(A3), etc.