SUMA 31
junio 1999, pp. 119-124

La inferencia estadística con Microsoft Excel

Julián Sainz Ruiz

IDEAS Y RECURSOS

En este artículo se dan algunas ideas prácticas para empezar a trabajar la inferencia estadística, por ejemplo, dentro del programa de Segundo de Bachillerato de Ciencias Sociales. Lo esencial del artículo es el uso de la hoja de cálculo *Microsoft Excel* para que, mediante procesos de simulación, se comprendan las ideas de la materia. El programa permite realizar cálculo que de otra forma podrían resultar largos y complicados

Los nuevos currículos de Bachillerato proponen un enfoque inferencial de la Estadística trabajando con muestras y extrapolando los resultados. Existe gran cantidad de programas informáticos para tratar la Estadística, algunos de ellos de una potencia que desbordan los contenidos curriculares. La hoja de cálculo *Microsoft Excel* dispone de un serie de herramientas que nos pueden facilitar la investigación de fenómenos y la resolución de los problemas de una forma rápida y sencilla. La *simulación de fenómenos* a través del ordenador va a constituir una herramienta útil para trabajar la Probabilidad y la Estadística. Mediante la simulación podemos simplificar y comprender mejor el problema y crearnos pautas para la teorización del mismo mas adelante.

Las funciones para hojas de cálculo son herramientas de cálculo que ayudan a tomar decisiones, llevar a cabo acciones y ejecutar operaciones que devuelven valores automáticamente. *Microsoft Excel* ofrece una amplia gama de funciones que permiten realizar diferentes tipos de cálculo.

Para el análisis estadístico de datos *Microsoft Excel* dispone de dos elementos fundamentales:

1. El Asistente para funciones.
2. Herramientas para el análisis.

El *Asistente para funciones* simplifica la introducción de fórmulas en la barra de fórmulas. Para iniciar el Asistente para funciones, elija el comando Función del menú Insertar o utilice el botón

de la barra de herramientas. Las funciones están agrupadas por categoría, tales como «Financieras», «Matemáticas y trigonométricas» o «Estadísticas». Cuando se selecciona

una función del cuadro de lista, la definición de la función y de sus argumentos aparecerá automáticamente, así como la posición correcta de los puntos y comas (;) y paréntesis [( )].

Microsoft Excel proporciona un juego de funciones especiales para el análisis de datos denominadas *Herramientas para el análisis*. Entre dichas funciones están las de análisis estadísticos que pueden ser utilizadas en varios tipos de datos. Para utilizar estas funciones es necesario proporcionar la datos y parámetros requeridos para cada análisis de manera adecuada. El programa hace entonces los cálculos necesarios y muestra los resultados obtenidos. Antes de utilizar una herramienta para análisis, se deben organizar los datos que s desea analizar en columnas o en filas dentro de la hoja de cálculo. Dicha disposición se denomina rango de entrada. También se pueden incluir títulos de texto en la primera celda de una fila o de una columna para identificar las variables. Cuando se utilizan herramientas para analizar datos de un rango de entrada, *Microsoft Excel* creará una tabla de resultados. El contenido de la tabla de resultados depende de la herramienta para análisis utilizada. Si se han incluido títulos en el rango de entrada, *Microsoft Excel* los utilizará en la tabla de resultados; de lo contrario *Microsoft Excel* creará automáticamente títulos para los resultados que figuran en la tabla de resultados.

Para utilizar una de estas herramientas para el análisis hay que seguir los siguientes pasos:

1. Seleccione *Análisis de datos* en el menú *Herramientas*.

   Si el comando *Análisis de datos* no aparece en el menú Herramientas, ejecute el programa *Instalar* para instalar las *Herramientas para análisis*. Lo que haremos ahora será instalar un macro automático, herramientas para el análisis, que incluye *Microsoft Excel*.

2. En el cuadro *Funciones para análisis*, seleccione la función que desea utilizar.
3. Elija el botón «Aceptar».
4. Escriba el rango de entrada, de salida y demás opciones deseadas.

   Podrá insertar rangos de celdas en los cuadros «Rango de entrada» y «Rango de salida», escribiendo referencias de celda en el cuadro, o seleccionando el contenido de cada cuadro, y luego el rango de celdas de la hoja de cálculo. También podrá introducir referencias en otras hojas de cálculo en los cuadros «Rango de entrada» y «Rango de salida».

5. Elija el botón «Aceptar».

   Los resultados del análisis aparecerán en el rango designado.

Microsoft Excel proporciona un juego de funciones especiales para el análisis de datos denominadas Herramientas para el análisis. Entre dichas funciones están las de análisis estadísticos que pueden ser utilizadas en varios tipos de datos.

Analizar cada una de estas herramientas que dispone *Microsoft Excel* sería bastante interesante, pero me centraré en poner algún ejemplo para trabajar la teoría de muestras en una clase de Segundo de Bachillerato de Ciencias Sociales.

Para introducirnos en el tema vamos a utilizar la herramienta *Generación de números aleatorios* que dispone *Excel* para obtener mediante un proceso de *simulación estadística* muestras aleatorias de determinadas poblaciones de manera muy rápida. Mediante este muestreo artificial vamos a comprobar que:

- La media muestral y la desviación típica corregida son buenos estimadores de la media y desviación típica poblacional. Sin embargo, la desviación típica muestral no estima bien a la poblacional.
- Las propiedades de la media muestral.

Con la ayuda de la función *Generación de números aleatorios*, construimos una tabla que contiene siete muestras de tamaño 5 de una población que sigue una distribución Normal de media 18,1 y desviación típica 0,4. En las tres últimas filas, para cada muestra, calculamos su media, desviación típica y desviación típica corregida. Obteniéndose los siguientes resultados:

Muestra	1	2	3	4	5	6	7
	17,1672295	18,8532362	17,6147073	17,8628474	19,3898818	17,2469449	17,3774865
	17,8156865	18,1946125	18,5550002	18,0178099	17,8757564	17,4872407	18,3255938
	18,2543249	18,2539875	18,1825548	18,2368981	18,6365764	18,0120005	18,2599569
	18,3322331	18,5330395	17,8296882	18,0390793	17,8808136	18,1457821	17,7256383
	18,1588111	18,3091256	18,0939423	17,9766266	17,6959988	18,3247675	18,3609986
Media	17,945657	18,4288003	18,0551786	18,0266523	18,2958054	17,8433471	18,0099348
σ_n	0,42731565	0,2411378	0,32012806	0,12146646	0,63585459	0,40845658	0,39146208
σ_{n-1}	0,47775342	0,26960026	0,35791405	0,13580363	0,71090705	0,45666834	0,43766791

Se observa fácilmente que la media muestral es un buen estimador de la media poblacional. Se ve que la desviación típica muestral, en la mayor parte de los casos, subestima a la desviación típica de la población y que es mejor estimador de esta la desviación típica corregida. Es conveniente repetir la simulación varias veces para comprobar que los resultados son ciertos. Incluso es aconsejable hacerlo tomando muchas más muestras y estas con un tamaño mayor. Esto, con ayuda de la hoja de cálculo se hace de una forma rápida y visual.

El siguiente paso va a ser estudiar que la serie estadística de las medias de las muestras constituyen una nueva variable, que llamamos media muestral $\bar{X}$, y que sigue una distribución Normal de media la media poblacional y desviación típica la poblacional dividida por $\sqrt{n}$.

Es decir, $\bar{X}$ es

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

La media de esta serie es 18,0864822 que prácticamente podemos decir que coincide con la media poblacional.

La desviación típica de esta serie es 0,18911949. Si hacemos

$$\frac{0,4}{\sqrt{5}} = 0,178885438$$

Para ver que los datos se ajustan a una distribución Normal podemos utilizar el papel probabilístico o aplicar el test de la χ^2 de Pearson.

Es conveniente repetir la simulación varias veces para comprobar que los resultados son ciertos. Incluso es aconsejable hacerlo tomando muchas más muestras y estas con un tamaño mayor. Esto, con ayuda de la hoja de cálculo se hace de una forma rápida y visual.

Teorización del problema

En primer lugar, consideremos X la variable aleatoria de una población de tamaño *N*. Si tomamos todas las posibles muestras de tamaño *n* de dicha población y para cada una de ellas calculamos su media, ésta se comporta como una nueva variable, que denotaremos $\bar{X}$ y llamaremos *Distribución de la media muestral*; $\bar{X}$ cumple:

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Si N es lo suficientemente grande

$$\sqrt{\frac{N-n}{N-1}} \rightarrow 1$$

En general si $N \geq 30$, $\bar{X}$ es

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

Denotando

$$f = \frac{n}{N}$$

fracción de muestreo tenemos que

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{1-f}$$

Cuando N es grande la fracción de muestreo es algo que no afecta significativamente la precesión de resultados.

Ejemplo

Supongamos que 8295 es el promedio de kilómetros en 6 meses recorridos por los automovilistas de una muestra de 1000 que se extrajo de una población de 100000. Además $\sigma = 3100$ km.

$$f = \frac{n}{N} = 0,01$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{1-f} = \frac{3100}{\sqrt{1000}} \cdot \sqrt{1-0,01} = 97,5$$

Si N lo hubiéramos tomado 800000,

$$\sqrt{1-f} = \sqrt{1-0,00125}$$

y

$$\sigma_{\bar{X}} = 97,9$$

Luego para todo propósito práctico el error estándar o desviación típica es el mismo para las dos muestras, pues en ambos casos el segundo término de la ecuación se aproxima a 1.

Consideremos, nuevamente, X la variable aleatoria de una población de tamaño N. Sea p la ocurrencia de un suceso. Si tomamos todas las posibles muestras de tamaño *n* de dicha población y para cada una de ellas calculamos la ocurrencia del suceso, ésta se comporta como una nueva variable, que denotaremos $\hat{p}$ y llamaremos *distribución de las proporciones*. Esta variable cumple

$$\mu_{\hat{p}} = p; \quad \mu_{\hat{p}} = \sqrt{\frac{p \cdot q(N-n)}{n \cdot (N-1)}}$$

Para N grandes tenemos que

$$\sqrt{\frac{N-n}{N-1}} \rightarrow 1$$

En general si $N \geq 30$, es

$$N(\mu_{\hat{p}}, \sigma_{\hat{p}})$$

con

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}}$$

Este resultado también es valido para poblaciones finitas con muestreo con reposición.

Margen de error e intervalo de confianza

Supongamos que se observa una muestra de 50 personas que realizan diariamente un trayecto, y que la duración media de dicho trayecto es de 30 minutos, con una desviación estándar de la población de 2,5. Se trata de obtener un intervalo de confianza del 95% para la media de la población.

Veamos como resolver este problema.

Nos están pidiendo que encontremos *a* y *b* de modo que la media μ esté entre estos valores con probabilidad del 95%, esto es, $p(a \leq \mu \leq b) = 0{,}95$. Este intervalo será:

$$\left(\overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Llamamos *margen de error* –y lo denotamos ε– al radio de este intervalo.

Hay que observar que los extremos de este intervalo no son fijos, ya que para cada muestra que tomemos tenemos un intervalo distinto. Sin embargo μ, aunque desconocida, si lo es. Es aquí donde se ve claramente la interpretación de lo que es un intervalo de confianza, en este caso para la media de la población. Los extremos del intervalo son unas variables aleatorias que varían con la muestra tomada. Si tomáramos muchas muestras cada una de ella originaría un intervalo de confianza y el 95% de ellos contendrían, en este caso, a la media.

La función INTERVALO.CONFIANZA( ), que *Excel* suministra en su asistentes de funciones, devuelve precisamente el margen de error. Dicha función posee la siguiente sintaxis INTERVALO.CONFIANZA(alfa; desv_estándar; tamaño) donde

- *Alfa:* es el nivel de significación empleado para calcular el nivel de confianza. El nivel de confianza es igual a 100(1 – alfa)%, o sea, un alfa de 0,05 indica un nivel de confianza de 95%.
- *Desv_estándar:* es la desviación estándar de la población y se asume que es conocida.
- *Tamaño:* es el tamaño de la muestra.

Los extremos del intervalo son unas variables aleatorias que varían con la muestra tomada. Si tomáramos muchas muestras cada una de ella originaría un intervalo de confianza y el 95% de ellos contendrían, en este caso, a la media.

Para este ejemplo el radio del intervalo o margen de error es igual al valor que devuelve la función INTERVALO.CONFIANZA(0,05;2,5;50).

INTERVALO.CONFIANZA(0,05;2,5;50)= = 0,692951

Luego el intervalo de confianza al 95% para μ es:

(30 – 0,692951,30 + 06951) = (29,3 ;30,7)

En ambos casos las unidades de los resultados son, obviamente, minutos

Determinación del tamaño de una muestra

Para ilustrar este tema vamos a empezar con un caso práctico.

Una empresa esta pensando utilizar una muestra aleatoria para determinar la vida media de las bombillas que ha recibido en un embarque de 20000 bombillas. Se sabe de una muestra que se extrajo de un embarque previo que la desviación estándar de la población es de 77 horas en la vida de la bombilla. Se desea tener un nivel de confianza de 98% en la precisión del estimado y que éste no fluctúe en más de 10 horas, por encima y por debajo, del verdadero valor de la vida promedio de las bombillas. ¿Cuál debe de ser el tamaño de la muestra?

$$\overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Es el intervalo de confianza para μ, luego

$$Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 10$$

Para un nivel de confianza de 98%,

$$1 - \alpha = 0{,}98$$

tenemos que $Z_{\alpha/2} = 2{,}33$

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{10} \right) = \left(\frac{2,33 \cdot 77}{10} \right)^2 = 321,879481$$

Por consiguiente, n = 322

Utilizando la función INTERVALO.CONFIANZA(0,05;2,5;n) se construye la siguiente tabla que muestra el margen de error para distintos tamaños de la muestra.

n	ε
200	12,6662856
250	11,3290702
300	10,3419789
318	10,0450179
319	10,029261
320	10,013578
321	9,99796834
322	9,98243147
350	9,5748119
400	8,95641642

Podemos utilizar la función INTERVALO.CONFIANZA(alfa; desv_estándar; tamaño) para determinar qué tamaño tendrá la muestra para un error y nivel de confianza determinado.

α	n	ε (σ=1)	ε (σ=2,5)
0,01	25	0,5151669	1,28791726
0,05	50	0,27718035	0,69295089
0,02	75	0,26862283	0,67155707
0,1	100	0,1644853	0,41121325
0,01	150	0,21031601	0,52579002
0,05	300	0,11315841	0,28289601
0,02	600	0,09497251	0,23743128
0,1	1200	0,04748282	0,11870704

Como se puede observar al aumentar el tamaño de la muestra disminuye el margen de error.

Como se puede observar al aumentar el tamaño de la muestra disminuye el margen de error.

Muestras de tamaño pequeño

Cuando la muestra es pequeña ($n \leq 30$) y la desviación típica es desconocida es necesario estimarla por S_{n-1} (corregida) y el estadístico:

$$\frac{\bar{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}}$$

sigue una distribución *t* de Student con n-1 grados de libertad.

Excel dispone de la función DISTR.T.INV( ) que devuelve, para una probabilidad dada, el valor de la variable aleatoria siguiendo una distribución *t* de Student para los grados de libertad especificados. Su sintaxis es DISTR.T.INV(probabilidad; grados_libertad), donde:

- *Probabilidad* es la probabilidad asociada con la distribución t de Student dos colas.
- *Grados_libertad* es el número de grados de libertad para diferenciar la distribución.

Utilizando esta función se ha construido la siguiente tabla:

α	n–1	$t_{\alpha,n-1}$
0,05	2	4,30265573
0,05	4	2,77645086
0,05	6	2,44691364
0,02	2	6,96454663
0,02	4	3,74693627
0,02	6	3,14266799
0,01	2	9,92498826
0,01	4	4,60408046
0,01	6	3,70742782

La tabla siguiente, construida con la hoja *Excel*, calcula los extremos *a* y *b* del intervalo de confianza al 95% para la media poblacional, a partir de cuatro muestras de tamaño 5, generadas aleatoriamente de una población Normal N(19,2).

Muestra	1	2	3	4
	18,3007123	17,3835141	18,8507678	19,6904884
	23,392241	18,0060451	20,7934872	15,352617
	16,4613593	19,3010996	18,2100265	18,8200474
	19,8266966	17,6805551	20,5027695	19,4973526
	16,1599027	17,6719467	17,5397062	16,044406
$t_{\alpha,n-1}$= 2,77645086				
$\bar{X}$	18,8281824	18,0086321	19,1793514	17,8809823
S_{n-1}	2,95073949	0,7553547	1,42239796	2,03312888
a	15,1643478	17,0707334	17,4132074	15,3565141
b	22,4920169	18,9465308	20,9454954	20,4054504

Bibliografía

CUADRAS, C. (1999): *Problemas de Probabilidades y Estadística*, PPU.

GARCÍA, R. C: *Métodos estadísticos: Teoría y práctica*, Scott, Foresman and Co.

QUESADA, V., A. ISIDORO, y L. A. LÓPEZ (1994): *Curso y Ejercicios de Estadística*, Alhambra Universidad.

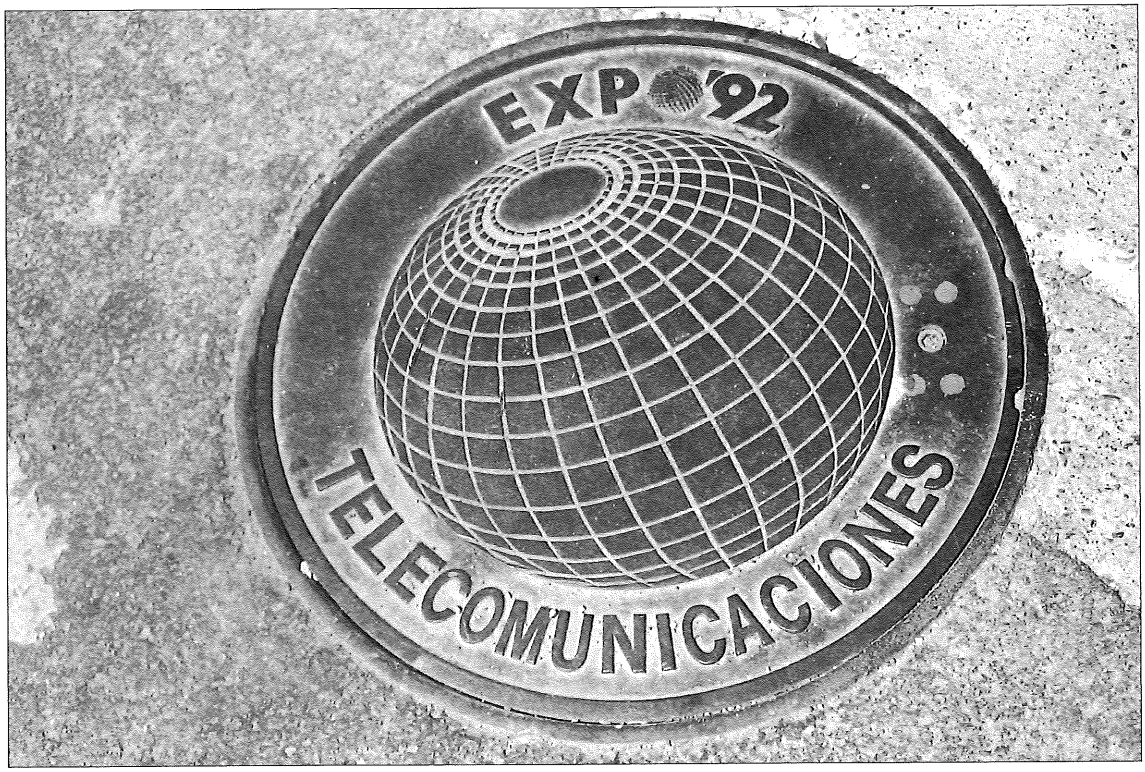
SPIEGEL, M. R. (1997): *Estadística*, McGraw-Hill.

STRUM, R. D. y E. K. DONALD (1988): *First Principles of Discrete Systems and digital Signal Processing*, Addison-Wesley Publishing Company.

VAN BUREN, C.: *Excel 5.0 para Windows*, Anaya.

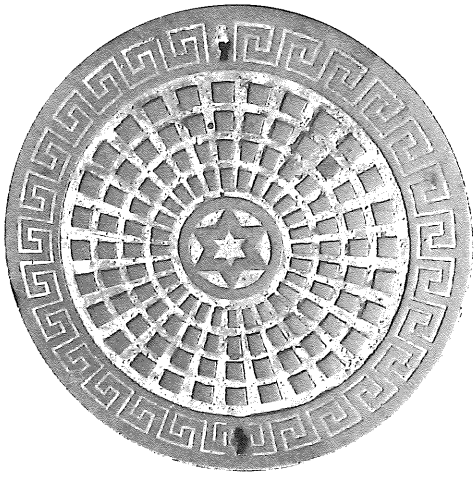
Manual del usuario de Microsoft Excel, Microsoft Corporation.

Julián Sainz
IES Santa Catalina
El Burgo de Osma (Soria)
Sociedad Castellano-Leonesa
de Profesores de Matemáticas

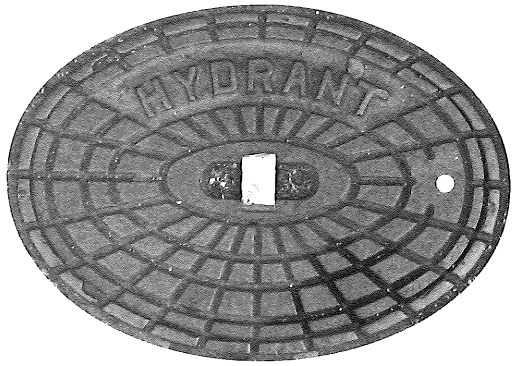


Sevilla

Fotos:
Luis Balbuena



Copenhague



Bremen