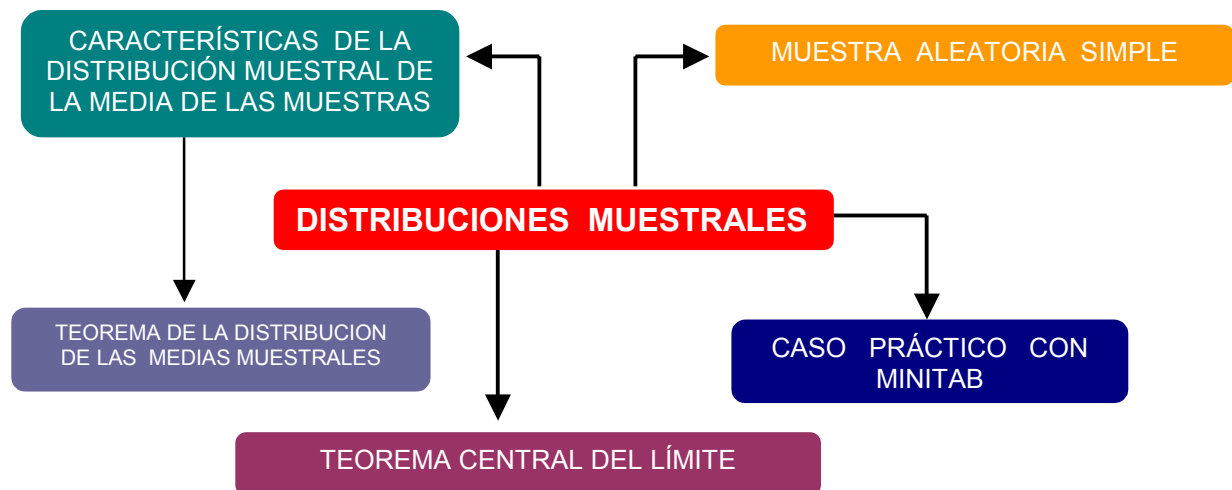


DISTRIBUCIONES MUESTRALES

Autores: Ángel A. Juan (ajuanp@uoc.edu), Máximo Sedano (msedanoh@uoc.edu), Alicia Vila (avilag@uoc.edu).

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

A menudo necesitamos estudiar las propiedades de una determinada población, pero nos encontramos con el inconveniente de que ésta es demasiado numerosa como para analizar a todos los individuos que la componen. Por tal motivo, recurrimos a extraer una muestra de la misma y a utilizar la información obtenida para hacer inferencias sobre toda la población. Estas estimaciones serán válidas sólo si la muestra tomada es “representativa” de la población.

Así, el muestreo es una técnica que utilizaremos para inferir algo respecto de una población mediante la selección de una muestra de esa población. En este math-block veremos, entre otras cosas, cómo es posible estimar la media de la población a partir de la distribución que siguen las medias de las diferentes muestras obtenidas. [2]

En muchos casos, el muestreo es la única manera de poder obtener alguna conclusión de una población, entre otras causas, por el coste económico y el tiempo empleado que supondría estudiar a todos los miembros de una población.

OBJETIVOS

- Entender la necesidad de porqué en numerosas ocasiones una muestra es la única forma factible de conocer una población.
- Explicar los métodos utilizados para seleccionar una muestra
- Entender cómo se diseña una distribución muestral para la media de la muestra
- Entender la importancia del Teorema Central del Límite, así como su aplicación

CONOCIMIENTOS PREVIOS

Sería conveniente revisar el *math-block* “La distribución normal” para tener asimilados los conceptos relacionados con las distribuciones de probabilidad y las definiciones de variables aleatorias, así como el *math-block* “La distribución binomial”, donde se introdujo el concepto de población y muestra. Por último, sería necesario consultar el manual de uso del Minitab.

CONCEPTOS FUNDAMENTALES

□ Definición muestra aleatoria simple

En principio, podríamos distinguir dos tipos de muestra: la *probabilística* y la *no probabilística*, en el sentido en que una **muestra probabilística** es una muestra seleccionada de tal forma que cada elemento de la población tiene la misma probabilidad de formar parte de la muestra.

De esta manera, si se utilizan métodos no probabilísticos, no todos los elementos de la población tienen la misma probabilidad de ser incluidos. En este caso, diríamos que los resultados están **sesgados**, lo cual quiere decir que tal vez los resultados de la muestra no sean representativos de la población.

Una forma de asegurarnos de que el subconjunto escogido es representativo de toda la población consiste en tomar una **muestra aleatoria simple**, la cual se caracteriza por:

1. Cada miembro de la población tiene la misma probabilidad de ser elegido, y
2. Las observaciones son elegidas siguiendo una secuencia aleatoria.

□ Error en el muestreo:

Tras entender la importancia de escoger una muestra representativa de la población, veamos que para lograr esto, podemos seleccionar, por ejemplo, una muestra aleatoria simple de la población, pero es muy improbable que la media de la muestra sea idéntica a la media de la población.

De la misma manera, tal vez la desviación estándar u otra medición que se calcule con base en la muestra no sea igual al valor correspondiente de la población

Por tanto, es posible que existan ciertas diferencias entre los estadísticos de la muestra (como la media o la desviación estándar), y los parámetros de población correspondientes. A dicha diferencia se la conoce como **error de muestreo**.

□ **Distribución muestral de la media de las muestras:**

Consistiría en una distribución de probabilidad de todas las medias posibles de las muestras de un tamaño de muestra dado.

Así pues, dada una población (a la cual representaremos por la v.a. X), podemos extraer de la misma k muestras, cada una de ellas de tamaño n . Para cada una de las k muestras podemos calcular un estadístico, p.e., la media de las n observaciones que la componen.

Así tendremos un total de k nuevos valores $\bar{x}_i, i = 1, \dots, k$. Podemos asociar estos valores a una nueva v.a. \bar{X} , cuya distribución llamaremos **distribución muestral**.

Una de las propiedades más importantes es la siguiente:

Teorema (Distribución de las Medias Muestrales):

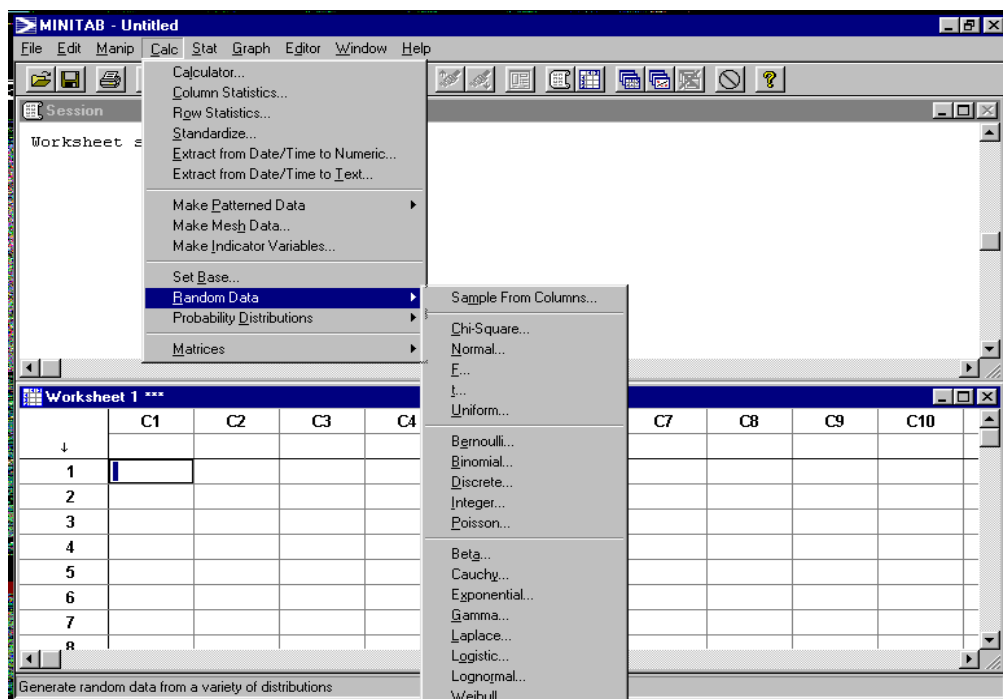
Sea X una v.a. **cualquiera** de media μ y desviación típica σ , entonces:

- Si consideramos **todas** las muestras aleatorias posibles, cada una de ellas de tamaño n , se cumplirá que $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
- Además, si X sigue una distribución normal, \bar{X} también será normal.

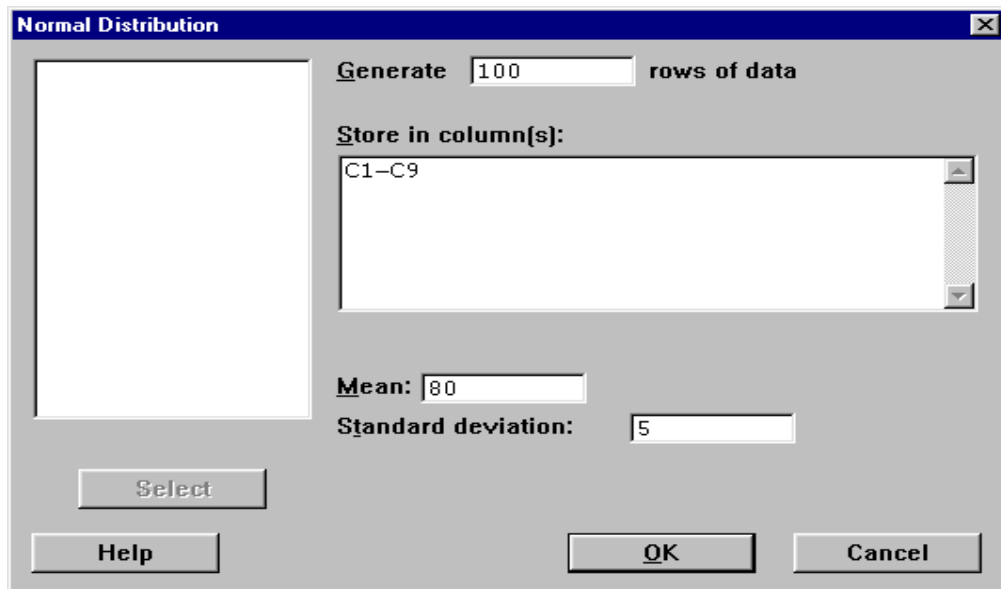
Ejemplo:

A fin de visualizar el *Teorema de Distribución de las Medias Muestrales*, vamos a “simular” la extracción de $k=100$ muestras de una variable normal con media 80 y desviación típica 5. Tomaremos como $n=9$ el tamaño de cada muestra:

Seleccionar *Calc > Random Data > Normal* :

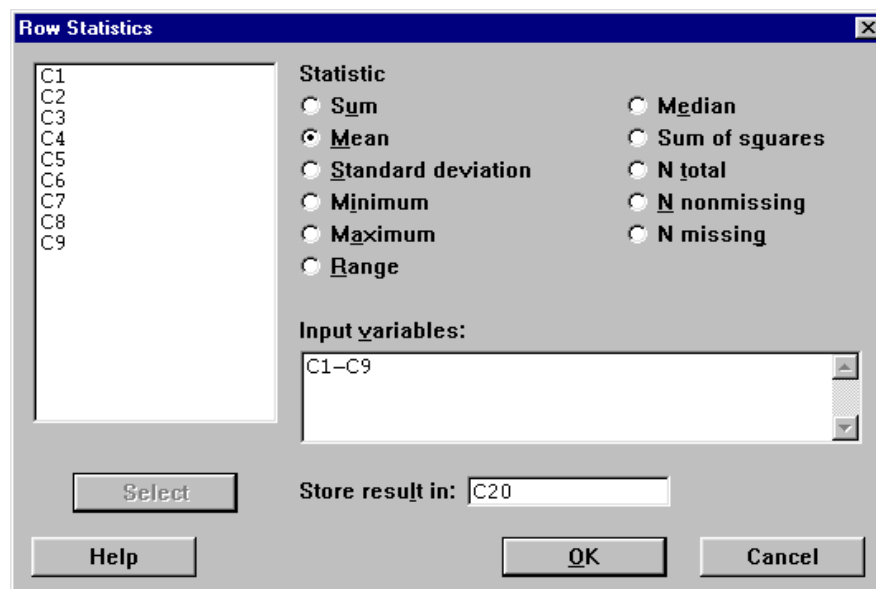


Rellenamos los campos según se indica en la imagen inferior:



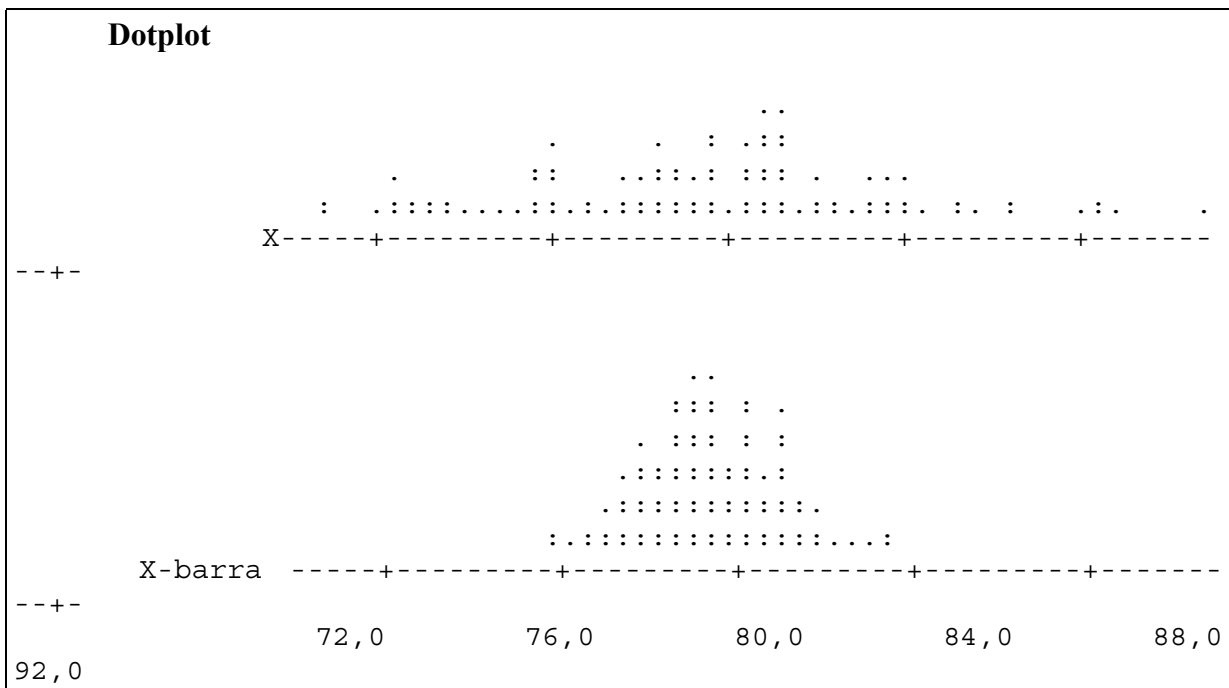
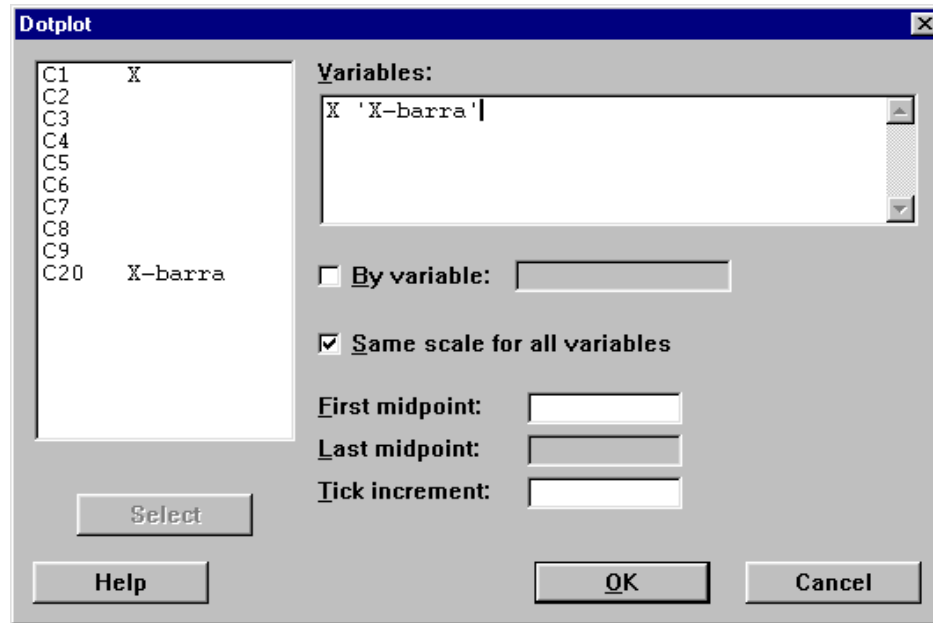
Habremos generado así una matriz de 9 columnas y 100 filas. Cada componente de esta matriz es una observación aleatoria proveniente de una distribución normal de media 80 y desviación estándar 5. Consideraremos que cada una de las filas obtenidas es una muestra, y lo que haremos ahora será calcular la media asociada a cada una de estas 100 muestras:

Seleccionar *Calc > Row Statistics* y rellenar los campos según se indica:



Disponemos ahora de 100 nuevos valores (las medias) situados en la columna 20. A continuación se muestran los "Dotplot" asociados a las columnas C1 (que representa 100 valores aleatorios obtenidos de una normal 80-5), y C20:

Seleccionar *Graph > Character graph > Dotplot*:



Finalmente, analizaremos también los estadísticos que describen la distribución de las medias muestrales:

Descriptive Statistics						
Variable	N	Mean	Median	TrMean	StDev	SE Mean
X-barra	100	79,566	79,446	79,543	1,596	0,160
Variable	Minimum	Maximum	Q1	Q3		
X-barra	76,035	83,799	78,459	80,627		

Observamos lo siguiente:

1. La distribución de la v.a. inicial X era normal y, según el gráfico de puntos anterior, parece que también la distribución de la v.a. \bar{X} es normal, de media muy similar y desviación estándar menor (los puntos de la \bar{X} están menos “dispersos” que los de la X).
2. Más concretamente, la media de los 100 valores contenidos en C20 (y que es una aproximación a la media de la v.a. \bar{X}) es de 79,566, valor muy similar a la media de X (que era de 80). Esto es coherente con lo que la teoría nos indica:

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

3. La desviación estándar de los 100 valores en C20 (que será una aproximación a la desviación estándar de \bar{X}) es de 1,596. Si tomamos la desviación estándar de X (que era de 5) y la dividimos por 3 (raíz de 9, el tamaño de la muestra), obtenemos el valor 1,667. Ambos valores son muy parecidos, tal y como la teoría predice:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

□ El Teorema Central del Límite:

Sea X una v.a. **cualquiera** de media μ y desviación típica σ , entonces:

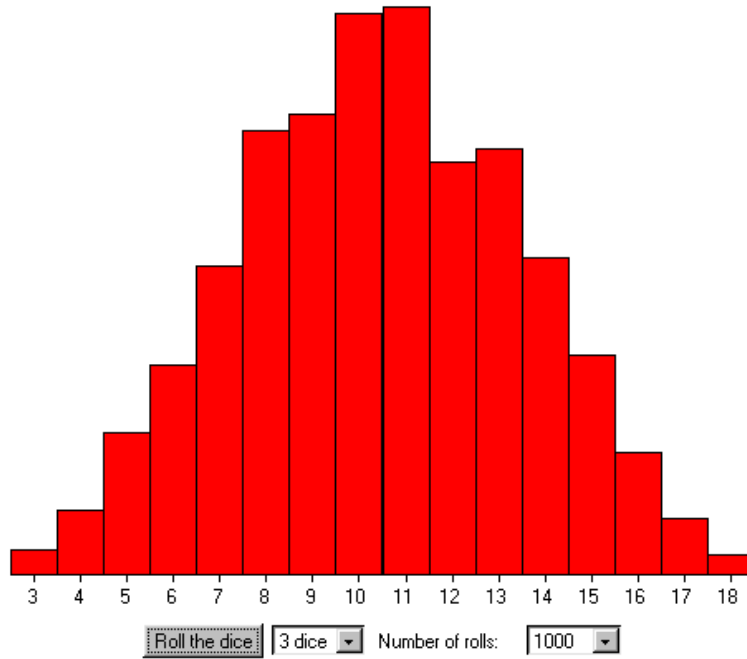
Si el tamaño muestral n es “suficientemente grande” (en la práctica suele valer $n > 30$), la distribución de las medias muestrales se aproxima a la de una normal, i.e.:

$$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

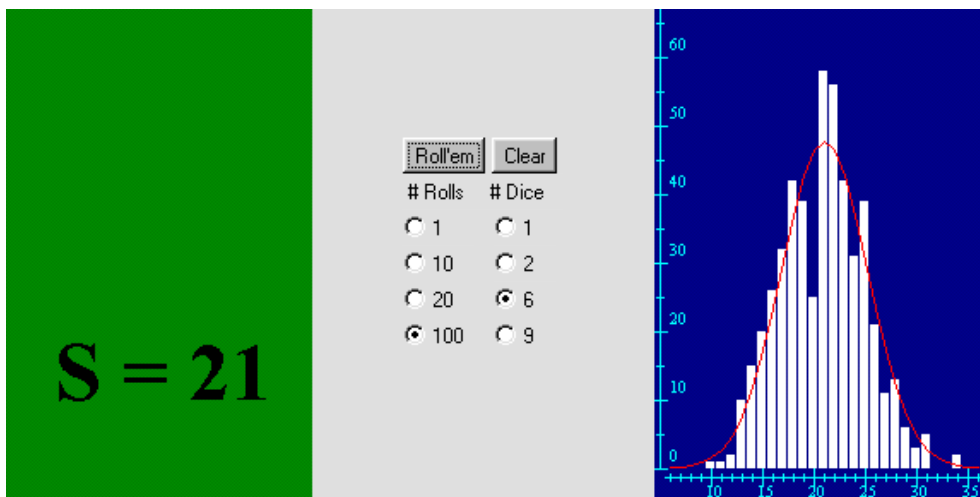
La importancia del TCL radica en que **sea cuál sea** la distribución de la población original (v.a. X), conforme el tamaño de las muestras (n) aumenta, la distribución de las medias se va aproximando a la de una normal (de la cual conocemos muchas propiedades).

Así, si la población tiene una distribución de probabilidad normal, entonces, para cualquier tamaño de muestra la distribución de la media también tendrá una distribución normal. Si la distribución de la población es simétrica (pero no normal), se verá que surge la forma normal como lo establece el TCL aún con muestras de al menos 30 para observar las características de normalidad.

Un ejemplo gráfico que muestra el Teorema Central del Límite, lo podemos encontrar en el siguiente enlace: http://www.unalmed.edu.co/~estadist/C.L.T/T_C_L.htm, de forma que cambiando el tamaño de la muestra veremos cómo va variando dicho gráfico, obtendremos representaciones similares a la siguiente:



Otro ejemplo similar al anterior, lo podemos encontrar en: http://www.ideamas.cl/cursoProb/javaEstat/central_limit_theorem/clk.html. Nuevamente, cambiando los datos veremos cómo la distribución resultante se va aproximando a una distribución normal:



CASOS PRÁCTICOS CON SOFTWARE

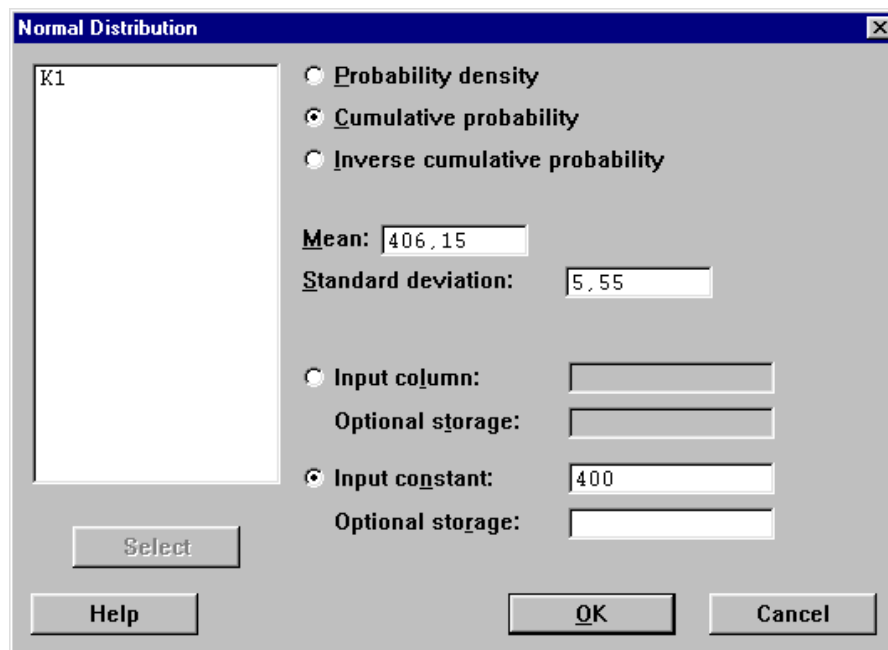
Según viene publicado en una prestigiosa revista de economía, el salario semanal medio de los profesores universitarios europeos es de 406,15 €. Se estima además que la desviación estándar de dichos salarios es de 55,50 €. Supongamos ahora que pretendemos tomar una muestra aleatoria de 100 profesores para estudiar sus salarios. Calcular las siguientes probabilidades referentes a la media de dicha muestra:

1. La probabilidad de que la media de la muestra sea menor de 400 €.

En primer lugar, observar lo siguiente: como $n = 100 \gg 30$, por el Teorema Central del Límite tendremos que la distribución de las medias muestrales \bar{X} se podrá aproximar por una normal con media 406,15 y desviación estándar 5,50.

Hemos de hallar $P(\bar{X} < 400)$:

Seleccionamos: *Calc > Probability Distributions > Normal* :



Cumulative Distribution Function	
Normal with mean = 406,150 and standard deviation = 5,55000	
x	P(X <= x)
400,0000	0,1339

2. La probabilidad de que la media de la muestra esté entre 400 y 410 € .

Sabemos que $P(400 < \bar{X} < 410) = P(\bar{X} < 410) - P(\bar{X} < 400)$. La segunda de éstas probabilidades ya la hemos calculado en el apartado anterior.

Para calcular la primera se razona análogamente, obteniendo que:

Cumulative Distribution Function	
Normal with mean = 406,150 and standard deviation = 5,55000	
x	P(X <= x)
410,0000	0,7561

Por tanto, tendremos: $P(400 < \bar{X} < 410) = P(\bar{X} < 410) - P(\bar{X} < 400) = 0,6222$

3. La probabilidad de que la media de la muestra sea mayor de 415 € .

En este caso, $P(\bar{X} > 415) = 1 - P(\bar{X} < 415)$. Hemos de calcular pues esta última probabilidad, lo cual haremos de forma análoga a los apartados anteriores.

Obtendremos lo siguiente:

Cumulative Distribution Function	
Normal with mean = 406,150 and standard deviation = 5,55000	
x	P(X <= x)
415,0000	0,9446

Por consiguiente, $P(\bar{X} > 415) = 1 - P(\bar{X} < 415) = 0,0554$

4. Hallar el valor del salario medio c tal que $P(X < c) = 0,95$

Seleccionamos nuevamente: *Calc > Probability Distributions > Normal* , pero ahora elegiremos la opción *Inverse Cumulative Probability* , con lo que obtendremos :

Inverse Cumulative Distribution Function	
Normal with mean = 406,150 and standard deviation = 5,55000	
P(X <= x)	x
0,9500	415,2789

BIBLIOGRAFÍA

- [1] Moya Anegón, F.; López Gijón, J.; García Caro, C. (1996): "Técnicas cuantitativas aplicadas a la biblioteconomía y documentación". Ed. Síntesis.
- [2] Lind, D.; Mason, R.; Marchal, W. (2001): "Estadística para Administración y Economía". Ed. Irwin McGraw-Hill.
- [3] Johnson, R. (1996): "Elementary Statistics". Ed. Duxbury.
- [4] Farber, E. (1995): "A Guide to Minitab". Ed. McGraw-Hill.

ENLACES

- ❑ http://www.unalmed.edu.co/~estadist/C.L.T/T_C_L.htm
Descripción y applet del Teorema Central del Límite
- ❑ http://www.ideamas.cl/cursoProb/javaEstat/central_limit_theorem/clt.html
Descripción y applet del Teorema Central del Límite.
- ❑ http://www.ruf.rice.edu/~lane/stat_sim/normal_approx/index.html
Teoría y applets, relacionados con la aproximación de una normal a una binomial.
- ❑ <http://www.udc.es/dep/mate/recursos.html>
Selección de recursos en Internet para la enseñanza-aprendizaje de la Estadística.
- ❑ <http://psych.colorado.edu/~mcclella/java/normal/accurateNormal.html>
Applets relacionados con la representación de una distribución normal.