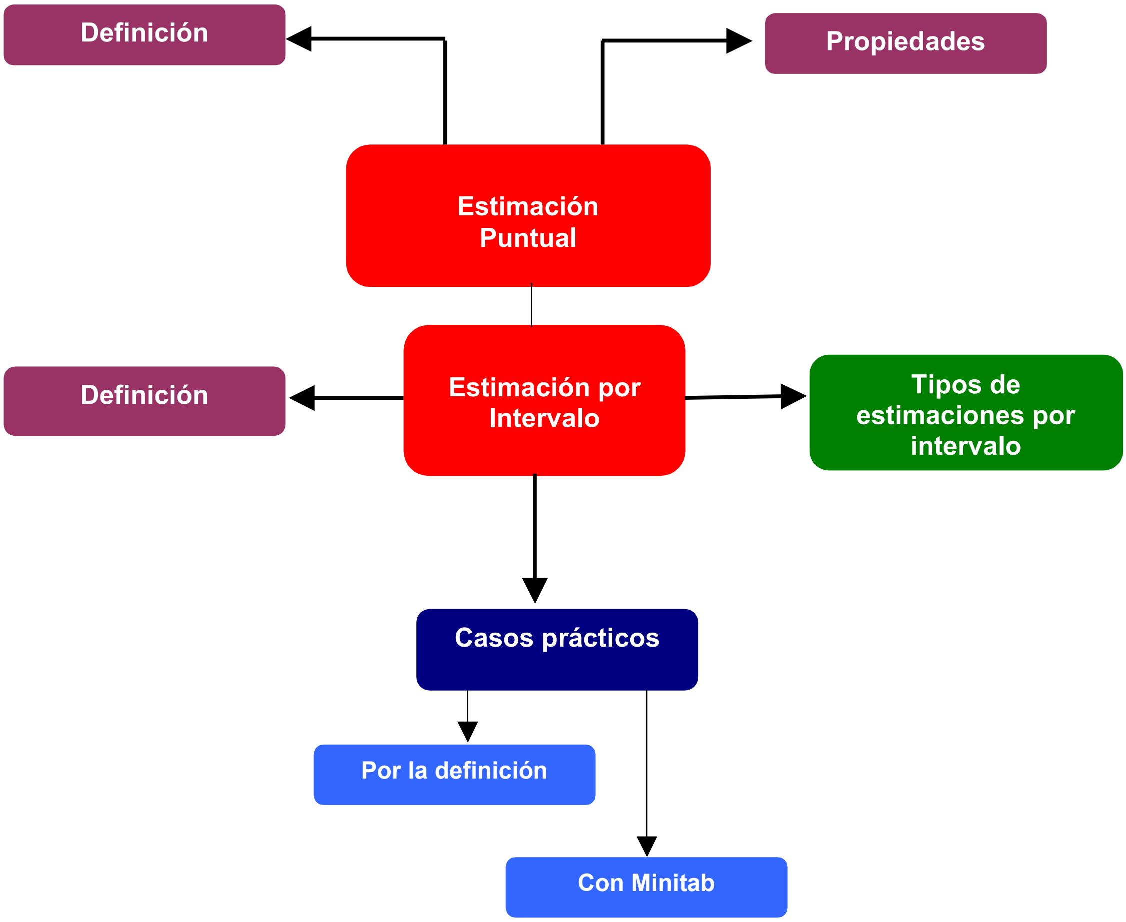
ESTIMACIÓN PUNTUAL Y ESTIMACIÓN POR INTERVALOS DE CONFIANZA

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

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

En este *math-block*, se pretende conocer y saber calcular las estimaciones puntuales y por intervalo para la media –ya sea conocida o no la desviación estándar poblacional-, así como las estimaciones para la probabilidad de éxito en una binomial.

En el caso en que conozcamos todos los elementos de una población, es sencillo calcular todos los parámetros asociados; sin embargo, en la mayoría de casos no será así, y necesitaremos estimar algunos de ellos a partir de los parámetros de la muestra.

OBJETIVOS

- Entender los conceptos de estimación puntual y estimación por intervalos.
- Calcular las estimaciones para la media poblacional, tanto en el caso en que la desviación estándar poblacional sea conocida como en el caso de que sea desconocida.
- Calcular las estimaciones (puntuales y por intervalos) para la probabilidad de éxito de una binomial.
- Saber interpretar correctamente los resultados de las estimaciones por intervalos.

CONOCIMIENTOS PREVIOS

Es recomendable haber leído, previamente los *math-blocks*: "Estadística Descriptiva con Minitab", "La distribución binomial" y "La distribución normal".

CONCEPTOS FUNDAMENTALES

❑ ¿Qué es una estimación?

Cuando queremos realizar un estudio de una población cualquiera de la que desconocemos sus parámetros, por ejemplo su media poblacional o la probabilidad de éxito si la población sigue una distribución binomial, debemos tomar una muestra aleatoria de dicha población a través de la cual calcular una aproximación a dichos parámetros que desconocemos y queremos estimar. Bien, pues esa aproximación se llama **estimación**.

Además, junto a esa estimación, y dado que muy probablemente no coincida con el valor real del parámetro, acompañaremos el error aproximado que se comete al realizarla.

❑ Estimación puntual

Una estimación puntual del valor de un parámetro poblacional desconocido (como puede ser la media μ , o la desviación estándar σ), es un número que se utiliza para aproximar el verdadero valor de dicho parámetro poblacional. A fin de realizar tal estimación, tomaremos una muestra de la población y calcularemos el parámetro muestral asociado ($\overline{x}$ para la media, *s* para la desviación estándar, etc.). El valor de este parámetro muestral será la **estimación puntual** del parámetro poblacional.

Por ejemplo, supongamos que la compañía Sonytron desea estimar la edad media de los compradores de equipos de alta fidelidad. Seleccionan una muestra de 100 compradores y calculan la media de esta muestra, este valor será un estimador puntual de la media de la población.

¿Qué propiedades debe cumplir todo buen estimador?

- **Insesgado:** Un estimador es **insesgado** cuando la media de su distribución muestral asociada coincide con la media de la población. Esto ocurre, por ejemplo, con el estimador $\bar{x}$, ya que $\mu_{\bar{x}} = \mu$ y con estimador p´ ya que $\mu_{p'} = p$

- **De varianza mínima:** La **variabilidad** de un estimador viene determinada por el cuadrado de su desviación estándar. En el caso del estimador $\bar{x}$, su desviación estándar es $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, también llamada **error estándar** de μ.

  En el caso del error estándar de p´, $\sigma_{p'} = \sqrt{\dfrac{p'*(1-p')}{n}}$

  Observar que cuanto mayor sea el tamaño de la muestra n, menor será la variabilidad del estimador $\bar{x}$ y de p´, por tanto, mejor serán nuestras estimaciones.

Estimación por intervalo

Dada una población *X*, que sigue una distribución cualquiera con media μ y desviación estándar σ.

1. Sabemos (por el TCL) que, para valores grandes de n, la media muestral $\bar{x}$ sigue una distribución aproximadamente normal con media $\mu_{\bar{x}} = \mu$ y desviación estándar $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.

2. Por otra parte, el Teorema de Chebyshev nos dice que, en una distribución normal, aproximadamente un 95% de los datos estaban situados a una distancia inferior a dos desviaciones estándar de la media.

   De lo anterior se deduce que: $P(\mu - 2\sigma_{\bar{x}} < \bar{x} < \mu + 2\sigma_{\bar{x}}) = 0{,}95$,

$$0{,}95 = P(\bar{x} < \mu + 2\sigma_{\bar{x}}) - P(\bar{x} < \mu - 2\sigma_{\bar{x}}) = P(\mu > \bar{x} - 2\sigma_{\bar{x}}) - P(\mu > \bar{x} + 2\sigma_{\bar{x}})$$

$$P(\bar{x} - 2\sigma_{\bar{x}} < \mu < \bar{x} + 2\sigma_{\bar{x}}) = 0{,}95$$

Por tanto, ésta última fórmula nos da un intervalo de valores tal que la probabilidad de que la media de la población μ esté contenida en él es de 0,95.

Este tipo de intervalos se llaman **intervalos de confianza** de un parámetro poblacional. El **nivel de confianza (1 - α)** del intervalo es la probabilidad de que éste contenga al parámetro poblacional. En el ejemplo anterior, el nivel de confianza era del 95% (α = 0,05).

Intervalos de confianza.

1. Intervalo de confianza para μ con σ conocida.

Un vendedor mayorista de partes automotrices necesita una estimación de la vida media que puede esperar de los limpiaparabrisas en condiciones normales de manejo. La administración de la empresa ya ha determinado que la desviación estándar de la vida útil de la población es de seis meses. Supongamos que se selecciona una sola muestra aleatoria de 100 limpiaparabrisas, y obtenemos que la vida media de estos 100 limpiaparabrisas es de 21 meses. Se pide calcular un intervalo de confianza del 95% para la vida media de la población de los limpiaparabrisas.

Tenemos X como la distribución de la vida útil en meses de la población de limpiaparabrisas, no sabemos qué distribución tiene, al igual que desconocemos su media. En este caso sí conocemos la desviación estándar poblacional.

$$X \approx (\mu, \sigma = 6)$$

La media muestral $\overline{X}$ por el teorema central del límite se va a aproximar la distribución normal:

$$\overline{X} \approx N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \sigma / \sqrt{n})$$

Por lo tanto, el intervalo de confianza del 95% para la vida media en meses de toda la población de limpiaparabrisas, es decir para μ

$$\overline{X} \pm Z_{0,05/2} * \frac{\sigma}{\sqrt{n}} = 21 \pm 1{,}96 * \frac{6}{\sqrt{100}} = 21 \pm 1{,}176 = [19{,}824\ ;\ 22{,}176]$$

$Z_{\alpha/2} = Z_{0,05/2} = Z_{0,025} = 1{,}96$, es decir que el valor Z de la tabla de la normal estándar que deja un área de 0,9 entre –Z Y +Z es Z=1,96. O de otro modo, como el nivel de confianza es 0,9, $\alpha = 0{,}05$, entonces el valor Z que deja su derecha un área de $\alpha/2 = 0{,}05/2 = 0{,}025$ y a la izquierda de –Z un área de $\alpha/2 = 0{,}05/2 = 0{,}025$ es Z=1,96

El **error máximo de estimación** es la mitad de la longitud del intervalo, $E = z(\alpha/2) * \frac{\sigma}{\sqrt{n}}$

Con una confianza del 95%, la vida media de la población de limpiaparabrisas que vende este mayorista está entre 19,824 meses y 22,176 meses.

Si extraemos varias muestras del mismo tamaño y calculamos un intervalo de confianza para cada muestra, el 95% de todos los intervalos van a incluir a la vida media poblacional en meses de todos los parabrisas que vende este mayorista.

2. Intervalo de confianza para μ con σ desconocida.

El administrador de una planta industrial generadora de energía desea estimar, por intervalo, la cantidad de carbón que se consumió por termino medio semanalmente durante año pasado. Para ello toma una muestra de 10 semanas. El consumo medio fue de 11.400 toneladas, la desviación estándar muestral 700 toneladas. ¿Cuál será el intervalo de confianza del 95% para el consumo medio semanal durante el año pasado?. (supongamos normalidad).

Tenemos X como la distribución de toneladas de carbón consumidas cada semana del año pasado por la planta de energía y su media y su desviación estándar desconocidas

$$X \approx (\mu, \sigma)$$

Aunque n < 30, suponemos que la media muestral, $\overline{X}$, sigue una distribución normal

$$\overline{X} \approx N(\mu_{\overline{x}} = \mu, S_{\overline{x}} = S / \sqrt{n})$$

Para estimar la desviación estándar poblacional σ vamos a utilizar la desviación estándar muestral S que es 700 toneladas.

Por lo tanto, el intervalo de confianza del 95% para el consumo promedio de toneladas de carbón en cada semana del año pasado, es decir para μ, será:

$$\overline{X} \pm t(n-1, \alpha/2) * \frac{S}{\sqrt{n}} = 11.400 \pm 2{,}262 * \frac{700}{\sqrt{10}} = 11.400 \pm 500{,}76 = (10.899\ ;\ 11.901)$$

Utilizamos la t-Student porque la desviación estándar poblacional σ es desconocida. En las tablas, $t(10-1, 0{,}05/2) = 2{,}262$, una t-Student con 10 – 1 = 9 grados de libertad que deja su derecha un área de 0,025. $\alpha = 0{,}05$ porque el nivel de confianza es de $1 - \alpha = 0{,}95$

Con una confianza del 95%, el consumo promedio semanal de carbón durante el año pasado por esta planta de energía estará entre 10.899 toneladas y 11.901 toneladas.

Si extraemos varias muestras del mismo tamaño y calculamos un intervalo de confianza para cada muestra, el 95% de todos los intervalos van a incluir al consumo promedio poblacional de toneladas de carbón por semana durante el año pasado por la planta de energía.

3. Intervalo de confianza para la probabilidad de éxito p en una binomial.

Durante un año y medio las ventas han estado disminuyendo de manera coherente en los 1.500 establecimientos de una cadena de comida rápida. Un empresa de consultoría ha determinado que el 30% de una muestra de 95 sucursales tiene claros signos de una mala administración. Construir un intervalo de confianza del 95% para esta porción.

A la población de todos los establecimientos de ésta cadena de comida rápida le vamos a llamar X que seguirá una binomial con probabilidad de éxito, probabilidad de tener signo de mala administración, p desconocida. A fin de estimar dicho parámetro, se toma una muestra de tamaño n = 95 y definimos p' como la proporción de éxitos en la muestra. En este caso p' es 0,3 y $1-p'$ = 0,7.

Como n > 20, $n * p' \geq 5$ y $n * (1-p') \geq 5$, entonces la distribución X es aproximadamente normal, i.e.:

$$X \approx N(np, \sqrt{np(1-p)})$$

Como p es desconocida, la aproximaremos por p' que es la estimación puntual de p.

Entonces, la proporción muestral de éxitos, que la hemos utilizado para estimar la proporción de la población tendrá la siguiente distribución:

$p' \approx N(p, \sqrt{\frac{p'(1-p')}{n}})$ con:

$$\sigma_{P'} = \sqrt{\frac{p'(1-p')}{n}} = \sqrt{\frac{0,3 * 0,7}{95}} = 0,047$$

Por lo tanto la estimación del error estándar de la proporción de establecimientos que tiene claros signos de mala será 0,057.

El intervalo de confianza del 95% para la probabilidad de éxito poblacional p viene dado por:

$p' \pm Z_{\alpha/2} * \sigma_{P'} = 0,3 \pm 1,96 * 0,047 = 0,3 \pm 0,0921 =$ [0,20788; 0,39212]

donde $Z_{\alpha/2} = Z_{0,05/2} = 1,96$ es el valor z*, de manera que el 95% del área bajo la curva normal se incluye entre –1,96 y 1,96.

Por lo tanto, con un nivel de confianza del 95%, la proporción de establecimientos de esta cadena de comida rápida que tiene mala administración estará entre 0,20788 y 0,39212.

Si extraemos varias muestras del mismo tamaño y calculamos un intervalo de confianza para cada muestra, el 95% de esos intervalos van a incluir a la verdadera proporción de establecimientos con mala administración

CASOS PRÁCTICOS CON SOFTWARE

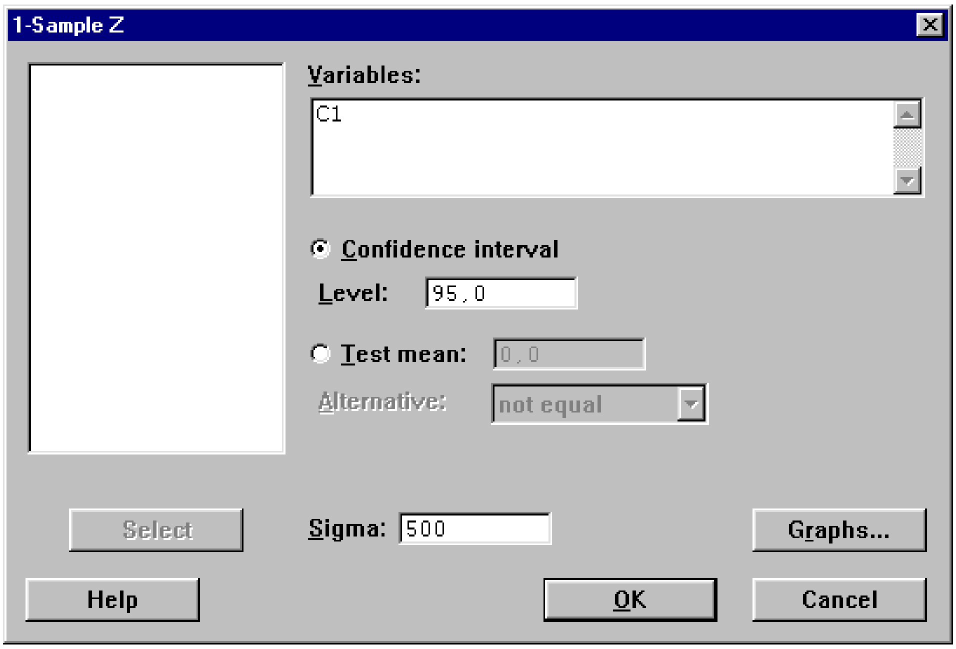
1. Imaginemos que trabajamos para una multinacional que se dedica a la venta de pantallas LCD. El departamento de ingeniería ha realizado pruebas de duración sobre una muestra aleatoria de 15 pantallas LCD, obteniendo los siguientes resultados (en horas de duración):

10014,8	8056,2	9166,1	8363,2	8869,7	8680,0	8930,4
8426,8	9488,3	8426,3	8924,6	7911,9	9667,2	8914,2
9220,2						

Supondremos que la duración (en horas de funcionamiento) de estas pantallas es una variable aleatoria que se distribuye de forma normal con desviación típica $\sigma = 500$ horas.

a) Hallar un intervalo de confianza, a nivel del 95% para la media poblacional μ (duración media de una pantalla LCD).

Seleccionamos *Stat* > *Basic Statistics* > *1-Sample Z*



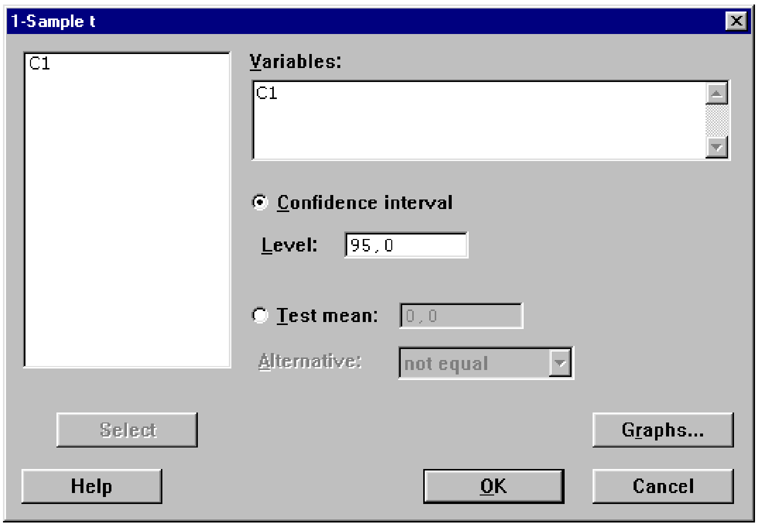

```
Z Confidence Intervals

The assumed sigma = 500

Variable      N      Mean     StDev  SE Mean         95,0 % CI
C1           15      8871      586       129   (     8618;     9124)
```

b) Suponiendo ahora que no conoces la desviación típica, halla un intervalo de confianza, a nivel del 95%, para µ. Compara este nuevo intervalo con el anterior.

Seleccionamos Stat > *Basic Statistics* > *1-Sample t*:




```
T Confidence Intervals

Variable      N      Mean     StDev  SE Mean        95,0 % CI
C1           15      8871      586      151    (    8546;    9195)
```

Observar que el primer intervalo está contenido en el segundo, i.e.: el segundo intervalo es menos “preciso” que el primero. Ello es lógico si tenemos en cuenta que para hallar el primer intervalo disponíamos de mayor información (conocíamos el valor de la desviación típica), por lo que el resultado es más preciso.

c) Suponiendo que no conoces la desviación típica, halla un intervalo de confianza, a nivel del 90%, para µ. Compara este intervalo con el obtenido en b).

```
T Confidence Intervals


Variable      N      Mean     StDev  SE Mean        90,0 % CI

C1           15      8871      586      151    (    8604;    9137)
```

Ahora, como somos menos exigentes por lo que al nivel de confianza se refiere (hemos pasado del 95 al 90%), lo que cabría esperar es que el intervalo obtenido esté contenido dentro del hallado en b). Observar que, en efecto, se cumple esta previsión.

2. Se quiere analizar el índice de productividad de los trabajadores de una empresa industrial, y se ha tomado una muestra aleatoria de 200 empleados y se ha observado que el 5% de ellos no alcanzan el nivel mínimo productivo que se quiere conseguir de cada uno de ellos. Calcular un intervalo de confianza del 95% para la proporción de empleados que no llegan al nivel de productividad fijado.

Nos interesa calcular un intervalo de confianza del 95% para la probabilidad p, de no alcanzar el nivel mínimo requerido.

Además, comprobamos que efectivamente se cumplen las hipótesis de normalidad:

n=200 >>30, n*p= 200*0.09 > 5 y n*p*(1-p) > 5

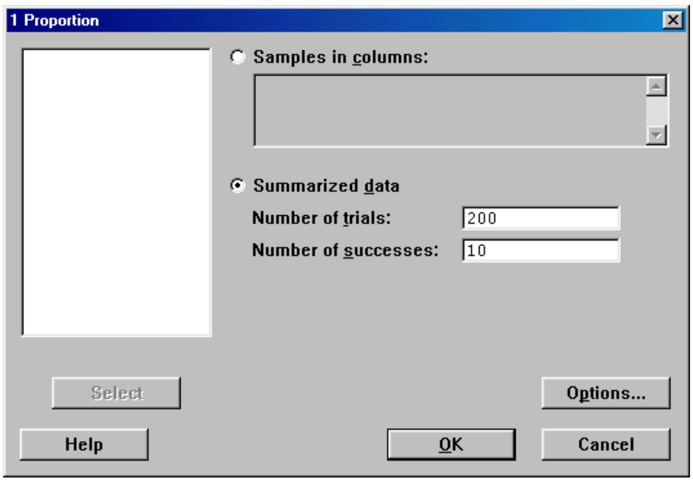
$$X \approx N(np, \sqrt{np(1-p)})$$

Como p es desconocida, la aproximaremos por p' que es la estimación puntual de p.

Entonces, la proporción muestral de éxitos, que la hemos utilizado para estimar la proporción de la población tendrá la siguiente distribución:

$$p' \approx N(p, \sqrt{\frac{p'(1-p')}{n}})$$

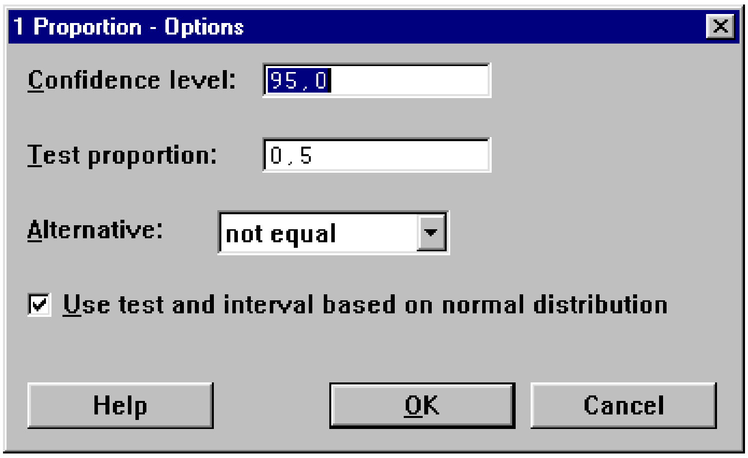
Para calcular el intervalo de confianza, seleccionamos: *Stat* > *Basic Statistics* > *1 Proportion*:



Seleccionamos *Options*, con las siguientes condiciones:

Ponemos el nivel de confianza del intervalo, la proporción del contraste que en este caso no nos interesa porque sólo queremos calcular el intervalo de confianza, por lo que en esta opción pondremos, o por omisión nos pondrá, 0,5.

En la alternativa ponemos lo que aparece como estándar, no igual y activamos la casilla de utilizar la normal para calcular el intervalo de confianza.



Confidence Interval for One Proportion

```
Test of p = 0,5 vs p not = 0,5

Sample       X      N  Sample p        95,0 % CI          Z-Value  P-Value
1           10    200  0,050000  (0,019795; 0,080205)    -12,73    0,000
```

Observamos que el intervalo de confianza está entre 0,0198 y 0,080. Por tanto, podemos concluir que con una confianza del 95%, la proporción de trabajadores de esta empresa que no alcanzan el nivel mínimo de productividad requerido estará entre el 2% y el 8%.

Si extraemos varias muestras del mismo tamaño y calculamos un intervalo de confianza para cada muestra, el 95% de esos intervalos van a incluir a la verdadera proporción de trabajadores que no alcanzan el nivel mínimo de productividad requerido.

BIBLIOGRAFÍA

[1] Lind, D.; Mason, R.; Marchal, W. (2001): “Estadística para Administración y Economía”. Ed. Irwin McGraw-Hill.F.

[2] Kvanli, A. (2000) “Introduction to Business Statistics” South-Western.

[3] Johnson, R. (1996): “Elementary Statistics”. Ed. Duxbury.

[4] Levin, R.; Rubin, D. (1996): “Estadística para Administradores”. Ed. Prentice Hall.

[5] Farber, E. (1995): “A Guide to Minitab”. Ed. McGraw-Hill

ENLACES

- http://oak.cats.ohiou.edu/~wallacd1/shyp.html
  Definición y ejemplos de contraste de hipótesis de una población conocida la media y la desviación estándar de la población.

- http://www.unalmed.edu.co/~estadist/C.L.T/T_C_L.htm
  Características y applet del Teorema Central del límite.

- http://www.unalmed.edu.co/~estadist/confinterval/intervalconf.htm
  Características y applet del concepto de Intervalo de confianza.

- http://e-stadistica.bio.ucm.es/mod_intervalos/intervalos_applet.html
  Características y applet del concepto de Intervalo de confianza.

- http://e-stadistica.bio.ucm.es/mod_contraste/contraste_applet.html
  Applet sobre contraste de hipótesis para muestras independientes.

- http://oak.cats.ohiou.edu/~wallacd1/ssample.html
  Teoria y ejemplos sobre distribuciones muestrales

- http://kitchen.stat.vt.edu/~sundar/java/applets/
  Aplicaciones estadísticas con JAVA

- http://e-stadistica.bio.ucm.es/mod_intervalos/intervalos_applet.html
  Applets sobre estimación por intervalos