Curso de Estadística Aplicada a las Ciencias Sociales

Tema 10. Estimación de una proporción

Cap. 20 del manual

Tema 10. Estimación de una proporción

Introducción

- Hasta ahora: estadística descriptiva (para describir datos)
- Ahora: estadística inferencial
- Para <u>inferir</u> datos sobre una población de la cual no tenemos todos los datos, a partir de los datos de una muestra, que sí tenemos

Introducción (2)

- En este tema: estimación de una proporción
- Cuál es la proporción de los elementos de una población que tiene una característica (votar a partido A, tener tal opinión, consumir tal producto, ir a la iglesia los domingos....)

Introducción (3)

- Procedimiento:
  - tomar una muestra aleatoria simple (lo vimos en el tema 8);
  - Calcular la proporción que tiene la característica en la muestra (estimador)
  - Calcular el valor del parámetro en la población, para lo que tenemos que tener una idea sobre la "precisión" del estimador

Introducción (4)

- El mismo estimador puede tener niveles de precisión muy diferentes
- Ejemplo, estimador del 30% (0,30) calculado
  - Encuestando a 2.000 personas de una población de varios millones (600 dicen votarán a partido A)
  - Encuestando a 60 de 90 estudiantes (18 aficionados al teatro)

1. Distribución en el muestreo de una proporción

- ¿Cómo calcular cuánto se puede alejar el estimador del parámetro?
- Imaginamos situación ficticia
  - Población muy grande (*N*)
  - Una cierto número de personas, n_1 tiene una característica
  - La proporción de personas que tienen la característica es $p=n_1/N$ (la frecuencia relativa). Supongamos p=0,4
  - La proporción de personas que no tienen la característica es $q=1-p$. Sería en este caso q=0,6

1. Distribución en el muestreo de una proporción

- En una urna (la urna A) ponemos una bola por cada persona
- Si tiene la característica: un 1. Si no la tiene: un 0
- Número de personas con la característica es igual al sumatorio de los valores de las bolas

$$\sum x_i = n_1 \cdot 1 + (N - n_1) \cdot 0 = n_1$$

- Media es igual a la proporción de personas que tienen la característica

$$\bar{x} = \frac{\sum x_i}{N} = \frac{n_1}{N} = p$$

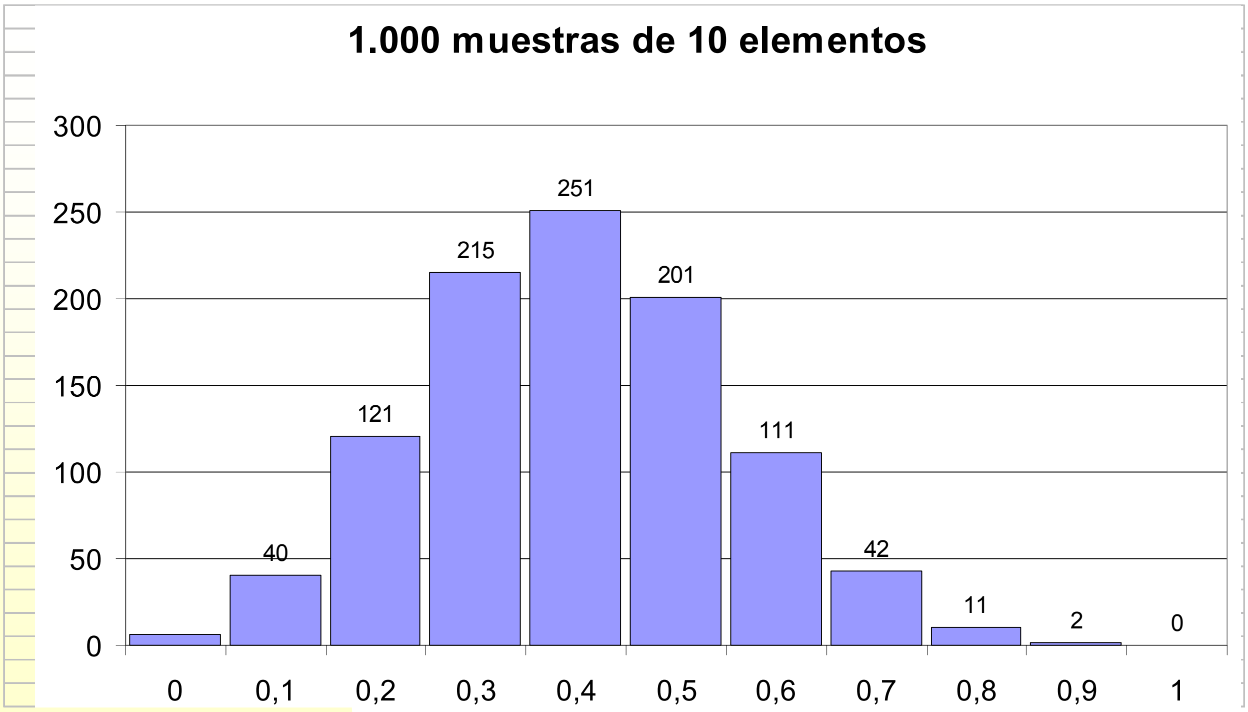
1. Distribución en el muestreo de una proporción

- Sacamos de la urna A una muestra de 10 bolas. Escribimos proporción en una papeleta: otra urna distinta (urna B)
- Devolvemos bolas a urna A y seguimos sacando muestras de 10 bolas, apuntando proporciones en papeletas en urna B
- Al final, urna B, llena de papeletas con resultados de muestreo

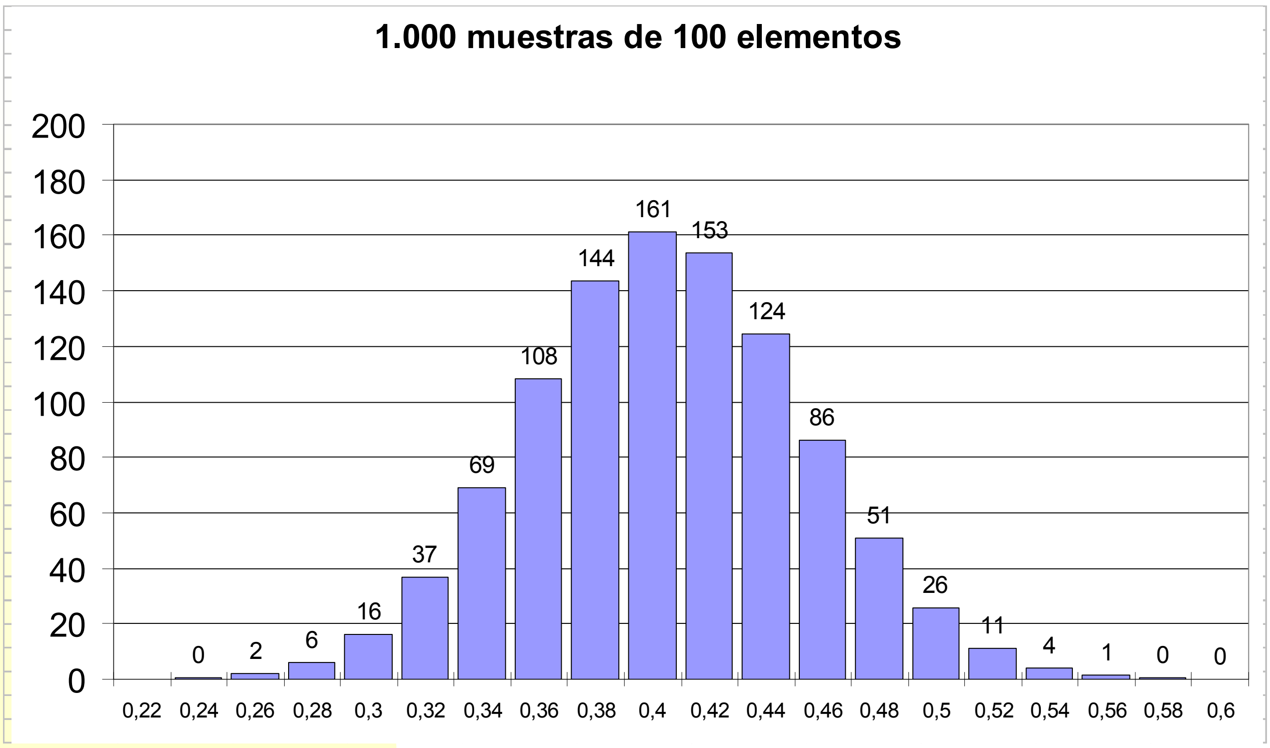
1. Distribución en el muestreo de una proporción

- ¿Qué tendríamos en urna B?
- Una variable aleatoria
- La mayor parte de los valores: en torno a p (0,4)
- Si las muestras han sido pequeñas (10, 15, 25): habrá algunas muestras con valores alejados de p
- Si muestras más grandes (más de 30 elementos): casi todas las muestras con poca variación de p
- Lo que es crucial: Distribución normal

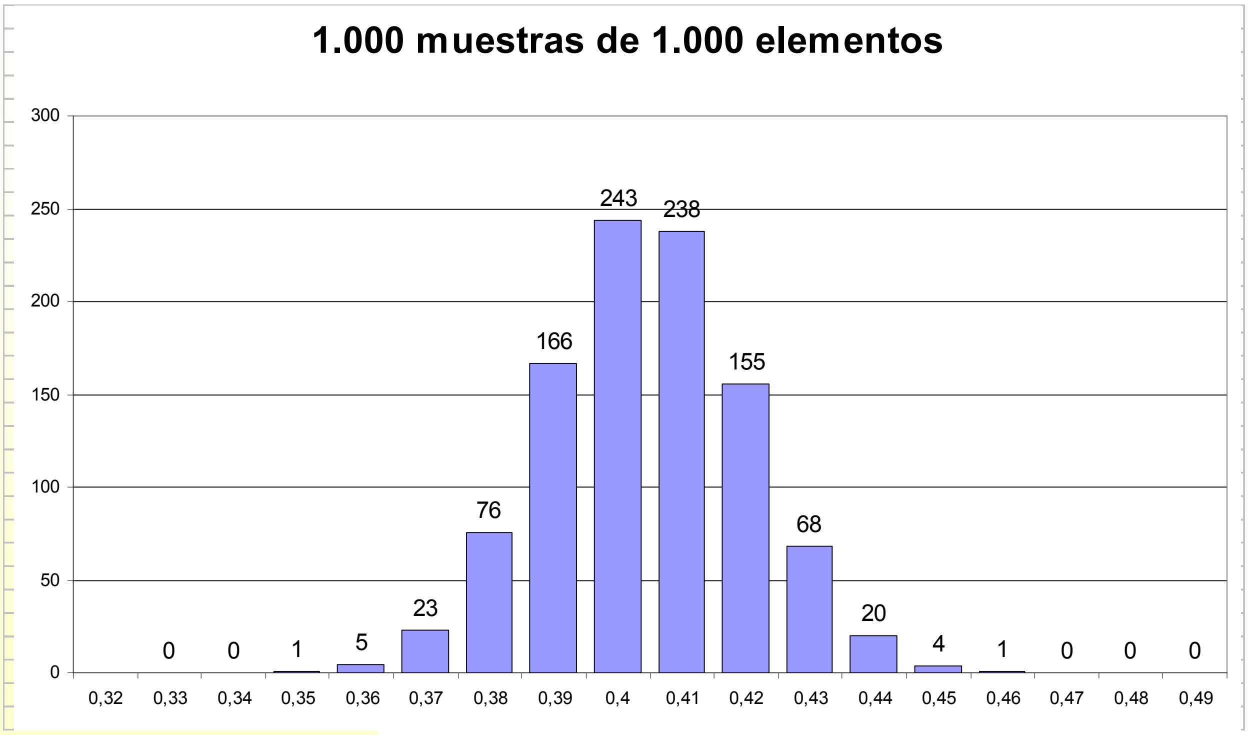
1. Distribución en el muestreo de una proporción



1. Distribución en el muestreo de una proporción



1. Distribución en el muestreo de una proporción



2. Estimadores centrados

- La variable proporción muestral estimada (la variable cuyos valores están en la urna B) es una variable aleatoria, como las vistas en el tema 9
- Media de la urna B = media de la urna A
- Lógico: media urna B = media valores urna A contados muchas veces
- Media de la urna B = proporción en la urna A (recordar cuando 0,1, media= proporción)

3. El error típico

- En la vida real no tenemos la “urna B”, para calcular su media (y eso nos daría la media=proporción de la población)
- Tenemos sólo un estimador (una papeleta de la urna B)
- ¿Cuánto puede alejarse el estimador de su media (y por tanto, del parámetro)?
- La respuesta: la desviación típica de los valores de la urna B
- **Error típico** de la estimación: la desviación que en promedio podemos esperar entre un estimador y el parámetro en la población (que es la media de los estimadores)

3. El error típico

- Intuitivamente: el error típico de la estimación será menor cuanto mayor sea la muestra
- Fórmula (aquí no demostramos). Si llamamos *n* al tamaño de la muestra:

$$ET = \frac{s_x}{\sqrt{n}}$$

- Es decir, el error típico es siempre menor que la desviación típica en la población
- Y será menor cuanto mayor sea la muestra
- Pero no disminuye proporcionalmente, sino proporcionalmente a la raíz de *n*

3. El error típico

- En una variable con valores 0,1 se puede demostrar que $s_x = \sqrt{p(1-p)} = \sqrt{pq}$

$$s_x = \sqrt{\sum (c_i - \bar{x}_c)^2 f_i} =$$
$$\sqrt{\left((1-p)^2 \cdot p\right) + \left((0-p)^2 \cdot (1-p)\right)} =$$
$$\sqrt{\left((1^2 + p^2 - 2p) \cdot p\right) + \left((0^2 + p^2 - 0p)(1-p)\right)} =$$
$$\sqrt{p + p^3 - 2p^2 + p^2 - p^3} = \sqrt{p - p^2} =$$
$$\sqrt{p(1-p)} = \sqrt{pq}$$

3. El error típico

- El error típico será entonces

$$ET = \sqrt{\frac{pq}{n}}$$

- Ejemplo: población con proporción de 0,4, y extraemos muestras de 10 elementos

$$ET = \sqrt{\frac{0{,}4(1-0{,}4)}{10}} = \sqrt{\frac{0{,}4 \cdot 0{,}6}{10}} = \sqrt{\frac{0{,}24}{10}} = \sqrt{0{,}024} = 0{,}15$$

- El ET será 0,15

3. El error típico

- Pero entonces, si la fórmula del ET incluye *p* y *q*, y si precisamente hemos hecho muestra para calcular *p* y *q* ¿círculo vicioso?
- No: podemos ver en tabla que *n* es mucho más importante que *p* y *q* para calcular ET

3. El error típico

Valores del ET según valores de *n* y *p*

		n						
		10	50	100	500	1.000	3.000	5.000
p	0,01	0,031	0,014	0,010	0,004	0,003	0,002	0,001
	0,05	0,069	0,031	0,022	0,010	0,007	0,004	0,003
	0,1	0,095	0,042	0,030	0,013	0,009	0,005	0,004
	0,2	0,126	0,057	0,040	0,018	0,013	0,007	0,006
	0,3	0,145	0,065	0,046	0,020	0,014	0,008	0,006
	0,4	0,155	0,069	0,049	0,022	0,015	0,009	0,007
	0,5	0,158	0,071	0,050	0,022	0,016	0,009	0,007
	0,6	0,155	0,069	0,049	0,022	0,015	0,009	0,007
	0,7	0,145	0,065	0,046	0,020	0,014	0,008	0,006
	0,8	0,126	0,057	0,040	0,018	0,013	0,007	0,006

3. El error típico

- Para un mismo valor de *p* error varía mucho según tamaño de *n*
- Para un mismo tamaño de *n* error no varía mucho con cambios en *p* (más alto error cuanto mayor es *p*, hasta máximo en 0,5)
- (Por cierto): ET es bastante bajo incluso para muestras de sólo 1.000 elementos

3. El error típico

- En resumen: podemos hacer estimación de ET con dos supuestos:
  - Suponiendo el caso más desfavorable, que es $p=q=0,5$
  - Suponer, para calcular el ET, que *p* fuera igual a la estimación

3. El error típico

- Ejemplo: muestra de 1.000 encuestados, 0,28 dicen que van a votar a partido A. Dos estimaciones de ET

$$ET = \sqrt{\frac{0{,}28 \cdot 0{,}72}{1000}} = \sqrt{\frac{0{,}2016}{1000}} = \sqrt{0{,}0002016} = 0{,}0142$$

$$ET = \sqrt{\frac{0{,}5 \cdot 0{,}5}{1000}} = \sqrt{\frac{0{,}25}{1000}} = \sqrt{0{,}00025} = 0{,}0158$$

4. Intervalos de confianza

- Dado un estimador, y un ET, podemos calcular en qué intervalo debe estar el parámetro en la población, con un cierto nivel de confianza
- Ejemplo: muestra de 1.000 personas, estimador de 0,20

$$ET = \sqrt{\frac{0{,}2 \cdot 0{,}8}{1000}} = \sqrt{\frac{0{,}16}{1000}} = \sqrt{0{,}00016} = 0{,}0126$$

4. Intervalos de confianza

- La distribución de las variables aleatorias derivadas de un experimento aleatorio con más de 30 elementos es una distribución normal
- Propiedad de las distribuciones normales (tema 9) :
  - 68,3% de los casos a menos de 1dt de la media
  - 95,5% de los casos a menos de 2dt de la media
  - 99,7% de los casos a menos de 3 dt de la media

4. Intervalos de confianza

- Por esa razón, cuando tenemos una muestra, en la que hemos calculado una proporción y un ET (que es la desviación típica de las muestras)
- Podemos decir que la proporción media de todas las muestras (y, por tanto, la de la población) estará con un
  - 68,3% de confianza (o de probabilidad) en el intervalo: estimador ± 1 ET
  - 95,5% de confianza (o de probabilidad) en el intervalo: estimador ± 2·ET
  - 99,7% de confianza (o de probabilidad) en el intervalo: estimador ± 3·ET

4. Intervalos de confianza

- El intervalo de confianza es un rango de valores que incluye el valor del parámetro en la población, con una determinada probabilidad
- En el ejemplo que poníamos (estimación es 0,20 y ET es 0,0126)
- Podemos decir que el verdadero parámetro, con un
  - 68,3% de confianza está en en el intervalo de 0,20 ± 0,0126 (es decir, entre 0,1874 y 0,2126)
  - 95,5% de confianza está en el intervalo de 0,20 ± 0,0252 (es decir, entre 0,1748 y 0,2252)
  - 99,7% de confianza está en el intervalo de 0,20 ± 0,0378 (es decir, entre 0,1622 y 0,2378)

4. Intervalos de confianza

- Dicho de otra forma, el ET nos permite calcular el error muestral (la diferencia entre el estimador y el parámetro)
- Podemos decir que el error muestral será, con:
  - 68,3% de confianza (o de probabilidad), menor o igual a 1ET
  - 95,5% de confianza (o de probabilidad), menor o igual a 2·ET
  - 99,7% de confianza (o de probabilidad), menor o igual a 3·ET

5. Estimación en poblaciones pequeñas

- Cuando tamaño población es pequeño comparado con tamaño de la muestra y muestreo sin reposición (60 alumnos de clase de 90): la fórmula del ET cambia
- La distribución de los errores de estimación sigue siendo aproximadamente normal
- El estimador sigue centrado (igual al parámetro)

5. Estimación en poblaciones pequeñas

- Pero: la desviación típica de la distribución de los estimadores muestrales es menor que el caso de población infinita
- Fórmula: siendo la fracción de muestreo, *f*= *n/N*

$$ET = \sqrt{\frac{pq}{n}} \text{ x } \sqrt{1-f}$$

5. Estimación en poblaciones pequeñas

- Ejemplo. Si en población de 500 personas tomamos muestra de 100 (*f* = 100/500=0,2) y obtenemos estimador de 0,3

$$ET = \sqrt{\frac{0{,}3 \cdot 0{,}7}{100}} \text{ x } \sqrt{1-0{,}2} = 0{,}0458 \cdot 0{,}894 = 0{,}040$$

- Correspondiente, los intervalos de confianza serán también más pequeños

6. Determinación del tamaño muestral

- Hemos visto que a partir de *n* se pueden estimar ET, y el intervalo de confianza
- Pero podemos operar también al revés: a partir del intervalo de confianza que queremos, podemos calcular el ET, y a partir de él podemos calcular qué tamaño debe tener la muestra

6. Determinación del tamaño muestral

- Ejemplo: queremos hacer encuesta con un intervalo no mayor de ± 0,04, con el 95,5% de confianza
- El intervalo con el 95,5% de confianza es el parámetro obtenido ± 2 ET
- Por tanto, si el intervalo que admitimos es ± 0,04, el ET será 0,02

6. Determinación del tamaño muestral

- Para calcular el tamaño de la muestra, despejamos su valor en la fórmula del error muestral o error típico

$$ET = \sqrt{\frac{pq}{n}}$$

$$ET^2 = \frac{pq}{n}$$

$$n = \frac{pq}{ET^2}$$

6. Determinación del tamaño muestral

- Cuando no sabemos los valores de *p* y *q* de antemano (normalmente)

$$n = \frac{(0{,}5)(0{,}5)}{ET^2} = \frac{0{,}25}{ET^2}$$

- Entonces, en el ejemplo puesto:

$$n = \frac{0{,}25}{0{,}02^2} = \frac{0{,}25}{0{,}0004} = 625$$

6. Determinación del tamaño muestral

- Otra manera de decir lo mismo
- Cuando buscamos un intervalo con una confianza del 95,5%: amplitud del intervalo=2ET
- Por tanto: ET=Amplitud intervalo/2
- Retomando la fórmula anterior:

$$n = \frac{0{,}25}{ET^2} = \frac{0{,}25}{\left(\frac{\text{amplitud del intervalo}}{2}\right)^2} = \frac{0{,}25}{\frac{(\text{amplitud del intervalo})^2}{4}} =$$

$$\frac{4 \cdot 0{,}25}{(\text{amplitud del intervalo})^2} = \frac{1}{(\text{amplitud del intervalo})^2}$$

Resumen

- Distribución en el muestreo
- Media = proporción
- Desviación típica = error típico
- Fórmula del ET
- Intervalos de confianza
- Estimación en poblaciones pequeñas
- Cálculo de muestra a partir de objetivo de error y nivel de confianza

Ejercicios recomendados

- Del manual:
  - 20.4
  - 20.5 b) y c)
  - 20.6
- De exámenes:
  - Feb02, Jun02: 10
  - Feb03, Sep03: 12
  - Feb04: 11
  - Jul04: 10
  - Feb05, Jul05: 12
  - Ene07, Ene08, Jun08: 8