

LA INFERENCIA ESTADÍSTICA EN EL NUEVO BACHILLERATO

José Ramón Vizmanos

I. CONTENIDOS DEL CURRÍCULO

Una de las novedades de los actuales currículos de matemáticas para el 2.º curso del Bachillerato de Ciencias Sociales es la inferencia estadística, dicha inclusión resulta lógica, dada la relevancia que dichos conocimientos tienen en la mayoría de las disciplinas científicas y en particular en las Ciencias Sociales.

En los contenidos para el 2º curso de Matemáticas Aplicadas a las Ciencias Sociales II para el territorio MEC, en el apartado correspondiente a Estadística y Probabilidad, publicados en el BOE, figura textualmente:

«Introducción al concepto, uso y alcance de la inferencia estadística: problemas relacionados con la elección de muestras, las condiciones de representatividad y análisis de las conclusiones que cabe extraer de ellas. Estudio de algún test de contraste de hipótesis basado en la distribución normal y aplicación a situaciones sencillas».

Estos párrafos son muy ambiguos hasta el extremo que, con sólo estas ideas, estaríamos perdidos a la hora de elaborar unidades didácticas que desarrollen estos contenidos.

Si observamos los currículos correspondientes a esta parte de otras Comunidades Autónomas, se puede observar que no existen prácticamente diferencias.

Ahora bien tanto el Ministerio como alguna Comunidad Autónoma, por ejemplo Galicia, han desarrollado ejemplificaciones donde se llega a detalles muy sutiles sobre el conocimiento de la Inferencia Estadística. Además el Ministerio en alguna ejemplificación asigna 5 semanas para el tratamiento de estos temas.

Por otra parte, si se tiene en cuenta que las pruebas de Selectividad sólo versarán sobre los contenidos de 2º de Bachillerato y sabiendo cómo se comportan algunas

Universidades en dichas pruebas, no sería de extrañar que pretendan de los alumnos un conocimiento bastante exhaustivo de la Inferencia Estadística.

Además hemos de reconocer aquí una dificultad añadida, estos contenidos no se han explicado nunca y algunos profesores con titulaciones diversas, es posible que nunca los hayan estudiado. De ahí la importancia a la hora de acertar en dar a estos contenidos un tratamiento equilibrado, de tal manera que no sobren muchos conceptos y que tampoco falten, ya que de todos es conocido como las pruebas de Selectividad para estas opciones son mucho mas disparatadas que para los alumnos de Ciencias.

2. POSIBLE SECUENCIACIÓN DE CONTENIDOS

Con el fin de clarificar un poco los contenidos que se deberían explicar, desde nuestro criterio, en este curso y a la vista de las distintas ejemplificaciones elaboradas tanto por el Ministerio como por alguna Comunidad Autónoma, nos parece que una posible secuenciación de contenidos sería la siguiente:

1. TEORÍA DE MUESTRAS

1.1 Muestra y población.

1.2 Tipos de muestreos.

1.2.1 Muestreo aleatorio simple.

1.2.2 Muestreo aleatorio estratificado.

1.2.3 Muestreo aleatorio sistemático.

1.2.4 Muestreo por conglomerados y áreas.

1.3 Distribución en el muestreo de una proporción.

1.4 Distribución en el muestreo de la media.

- 1.5 Distribución en el muestreo de las sumas muestrales.
- 1.6 Teorema central del límite.
- 1.7 Distribución en el muestreo de la diferencia de medias.

2. INTERVALOS DE CONFIANZA

- 2.1 Estimación puntual.
- 2.2 Propiedades de los estimadores.
- 2.3 Estimación por intervalo.
- 2.4 Intervalo de confianza para el parámetro p de una binomial.
- 2.5 Intervalo de confianza para la media poblacional.
- 2.6 Intervalo de confianza para la diferencia de medias.
- 2.7 Tamaño de la muestra.

3. CONTRASTE DE HIPÓTESIS

- 3.1 Contraste de hipótesis.
- 3.2 Errores de tipo I y de tipo II.
- 3.3 Contraste para el parámetro p de una población binomial.
- 3.4 Contraste para la media de una población normal.
- 3.5 Analogías entre el contraste de hipótesis e intervalos de confianza.

3. ORIENTACIONES DIDÁCTICAS

En el tema de teoría de muestras, conviene introducir al alumno en la necesidad de la elección de muestras, para poder realizar cualquier tipo de inferencia estadística, para ello se analizarán distintos tipos de muestreos, por ejemplo: el muestreo aleatorio simple, el estratificado, el sistemático y por conglomerados.

Uno de los conceptos más importantes y que más aplicación tendrá en los siguientes temas es el de distribución muestral de un determinado parámetro, por ejemplo: el parámetro p de la binomial, la media de una normal, la diferencia de medias de dos poblaciones normales, etc. Para poder llegar a la idea de distribución en el muestreo es importante ver el teorema central del límite, al menos de una forma intuitiva, con el fin de que el alumno comprenda como, a medida que el tamaño de la muestra crece, la distribución de

las medias muestrales se aproxima a una distribución normal y la desviación típica disminuye.

En el tema de intervalos de confianza, comenzaremos por utilizar una terminología precisa, para lo cual es conveniente distinguir entre los términos: parámetro, estadístico, estimador puntual, estimación puntual. De entre las propiedades de los estimadores parece suficiente con distinguir los estimadores insesgados y los estimadores eficientes. Conviene hacer ver que la estimación puntual es poco útil, pues sólo obtenemos una cierta aproximación al valor que tratamos de estimar, y en consecuencia resulta mucho más interesante tratar de obtener un intervalo dentro del cual se tenga una cierta confianza de que se encuentra el parámetro a estimar, con lo que llegamos a la idea de estimación por intervalo.

A continuación se aplicará la idea de estimación por intervalo a los parámetros cuyas distribuciones en el muestreo han sido estudiadas en el tema de teoría de muestras. Por último, conviene estudiar el tamaño de la muestra como método para aumentar la confianza, tratando de dar respuesta a preguntas del tipo, ¿cómo deberá ser de grande la muestra para tener una confianza, por ejemplo, del 99%?

En el tema de contraste de hipótesis, se puede comenzar dando una idea general de lo que se persigue al contrastar una hipótesis, analizando los errores que se cometen al aceptar una hipótesis siendo falsa, o al rechazar una hipótesis siendo verdadera. También aquí debe usarse una terminología precisa sobre: hipótesis nula, de alternativa, estadístico del contraste, región de aceptación, región de rechazo, contraste bilateral, contraste unilateral, error de tipo I, error de tipo II, nivel de significación, potencia de un contraste, etc.

A continuación se aplicará la idea general del contraste de hipótesis para el contraste de los parámetros cuyas distribuciones en el muestreo han sido estudiadas con anterioridad.

Parece conveniente estudiar las analogías que existen entre el contraste de hipótesis y los intervalos de confianza.

4. ALGUNOS EJEMPLOS

1. DISTRIBUCIÓN EN EL MUESTREO DE UNA PROPORCIÓN DE UNA POBLACIÓN BINOMIAL

Un candidato a diputado en unas elecciones legislativas desea saber qué tanto por ciento de personas le otor-

garán su confianza votándole, para ello encarga un estudio estadístico a una empresa de sondeos electorales.

La población de ciudadanos con derecho a voto está formada por 8 millones de personas. La empresa elige una muestra aleatoria de la población formada por 1.000 personas y obtiene que la proporción de votantes al candidato es del 52%.

Si p es la proporción de votantes en la población, al valor 0,52 lo representaremos por \hat{p} , ya que no es propiamente p , pero sí da la proporción de votantes en la muestra elegida.

Si elegimos otras muestras de tamaño 1.000, evidentemente el valor de p variará. Los distintos valores de \hat{p} dan lugar a una variable aleatoria que representaremos por \hat{P} que llamaremos **estadístico**.

La distribución de los valores de \hat{P} como depende de las muestras se llama **distribución muestral** o **distribución en el muestreo de una proporción** y se demuestra que:

La variable aleatoria \hat{P} tiene:

1º) Media: $\mu = p$

2º) Desviación típica: $s = \sqrt{\frac{p(1-p)}{n}}$

3º) A medida que n crece, la distribución de \hat{P} se aproxima a la normal siempre que p no se acerque ni a 0 ni a 1.

Por ejemplo, estudiar las distribuciones en el muestreo de la proporción de votantes para muestras de tamaño 40, 70, 100 y 1.000 personas.

Como no conocemos el valor del parámetro poblacional p , lo aproximaremos mediante el valor del estadístico $\hat{p} = 0,52$ obtenido en la muestra. Por tanto, supondre-

mos que $p = 0,52$ y sustituyendo para los distintos valores del tamaño de la muestra n obtenemos la tabla que se encuentra en esta página, abajo, si disponemos de una calculadora gráfica del tipo TI-83.

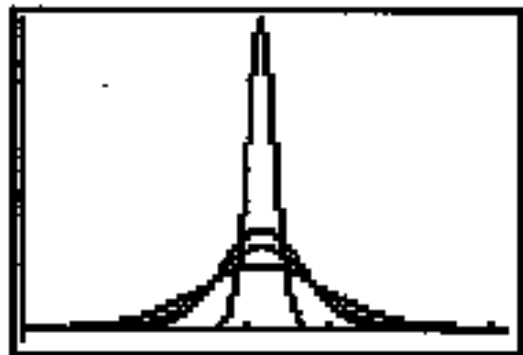
Introducimos, en el editor de funciones, las funciones de densidad de las distribuciones correspondientes a la cuarta columna.

Elegimos una buena pantalla de visualización y obtenemos las gráficas de las funciones de densidad. Conviene observar cómo a medida que el tamaño de la muestra crece la desviación típica disminuye y la distribución se aproxima más a la normal.

```

30001 Plot2 Plot3
\Y1=normalpdf(X,
.52,.079)
\Y2=normalpdf(X,
.52,.06)
\Y3=normalpdf(X,
.52,.05)
\Y4=normalpdf(X,

```



Tamaño de la muestra n	Media $\mu = p$	Desviación típica $s = \sqrt{\frac{p(1-p)}{n}}$	Distribución muestral $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
40	0,52	$\sqrt{\frac{0,52 \cdot 0,48}{40}} = 0,079$	$N(0,52; 0,079)$
70	0,52	$\sqrt{\frac{0,52 \cdot 0,48}{70}} = 0,060$	$N(0,52; 0,060)$
100	0,52	$\sqrt{\frac{0,52 \cdot 0,48}{100}} = 0,05$	$N(0,52; 0,05)$
1.000	0,52	$\sqrt{\frac{0,52 \cdot 0,48}{1000}} = 0,016$	$N(0,52; 0,016)$

2. IDEA INTUITIVA DE LA ESTIMACIÓN POR INTERVALO

En el apartado anterior hemos obtenido para una determinada muestra que la proporción de votantes al candidato era $\hat{p} = 0,52$. Esto es una estimación puntual, pero es poco útil, pues solo obtenemos una cierta aproximación al valor que tratamos de estimar, ya se comprende que es mucho más interesante obtener un intervalo dentro del cual se tiene una cierta confianza de que se encuentre el parámetro que tratamos de estimar.

Sabemos que el estimador \hat{p} sigue una distribución normal de parámetros: $N(p, \frac{p(1-p)}{n})$ donde n es el tamaño

de la muestra y p es el parámetro poblacional que queremos estimar. Ahora bien el problema es que no conocemos p , lo que podemos hacer es aproximarlos mediante el valor del estadístico $\hat{p} = 0,52$ obtenido en la muestra. Así pues, sustituyendo \hat{p} por 0,52 se tiene:

$$N(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}) = N(0,52; \frac{0,52 \cdot 0,48}{1.000}) = N(0,52; 0,016)$$

Entonces el parámetro p que queremos estimar sigue una $N(0,52; 0,016)$, tipificando resulta que la variable: $Z = \frac{p - 0,52}{0,016}$ sigue una $N(0,1)$.

© Supongamos ahora que el diputado quiere saber con una probabilidad del 95% entre qué valores se encontrará la proporción poblacional de las personas que le votarán.

Si Z es una variable $N(0,1)$ de las tablas de la normal sabemos que:

$$\Pr(-1,96 < Z < 1,96) = 0,95$$

Sustituyendo Z por su valor: $\Pr(-1,96 < \frac{p - 0,52}{0,016} < 1,96) = 0,95$

operando se deduce:

$$\Pr(-1,96 \cdot 0,016 < p - 0,52 < 1,96 \cdot 0,016) = 0,95$$

sumando 0,52:

$$\Pr(0,52 - 1,96 \cdot 0,016 < p < 0,52 + 1,96 \cdot 0,016) = 0,95$$

operando resulta:

$$\Pr(0,489 < p < 0,551) = 0,95$$

Así pues, la proporción de votantes estará en el intervalo (0,489; 0,551) con una probabilidad del 95%. Al intervalo obtenido lo llamaremos **intervalo de confianza** y a la probabilidad 0,95 la llamaremos **coeficiente de confianza** o **nivel de confianza**.

El margen de error vendrá dado por la amplitud del intervalo, por ejemplo en nuestro caso el margen de error será:

$$\text{margen de error} = 0,551 - 0,489 = 0,062 = 6,2\%$$

Es totalmente incorrecto interpretar el intervalo de confianza diciendo que en el 95% de los casos el parámetro a estimar está contenido en el intervalo, ya que lo que es variable es el propio intervalo. Por tanto, hay que entenderlo en el sentido que se acaba de exponer, es decir que para cada muestra obtendríamos un intervalo de confianza que, en general, sería distinto de unas muestras a otras, pero si hiciéramos un larga serie de determinaciones de estos intervalos para diferentes muestras se cumpliría que el 95% de los intervalos contendrían el valor del parámetro a estimar.

© Si el diputado fuera más exigente y quisiera obtener un intervalo de confianza para el parámetro p con un coeficiente de confianza mayor, por ejemplo del 99%, lo único que tendremos que hacer será efectuar nuevamente los cálculos.

Si Z es una variable $N(0,1)$ de las tablas de la normal sabemos que:

$$\Pr(-2,58 < Z < 2,58) = 0,99$$

Sustituyendo Z por su valor:

$$\Pr(-2,58 < \frac{p - 0,52}{0,016} < 2,58) = 0,99$$

operando se deduce:

$$\Pr(-2,58 \cdot 0,016 < p - 0,52 < 2,58 \cdot 0,016) = 0,99$$

sumando 0,52:

$$\Pr(0,52 - 2,58 \cdot 0,016 < p < 0,52 + 2,58 \cdot 0,016) = 0,99$$

operando resulta:

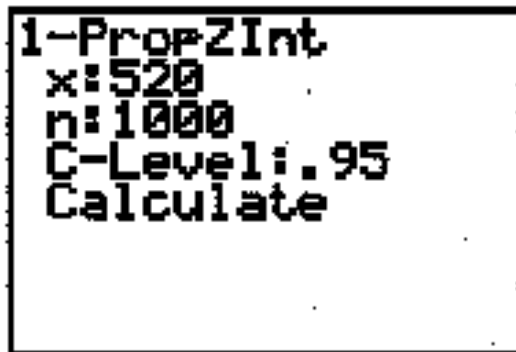
$$\Pr(0,479 < p < 0,561) = 0,99$$

Cuando aumenta el coeficiente de confianza la amplitud del intervalo es mayor y en consecuencia el margen de error también aumenta.

$$\text{margen de error} = 0,561 - 0,479 = 0,082 = 8,2\%$$

Hoy en día existen modelos de calculadoras gráficas que permiten obtener los intervalos de confianza sin más que introducir los datos. Hallemos con calculadora el intervalo de confianza de nuestro ejemplo.

Introducimos los datos $x = 520$; $n = 1000$ y el coeficiente de confianza 0,95



Pulsamos Calculate y obtenemos el intervalo de confianza al nivel de confianza deseado.

```
1-PropZInt
(.48904, .55096)
p=.52
n=1000
```

Para el nivel de confianza 0,99 repetimos los pasos anteriores.

```
1-PropZInt
x:520
n:1000
C-Level:.99
Calculate
```

```
1-PropZInt
(.47931, .56069)
p=.52
n=1000
```

3. IDEA INTUITIVA DEL CONTRASTE DE HIPÓTESIS

Ya hemos visto que en el sondeo realizado por la empresa se ha estimado que nuestro candidato obtendría el 52% de los votos emitidos.

Queremos contrastar esta hipótesis que llamaremos **hipótesis nula** y representaremos por H_0 . Para ello establecemos la hipótesis contraria que llamaremos

hipótesis de alternativa y representaremos por H_a .

Así pues:

Hipótesis nula H_0 : $p = 0,52$

Hipótesis alternativa H_a : $p \neq 0,52$

Para contrastar el parámetro p de la población tomaremos el valor de la muestra que llamaremos **estadístico del contraste**. Este estadístico, como ya se vio en el epígrafe 1, es una variable aleatoria que tiene por distribución en el muestreo una

$$N\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right) = N\left(0,52; \frac{0,52 \cdot 0,48}{1000}\right) = N(0,52; 0,016)$$

Para cada muestra el estadístico del contraste \hat{P} toma un valor particular \hat{p} , supongamos que para una determinada muestra aleatoria de tamaño 1000 personas el valor de $\hat{p} = 0,49$.

¿Cómo podemos medir hasta qué punto la diferencia entre $p = 0,52$ y $\hat{p} = 0,49$ es significativa?

Para ello fijaremos un nivel de confianza, por ejemplo $1 - \alpha = 0,99$ y entonces aceptaremos la hipótesis nula si el estadístico del contraste, una vez tipificado cae dentro del intervalo $(-z_{\alpha/2}, z_{\alpha/2})$, es decir, $(-2,58, 2,58)$ que llamaremos **región de aceptación**.

En caso contrario rechazaremos la hipótesis nula, ya que el estadístico del contraste una vez tipificado caerá en la región contraria que llamaremos **región crítica o de rechazo**.

En nuestro caso:

\hat{P} sigue una $N(0,52, 0,016)$ tipificando $\frac{\hat{P}-0,52}{0,016}$ sigue una $N(0, 1)$

para el valor particular de $\hat{p} = 0,49$ sustituimos y se obtiene:

$$\frac{0,49-0,52}{0,016} = -1,87$$

Como $-1,87 \notin (-2,58, 2,58)$ aceptaremos la hipótesis nula. Es decir, la muestra es realmente compatible con la población en el 90% de los casos. O también, a partir de los datos muestrales se acepta la hipótesis de que el tanto por ciento de votantes del candidato es de 52% con un nivel de confianza del 99%.

Si el nivel de confianza fuera del 90% por la tabla de la normal sabemos que la región de aceptación es

$$(-1,64, 1,64)$$

Como el valor de estadístico del contraste es $-1,87$ y cae fuera de la región de aceptación, entonces rechazaremos la hipótesis nula. En este caso diremos que a partir de los datos muestrales se rechaza la hipótesis de que el tanto por ciento de votantes al candidato es de