Inferencia estadística en los bachilleratos

Manuel Sobrino Reyes

Introducción

Los cambios introducidos en el Sistema Educativo español con la paulatina entrada en vigor de la LOGSE y las distintas disposiciones legales que la desarrollan son sustanciales.

De todas las importantes modificaciones voy a centrarme en una pequeña parcela del saber matemático, la Estadística, y más aún, en la Estadística Inductiva de los últimos cursos de la Enseñanza Secundaria que utiliza el Cálculo de Probabilidades como soporte y fundamentación.

Posiblemente sean la Estadística y la Geometría las partes de la Matemática más beneficiadas con su mayor presencia en los currículos futuros de toda la enseñanza obligatoria. No obstante, los motivos son diferentes.

En lo que a los contenidos se refiere, en Geometría se trata de volver a beber, aunque sea de manera metodológica distinta, de unas fuentes cerradas por la Ley General de Educación donde se apostaba por una Geometría Algebráica que posiblemente no tenga en los niveles obligatorios el valor formativo y de desarrollo de estrategias generales de pensamiento como el razonamiento, la intuición y la visión espacial que aporta la Geometría sintética.

Sin embargo, en Estadística se trata de incorporar a los curricula conocimientos cuya necesidad y utilidad son incuestionables en un mundo cuyo volumen de información suministrada va en aumento y en el que el progreso de casi todas las Ciencias va ligado a la aplicación y desarrollo de los potentes métodos estadísticos. Es imprescindible, por tanto, capacitar a los ciudadanos y dotarlos de unas herramientas básicas que, por una parte, les permitan asimilar, criticar y contrastar la información recibida y, por otra, les permitan aplicarlas a los campos del saber en que desarrollarán su trabajo.

La respuesta de la Escuela a una demanda social de este tipo no se debe demorar. Los profesores de Matemáticas además de actualizarnos en diversos aspectos de la didáctica también debemos adquirir conocimientos que posiblemente en algún caso no hayamos recibido en nuestra formación inicial ni en la permanente, como suele ser habitual en Estadística.

El carácter abierto y escueto de los curricula de Bachillerato obliga a los profesores de cualquier Centro a matizar y completar la propuesta atendiendo a las características del alumnado y al entorno en que se desenvuelve.

Tenemos en nuestras manos la posibilidad y la responsabilidad de marcar el rumbo de la nave y para ello es necesario tener muy claro cuál es el puerto de llegada o mejor, cuál es el puerto donde nuestros alumnos dispondrán de mayores y mejores comunicaciones con el exterior.

Seguidamente me limitaré a mencionar de manera lacónica algunos conceptos e ideas fundamentales ambientándolas o desarrollando en paralelo diversos ejemplos que no pretenden ser un modelo sino una excusa para empezar a preguntarnos qué podemos y debemos hacer en el aula. Es necesario, en cualquier caso, establecer el entramado conceptual necesario para que la aprehensión de la Estadística y sus métodos sea significativa para los alumnos. Los siguientes epígrafes de contenido no constituyen una secuencia ordenada y los ejemplos menos aún, máxime cuando ponen en marcha toda una serie de contenidos que es impensable puedan desarrollarse en un breve artículo.

Variable aleatoria

Existe diferencia entre variable estadística y variable aleatoria. El primer término, muy utilizado en Estadística descriptiva o elemental, es la característica en estudio de la población. La variable aleatoria -cuya denominación quizá no sea muy acertada- es una función del espacio muestral en el conjunto de los números reales; la variable independiente está formada por todos los posibles resultados de un experimento aleatorio (lanzamiento de tres monedas, estaturas de una muestra de cinco individuos,...).

Unas variables aleatorias especialmente interesantes en Inferencia son los estimadores cuya idea intuitiva es muy comprensible.

De manera natural por extensión de la idea de media y varianza de una variable estadística (estudio descriptivo) surge la definición de esperanza (valor medio o valor esperado) y varianza de una variable aleatoria donde los valores que toma la variable se ponderan por las probabilidades en vez de por las frecuencias.

Es importante conocer las propiedades básicas de la esperanza y varianza de la suma y diferencia de variables -independientes, al menos- de cara a utilizarlas por ejemplo para determinar la distribución del estadístico más usado en estos niveles que es la «media muestral».

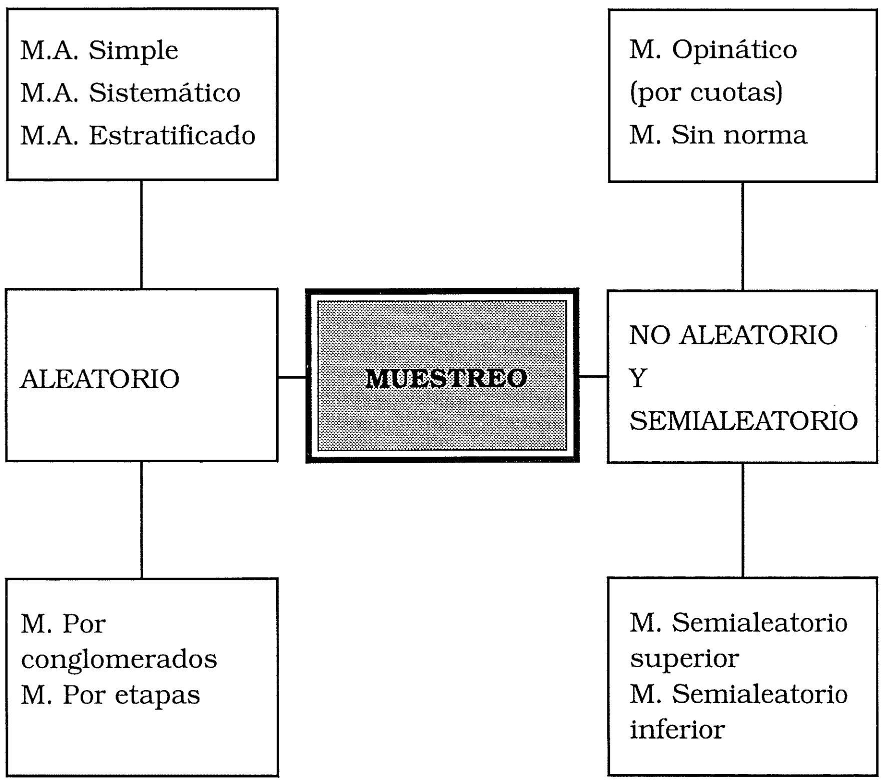
Población y muestra

Cuando el alumno llega a Bachillerato ya ha trabajado estos conceptos en la etapa obligatoria. Ahora debe notar paulatinamente que la información obtenida de una muestra representativa de la población aunque no es exacta -la Estadística no hace milagros- si es rigurosa desde el punto de vista científico (tendrá que dar sentido a conclusiones tales como «de la muestra deducimos que la duración media de las bombillas «Luce-bien» está comprendida entre 980 h y 1210 h con una probabilidad del 95%»).

Aunque sólo se trabaje con muestras aleatoria simples, se pueden mencionar otros tipos de muestreos.

Presentemos a continuación un ejemplo, quizá de los más difíciles que se pueden hacer en estos niveles pero no rehusemos a otros más sencillos de intervalos de confianza.

Ejemplo 1.
A un Centro de Salud están adscritos 100.000 niños y se quiere averiguar la proporción de vacunados pero sin hacer una observación exhaustiva[1].



1. Corresponde a una pregunta hecha por mi compañero y amigo Florencio Pascual. Se ha aumentado el tamaño de la población que en realidad era de 1200 para poder considerarla infinita y facilitar la tarea.

Mediante la elección de una muestra ganaremos tiempo y dinero, ¿qué estamos dispuestos a perder?

Fijaremos un grado de confianza, p_k=0.90, y un grado de precisión o error máximo admisible, e=2%. Esto quiere decir que el porcentaje de vacunados que obtengamos con la muestra no diferirá en más del 2% del verdadero porcentaje de la población con una probabilidad del 90%.

Sea **n** el tamaño de la muestra –nuestra incógnita– y sea **p** la verdadera proporción de vacunados –parámetro desconocido–.

La proporción de niños vacunados de cualquier muestra de tamaño n es una variable aleatoria –llamémosla $\overline{\mathbf{X}}$ ya que en realidad es la media de variables de Bernoulli– cuya distribución es aproximadamente

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

De acuerdo con las condiciones impuestas se debe cumplir:

$$p\left[\,|\overline{X}-p| \le 0.02\,\right] \ge 0.90 \leftrightarrow$$

$$\leftrightarrow p\left[\left|\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \le \frac{0.02}{\sqrt{\frac{p(1-p)}{n}}}\right] \ge 0.90$$

Como aproximadamente $\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \rightarrow N(0,1) \rightarrow \frac{0.02}{\sqrt{\frac{p(1-p)}{n}}} \ge 1.64$

Sabiendo que p(1-p) es máximo cuando p=1/2, se deduce fácilmente que a partir de n=1681 se cumplen las condiciones impuestas y, por tanto, ese debe ser el tamaño de la muestra.

Idea de independencia

El alumno ya ha lanzado monedas y dados, ha extraído bolas de urnas, ... y, por tanto, tiene adquirida la idea intuitiva de experimentos o pruebas dependientes e independientes. Cuando se introduzca la definición de probabilidad condicional se formalizará de manera natural la noción de independencia de sucesos y ello se utilizará para saber cuándo las variables aleatorias son independientes.

La idea de independencia es muy importante, según Kolmogorov en ella estriba la diferencia entre el Cálculo de Probabilidades y la Teoría de la Medida.

Desearía haber introducido aquí un contraste no paramétrico para la independencia de dos variables de la población mediante el test de la χ^2 pero en el artículo, como en nuestras aulas, debemos establecer prioridades.

Distribuciones binomial y normal

La ley Binomial aparece siempre que una prueba alaetoria con sólo dos resultados posibles (éxito y fracaso) se repite varias veces de manera independiente. Veamos un nuevo ejemplo.

Ejemplo 2. *El 25% de los conejos contraen cierta enfermedad. Se quiere probar una vacuna para lo cual se inyecta a 24 animales sanos y se comprueba que, al cabo de un tiempo, 3 han enfermado. ¿Es efectivo el tratamiento?*

Una primera respuesta intuitiva nos lleva a pensar que la vacuna es efectiva -el número de animales enfermos es la mitad de la cuarta parte del total-. Planteemos un test y veamos lo que sucede.

La probabilidad de que un animal enferme es p=0.25 y supongamos que se mantiene aun cuando se le aplique la vacuna. Esta es nuestra hipótesis nula H_0: p=0.25.

La hipótesis alternativa será p<0.25, es decir, la vacuna es efectiva.

Llamemos X a la variable «número de animales que enferman» y observemos que supuesta cierta H_0, $X \rightarrow B(24, 0.25)$. Fijemos α=0.05 como nivel de significación o error de primera especie (probabilidad de rechazar la hipótesis nula siendo cierta) y hallemos la región de rechazo, [0, c], de modo que

$$p_{H_0}\ (\ X \in [\ 0,\ c\ ]\ )\ \leq\ 0.05.$$

$$p_{H_0}\ (\ X \leq 2\ ) = \binom{24}{0} 0.25^0\ 0.75^{24} +$$

$$+ \binom{24}{1} 0.25^1\ 0.75^{23} + \binom{24}{2} 0.25^2\ 0.75^{22} =$$

$$= 0.0398 \leq 0.05$$

Sin embargo,

$$p_{H_0}\ (\ X \leq 3\ ) = 0.115 \nleq 0.05$$

con lo cual c=2 y la región de rechazo será el intervalo cerrado [0, 2]. Entonces, como son 3 los animales que han enfermado, no existe la evidencia suficiente como para rechazar la hipótesis nula y, por tanto, no podemos afirmar que la vacuna sea efectiva en contra de lo que a priori se pudiera pensar.

La distribución Normal es la más importante y su conocimiento y manejo son incuestionables.

Conocer las propiedades y aplicaciones no lleva implícito grandes desarrollos teóricos. ¿No podemos prescindir, incluso, de su función de densidad? Pero sí es necesario calcular probabilidades manejando las tablas, saber que la media y la varianza caracterizan totalmente la distribución y admiten una interpretación conjunta muy clara -extensible a gran cantidad de distribuciones-, conocer su forma y tipificar cualquier conjunto de datos.

Por último, la importancia de la Normal se manifestará de manera rotunda con el teorema Central del Límite y ello hará que en todos los casos de Inferencia Estadística estudiados con posterioridad sea «razonable» trabajar bajo la hipótesis de normalidad.

Del teorema Central del Límite los alumnos deben sacar una idea intuitiva: cuando los resultados de un experimento son debidos a un número *grande* de causas *independientes* que actúan *sumando* sus efectos -siendo cada efecto individual de poca importancia respecto al conjunto- es esperable que los resultados sigan una distribución Normal.

Podría a determinados alumnos introducírsele la distribución de Poisson aunque no es aconsejable cargar los programas de contenido en detrimento de otras partes que pudieran quedar incompletas. El siguiente cuadro representa las aproximaciones entre las distribuciones Binomial, de Poisson y Normal.

Desarrollemos el ejemplo siguiente.

Ejemplo 3. *En un Centro de Primaria[2] se han llevado a cabo sendos programas de intervención en alumnos del Ciclo Medio con dificultades en Lengua Castellana y Matemáticas. Para intentar medir la eficacia de los mismos se han pasado dos pruebas, una inicial y otra final, a los alumnos sobre los que se ha intervenido, siendo éstos los resultados:*

Lengua Castellana

Alumno	*P. Inicial: X*	*P. Final: Y*
1	*17*	*20*
2	*21*	*40*
3	*29*	*30*
4	*22*	*20*
5	*33*	*40*
6	*32*	*30*
7	*23*	*40*
8	*20*	*15*
9	*36*	*40*
10	*30*	*35*
11	*21*	*40*
12	*19*	*20*
13	*30*	*30*

APROXIMACIONES ENTRE DISTRIBUCIONES

BINOMIAL (n,p)	$1 < np = \lambda < 5$ $p < 0.1$ $n > 30$	POISSON (λ)
$np > 5$ $y\ \ nq > 5$		$\lambda > 5$
	NORMAL (μ, σ)	

2. Los datos han sido facilitados por Mª Teresa García y corresponden a un programa de compensación educativa llevado a cabo en el curso 1992-93.

Matemáticas

Alumno	*P. Inicial: X*	*P. Final: Y*
1	*23*	*35*
2	*10*	*40*
3	*50*	*60*
4	*28*	*70*
5	*25*	*50*
6	*15*	*40*
7	*21*	*60*
8	*34*	*45*
9	*45*	*75*
10	*25*	*75*
11	*39*	*65*
12	*35*	*40*
13	*25*	*45*
14	*0*	*25*
15	*15*	*35*

Fijémonos en los resultados de Lengua ya que los de Matemáticas parecen contundentes a simple vista.

Supondremos que las variables son Normales[3] y que tienen la misma variabilidad.

En Lengua se aprecia progreso, pero ¿es lo suficientemente considerable -significativo- como para atribuirlo al programa y no simplemente a estos alumnos?

En principio se puede pensar en atacar la cuestión de dos maneras:

¿De manera global el rendimiento en la prueba final es superior al de la inicial?

O, si consideramos la diferencia entre las puntuaciones de todos y cada uno de los alumnos, ¿son positivas?

Estas dos formas a priori parecen similares pero no lo son. Plantear la primera pregunta supone ignorar el carácter dependiente entre las puntuaciones iniciales y finales. Debemos, pués, hallar en cada alumno la diferencia de sus puntuaciones obteniendo así una nueva variable, D, que con nuestros datos toma los valores: 3, 19, 1, -2, 7, -2, 17, -5, 4, 5, 19, 1, 0.

Como decíamos anteriormente estos datos mayoritariamente positivos parecen indicar la eficacia del programa, pero ¿no podría haber sido esta situación fruto del azar? Démosle al azar un margen del 5%.

Vamos a suponer mientras no se demuestre lo contrario que el programa no produce mejora; ésa será la hipótesis nula (H_0: $\mu=0$).

La hipótesis alternativa, H_1, será $\mu>0$.

Bajo H_0 la media muestral de D, $\bar{d}$ sigue aproximadamente[4] una distribución Normal de media 0 y varianza s_{n-1}^2/n.

Entonces el estadístico $\dfrac{\bar{d}}{\frac{s_{n-1}}{\sqrt{n}}}$

es aproximadamente N(0, 1), por tanto la región de rechazo será: [1.64, +∞).

En el caso de nuestros datos, el valor de la variable $\dfrac{\bar{d}}{\frac{s_{n-1}}{\sqrt{n}}}$ es

$$\frac{5.1538}{\frac{8.1634}{\sqrt{13}}} = 2.187 \; \varepsilon \; [\, 1{,}64 \,,\, +\infty \,).$$

Se halla en la región de rechazo y, por ello, el programa es efectivo.

En el caso de Matemáticas el valor que toma la variable es 7.62, muy superior al valor crítico 1.64, luego el programa también se muestra eficaz.

Observar que s_{n-1} no es la desviación típica de la muestra, ya que ésta subestima a la de la población por término medio, sino la desviación típica corregida. Como los alumnos utilizarán la calculadora ya habrán apreciado dos teclas diferentes, s_n y s_{n-1}, y tal vez nos pregunten la diferencia entre ellas. El momento de aclarar la cuestión puede ser cuando trabajen la propiedad de insesgadez

3. La utilización del papel probabilístico Normal da una idea muy gráfica sobre la suposición de normalidad de los datos -véase el libro de Peña-.
4. En realidad es una distribución t de Student con 12 grados de libertad pero si ésta no se ha introducido se puede considerar como buena la aproximación Normal.

de los estimadores (la media del estimador es igual al parámetro que estima).

Números Aleatorios

Las tablas de números aleatorios se utilizan para hacer cualquier tipo de sorteo o juego o simular cualquier experimento. Con ellas podemos simular lanzamientos de monedas o dados, extracciones de naipes, determinar el número de monedas que esconderemos en la mano en cada partida de «chinos» para evitar que un hábil contrincante halle la tendencia involuntaria, o averiguar la distribución de los 100 primeros copos de nieve que caen en un patio embaldosado[5].

Es deseable que los niños en las etapas primaria y secundaria obligatoria utilicen e incluso construyan estas tablas. Si han adquirido soltura en el manejo tanto de esas -que corresponden a una distribución Uniforme- como de la Normal pueden generar por el método de Montecarlo una distribución χ^2 de Pearson. La tabla siguiente permite anotar de manera clara y ordenada los valores obtenidos artificialmente de una χ^2_4 (suma de los cuadrados de 4 distribuciones N(0, 1) independientes).

Otro ejercicio interesante consiste en comprobar la alaetoriedad de una tabla de números supuestamente aleatorios mediante el test de Poker[6]. Para ver la bondad del ajuste de las frecuencias observadas a las probabilidades teóricas se utiliza la prueba de la χ^2. Rechazar la hipótesis nula significaría que la tabla de números no era aleatoria.

A modo de final

Es difícil escribir el final de un artículo cuando éste se considera inconcluso, pero todo tiene un límite y seguro que éste ha sido pasado con creces.

Seguramente estáis pensando en muchas partes de la Estadística Inductiva unidimensional que podrían tratarse en el Bachillerato y aquí no se han mencionado ni siquiera de manera implícita; de hecho detrás de todos los ejemplos desarrollados tan solo están subyacentes la estimación puntual y por intervalos de confianza y el contraste de hipótesis paramétricos.

¡Ojalá sean esos vuestros pensamientos!

Ello querría decir que para vosotros la Estadística ha dejado de ser la parte de las Matemáticas relegada en vuestros programas o de tan tardía aparición en los mismos como lo ha hecho en la propia Historia.

Si no lo hemos hecho aún, entremos con cautela y decisión en un mundo más próximo a la realidad de lo que no hace mucho tiempo se pudiera pensar.

Bibliografía

* PEÑA SÁNCHEZ DE RIVERA, D.(1991): **Estadística. Modelos y métodos. 1. Fundamentos**. Madrid,Alianza.

* ENGEL, A.(1988): **Probabilidad y Estadística. Vol. 1 y 2.** Consorci d'Editors Valencians S.A., Valencia.

* AZORÍN F.y SÁNCHEZ-CRESPO J.L.(1986): **Métodos y aplicaciones del muestreo**. Alianza, Madrid.

* TURNER J.C.(1986): **Matemática moderna aplicada. Probabilidades, Estadística e Investigación Operativa**. Alianza, Madrid.

* FELLER W.(1975): **Introducción a la Teoría de Probabilidades y sus aplicaciones. Vol. I.** Limusa, Méjico.

Manuel Sobrino Reyes

M. a. s. U(0,1)				M. a. s. N(0,1)				χ_4^2
x_1	x_2	x_3	x_4	$z_1=\phi^{-1}(x_1)$	$z_2=\phi^{-1}(x_2)$	$z_3=\phi^{-1}(x_3)$	$z_4=\phi^{-1}(x_4)$	Σz_i^2

5. Un ejemplo menos ecológico que ese pero muy interesante sobre el bombardeo aéreo sobre Londres en la Segunda Guerra Mundial puede verse en los libros de Feller y Peña.
6. Ver los libros de Engel y de Azorín citados en la bibliografía.