10

METODOS ESTADISTICOS

José Jiménez

La estadística puede definirse como un método de razonamiento que permite interpretar datos cuyo carácter esencial es la variabilidad. Está presente en la práctica médica cada vez con más frecuencia y en muy diversas formas, desde las estadísticas de actividad de un hospital o los resultados de auditorías, por ejemplo, hasta los hallazgos de estudios de investigación que aparecen en la literatura médica.

En investigación, la finalidad de la estadística es utilizar datos obtenidos en una *muestra* de sujetos para realizar inferencias válidas para una población más amplia de individuos de características similares. La validez y utilidad de estas inferencias dependen de cómo el estudio ha sido diseñado y ejecutado, por lo que la estadística debe considerarse como una parte integrante del método científico. Muchos profesionales creen que se trata simplemente de un conjunto de fórmulas y cálculos matemáticos que se aplican a un conjunto de datos. Si bien el análisis de datos es la parte más visible de la estadística, deben tenerse en cuenta los aspectos metodológicos relacionados con el estudio. La justificación del análisis no radica en los datos, sino en la forma en que han sido recogidos.

Habitualmente se distingue entre estadística descriptiva, que comprende la organización, presentación y síntesis de datos de una manera científica, y estadística inferencial, que comprende las bases lógicas mediante las cuales se establecen conclusiones relacionadas con poblaciones a partir de los resultados obtenidos en muestras. Las técnicas estadísticas pueden utilizarse para confirmar *hipótesis* de trabajo o bien para explorar conjuntos de datos sin hipótesis previas. Ambas finalidades, la confirmación y la exploración, están vinculadas a la naturaleza de los objetivos del estudio, a la actitud con que el investigador se enfrenta a los datos y a los términos en que deberán interpretarse los resultados. Una hipótesis se confirma cuando se diseña un estudio con el propósito de hacerlo. Se explora cuando se rastrean datos en busca de información, sin objetivos concretos y formales que hayan gobernado el diseño del estudio. La exploración puede servir para sugerir nuevas hipótesis, pero de ningún modo para contrastarlas, sino que la confirmación deberá obtenerse en un nuevo estudio diseñado *específicamente* para ello.

Para las finalidades de este capítulo, consideraremos que existen dos grandes tipos de estudio: los que tienen por objetivo estimar un parámetro a partir de observaciones obtenidas en una muestra (por ejemplo, determinar el porcentaje de errores de medicación en

un hospital), y los que contrastan hipótesis mediante la comparación de dos o más grupos (por ejemplo, determinar cuál de dos estrategias es más eficaz para reducir el porcentaje de infecciones quirúrgicas).

ESTUDIOS DE ESTIMACION DE UN PARAMETRO

Principio de representatividad

En estadística, el término población se utiliza para describir todas las posibles observaciones de una determinada variable o todas las unidades sobre las que podría haberse realizado una observación. Puede tratarse de pacientes, de profesionales o de prescripciones terapéuticas, por ejemplo. Habitualmente se estudian muestras en lugar de poblaciones por criterios de eficiencia. El término muestra se refiere a cualquier conjunto específico de sujetos u observaciones procedentes de una población determinada. Para que sea útil y la estadística aplicable, se requiere que la muestra tenga un tamaño razonable y sea representativa de la población de la que procede. Un tamaño elevado no asegura la representatividad, sino que ésta radica básicamente en que la muestra haya sido escogida adecuadamente y esté libre de sesgos.

En cualquier estudio pueden considerarse tres niveles de población:

Población diana, a la que hace referencia el objetivo del estudio, y a la que se desearía generalizar los resultados.

Población de estudio, a la que se tiene la intención de estudiar, definida por los criterios de selección establecidos en el protocolo del estudio.

Muestra o conjunto de individuos realmente estudiados.

La validez de las conclusiones de un estudio dependen de cómo haya sido diseñado, de si la muestra es representativa, de si no se han producido pérdidas o no respuestas, de si las mediciones se han realizado correctamente y son de calidad, etc. (validez interna). Por otro lado, la capacidad para generalizar las conclusiones o extrapolarlas a otras poblaciones diferentes de la estudiada dependen de las diferencias entre la población diana y la de estudio, y entre éstas y la población a la que se quiera aplicar los resultados (validez externa).

Para que los resultados de un estudio tengan validez interna, la muestra de sujetos estudiada debe ser representativa de la población de estudio (principio de representatividad). Este principio puede verse comprometido cuando la muestra inicial ha sido mal seleccionada, cuando, aunque se haya utilizado una técnica de muestreo adecuada, la variabilidad aleatoria (el azar) ha hecho que se obtenga una muestra no representativa, o bien cuando la muestra de sujetos finalmente analizados está sesgada debido a las no respuestas (sujetos de los que no se ha podido obtener la información deseada).

Intervalos de confianza

En un estudio, tan sólo se estudia una de las múltiples muestras que podrían haberse obtenido de la población de referencia. Si se estudiara más de una,

en cada una de ellas el resultado podría presentar valores diferentes simplemente por azar. Las diferentes técnicas de la estadística inferencial se fundamentan en que esta variabilidad inherente al proceso de muestreo sigue unas leyes conocidas y puede ser cuantificada.

Si la variable es cuantitativa, la media *m* y la desviación estándar *s* observadas en la muestra son la mejor estimación que se dispone de los verdaderos valores de los parámetros poblacionales. Pero ¿cuáles serían los resultados si se repitiera el estudio en múltiples ocasiones?

Supongamos que en una muestra de 60 sujetos se observa una media de tensión arterial sistólica (TAS) de 150 mmHg con una desviación estándar de 20 mmHg. Se desea conocer el verdadero valor de la TAS media en la población de referencia. El valor más probable es el observado en la muestra (150 mmHg), conocido por ello como estimación puntual. Pero éste no es más que el resultado observado en una de las múltiples muestras que hubieran podido obtenerse de la misma población. Dado que diferentes muestras podrían conducir a diferentes resultados, se necesita una medida de la precisión de esta estimación, lo que se hace mediante el cálculo del llamado intervalo de confianza (IC). Por ello, siempre que se estimen parámetros poblacionales a partir de estadísticos muestrales, los resultados deben expresarse como IC, y no sólo como estimaciones puntuales.

Si se desea una confianza del 95% en la estimación, se trabaja con un valor α del 5%, que corresponde a un valor Z (distribución normal tipificada) de 1.96. En el ejemplo, aplicando la fórmula de la tabla 1, se obtendría un IC del 95% que sería aproximadamente de 150 ± 5 mmHg, lo que significa que la TAS media de la población de referencia está situada entre 145 y 155 mmHg con un 95% de confianza. De forma similar se calcularía el IC en el caso de una variable cualitativa (tabla 1).

El cálculo del IC proporciona mucha más información que la simple estimación puntual, ya que permite evaluar la

Tabla 1. Cálculo del intervalo de confianza (IC) en la estimación de un parámetro poblacional.

IC DE UNA MEDIA (variable cuantitativa)*: $m \pm (Z \cdot ESM)$	siendo	$ESM = s\sqrt{n}$
IC DE UNA PROPORCION (variable cualitativa)**: $p \pm (Z \cdot ESP)$	siendo	$ESP = \sqrt{p\ (1-p)/n}$

m: Media observada en la muestra; s: Desviación estándar observada en la muestra; n: Número de individuos de la muestra; ESM: Error estándar de la media; p: Proporción observada en la muestra, ESP: Error estándar de la proporción; Z: Valor de la variable normal tipificada correspondiente al valor α para un nivel de confianza (1-α).

(*) Este cálculo se basa en la distribución normal. El valor de Z para un IC del 95% es 1.96. Para muestras de tamaño inferior a 30 individuos, este valor debe sustituirse por el de la distribución de la t de Student-Fisher para (n-1) grados de libertad.

(**) Las variables cualitativas no presentan una distribución normal. Las fórmulas de la tabla se basan en una aproximación a la normalidad, aplicable cuando los productos n·p y n·(1-p) son mayores de 5.

precisión con que el parámetro poblacional ha sido estimado, es decir, entre qué límites se tiene una determinada confianza de que esté situado su verdadero pero desconocido valor. Si se repitiera el estudio en 100 ocasiones, el IC incluiría el verdadero valor en 95 de ellas.

De las fórmulas se deduce que un aumento del número de sujetos produce un estrechamiento del intervalo, aumentando así la precisión de la estimación. Su amplitud depende también del nivel de confianza que se utilice, aumentando si se incrementa su valor convencional del 95% al 99%, por ejemplo.

En el cálculo del IC se asume que se ha estudiado una muestra aleatoria de la población de referencia. Al interpretarlo, hay que tener siempre en cuenta la posibilidad de existencia de otras fuentes de error no debidas al azar (errores sistemáticos o sesgos). Si éstos existen, o si la muestra no es aleatoria, el error de la estimación puede ser mayor que el sugerido por la amplitud del intervalo.

Tamaño de la muestra

En cualquier estudio, es importante determinar a priori el número de sujetos que es necesario incluir, aunque el resultado de este cálculo debe considerarse como orientativo, ya que se basa en asunciones que pueden ser incorrectas. La inclusión de un número excesivo de sujetos encarece el estudio, tanto desde el punto de vista económico como de los recursos humanos y físicos necesarios. Por otra parte, un estudio con un tamaño insuficiente estimará un parámetro con poca precisión. La amplitud del IC, es decir, la precisión de la estimación, depende del nivel de confianza utilizado, de la variabilidad del parámetro de interés y del número de sujetos estudiados. Cuanto menor sea la variabilidad del parámetro y mayor el número de sujetos, mayor precisión existirá en la estimación para un nivel de confianza determinado.

Para el cálculo del tamaño de la muestra debe conocerse:

La variabilidad del parámetro que se desea estimar. Si no se conoce, puede obtenerse una aproximación a partir de datos propios o de otras investigaciones, o un estudio piloto. En el caso de las variables cuantitativas se mide por la variancia, y en el de las cualitativas, por el producto $p \cdot (1-p)$.

La precisión con que se desea obtener la estimación, es decir, la amplitud deseada del IC. Cuanto mayor precisión se desee, más estrecho deberá ser este intervalo, y más sujetos deberán ser estudiados.

El nivel de confianza deseado. Habitualmente se fija en el 95%. Este valor indica el grado de confianza que se tendrá de que el verdadero valor del parámetro en la población se sitúe en el intervalo obtenido. Cuanto más confianza se desee, mayor será el número de sujetos necesario.

De estos tres elementos, sólo debe conocerse la variabilidad del parámetro, ya que tanto la precisión como el nivel de confianza son fijados en función de los intereses del investigador.

Estimación de una proporción

La fórmula para el cálculo del número de sujetos necesarios para estimar una proporción se presenta en la tabla 2. Supongamos que se desea estimar el porcentaje de pacientes ingresados en un servicio que requieren una dieta determinada. A partir de datos previos se supone que debe estar situado alrededor del 40% (p=0,40). Se quiere realizar la estimación con una precisión de ±4 % (i=0,04) y una confianza del 95% (1-α =0,95; Z =1,96). Aplicando la fórmula, puede determinarse que serían necesarios 576 sujetos. Esta cifra se convierte en 9.220 cuando se desea una precisión muy alta (i=0,01), o en tan sólo 92 si se es menos exigente (i=0,1).

Modificando cualquier valor, puede obtenerse un número de sujetos que se aproxime al "deseado" o al disponible. Debe evitarse esta manipulación del cálculo ya que, al reducir el número de sujetos que se van a estudiar, también disminuye el grado de precisión con que el parámetro va a ser estimado y aumenta la amplitud del IC.

En el cálculo del tamaño de la muestra debe tenerse en cuenta también la estrategia de análisis y cómo se presentarán los resultados. Así, por ejemplo, si los investigadores desean presentar el resultado en función del sexo, la estratificación hará que la estimación se haya obtenido en un número menor de sujetos por lo que la precisión en cada estrato será menor de la deseada.

En algunas ocasiones no se conoce el valor aproximado del parámetro que se está buscando. Si no existen datos de la literatura que resulten útiles, o si no puede realizarse una prueba piloto para obtener una primera aproximación a dicho valor, puede adoptarse la postura de la máxima indeterminación, que consiste en suponer que el porcentaje que se desea estimar se sitúa alrededor del 50%, ya que es el valor que requiere una mayor cantidad de individuos para una precisión determinada.

Estimación de una media

Cuando el objetivo del estudio es estimar una media, el cálculo del número de sujetos necesario es similar (tabla 2), con la diferencia que la medida de la variabilidad es la variancia de la distribución de la variable en la población. Supongamos que desea estimar la tensión arterial diastólica (TAD) de los pacientes diabéticos ingresados en un servicio. Por estudios previos, se conoce que la desviación estándar de la TAD

Tabla 2. Fórmulas para el cálculo del número de sujetos necesarios para la realización de un estudio cuyo objetivo es la estimación de una media o una proporción.

ESTIMACION DE UNA PROPORCION (Variable cualitativa) $N = (Z^2 \cdot P \cdot (1-P))/i^2$

ESTIMACION DE UNA MEDIA (Variable cuantitativa) $N = (Z^2 \cdot s^2)/i^2$

N: Número de sujetos necesarios; Z:Valor de Z correspondiente al riesgo α fijado (cuando α=0,05, Z=1,96); P: Valor de la proporción que se supone existe en la población; s^2:Variancia de la distribución de la variable cuantitativa que se supone que existe en la población.
i: Precisión con que se desea estimar el parámetro (2i es la amplitud del intervalo de confianza).

en sujetos diabéticos es de 25 mmHg (s=25 mmHg; s^2=625 mmHg). Se desea realizar la estimación con una confianza del 95% (1-α =0,95) y una precisión de ±5 mmHg (i=5). Aplicando la fórmula, se puede determinar que son necesarios 96 sujetos.

Corrección para poblaciones finitas

En los cálculos anteriores no ha intervenido el tamaño de la población, ya que se ha asumido que es infinito. Sin embargo, en muchas ocasiones, desea obtenerse una muestra de una población de tamaño conocido (finito). En esta situación, puede aplicarse la siguiente fórmula que ajusta el número de sujetos necesarios en función del tamaño de la población:

$$n_a = \frac{n}{1+\frac{n}{N}}$$

donde n_a es el número de sujetos necesarios, n es el número de sujetos calculado para poblaciones infinitas y N es el tamaño de la población de referencia.

En el ejemplo en que se había calculado que eran necesarios 576 sujetos para estimar el porcentaje de pacientes ingresados que requerían una dieta, si la población de referencia fuera de 1000 sujetos, aplicando la fórmula anterior podría determinarse que son necesarios 365 pacientes pacientes.

Corrección según el porcentaje esperado de no respuestas

El número de sujetos calculado debe ser ampliado en función del porcentaje de no respuestas que se espera que se produzcan, de forma que se asegure que se obtendrá información del número de pacientes deseado. Una fórmula para hacerlo es la siguiente:

$$Na = N\cdot(1/(1-R))$$

donde N representa el número de sujetos teórico, Na el número de sujetos ajustado y R la proporción esperada de no respuestas.

Supongamos que para realizar un estudio se ha calculado que son necesarios 300 sujetos (N=300) y que se espera un 20% de no respuestas (R=0,20). El número de sujetos que deberían iniciar el estudio sería Na=300(1/(1-0,2))=375 sujetos.

La utilización de esta fórmula asegura que el estudio mantenga la potencia estadística deseada pero no evita que se puedan producir sesgos si las no respuestas no se han producido aleatoriamente, es decir, si los sujetos de los que no se obtiene información son diferentes de aquellos de los que sí se obtiene (lo que suele ser lo habitual).

Muestreo

Para que se cumpla el principio de representatividad, debe prestarse atención al proceso de selección de los sujetos, utilizando una técnica de muestreo adecuada que aumente la probabilidad de obtener una muestra representativa.

El muestreo probabilístico se define como el proceso de selección en que todos los individuos candidatos tienen una probabilidad conocida, distinta de

cero, de ser incluidos en la muestra, utilizándose alguna forma de selección aleatoria para obtener las unidades que serán estudiadas. Tiende a asegurar que se obtendrá una muestra representativa, especialmente si la población y la muestra son de gran tamaño, pero también puede ocurrir que no sea así, ya que el propio azar puede conducir a una muestra que no tenga la misma distribución de las variables de interés que la población de referencia, especialmente si su tamaño es reducido.

La unidad de muestreo es el elemento sobre el que se aplica la técnica de selección, ya sean personas, servicios u hospitales. La unidad de muestreo no tiene por qué coincidir con la unidad de análisis. En un estudio para conocer la frecuencia de errores de medicación en un hospital, la unidad de muestreo pueden ser los servicios, y analizar en una muestra de ellos las prescripciones terapéuticas.

En las técnicas probabilísticas la selección de las unidades se realiza al azar, evitando la posible parcialidad, consciente o inconsciente, de los investigadores. Por esta razón, es más probable que las muestras tiendan a ser representativas de la población de referencia. En el muestreo aleatorio simple, se prepara un listado de las unidades de muestreo, numerándolas, por ejemplo, secuencialmente, y a continuación, se seleccionan tantos números aleatorios como elementos debe tener la muestra. El muestreo aleatorio estratificado es una modificación que intenta asegurar que la muestra presenta la misma distribución que la población en relación a determinadas variables, previniendo la aparición de sesgos debidos a las mismas. La población se divide en estratos en función de las categorías de las variables por las que se desea estratificar, es decir, se forman subgrupos de población que comparten alguna característica en común y son mutuamente excluyentes. A continuación, se escoge una muestra al azar en cada estrato, habitualmente manteniendo las proporciones observadas en la población de referencia (muestreo aleatorio estratificado proporcional). Es preciso que los estratos se delimiten en función de variables que puedan influir sobre los resultados.

El muestreo en múltiples etapas consiste en seleccionar unidades de muestreo de una población (unidades primarias, por ejemplo, servicios), y, en una segunda etapa, obtener una muestra de cada una de las unidades primarias seleccionadas (unidades secundarias, por ejemplo, pacientes ingresados). Se pueden usar el número de etapas que sean necesario y, en cada una de ellas, un método diferente de muestreo (simple, estratificado, sistemático). Cuando se incluyen todas las unidades secundarias, se denomina muestreo en conglomerados.

El muestreo sistemático se basa en aplicar alguna regla sistemática simple, como elegir uno de cada n individuos. En primer lugar, se calcula la constante de muestreo k, dividiendo el tamaño de la población candidata por el de la muestra. A continuación, se extrae la primera unidad al azar entre las k primeras unidades de muestreo y se le suma la constante sucesivamente hasta completar el tamaño de la muestra.

Tiene la ventaja de que es más cómodo y práctico que el muestreo aleatorio simple, y de que no siempre es necesario tener de antemano una lista completa y exhaustiva de toda la población. Además, cuando la población de referencia está ordenada siguiendo una tendencia conocida (de mayor a menor, de más viejo a más joven...), el muestreo sistemático asegura una cobertura de unidades de todos los tipos.

En muchos estudios, bien porque no se dispone de un listado con los miembros que forman la población de estudio o bien porque ésta es dinámica, la muestra de sujetos se selecciona por otros métodos no probabilísticos (por ejemplo, incluyendo consecutivamente a los pacientes que acuden a la consulta y cumplen los criterios de selección, o a voluntarios). En estos casos, para poder realizar inferencias válidas, debe poderse asumir que la muestra seleccionada es representativa de la población de estudio.

ESTUDIOS DE CONTRASTE DE HIPOTESIS

Principio de comparabilidad

En los estudios analíticos, además del principio de representatividad, debe cumplirse el de comparabilidad de los grupos. Estos estudios se basan en que los grupos son comparables por todos los factores pronósticos y en que se ha obtenido la información de la misma forma en todos ellos, de manera que las diferencias en los resultados observados puedan atribuirse al factor que se está estudiando. La función del grupo control es proporcionar una estimación del valor de la variable de respuesta en ausencia del factor de estudio. En otras palabras, debe permitir aislar el efecto del factor de estudio del debido a otros factores, por lo que el grupo control debe ser comparable al de estudio en todas aquellas variables que puedan influir sobre la respuesta o su medición.

El proceso de formación de los grupos depende del tipo de estudio. En los diseños observacionales, se realiza en función de la existencia o no de la enfermedad de interés (estudios de casos y controles) o de la presencia o no de la exposición (estudios de cohortes). En los estudios experimentales, los sujetos son asignados a los diferentes grupos que se desea comparar por un procedimiento aleatorio.

Contraste de hipótesis

La aplicación más frecuente de la inferencia estadística en investigación médica son las llamadas pruebas de contraste de hipótesis o de significación estadística. Supongamos que existe interés en comparar dos tratamientos (un diurético D y el tratamiento estándar E), y determinar cuál de ellos es el más eficaz en el control de las cifras tensionales. Para ello, se diseña un ensayo clínico controlado, distribuyendo aleatoriamente una muestra de pacientes hipertensos en dos grupos, cada uno de los cuales recibe uno de los tratamientos. A los tres meses, el porcentaje de individuos controlados en cada grupo es del 70 y 50%, respectivamente. ¿Qué conclusión puede obtenerse a la vista de estos resultados?

Lo que se quiere determinar es hasta qué punto es posible que la diferencia observada sea debida exclusivamente al azar (variaciones del muestreo).

Hipótesis nula e hipótesis alternativa

La hipótesis que en realidad se va a contrastar estadísticamente es la de que no existen diferencias entre los porcentajes de hipertensos controlados observados en ambos grupos. La prueba de significación estadística intentará rechazar esta hipótesis, conocida como hipótesis nula *Ho*. Si lo consigue, se aceptará la hipótesis alternativa *Ha* de que existen diferencias entre ambos grupos.

El primer paso es, pues, formular la Ho. A continuación, se calcula, mediante la prueba estadística más adecuada, la probabilidad de que los resultados observados puedan ser debidos al azar, en el supuesto de que Ho sea cierta. En otras palabras, la probabilidad de que, a partir de una población de referencia, puedan obtenerse dos muestras que presenten unos porcentajes tan diferentes como los observados. Esta probabilidad es el grado de significación estadística, y suele representarse con la letra p. Basándose en su valor, se decide si se rechaza o no Ho. Cuanto menor sea la p, es decir, cuanto menor sea la probabilidad de que el azar pueda haber producido los resultados observados, mayor será la evidencia en contra de Ho, y, por lo tanto, mayor será la tendencia a concluir que la diferencia existe en la realidad. El valor de p por debajo del cual se considerará que se dispone de la suficiente evidencia en contra de Ho para rechazarla, conocido como el nivel de significación estadística, debe fijarse previamente. De forma arbitraria, y por convenio, suele fijarse este valor en el 5% (0,05).

Supongamos que en el ejemplo se obtiene un valor de p de 0,10. Esto significa que, si Ho fuera cierta, la probabilidad de que el azar pueda producir unos resultados como los observados es del 10%, o bien, que existe un 10% de probabilidad de que dos muestras del tamaño de las estudiadas obtenidas de una misma población presenten unos porcentajes del 70 y 50% sólo por variabilidad aleatoria. Si se había prefijado el valor 0,05 para el nivel de significación, dado que el valor de p obtenido es superior, se considerará que la probabilidad de haber obtenido estos resultados por azar es demasiado elevada y que, por tanto, no se dispone de la suficiente evidencia para rechazar la Ho. Se concluye que no se han encontrado diferencias estadísticamente significativas en el porcentaje de pacientes controlados en ambos grupos. No se concluye que ambos grupos son iguales, sino que no se ha encontrado la suficiente evidencia para decir que son diferentes.

Supongamos que se hubiera obtenido un valor de p de 0,02. Como este valor es inferior al nivel de significación del 0,05, se considerará que la diferencia observada es estadísticamente significativa, ya que es poco probable ($p<5\%$) que el azar pueda haber producido estos resultados si la Ho fuera cierta. Se concluye por tanto que existe una diferencia entre los grupos. La respuesta a la pregunta de si esta diferencia es debida al nuevo tratamiento dependerá del diseño y ejecución correctas del estudio.

El verdadero interés de la p es el de permitir descartar que la diferencia observada es fruto de la variabilidad aleatoria. No es una medida de la fuerza de la asociación. Un estudio en el que se obtenga una p<0,001 no quiere decir que la asociación encontrada sea más fuerte (o la diferencia más importante) que otro estudio en que la p sea del 0,05. Sólo quiere decir que es más improbable que su resultado sea debido al azar. Por ello, no hay que ser excesivamente rígido en el límite del nivel de significación. Un valor p de 0,048 es estadísticamente significativo al nivel del 5%, y uno de 0,052, en cambio, no lo es, pero en ambos casos la probabilidad de observar el resultado por azar es prácticamente la misma, y muy próxima al 5%.

Pruebas unilaterales y pruebas bilaterales

En ocasiones, lo que interesa no es determinar si existen o no diferencias entre dos tratamientos, sino evaluar si un nuevo fármaco es mejor que otro. En este caso, la hipótesis alternativa no es que D y E difieren, sino que D es mejor que E. Por tanto, la Ho que se va a contrastar es que D no difiere o es peor que E. Dado que sólo interesa un sentido de la comparación, se habla de pruebas unilaterales o de una cola.

Este hecho no afecta al cálculo del estadístico, sino que modifica el grado de significación. Como la distribución de Z sigue la ley normal, y por lo tanto es simétrica, en las pruebas unilaterales el verdadero valor de p es la mitad del valor α, dado que sólo se está interesado en uno de los extremos.

Error α y error β

En estadística no puede hablarse de certeza absoluta, sino de mayor o menor probabilidad. Sea cual sea la decisión que se tome respecto a la Ho, se corre un cierto riesgo de equivocarse (tabla 3). La realidad no es conocida, ya que, si lo fuera, no sería necesario realizar el estudio. Si no se rechaza la Ho, y ésta es cierta, no se comete ningún error. Si se rechaza y es falsa, tampoco se comete un error. Pero, ¿qué pasa en las otras situaciones?

En un estudio, puede concluirse que existen diferencias cuando de hecho no las hay. Es decir, puede rechazarse Ho cuando es cierta. Si ésto ocurre, la decisión es incorrecta y se comete un error, conocido como error tipo I o error α. La probabilidad de cometer este tipo de error es la de que se concluya que existen diferencias significativas cuando en realidad son debidas al azar. Si se hace un símil entre una prueba estadística y una diagnóstica, equivale a la probabilidad de obtener un resultado falso positivo. Esto es precisamente lo que mide el valor de p o grado de significación estadística.

Si, por el contrario, se concluye que no existen diferencias estadísticamente significativas, es decir, si no puede rechazarse la Ho, puede ocurrir que en realidad ésta sea falsa y sí existan diferencias entre ambos grupos, en cuyo caso se comete otro tipo de error, llamado error β o tipo II. Utilizando el símil con la prueba diagnóstica, equivale a la probabilidad de obtener un resultado falso negativo. Su valor complementario (1-β), denominado potencia o

Tabla 3. Tipos de error aleatorio en una prueba estadística de contraste de hipótesis.

		REALIDAD (POBLACION)	
		EXISTE DIFERENCIA O ASOCIACION(Ho falsa)	NO EXISTE DIFERENCIA O ASOCIACION(Ho cierta)
RESULTADO DE LA PRUEBA (MUESTRA)	DIFERENCIA O ASOCIACION SIGNIFICATIVA (Rechazo de Ho)	NO ERROR	ERROR TIPO I α
	DIFERENCIA O ASOCIACION NO SIGNIFICATIVA (No rechazo de Ho)	ERROR TIPO II β	NO ERROR

Ho: Hipótesis nula

poder estadístico, indica la capacidad que tiene la prueba para detectar una diferencia cuando ésta ya existe en la realidad.

Lógicamente, cuanto mayor es la diferencia existente entre dos poblaciones y mayor el número de individuos estudiados, mayor capacidad existe para detectarla, es decir, el poder estadístico es mayor y, por lo tanto, la probabilidad de cometer un error tipo II es menor. No es lo mismo concluir que no se ha encontrado una diferencia estadísticamente significativa cuando se tiene una probabilidad del 90% de haberla detectado si hubiera existido (β=0,10), que cuando esta probabilidad es sólo del 50% (β=0,50).

¿Diferencia estadísticamente significativa o clínicamente relevante?

Un resultado estadísticamente significativo no implica necesariamente que sea clínicamente relevante. El valor de p no mide la fuerza de la asociación. Pueden obtenerse valores pequeños de p (y, por lo tanto, resultados estadísticamente significativos), simplemente estudiando un elevado número de sujetos ya que al aumentar el tamaño de la muestra, se incrementa el poder estadístico para detectar incluso pequeñas diferencias.

La diferencia que se considera clínicamente relevante depende de su magnitud y de otros factores, tales como la frecuencia y gravedad de los efectos secundarios, la facilidad de administración o su coste económico, por ejemplo, cuando se trata de comparar la eficacia de dos fármacos.

Elección de la prueba estadística

La elección de la prueba estadística depende de:

La escala de medida de la variable de respuesta. Las pruebas estadísticas tienen una mayor potencia si la variable de respuesta es cuantitativa, ya que contiene más información que si fuera cualitativa.

La escala de medida del factor de estudio. Puede ser cualitativa dicotómica (tratamiento activo/placebo, exposición/no exposición), cualitativa con más de dos categorías (tres pautas terapéuticas, o diferentes niveles de exposición a un factor de riesgo) o cuantitativa (valores de la colesterolemia o la presión arterial).

El carácter apareado o independiente de los datos. Desde el punto de vista estadístico, se habla de medidas repetidas o apareadas cuando han sido realizadas sobre los mismos sujetos (por ejemplo, comparación de las cifras de presión arterial obtenidas en los individuos de una muestra al inicio y al final de un determinado período de tiempo). Dado que los sujetos son los mismos, existe una menor variabilidad en las mediciones, lo que permite utilizar pruebas más potentes que tengan en cuenta este fenómeno. En caso de que los grupos que se comparan estén formados por individuos diferentes, se habla de datos independientes.

Las condiciones de aplicación específicas de cada prueba. Las pruebas estadísticas que utilizan datos cuantitativos suelen realizar determinadas asunciones sobre la distribución de las variables en las poblaciones que están siendo comparadas. Estas pruebas son conocidas como pruebas paramétricas. La mayoría son robustas, es decir, que toleran relativamente violaciones de estas asunciones, especialmente si el número de sujetos estudiado es elevado. En muchas situaciones, especialmente cuando las muestras son de pequeño tamaño, no se puede determinar si se cumplen dichas asunciones. En estos casos, se recurre a otras pruebas estadísticas menos potentes, que no requieren asunciones para su aplicabilidad, conocidas como pruebas no paramétricas. Este mismo tipo de pruebas es aplicable cuando se trata de analizar datos ordinales.

En la tabla 4 se resumen las pruebas estadísticas que se utilizan en las situaciones más frecuentes. Cuando tanto el factor de estudio como la variable de respuesta son variables cualitativas, la prueba estadística más apropiada para determinar si existe asociación entre ellas es la ji al cuadrado, siempre que exista un número suficiente de sujetos en cada una de las casillas de la tabla de contingencia. Cuando se comparan dos grupos (factor de estudio dicotómico) respecto a una variable cuantitativa, la prueba estadística más adecuada es la t de Student-Fisher, si se cumplen las condiciones necesarias para su aplicación. En caso contrario, debe recurrirse a una prueba no paramétrica equivalente, como la U de Mann-Whitney.

Si se comparan más de dos grupos (factor de estudio con más de dos categorías) respecto a una variable cuantitativa, debe utilizarse el análisis de la variancia (ANOVA). Si no se cumplen los criterios de aplicación del análisis de la variancia, debe recurrirse a la prueba de Kruskal-Wallis. Si se trata de determinar la posible asociación entre un factor de estudio y una variable de respuesta cuantitativos, la prueba adecuada es la correlación de Pearson, o, si no se cumplen las condiciones de aplicación, la correlación de Spearman. En el caso de que pueda asumirse una relación de dependencia lineal de una de las varia-

Tabla 4. Pruebas bivariantes de significación estadística utilizadas con mayor frecuencia.

Factor de estudio		Variable de respuesta			
		Cualitativa nominal		Cualitativa ordinal	Cuantitativa (*)
		2 categorías	>2 categorías		
Cualitativo (dos grupos)	Independientes	Ji al cuadrado Prueba de Fisher	Ji al cuadrado	U de Mann-Witney	t de Student-Fisher Prueba de Welch
	Apareados	Prueba de McNemar Prueba de Fisher	Q de Cochran	Prueba de los signos Prueba de los rangos signados de Wilcoxon	t de Student-Fisher datos apareados
Cualitativo (más de dos grupos)	Independientes	Ji al cuadrado	Ji al cuadrado	Prueba de Kruskal-Wallis	Análisis de la variancia
	Apareados	Q de Cochran	Q de Cochran	Prueba de Friedman	Análisis de la variancia de medidas repetidas
Cuantitativo		t de Student-Fisher	Análisis de la variancia	Correlación de Spearman	Correlación de Pearson Regresión lineal simple

(*) Cuando las pruebas estadísticas aplicables a las variables cuantitativas no cumplen las asunciones necesarias para su uso, se recurre a las pruebas correspondientes como si la variable de respuesta fuera ordinal (pruebas no paramétricas).

bles respecto a la otra, se habla de regresión lineal simple.

Tamaño de la muestra

Para realizar el cálculo del tamaño de la muestra necesario para comparar dos grupos, deben utilizarse los siguientes elementos:

Definir la hipótesis que se va a contrastar, precisando si es unilateral o bien bilateral.

Establecer el riesgo de cometer un error α que se está dispuesto a aceptar. Habitualmente suele aceptarse un 5%, y preferiblemente con hipótesis bilaterales, ya que son más conservadoras.

Establecer, asimismo, **el riesgo que se acepta de cometer un error β.** Habitualmente se sitúa entre el 5 y el 20%. A menudo, es más fácil enfrentar esta decisión a partir del concepto de poder o potencia estadística (1-β), que es la capacidad del estudio para detectar una determinada diferencia. Aceptar un riesgo de cometer un error β del 20%, significa que, si la diferencia que se busca existe en la realidad, el estudio tiene un 80% de probabilidades de detectarla.

Definir la mínima magnitud de la diferencia, efecto o asociación, que se desea ser capaz de detectar. Debe estar basada en datos de estudios previos o de la literatura que definan el rango de valores esperables, y en la mínima magnitud que se considera de relevancia clínica.

Es necesario, también, disponer de alguna **medida de la variabilidad de la variable de respuesta** en la población o grupo de referencia.

De estos cinco elementos, sólo el último debe ser conocido, ya que los otros cuatro son fijados por el investigador. A continuación, se aplica la fórmula correspondiente (tabla 5).

Supongamos un estudio que tiene por objetivo determinar si un nuevo tratamiento T consigue un mayor porcentaje de éxitos en las sobreinfecciones respiratorias que el tratamiento estándar E. Lo primero que debe conocerse es el porcentaje de curaciones en pacientes de características similares a los que van a ser estudiados obtenido con el tratamiento estándar E. Supongamos que esta cifra se sitúa alrededor del 40% (P1=0,4). El siguiente paso es determinar la diferencia mínima que se desea detectar, es decir, responder a la siguiente pregunta: ¿A partir de qué porcentaje de éxitos con el nuevo tratamiento se considerará que éste es mejor que E, y, por lo tanto, se estará dispuesto a modificar la pauta terapéutica habitual? Es decir, si el porcentaje de indi-

Tabla 5. Fórmulas para el cálculo del número de sujetos necesarios por grupo en un estudio cuyo objetivo es la comparación de dos muestras del mismo tamaño.

COMPARACION DE DOS PROPORCIONES (Variable cualitativa)

$$N = \frac{\left[Z\alpha \sqrt{2\ P\ (1-P)} + Z\beta \cdot \sqrt{P1 \cdot (1-P1) + P2\ (1-P2)}\right]^2}{(P1-P2)^2}$$

COMPARACION DE DOS MEDIAS (Variable cuantitativa)

$$N = [2 \cdot (Z\alpha + Z\beta)^2 \cdot s^2]/d^2$$

N: Número de sujetos necesarios en cada uno de los grupos; Zα: Valor de Z correspondiente al riesgo α fijado (cuando α=0,05, Zα=1,96 en hipótesis bilateral y Zα=1,645 en unilateral); Zβ: Valor de Z correspondiente al riesgo β fijado (cuando β=0,20, Zβ=0,842; cuando β=0,10, Zβ=1,282; cuando β=0,05, Zβ=1,645); P1: Valor de la proporción que se supone que existe en el grupo de referencia; P2· Valor de la proporción que se supone que existe en el grupo de estudio; P2-P1: Valor mínimo de la diferencia que se desea detectar (variable cualitativa); P: Media ponderada de las proporciones P1 y P2; s^2: Variancia de la distribución de la variable cuantitativa que se supone que existe en el grupo de referencia; d: Valor mínimo de la diferencia que se desea detectar (variable cuantitativa).

viduos curados con T es del 41%, ¿puede considerarse que esta diferencia del 1% es un resultado lo suficientemente importante para modificar la pauta terapeútica? ¿O se exigirá un mínimo, por ejemplo, del 50% de éxitos? La respuesta a esta pregunta depende de muchos factores, tales como la seguridad del fármaco, la facilidad de administración o el coste, entre otros.

Supongamos que los investigadores consideran que, si se cura el 50 % de pacientes con T (P2=0,5), se aceptará como la elección terapeútica. A continuación, sólo falta determinar los niveles de riesgo de cometer algún tipo de error aleatorio que se está dispuesto a asumir. Supongamos que se acepta el nivel de riesgo α habitual del 5% con una hipótesis bilateral y un riesgo β del 20% (potencia: 1- β=0,80). Aplicando la fórmula puede calcularse que son necesarios 387 sujetos por grupo de estudio. Esta cifra indica el número de sujetos que deben finalizar el estudio para tener un 80% de probabilidades de detectar una diferencia igual o superior a la fijada, con un nivel de error α del 5%. Por lo tanto, hay que incrementarlo en función del número de pérdidas de seguimiento y de abandonos que se prevea que ocurrirán durante el estudio, aplicando la misma fórmula que se ha presentado en el caso de la estimación de parámetros.

Estimación frente a significación estadística

En realidad, cuando analizan los resultados de un estudio, los investigadores están interesados no sólo en saber si una diferencia o asociación es estadísti-

Tabla 6. Cálculo del intervalo de confianza (IC) de la diferencia entre dos proporciones.

IC DE LA DIFERENCIA DE DOS PROPORCIONES (*)

a) MUESTRAS INDEPENDIENTES: $(P_A - P_B) \pm Z \cdot ESD$

$$ESD = \sqrt{\frac{P_A \cdot (1-P_A)}{n_A} + \frac{P_B\ (1-P_B)}{n_B}}$$

b) MUESTRAS APAREADAS $(P_A - P_B) \pm Z \cdot ESD$

$$ESD = \frac{1}{n}\sqrt{b+c-\frac{(b-c)^2}{n}}$$

P_A, P_B: Proporciones observadas en las muestras A y B; n_A, n_B: Número de sujetos de las muestras A y B; b, c: Número de casos que presentan valores diferentes en ambas mediciones (series apareadas); n: Número total de casos; ESD: Error estándar de la diferencia; Z : Valor de la variable normal tipificada correspondiente al valor α, para un nivel de confianza (1-α).

(*) Las variables cualitativas no presentan una distribución normal. Las fórmulas de la tabla corresponden a una aproximación a la normalidad, aplicable cuando todos los productos $n \cdot P_A$, $n \cdot (1-P_A)$, $n \cdot P_B$ y $n \cdot (1-P_B)$ son mayores de 5.

Tabla 7. Cálculo del intervalo de confianza (IC) de la diferencia entre dos medias.

IC DE LA DIFERENCIA DE DOS MEDIAS (*)

a) MUESTRAS INDEPENDIENTES: $(m_A - m_B) \pm Z \cdot ESD$

$$ESD = s \cdot \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \qquad s = \sqrt{\frac{s_A^2 \cdot (n_A - 1) + s_B^2 \cdot (n_B - 1)}{n_A + n_B - 2}}$$

b) MUESTRAS APAREADAS $m_d \pm Z \cdot ESm_d$

m_A, m_B: Medias observadas en las muestras A y B; s_A, s_B: desviaciones estándar observadas en las muestras A y B; n_A, n_B: Número de sujetos de las muestras A y B; ESD: Error estándar de la diferencia; m_d: Media de las diferencias de las dos mediciones en cada individuo (series apareadas); ESm_d: Error estándar de la media de las diferencias individuales; Z: Valor de la variable normal tipificada correspondiente al valor α, para un nivel de confianza $(1-\alpha)$.

(*) El cálculo se basa en la distribución normal. El valor de Z para un IC del 95% es 1,96. Para muestras de tamaño inferior a 30 individuos, este valor debe sustituirse por el de la t de Student para (n-1) grados de libertad. Asimismo, el cálculo requiere que no existan diferencias significativas entre las desviaciones estándar de ambas muestras.

camente significativa, sino también en determinar su magnitud. El valor observado en el estudio es la mejor estimación puntual de dicha magnitud. Si se repitiera el estudio con otras muestras, podrían observarse resultados de diferente magnitud. Por tanto, hay que calcular un IC que contenga, con una determinada confianza, la verdadera magnitud de interés. Las tablas 6 y 7 presentan las fórmulas para el cálculo del IC de la diferencia entre dos proporciones y entre dos medias, respectivamente.

Cuando se utiliza como medida del efecto una diferencia, si el IC del 95% incluye el valor 0, que es el valor correspondiente a la Ho de que no existe diferencia entre ambos grupos, se concluirá que el resultado no es estadísticamente significativo. Si, por el contrario, el IC del 95% excluye este valor 0, se concluirá que la diferencia observada es estadísticamente significativa. Además de saber si la diferencia es o no estadísticamente significativa, el IC permite conocer entre qué límites es probable que se encuentre la verdadera diferencia, lo que es muy útil en la interpretación de los resultados.

Supongamos un estudio que compara la eficacia de dos tratamientos A y B en dos grupos de 30 pacientes. Se observa una diferencia en el porcentaje de éxitos del 20% (70% - 50%) a favor del tratamiento B, que no es estadísticamente significativa (p=0,12). El IC del 95% de la diferencia entre los dos tratamientos es 0,2 ± 0,24, es decir, de 4% a 44%. La verdadera magnitud de la diferencia está en un intervalo que va desde un 4% a favor del tratamiento A hasta un 44% a favor de B. Dado que una diferencia del 0 % también es posible, no puede descartarse que éste sea su verdadero valor, por lo que la prueba esta-

dística da un valor no significativo. En cambio, el IC informa además que también son posibles grandes diferencias a favor de B, y que son improbables grandes diferencias a favor de A. Aunque los resultados siguen sin ser concluyentes, se dispone de más información para interpretarlos adecuadamente. El IC cuantifica el resultado encontrado y provee un rango donde es muy probable que se encuentre el valor real que se está buscando.

Los IC tienen otra ventaja adicional, y es la de expresar los resultados en las unidades en que se han realizado las mediciones, lo que permite al lector considerar críticamente la relevancia clínica de los mismos.

Aunque las pruebas de significación continúan siendo los procedimientos estadísticos utilizados con mayor frecuencia, las ventajas de la utilización de los IC en el análisis e interpretación de los resultados, tanto si el objetivo es la estimación de parámetros como el contraste de una hipótesis, hacen que cada vez más revistas recomienden a los autores la utilización de los mismos.

Análisis multivariante

En muchas ocasiones, interesa considerar la influencia de más de dos variables simultáneamente. Ello requiere técnicas sofisticadas, basadas en modelos matemáticos complejos, agrupadas bajo el nombre genérico de análisis multivariante.

Existen múltiples técnicas estadísticas multivariantes. En investigación clínica y epidemiológica las más utilizadas son las que analizan la relación entre una variable dependiente (variable de respuesta) y un grupo de variables independientes (factor de estudio y variables a controlar). Estas técnicas implican la construcción de un modelo matemático. La elección de un modelo u otro dependerá del diseño empleado en el estudio, la naturaleza de las variables y de las interrelaciones entre el factor de estudio, la variable de respuesta y las restantes variables incluidas en el modelo (variables a controlar). Los utilizados con más frecuencia son la regresión lineal múltiple cuando la variable dependiente es cuantitativa, y la regresión logística cuando es dicotómica.

BIBLIOGRAFIA

1. Altman DG. Practical statistics for medical research. London: Chapman & Hall, 1991.
2. Andersen B. Methodological errors in medical research. Oxford: Blackwell Scientific Publications, 1990.
3. Argimon Pallás JM, Jiménez Villa J. Métodos de investigación clínica y epidemiológica. Madrid: Harcourt Internacional, 2000.
4. Armitage P, Berry G. Estadística para la investigación biomédica. Barcelona: Doyma, 1992.
5. Campbell MJ, Julious SA, Altman DG. Estimating sample size for binary, ordered categorial, and continuous outcomes in two group comparison. BMJ 1995; 311: 1145-1148.
6. Dawson-Saunders E, Trapp RG. Bioestadística médica. México: El Manual Moderno, 1993.
7. Essex-Sorlie D. Medical bioestatistics & epidemiology. East Norwalk: Appleton & Lange, 1995.
8. Everitt BS. Statistical methods for medical investigations. New York: Oxford

University Press, 1989.
9. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & sons, 1981.
10. Florey CV. Sample size for beginners. BMJ 1993; 306: 1181-1184.
11. Gardner MJ, Altman DG. Confidence intervals rather than p values: estimation rather than hypotesis testing. BMJ 1986; 292: 746-750.
12. Gardner MJ, Altman DG. Statistics with confidence: confidence intervals and statistical guidelines. Londres: Bristish Medical Journal, 1989.
13. Kelsey JL, Thompson WD, Evans A. Methods in observational epidemiology. Nueva York, Oxford University Press; 1986.
14. Kleinbaum D, Kupper L, Morgenstern H. Epidemiologic Research. Belmont, Lifetime Learning Publications 1982.
15. Marrugat J, Vila J, Pavesi M, Sanz F. Estimación del tamaño de la muestra en la investigación clínica y epidemiológica. Med Clin (Barc) 1998; 111: 267-76.
16. Martín Andrés A, Luna del Castillo J de D. Bioestadística para las ciencias de la salud. 2ª edición. Madrid: Norma, 1989.
17. Norman GR, Streiner DL. Bioestadística. Madrid: Mosby/Doyma Libros, 1996;
18. Plasencia A, Porta M. La calidad de la información clínica (II): significación estadística. Med Clin (Barc) 1988; 90: 122-126.
19. Porta M, Plasencia A, Sanz F. La calidad de la información clínica (III): ¿estadísticamente significativo o clínicamente importante? Med Clin (Barc) 1988; 90: 463468.
20. Sahai H, Khurshid A. Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review. Stat Med 1996; 15: 1-21.
21. Silva Ayçaguer LC. Muestreo para la investigación en ciencias de la salud. Madrid: Díaz de Santos; 1993.