ARTICULO ESPECIAL

NORMAS ESTADISTICAS PARA LOS COLABORADORES DE REVISTAS DE MEDICINA

D.G. ALTMAN, S.M. GORE, M.J. GARDNER y J. POCOCK

Con el ánimo de facilitar una orientación a los autores potenciales sobre la forma de analizar los datos de los trabajos que deseen publicar en Archivos de Bronconeumología, la redacción de la revista cree conveniente ofrecer las normas estadísticas reunidas por Altman et al y aparecidas originalmente en el British Medical Journal (1983; 1:1.489-1.493). Las normas han recibido una favorable acogida y ya son varias las publicaciones de prestigio que las han hecho suyas.

Archivos de Bronconeumología agradece a los autores y al British Medical Journal su autorización para publicarlas, desea acoger las normas como suyas y recomienda a los futuros colaboradores su atenta lectura y observancia.

El Redactor

La mayoría de artículos publicados en revistas de medicina contienen análisis que han sido realizados sin la ayuda de un estadístico. Aunque casi todos los investigadores médicos están familiarizados con la estadística básica, no tienen ningún método sencillo que les permita discernir sobre conceptos y principios estadísticos importantes. También existe poca ayuda con respecto a cómo diseñar, analizar y escribir un proyecto completo. Debido en parte a estas razones, la mayoría de lo que se publica en revistas médicas es pobre estadísticamente hablando o incluso equivocado[1]. Se ha detectado un alto nivel de errores estadísticos en algunas revisiones de artículos de revistas, lo que ha provocado una gran preocupación.

Algunas revistas incluso ofrecen consejos estadísticos rudimentarios a sus colaboradores. Se ha sugerido[1,2] que unas normas estadísticas amplias podrían ser de utilidad, haciendo más conscientes a los investigadores médicos de los principios estadísticos importantes e indicando qué información debería proporcionarse en un artículo. Presentamos a continuación un intento de llevar esto a la práctica.

Ha resultado problemático decidir qué se debe incluir en las normas, cuántos detalles dar y cómo tratar los temas en los que no hay consenso. Estas normas deben, por lo tanto, considerarse como una visión de lo que es importante más que como un documento definitivo. No ha sido nuestra intención proporcionar una serie de reglas sino más bien dar una información y unos consejos generales sobre aspectos importantes de diseño, análisis y presentación estadística. Las recomendaciones específicas que hemos hecho son en su mayor parte firmes consejos contra ciertas prácticas.

Se supone una cierta familiaridad con la estadística puesto que son necesarios unos conocimientos de estadística antes de llevar a cabo los análisis estadísticos. Para aquellos que sólo tienen una limitada familiaridad con la estadística, las normas deben mostrarles que el tema es mucho más amplio que el simple análisis de significación e ilustrar lo importante que es la interpretación correcta. La falta de recomendaciones precisas indica que un buen análisis estadístico requiere sentido común y capacidad de decisión; así como un repertorio de técnicas formales de manera que hay un arte de estadística, igual que en medicina. Creemos que las normas representen una perspectiva indiscutible de los procedimientos estadísticos utilizados y aceptados con mayor frecuencia. Hemos limitado deliberadamente el alcance de las normas para cubrir los procedimientos estadísticos más comunes.

Los lectores pueden encontrar que una sección pertinente presenta información o consejos que no les resultan familiares o que no comprenden. En tales circunstancias, aunque casi todos los temas tratados pueden encontrarse en los libros de texto de estadística médica de mayor extensión[3,4], recomendamos enfáticamente solicitar el consejo de un estadístico. La ausencia de referencias específicas en las normas es deliberada: es mejor obtener el consejo personal experto cuando se precisa mayor discernimiento. Además, como los errores en el diseño no pueden rectificarse más tarde, cuando se planee un proyecto de investigación, debe obtenerse en primer lugar el consejo de un profesional en lugar de hacerlo al final cuando se analizan los datos.

Queremos dar las gracias a un gran número de personas que leyeron las versiones previas de las normas, por sus comentarios constructivos y provechosos.

En todo este artículo hemos seguido la convención de Vancouver utilizando P para probabilidad, aunque la notación estadística apoya el uso de P.

1. Introducción

Estas normas tienen el propósito de ayudar a los autores a saber lo que es estadísticamente importante y cómo presentarlo en sus artículos. Remarcan que tales formas de presentación están estrechamente ligadas a consideraciones más generales de principios estadísticos. No se discute en detalle cómo elegir un método estadístico apropiado; esta información se obtiene mejor consultando un estadístico. Llamamos la atención, sin embargo, sobre ciertos usos erróneos de los métodos estadísticos.

Estas normas siguen la estructura usual en los artículos de investigación médica: métodos, resultados (análisis y presentación),y discusión (interpretación). Como resultado, algunos temas aparecen en más de un lugar y son mencionados de forma cruzada siempre que es oportuno.

2. Sección de métodos

2.1 Principios generales

Es muy importante describir claramente lo que se ha hecho, incluyendo el diseño de la investigación (ya sea un experimento, un ensayo o una revisión) y la recogida de los datos. La finalidad debe ser dar suficiente información para permitir que los métodos sean totalmente comprendidos y repetibles por otros cuando lo deseen. Los autores deben incluir información sobre los siguientes aspectos del diseño de su investigación:

- objetivo de la investigación y principales hipótesis;
- tipo de individuos, establecimiento de criterios para inclusión y exclusión;
- la procedencia de los individuos y cómo fueron seleccionados;
- el número de sujetos estudiados y el porqué se utilizó ese número;
- los tipos de observación y las técnicas de medición utilizadas.

Cada tipo de estudio —por ejemplo revisiones y ensayos clínicos— requerirá cierta información adicional.

2.2 Estudios de observación

El diseño del estudio debería explicarse claramente. Por ejemplo, la selección de un grupo control y el procedimiento de emparejamiento precisan una descripción detallada. Debe también establecerse claramente si el estudio es retrospectivo, transversal o prospectivo. El procedimiento para la selección de los individuos y la consecución de una alta tasa de participación son particularmente importantes, ya que los hallazgos son por regla general inferidos desde la muestra a una población general. Es útil informar de cualquier paso seguido para estimular la participación en el estudio.

2.3 Ensayos clínicos

Los regímenes de tratamiento (incluyendo la atención auxiliar del paciente y los criterios para modificar o parar el tratamiento) precisan una detallada definición. El método de la distribución de los tratamientos a los sujetos debe describirse de forma explícita. En particular, debe explicarse el método específico de sorteo o distribución aleatoria (incluyendo cualquier estratificación) y cómo fue llevada a cabo. Cualquier falta de aleatoriedad debería señalarse como una deficiencia en el diseño y explicarse las razones.

Debe describirse la utilización de técnicas de «ciego» y otras precauciones tomadas para asegurar la evaluación imparcial de la respuesta del paciente. Deben enumerarse los principales criterios para la comparación de los tratamientos, como se acordó en el protocolo de ensayos. Para los ensayos con tratamientos cruzados debe explicarse la pauta exacta del orden de los tratamientos, y cualquier período de tratamiento en blanco.

2.4 Métodos estadísticos

Deberán identificarse todos los métodos estadísticos utilizados en un artículo. Cuando se usan varias técnicas deberá estar absolutamente claro qué método se utiliza en cada caso y esto puede hacer necesaria la clarificación en la sección de resultados. Las técnicas muy corrientes tales como los tests de «t» de X^2, de Wilcoxon y Mann-Whitney, la correlación (r) y la regresión lineal, no necesitan ser descritas, pero los métodos con más de una acepción, tales como los tests «t» (apareados o independientes), análisis de varianza y la correlación de rangos, deberían identificarse inequívocamente. Los métodos más complejos necesitan alguna explicación y, si los métodos son infrecuentes, deberá darse una referencia precisa. Podría ayudar el incluir breves comentarios sobre porqué se utilizó el método particular de análisis, especialmente cuando no se ha escogido un planteamiento más familiar. Podría ser útil proporcionar el nombre del programa o conjunto de programas informáticos utilizados —por ejemplo, el conjunto de programas estadísticos para ciencias sociales (SPSSA), pero aún así deben identificarse los métodos estadísticos específicos usados.

3. Sección de resultados: análisis estadísticos

3.1 Información descriptiva

La adecuada descripción de los datos debería preceder y complementar los análisis estadísticos

formales. En general, deberían ser descritas con el máximo detalle las variables que son importantes para la validez e interpretación de los análisis estadísticos subsiguientes. Esto puede llevarse a cabo por métodos gráficos, tales como diagramas de dispersión (puntos) e histogramas, o utilizando resúmenes estadísticos. Las variables continuas (tales como peso o presión sanguínea) pueden ser resumidas utilizando la media y la desviación típica o la media y un rango de percentil, por ejemplo, el rango intercuartil (del percentil 25° al 75°). El último planteamiento es preferible cuando las mediciones continuas tienen una distribución asimétrica. Para datos cualitativos ordenados (tales como etapas de enfermedad, de I a IV) es incorrecto el cálculo de las medias y desviaciones típicas; en su lugar deben ofrecerse las proporciones.

Deben describirse las desviaciones respecto al diseño del estudio proyectado. Por ejemplo, en los ensayos clínicos es particularmente importante señalar los pacientes rechazados del estudio, con las razones si se conocen y el tratamiento asignado. Para encuestas, donde la tasa de respuesta es de fundamental importancia, es valioso dar información sobre las características de los que no respondieron en comparación con los que participaron. Deberá investigarse la representatividad de la muestra del estudio cuando la inferencia de los resultados a alguna población apropiada es una intención primordial.

Es útil comparar la distribución de las características basales en diferentes grupos, tales como los grupos de tratamiento en un ensayo clínico. Las diferencias que existan, incluso si no son estadísticamente significativas, son reales y deberían ser adecuadamente tratadas en los análisis (ver sección 3.12).

3.2 Hipótesis subyacentes

Los métodos de análisis tales como los tests de «t», la correlación, la regresión y los análisis de varianza dependen todos en cierta medida de algunas hipótesis o condiciones sobre la distribución de las variables que están siendo analizadas. Técnicamente, las hipótesis son que, en algún aspecto, los datos proceden de una distribución normal y, que si se comparan dos o más grupos, se da por hecho que la variabilidad dentro de cada grupo es la misma.

No es posible dar de forma categórica el grado hasta el que estas hipótesis pueden ser violadas sin invalidar el análisis. Pero los datos cuya distribución es muy sesgada (asimétrica), o en los que la variabilidad es considerablemente diferente de unos grupos a otros, pueden requerir cierta transformación antes del análisis (ver sección 3.7), o el uso de métodos alternativos que no dependan de las hipótesis sobre la distribución, a menudo llamados métodos no paramétricos. Por ejemplo, el test U de Mann-Whitney es un test independiente de la distribución, equivalente del test de «t» para dos muestras. Los métodos no dependientes de la distribución pueden ser apropiados también para series pequeñas de datos, para las cuales las hipótesis no pueden ser verificadas adecuadamente.

En ocasiones, presuponer que la distribución de la muestra es gaussiana puede tener consecuencias especialmente importantes —por ejemplo, cuando el intervalo de valores calculado como dos desviaciones típicas a ambos lados de la media es tomado como el 95 % del intervalo «normal» o de referencia. En tales casos tiene que demostrarse que la hipótesis de distribución está justificada.

3.3 Análisis de significación

El principal objeto de analizar la significación es evaluar un número limitado de hipótesis preformuladas. Otros tests de significación, hechos porque fueron sugeridos durante la inspección preliminar de los datos, darán una falsa impresión, porque en tales circunstancias el valor calculado de «p» es demasiado pequeño. Por ejemplo, no es válido analizar la diferencia entre la menor y la mayor de un conjunto de medias sin justificar la razón para analizar esta particular diferencia. Se dispone de técnicas especiales para hacer comparaciones apareadas entre varios grupos.

Es costumbre llevar a cabo análisis bilaterales de significación. Si se utiliza un análisis unilateral debería indicarse y justificarse para el problema de que se trata.

La presentación e interpretación de los resultados de los análisis de significación se discuten en las secciones 4.3, 5.1 y 5.2.

3.4 Intervalos de confianza

La mayoría de los estudios se ocupan de estimar alguna cantidad, tal como una diferencia media o riesgo relativo. Es deseable calcular el intervalo de confianza alrededor de tal estimación. Este es un intervalo de valores sobre el que estamos, digamos un 95 %, de confiados en que incluye el valor verdadero. Hay una estrecha relación entre el resultado de un test de significación y el intervalo de confianza asociado: si la diferencia entre los tratamientos es significativa al nivel del 5 % entonces el intervalo de confianza asociado del 95 % excluye la diferencia cero. El intervalo de confianza expresa más información porque indica el verdadero efecto mayor y menor que es probablemente compatible con las observaciones de la muestra (ver también sección 5.1).

Los intervalos de confianza revelan la precisión de una estimación. Un amplio intervalo de confianza indica falta de información, tanto si la diferencia es estadísticamente significativa como si no lo es, y es una advertencia contra resultados sobreinterpretados a partir de estudios pequeños.

3.5 Observaciones apareadas

Es esencial distinguir el caso de las observaciones independientes, en las que la comparación se realiza entre mediciones en dos grupos diferentes (por ejemplo, sujetos que reciben tratamientos alternativos), de las observaciones pareadas, en las que la comparación es entre dos mediciones hechas en los mismos individuos en diferentes circunstancias (tales como antes y después del tratamiento). En las que con los datos independientes se utilizaría, por ejemplo, el test de «t» para dos muestras, con los datos apareados se utilizaría en su lugar el test «t» apareado. Del mismo modo, el test U de Mann-Whitney para los datos independientes es reemplazado por el test apareado de Wilcoxon y el habitual test de la X^2 para tablas de 2×2 es reemplazado por el test de McNemar. Debería aclararse siempre qué forma de test se utilizó.

La misma distinción debe hacerse cuando hay tres o más series de observaciones. Todos los métodos estadísticos mencionados en esta sección pueden generalizarse a más de dos grupos; en particular los test «t» apareados y de dos muestras se generalizan a diferentes formas de análisis de varianza.

3.6 Mediciones repetidas

Un diseño común de estudio supone registrar mediciones en serie de la(s) misma(s) variable(s) en los mismos individuos en distintos momentos. Tales datos son a menudo analizados calculando las medias y desviaciones típicas en cada momento y presentándolas gráficamente mediante una línea que une estas medias. La forma de esta curva de medias puede no dar una buena idea de las formas de las curvas individuales. A menos que las respuestas individuales sean muy similares, puede tener más valor analizar algunas características de los perfiles individuales, tales como el tiempo tardado en alcanzar un máximo o el período de tiempo por encima de un nivel dado. Esto ayudaría también a evitar los problemas asociados con los análisis de significación múltiple (ver la sección 5.2).

Las mediciones repetidas de la misma variable en un individuo bajo las mismas condiciones experimentales, conocidas como lecturas repetidas, no deberían tratarse como observaciones independientes cuando se comparan grupos de individuos. En los casos en que el número de repeticiones es el mismo para todos los sujetos, el análisis no es difícil; en particular, se utilizan análisis de varianza cuando se habrían aplicado tests de «t» a los datos no repetidos. Si el número de repeticiones varía entre los individuos, un análisis completo podría resultar muy complejo. El uso de la mayor o la menor de una serie de mediciones (tal como la presión sanguínea máxima durante el embarazo) puede ser erróneo si el número de observaciones varía ampliamente entre los individuos.

3.7 Transformación de los datos

Muchas variables biomédicas son sesgadas positivamente, con algunos valores muy altos, y pueden requerir la transformación matemática para hacer los datos apropiados para el análisis. En tales circunstancias es aplicable a menudo la transformación logarítmica, aunque ocasionalmente pueden ser más apropiadas otras transformaciones (tales como la raíz cuadrada o la recíproca).

Después del análisis es deseable convertir los resultados de nuevo en la escala original para su informe. En el caso frecuente de la transformación logarítmica debería usarse el antilogaritmo de la media de los datos logarítmicos (conocido como la media geométrica). Sin embargo, no debe hacerse el antilogaritmo de la desviación típica o del error típico; en cambio, puede hallarse el antilogaritmo de los límites de confianza en la escala original. Un procedimiento similar se adopta con otras transformaciones.

Si se utiliza una transformación, es importante comprobar que el efecto deseado se ha conseguido (tal como una distribución aproximadamente normal). No debe suponerse que la transformación logarítmica, por ejemplo, es necesariamente adecuada para todas las variables sesgadas positivamente.

3.8 Valores aberrantes

Las observaciones muy discordantes respecto al conjunto principal de los datos no deberían ser excluidas del análisis a menos que haya razones adicionales para dudar de su credibilidad. Cualquier omisión de tales aberrantes debe ser informada. Puesto que los aberrantes pueden tener un efecto pronunciado sobre un análisis estadístico, es útil analizar los datos con y sin estas observaciones, para valorar en qué medida dependen algunas conclusiones de estos valores.

3.9 Correlación

Es preferible incluir un diagrama de dispersión de los datos para cada coeficiente de correlación presentado, aunque esto puede no ser posible si hay distintas variables. Cuando están siendo investigadas muchas variables, es útil mostrar las correlaciones entre todos los pares de variables en una tabla (matriz de correlación), en lugar de simplemente citar los valores mayores o significativos.

Para los datos que están distribuidos irregularmente puede calcularse la correlación de rango en lugar de la habitual correlación «momento del producto» de Pearson (r). La correlación de rango

puede ser utilizada también para variables que están restringidas a estar por encima o por debajo de ciertos valores —por ejemplo pesos al nacer por debajo de 2.500 g— o para variables absolutas ordenadas. La correlación de rango es también preferible cuando la relación entre las variables no es lineal o cuando los valores de una variable han sido escogidos por el experimentador en lugar de ser libres.

El coeficiente de correlación es un resumen útil del grado de asociación lineal entre dos variables cuantitativas, pero es uno de los métodos estadísticos peor empleados. Hay varias circunstancias en las que no debería utilizarse la correlación. Es incorrecto calcular un simple coeficiente de correlación para los datos que incluyen más de una observación en algunos o todos los sujetos, debido a que tales observaciones no son independientes. La correlación es inapropiada para comparar métodos alternativos de medición de la misma variable, debido a que valora asociación, no concordancia. El uso de la correlación para relacionar el cambio a lo largo del tiempo con el valor inicial puede dar resultados excesivamente erróneos.

Puede ser erróneo calcular el coeficiente de correlación para subgrupos que incluyen datos que se sabe que difieren en sus niveles medios de una o las dos variables —por ejemplo, datos combinados para el hombre y la mujer cuando una de las variables es la talla.

La regresión y la correlación son técnicas separadas, que sirven para diferentes propósitos, y no precisan acompañar automáticamente una a la otra. La interpretación de los coeficientes de correlación se discute en la sección 5.3.

3.10 Regresión

Es altamente deseable presentar una línea de regresión ajustada junto con un diagrama de dispersión de los datos no elaborados. Una representación de la línea ajustada sin los datos da menos información adicional que la propia ecuación de regresión. Es útil dar los valores de la pendiente (con su error típico) e intersección y una medida de la dispersión de los puntos alrededor de línea ajustada (la desviación típica residual). Pueden construirse alrededor de una línea de regresión los límites de confianza para mostrar la incertidumbre de las predicciones basadas en la relación ajustada. Estos límites no son paralelos a la línea sino curvados, mostrando que la mayor incertidumbre de la predicción corresponde a los valores sobre el eje horizontal (x), lejos del grueso de las observaciones.

La regresión sobre datos que incluyen distintos subgrupos puede dar resultados erróneos, en particular si los grupos difieren en su nivel medio de la variable dependiente (y). Pueden obtenerse resultados más fidedignos utilizando análisis de covarianza.

La regresión y la correlación son técnicas independientes que sirven a diferentes propósitos y no necesitan ir acompañadas automáticamente una de la otra. La interpretación de los análisis de regresión se discute en la sección 5.4.

3.11 Datos de supervivencia

El informe de los datos de supervivencia debería incluir la representación gráfica o tabular de las tablas de vida, con detalles de cuántos pacientes corrían el riesgo (de muerte, por ejemplo) en diferentes momentos del seguimiento. La estadística de promedio de vida trata de manera eficaz los tiempos de supervivencia que aparecen cuando los pacientes se pierden para el seguimiento o están aún vivos; se sabe que su tiempo de supervivencia es de por lo menos de tantos días. El cálculo del tiempo medio de supervivencia es desaconsejable si hay «censoring» y debido a la distribución de los tiempos de supervivencia es en general sesgado positivamente.

La comparación entre grupos de tratamiento de la proporción de sobrevivientes a tiempos fijados arbitrariamente puede ser errónea y es generalmente menos eficiente que la comparación de estadísticas de promedio de vida por un método tal como el test «logrank».

Cuando hay suficientes muertes, puede mostrarse cómo el riesgo de muerte varía con el tiempo representando, para apropiados intervalos iguales iguales de tiempo, la proporción de los que estaban vivos al principio de cada intervalo de tiempo y murieron durante ese intervalo. Es posible el ajuste respecto a los factores del paciente que pueden influir sobre el pronóstico utilizando modelos de regresión apropiados para los datos de supervivencia (ver sección 3.12).

3.12 Análisis complejos

En muchos estudios las observaciones de interés principal pueden estar influidas por algunas otras variables. Estas podrían ser cualquier cosa que varíe entre los sujetos y que pueda haber afectado el resultado que se está observando. Por ejemplo, en los ensayos clínicos pueden incluirse características o signos y síntomas del paciente. Algunas o todas las covariantes pueden combinarse mediante técnicas apropiadas de regresión múltiple para explicar o predecir una variable resultante, ya sea una variable continua (presión sanguínea), una variable cualitativa (trombosis postoperatoria), o la duración de la supervivencia. Incluso en ensayos clínicos aleatorios los investigadores necesitan asegurarse de que aún está presente el efecto del tratamiento después del ajuste simultáneo para distintos factores de riesgo.

Las técnicas multivariantes, para tratar más de una variable resultante simultáneamente, requieren realmente la ayuda de un experto y están fuera del ámbito de estas normas.

Cualquier método estadístico complejo ha de ser expuesto de manera que sea comprensible para el lector.

4. Sección de resultados: presentación de los resultados

4.1 Presentación de estadísticas sumariales

Los valores medios no deben proporcionarse sin una cierta medida de variabilidad o de precisión. Debe usarse la desviación típica (SD) para mostrar la variabilidad entre los individuos y el error típico de la media (SE) para mostrar la precisión de la media de la muestra. Debe aclararse cuál se presenta.

El uso del símbolo ± para añadir el error típico o la desviación típica a la media (como en 14,2 ± 1,9) produce confusión y debe evitarse. Es preferible la presentación de las medias como, por ejemplo, 14,2 (1,9 SE) o 14,2 (7,4 SD). Los intervalos de confianza son una buena manera de presentar las medidas junto con límites razonables de incertidumbre —un intervalo de un 95 % de confianza para la verdadera media es desde aproximadamente dos errores típicos por debajo de la media observada hasta dos errores típicos por encima de ella. Los intervalos de confianza se presentan de una manera más clara cuando se dan los límites, tales como 10,4; 18,0), que cuando se usa el símbolo ±.

Al hacer comparaciones con datos apareados, como ocurre cuando se utiliza el test «t» apareado, es deseable dar la media y el error típico (o desviación típica) de las diferencias entre las observaciones.

Para datos que han sido analizados con métodos independientes de la distribución, es más apropiado dar la mediana y un rango central que cubra, por ejemplo, el 95 % de las observaciones, que utilizar la media y la desviación típica (ver sección 3.1). Asimismo, si los análisis han sido realizados sobre datos transformados, la media y la desviación típica de los datos en bruto no serán probablemente buenas mediciones del centro y de la dispersión de los datos y no debieran presentarse.

Cuando se dan porcentajes debe presentarse siempre claramente el denominador. Para muestras pequeñas es inútil el uso de porcentajes. Cuando se comparan porcentajes, es importante distinguir una diferencia absoluta de una relativa. Por ejemplo, una reducción desde el 25 % al 20 % puede ser expresada como 5 % o como 20 %.

4.2 Resultados por individuos

El rango global no es un buen indicador de la variabilidad de una serie de observaciones, ya que puede estar fuertemente afectado por un único valor extremo y aumenta con el tamaño de la muestra. Si los datos tienen una distribución razonablemente Normal, el intervalo de dos desviaciones típicas a ambos lados de la media cubrirá aproximadamente el 95 % de las observaciones; pero un rango expresado en percentiles es más ampliamente aplicable a otras distribuciones (ver sección 3.1).

Aunque el análisis estadístico está relacionado con efectos promedio, en muchas circunstancias es importante también considerar como responden los sujetos individualmente. Así, por ejemplo, es a menudo muy importante clínicamente saber cuántos pacientes no mejoraron con un tratamiento además del promedio de mejoría. Un efecto promedio no debe interpretarse como aplicable a todos los individuos (ver también sección 3.6).

4.3 Presentación de los resultados de los tests de significación

Los tests de significación producen valores que se comparan con los valores tabulados para la distribución apropiada (Normal, t, X^2, etc.) para obtener los correspondientes valores de «p». Es deseable informar de los valores observados de las estadísticas en estudio y no simplemente los valores de «p». Los resultados cuantitativos analizados, tales como los valores medios, las proporciones o los coeficientes de correlación, deben ser dados tanto si el test fue significativo como si no lo fue. Debe aclararse precisamente qué datos han sido analizados. Si se utilizan símbolos, tales como asteriscos, para indicar los niveles de probabilidad éstos deben ser definidos y conviene que sean los mismos durante todo el artículo.

Los valores de «p» se dan convencionalmente como $< 0{,}05$, $< 0{,}01$, o $< 0{,}001$, pero no hay ninguna otra razón que la de la familiaridad para utilizar estos valores en particular. Valores de «p» exactos (para no más de dos cifras significativas), tales como $p = 0{,}18$ o $0{,}03$, son muy útiles. Es improbable que sea preciso especificar niveles de «p» inferores a 0,0001. No se recomienda denominar «no significativo» a cualquier valor con $p > 0{,}05$, ya que puede obscurecer los resultados que no son bastante significativos estadísticamente pero que sugieren un efecto real (ver sección 5.1). Cuando se citan valores de «p» es importante distinguir $<$ (menor de) de $>$ (mayor de). Los valores de «p» entre dos límites deben expresarse en el orden lógico —por ejemplo, $0{,}01 < p < 0{,}05$ donde «p» queda entre 0,01 y 0,05. Los valores dispuestos en tablas no es necesario repetirlos en el texto.

La interpretación de los tests de significación y de los valores de «p» se discute en la sección 5.1.

4.4 Representación gráfica

La exposición gráfica de los resultados es útil para los lectores y son de aconsejar las figuras que muestran las observaciones individuales. Los puntos de una gráfica que pertenecen a un mismo individuo en diferentes ocasiones es preferible unirlos, o usar símbolos para indicar los puntos que están relacionados. Una alternativa útil es representar la diferencia entre las ocasiones para cada individuo.

Las acostumbradas «barras de error» de un error típico por encima y por debajo de la media representan únicamente un intervalo de confianza del 67 % y son, por lo tanto, susceptibles de interpretación equívoca; son preferibles los intervalos de confianza del 95 %. La representación de tal información en figura está sujeta a las mismas consideraciones discutidas en la sección 4.1.

Los diagramas de dispersión que relacionan dos variables deben mostrar todas las observaciones, incluso si ello significa un ligero ajuste para acomodar puntos que están duplicados. Esto puede también resolverse reemplazando el símbolo que los representa por el número real de puntos coincidentes.

4.5 Tablas

Es mucho más fácil leer los resultados numéricos en columnas descendentes que en filas horizontales, y por lo tanto es mejor tener los diferentes tipos de información (tales como las medias y los errores típicos) en columnas separadas. El número de observaciones para cada resultado debe hacerse explícito en la tabla. Las tablas que dan información sobre los pacientes individuales, áreas geográficas, etc., son más fáciles de leer si las filas se ordenan según el nivel de una de las variables presentadas.

4.6 Precisión numérica

La precisión espúrea no añade valor a un artículo e incluso disminuye su interés y credibilidad. Los resultados obtenidos a partir de una calculadora o un ordenador generalmente necesitan ser redondeados. Cuando se presentan medias, desviaciones típicas y otros estadísticos el autor debe tener presente la precisión de los datos originales. Las medias no deben darse, en general, con más de un decimal más que los datos sin elaborar, pero las desviaciones típicas o los errores típicos puede ser preciso darlos con un decimal extra. Raramente es necesario citar porcentajes con más de un decimal, e incluso a menudo no se precisa ni siquiera uno. Con muestras de menos de 100 el uso de un decimal implica que la precisión no está garantizada y debe evitarse. Nótese que estas observaciones corresponden únicamente a la presentación de los resultados; el redondeo no debe utilizarse antes o durante el análisis. Es suficiente citar valores de t, X^2 y r con dos decimales.

4.7 Términos técnicos diversos

Es imposible definir aquí todos los términos estadísticos. Los siguientes comentarios se refieren a algunos términos que son utilizados frecuentemente de una manera incorrecta o confusa.

La *correlación* es mejor no utilizarla como un término general para describir cualquier relación. Tiene un significado técnico específico como medida de asociación, para la cual debe reservarse en el trabajo estadístico.

La *incidencia* debe usarse para describir la tasa de presentación de nuevos casos de una característica dada en una muestra o población de estudio, tal como el número de nuevas notificaciones de cáncer en un año. La proporción de una muestra que ya tiene una característica es la *prevalencia.*

No paramétrico se refiere a ciertos análisis estadísticos tales como el test U de Mann-Whitney; no es una característica de las propias observaciones.

Parámetro no debe utilizarse en lugar de «variable» para referirse a una medición o atributo sobre la cual se hacen las observaciones. Los parámetros son características de las distribuciones o relaciones en la población que son estimadas mediante análisis estadísticos de una muestra de observaciones.

Percentiles. Cuando el rango de los valores de una variable se divide en grupos iguales, los puntos de corte son la mediana, terciles, cuartiles, quintiles, etc.; los grupos mismos deben ser denominados mitades, tercios, cuartos, quintos, etc.

Sensibilidad es la capacidad de un test para identificar una enfermedad cuando realmente está presente; esto es, la proporción de positivos entre los que tienen la enfermedad. Especificidad es la capacidad de un test para identificar la ausencia de una enfermedad cuando realmente no está presente; es decir, la proporción de negativos entre los que no tienen la enfermedad (ver también la sección 5.4).

5. Sección de discusión: interpretación

5.1 Interpretación de los tests de significación

Un test de significación valora, por medio de la probabilidad «p», la plausibilidad de los datos observados cuando es cierta la «hipótesis nula» (tal como no haber diferencia entre los grupos). El valor de «p» es la probabilidad de que los datos observados, o un resultado más extremo, hubieran ocurrido por casualidad— cuando la hipótesis nula es cierta. Si «p» es pequeña se duda de la hipótesis nula. Si «p» es grande los datos son verosímil-

mente consistentes con la hipótesis nula, que por lo tanto no puede ser rechazada. El valor «p» no es por lo tanto la probabilidad de que no haya un efecto real.

Incluso si hay un gran efecto real, es bastante probable que el resultado sea no significativo si el número de observaciones es pequeño. A la inversa, si el tamaño de la muestra es muy grande puede producirse un resultado estadísticamente significativo cuando hay sólo un pequeño efecto real. Por lo tanto, la significación estadística no debería ser tomada como sinónimo de importancia clínica.

La interpretación de los resultados de los tests de significación se deduce ampliamente de lo anterior. Un resultado significativo no indica necesariamente un efecto real. Siempre hay algún riesgo de un hallazgo falso positivo; este riesgo disminuye para los valores de «p» más pequeños. Es más, un resultado, no significativo (convencionalmente $p > 0,05$) no quiere decir que no hay efecto sino, únicamente, que los datos son compatibles con que el efecto no exista. Es deseable una cierta flexibilidad en la interpretación de los valores de «p». El nivel 0,05 es un punto de separación conveniente, pero los valores de «p» de 0,04 y 0,06 que no son muy diferentes, deben conducir a interpretaciones análogas más que a interpretaciones radicalmente opuestas. La designación de cualquier resultado con $p > 0,05$ como no significativo puede por lo tanto confundir al lector (y a los autores); de ahí la sugerencia en la sección 4.3 de citar los valores de «p» reales.

Los intervalos de confianza son extremadamente útiles en la interpretación, en particular para estudios pequeños, cuando muestran el grado de incertidumbre relacionado con un resultado —como la diferencia entre dos medias— tanto si era estadísticamente significativa como si no lo era. Su uso junto a los resultados no significativos puede ser especialmente esclarecedor.

5.2 *Muchos tests de significación*

En muchos proyectos de investigación algunos tests de significación están relacionados con importantes comparaciones que fueron consideradas cuando se inició la investigación. Los tests de las hipótesis que no fueron decididas por adelantado son secundarios, especialmente si fueron sugeridos por los resultados. Es importante distinguir estos dos casos y dar mucho mayor peso a los tests de aquéllas hipótesis que fueron formuladas inicialmente. Los otros tests deben considerarse como únicamente exploratorios —para formular nuevas hipótesis a investigar en estudios subsiguientes. Una razón para esto es que cuando se realizan muchos tests de significación en el análisis de un estudio, quizá comparando muchos subgrupos o considerando muchas variables, puede esperarse que aparezca, sólo por casualidad, cierto número de resultados positivos espúreos, que pueden plantear importantes problemas de interpretación. Evidentemente, cuantos más tests se lleven a cabo, mayor es la probabilidad de encontrar algunos resultados significativos, pero el número esperado de falsos positivos también se incrementará. Una manera de soslayar el riesgo de resultados positivos falsos es establecer un bajo nivel de «p» como criterio de significación estadística.

Un problema más complejo se plantea cuando los tests de significación se realizan sobre datos dependientes (correlacionados). Un ejemplo de esto lo encontramos en el análisis de datos apareados (discutidos en la sección 3.6) cuando se realiza el mismo test para la misma variable sobre los datos recogidos en diferentes momentos. Otro ejemplo es cuando se llevan a cabo los análisis separados de dos o más variables correlacionadas como si fueran independientes; cualquier corroboración puede no incrementar grandemente el peso de la evidencia porque los tests se refieren a datos muy similares. Por ejemplo, las presiones sanguíneas diastólicas y sistólicas se comportan de manera muy similar, como lo hacen las formas alternativas de valorar la respuesta de pacientes en general. En tales casos se precisa la interpretación muy cuidadosa de los resultados.

5.3 *Asociación y causalidad*

Una asociación estadísticamente significativa (obtenida a partir de la correlación o del análisis de X^2) no proporciona por sí misma evidencia directa de una relación causal entre las variables implicadas. En estudios observacionales puede establecerse la causalidad únicamente sobre bases no estadísticas; es más fácil inferir causalidad en ensayos al azar. Debe tenerse mucho cuidado al comparar variables que varían con el tiempo porque es fácil obtener una aparente asociación que es falsa.

5.4 *Predicción y análisis diagnóstico*

Incluso cuando el análisis de regresión ha indicado una relación estadística significativa entre dos variables, puede haber una imprecisión considerable cuando se utiliza la ecuación de regresión para predecir el valor numérico de una variable (y) a partir de la otra (x) para los casos individuales. La exactitud de tales predicciones no puede ser valorada a partir de la correlación o del coeficiente de regresión sino que requiere el cálculo del intervalo de confianza para el valor esperado de y que corresponde a un valor específico x (ver sección 3.10). La línea de regresión debe utilizarse únicamente para predecir la variable y a partir de la x, y no a la inversa.

Un test diagnóstico con alta sensibilidad y especificidad puede no ser necesariamente útil para los

propósitos diagnósticos, especialmente cuando se aplica a una población en que la frecuencia de la enfermedad es muy baja. Es útil en este caso calcular la proporción de sujetos con resultados del test positivos que realmente tienen la enfermedad. Recuérdese que no hay consenso en la definición del «porcentaje de falsos positivos» o «porcentaje de falsos negativos»; debe aclararse siempre exactamente qué es lo que se está calculando. Este punto puede ilustrarse mejor mediante una tabla de 2 × 2 que relacione los resultados del test con el estado real de enfermedad del paciente.

Un problema diagnóstico similar se presenta con las variables continuas. Es frecuente para una variable la clasificación como «anormal» cuando su valor está fuera de los «límites normales». Pero si la prevalencia de anormalidad verdadera es baja, la mayoría de los valores fuera del «rango normal» serían normales. La definición de anormalidad debe basarse en criterios tanto clínicos como estadísticos.

5.5 Debilidades

Es mejor tratar las debilidades en el diseño de la investigación y ejecución, si uno se da cuenta de ellas, y considerar sus posibles efectos sobre los resultados y su interpretación, que ignorarlas con la esperanza de que no se notarán.

6. Observaciones finales

El objetivo de los métodos estadísticos es proporcionar un sencillo y objetivo informe de la evidencia científica obtenida a partir de un trabajo de investigación. La capacidad y experiencia necesarias para diseñar estudios adecuados, llevar a cabo análisis estadísticos con sentido y expresar los hallazgos de una manera clara y objetiva no son fáciles de adquirir. Esperamos que estas normas puedan contribuir a una mejora en el nivel del trabajo estadístico utilizado en las publicaciones médicas.

BASES PREMIO BOI

1. El premio BOI, de periodicidad bienal, está dotado con 300.000 pesetas, además de cuadro Accésit de 100.000 pesetas cada uno.
2. Se considerarán con opción al premio todos los trabajos ORIGINALES, de autores españoles, realizados en España y publicados en Archivos de Bronconeumología entre el n.º 3 del vol. 22 y el n.º 2 del vol. 24, ambos inclusive.
3. Los autores que publiquen sus trabajos en dichos números y no deseen concurrir al Premio, lo manifestarán previamente.
4. Serán eliminados los trabajos relacionados con nuevos medicamentos y los que se refieran a preparados ya comercializados.
5. Por acuerdo de la Junta directiva de SEPAR, el Tribunal Calificador queda constituido como sigue:
   Presidente de SEPAR
   Director de la Revista «Archivos de Bronconeulogía»
   Presidente de las cinco Secciones de SEPAR.
6. La Secretaría de Redacción de la Revista se encargará de comunicar a los miembros del Tribunal de los trabajos que participan en la Convocatoria, tres meses antes de la concesión del Premio y asimismo de reclamar la calificación de los trabajos dos meses después para proceder a la selección de los 10 mejores con arreglo a la media aritmética alcanzada. Asimismo, la Secretaría de SEPAR cuidará de convocar a los componentes del Tribunal.
   Los miembros del Tribunal enviarán a la Secretaría de SEPAR en sobre cerrado su calificación en números enteros de uno a diez, de todos y cada uno de los trabajos que participan en la Convocatoria del Premio BOI.
7. Mediante votaciones secretas el Tribunal Calificador, reunido en la Sede del Congreso, irá eliminando un trabajo en cada votación hasta que solamente quede el ganador, utilizándose después el mismo procedimiento para la adjudiación de los Accésit.
8. La decisión del Tribunal Calificador se anunciará en uno de los actos oficiales del Congreso, e inmediatamente se procederá a la entrega del premio a los autores, o a sus representantes.
9. Estas bases sólo podrán ser modificadas por acuerdo conjunto entre el Laboratorio patrocinador y la Junta de Gobierno de SEPAR.

Diciembre 1987.

XXVI CURSO TEORICO-PRACTICO DE BRONCOLOGIA. CURSO DE ENDOSCOPIA RESPIRATORIA

(Curso Nacional SEPAR)

18-22 de abril de 1988

Información: Hospital de la Sta. Creu i St. Pau.
Av. St. Antoni M.ª Claret, 167
08025 Barcelona
Sección de Broncología del Servicio de Aparato Respiratorio.
Secretaria: Dra. M.C. Puzo.
Tel. 348 12 18