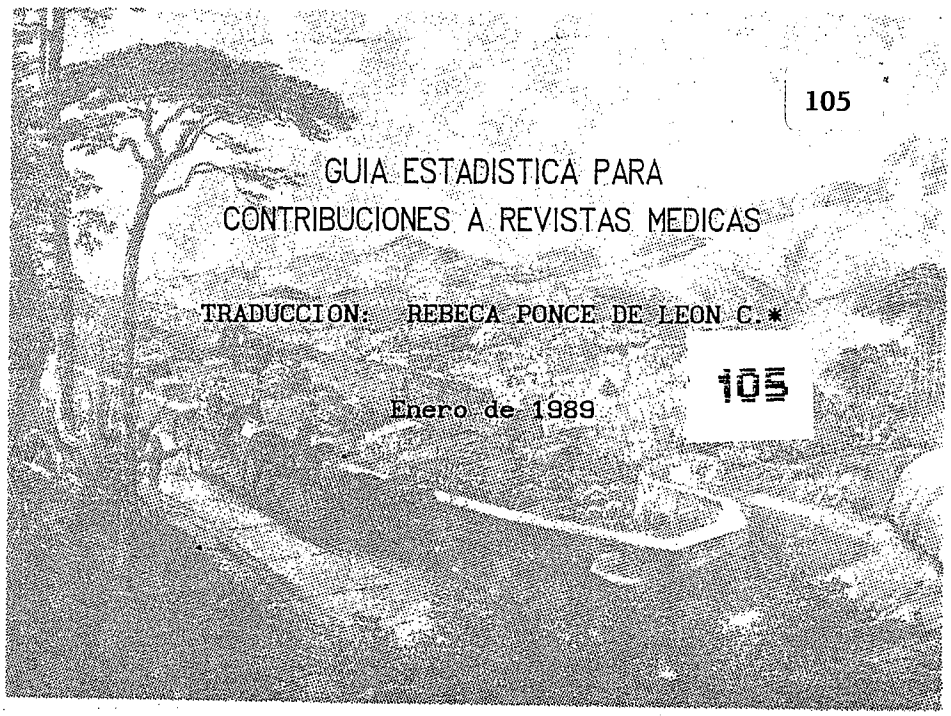
COMUNICACIONES DEL CIMAT



* Investigadora del

CENTRO DE INVESTIGACION EN MATEMATICAS

Apartado Postal 402
Guanajuato, Gto.
México
Tels. (473) 2-25-50
2-02-58

GUIA ESTADISTICA PARA CONTRIBUCIONES A REVISTAS MEDICAS

Douglas G. Altman, Sheila M. Gore, Martin J Gardner y Stuart J Pocock.

Traducido por: Rebeca Ponce de León, Investigadora del CIMAT, Guanajuato, Gto.

La mayoría de los artículos publicados en revistas médicas contienen análisis que fueron llevados a cabo sin ninguna ayuda por parte de un estadístico. Aún cuando casi todos los investigadores médicos están familiarizados con la estadística básica, no hay un camino fácil por el que se puedan adentrar a los conceptos y principios estadísticos importantes. Además, hay poca ayuda disponible para diseñar, analizar y escribir un proyecto de investigación completo. En parte por las razones antes expuestas mucho de lo que se publica en las revistas médicas es estadísticamente pobre o aún equivocado (1). En varias revisiones de artículos publicados en revistas médicas se ha detectado un alto nivel de errores estadísticos y esto ha generado preocupación.

Muy pocas son las revistas que ofrecen consejo estadístico a sus autores potenciales. Se ha sugerido (1,2) que la existencia de guías al respecto podrían ayudar a que los investigadores médicos estuvieran más concientes de la importancia de los principios estadísticos, y a que los orientaran en cuanto a que información deben de proporcionar en un artículo. Nosotros presentamos a continuación un intento en este sentido.

Division of Computing and Statistics, MRC Clinical Research Center, Harrow, Middlesex HA1 3UJ

DOUGLAS G ALTMAN, BSC, medical statistician

MRC Biostatistics Unit, Medical Research Council Center, Cambridge CB2 2QH

SHEILA M GORE, MA PHD, medical statistician

MRC Environmental Epidemiology Unit, Southampton General Hospital, Southampton SO9 4XY

MARTIN J GARDNER, BSC, PHD, reader in medical statistics

Department of Clinical Epidemiology and General Practice, Royal Free Hospital School of Medicine, London NW3 2PF

STUART J POCOCK MSC, PHD, senior lecturer in medical statistics

Correspondence to: Dr S M Gore

Ha sido problemático decidir sobre que incluir en la guía, que tan detallado, y como manejar aspectos respecto a los cuales no hay consenso. Por tanto, esta guía deberá tomarse como un punto de vista, más que como un documento definitivo. La idea no es proporcionar un conjunto de reglas sino aportar información general y consejo sobre los aspectos relevantes del diseño, análisis y presentación estadística. Estas recomendaciones son principalmente un fuerte señalamiento para evitar ciertas prácticas.

Se asume cierta familiaridad con las ideas y los métodos estadísticos ya que se requiere tener ciertos conocimientos de estadística antes de intentar hacer un análisis estadístico. Para aquellos que tengan un conocimiento limitado de lo que es la estadística, estos lineamientos les van a mostrar que el tema es mucho más amplio que las meras pruebas de significancia, e ilustran que tan importante es la correcta interpretación de los resultados. La falta de recomendaciones específicas indica que un buen análisis estadístico requiere de juicio y sentido común, así como de un repertorio de técnicas formales, hay un arte en la estadística así como en la medicina. Esperamos que estos lineamientos representen una visión no polémica de los procedimientos estadísticos más usados y aceptados. Deliberadamente hemos limitado la extensión de estos lineamientos para cubrir unicamente los procedimientos estadísticos más comunes.

Es probable que algunos lectores encuentren que alguna sección relevante presente información con la que no estén familiarizados o que no sea entendible. En estos casos aunque la mayoría de los aspectos que se cubren pueden consultarse en un texto de estadística médica (3,4), le recomendamos fuertemente que busque el apoyo de un estadístico. La ausencia de referencias específicas es intencional: es mejor buscar el apoyo de personas expertas en caso de requerirse mayor explicación. Más aún, como los errores en el diseño no pueden ser rescatados posteriormente, la ayuda de profesionales debe buscarse durante la fase de planeación de un proyecto de investigación y no durante el análisis de la información.

Queremos agradecer al gran número de personas que leyeron las versiones previas de estos lineamientos por sus comentarios constructivos y útiles.

A lo largo del artículo hemos seguido la convención de Vancouver usando "p" para probabilidad, aun cuando la notación estadística prefiere "P".

(I) Introducción

Estos lineamientos pretenden ayudar a los autores para que sepan que es importante desde el punto de vista estadístico y como presentar estos aspectos en sus trabajos. Enfatizan que estos aspectos de presentación se ligan estrechamente con consideraciones de principios estadísticos más generales. No se presenta una discusión detallada sobre como seleccionar un método

estadístico apropiado; la mejor manera de obtener este tipo de información es consultando a un estadístico. Nosotros, sin embargo, llamaremos la atención sobre ciertos usos inadecuados de los métodos estadísticos.

Los lineamientos siguen la estructura usual de un artículo médico: métodos, resultados (análisis y presentación), y discusión (interpretación). Algunos rubros aparecen en más de una sección y se hacen las respectivas referencias.

(2) Sección de Métodos

2.1 Principios Generales

Lo más importante es describir claramente que es lo que se hizo, incluyendo el tipo de diseño de la investigación (sea un experimento, ensayo o encuesta) y la recolección de los datos. El objetivo es presentar la información suficiente para que quede clara la metodología y, si así se desea, pueda ser repetida por otros. Los autores deben incluir la siguiente información relativa al diseño de su investigación:

el objetivo de la investigación, y sus principales hipótesis;

el tipo de sujetos estudiados, mencionando los criterios de inclusión y de exclusión;

la fuente de donde se obtuvieron dichos sujetos y la manera como fueron seleccionados;

el número de sujetos estudiados y la razón por la que se estudió ese número;

los tipos de técnicas de observación y medición que fueron usadas.

Según el tipo de estudio, por ejemplo, encuestas y ensayos clínicos, se va a requerir proporcionar cierta información adicional.

2.2 Encuestas (estudios observacionales)

El tipo de diseño del estudio debe quedar claramente explicado. Por ejemplo, se requiere hacer una descripción detallada del procedimiento que se siguió para seleccionar al grupo control y de cualquier procedimiento de apareamiento. También debe quedar claramente establecido si el estudio fue retrospectivo, transversal o prospectivo. Los procedimientos para seleccionar a los sujetos que van a ser estudiados y el que se alcance una alta tasa de participación son particularmente importantes ya que por lo general los hallazgos se extrapolan a partir de la muestra a alguna población más general. Es útil mencionar si se tomó alguna medida para motivar la participación en la encuesta.

2.3 Ensayos Clínicos

Se requiere de una definición detallada de los regímenes de tratamiento (esto incluye atención médica auxiliar y los criterios para modificar o interrumpir el tratamiento). Se debe señalar explicitamente el método que se siguió para asignar los tratamientos a los sujetos, en particular, el método específico de asignación aleatoria (incluyendo cualquier estratificación) y como éste fue instrumentado. Si se omitió la asignación aleatoria deberá ser señalada esta omisión como un error en el diseño y se deberán dar las razones de la misma.

Se debe describir el uso de técnicas de cegado y de otras precauciones que hayan sido tomadas para asegurar una evaluación insesgada de la respuesta de los pacientes. Hay que enlistar los principales criterios que se siguieron para comparar los tratamientos, tal cual se establecieron en el protocolo del ensayo. Si se trató de un ensayo cruzado habrá que explicar en forma precisa el patrón de tratamiento (y cualquier periodo de "run in" y de lavado).

2.4 Métodos Estadísticos

Todos los métodos estadísticos usados en un artículo deben ser identificados. Cuando se usan numerosas técnicas deberá quedar absolutamente claro que método se usó donde, y quizá esto requiera ser clarificado nuevamente en la sección de resultados. Algunas técnicas de uso muy común, como la prueba de t, la ji-cuadrada, pruebas de Wilcoxon y Mann-Whitney, correlación (r), y regresión lineal, no requieren ser descritas, pero los métodos con más de una forma, como la prueba de t (pareada y no pareada), análisis de varianza y correlación de rangos deberán ser claramente identificadas. Los métodos más complicados requieren ser explicados, y si los métodos son poco usuales entonces habrá que dar una referencia específica. Puede ser de utilidad incluir algún comentario sobre cuál es la razón por la que se usó ese método en particular, sobre todo si sustituye a un método más familiar. También es recomendable proporcionar el nombre del programa o paquete de computo que se usó, por exemplo, el "Statistical Package for the Social Sciences" (SPSS), pero como información adicional a los métodos estadísticos específicos que se usaron.

(3) Sección de resultados: análisis estadístico

3.1 Información descriptiva

Una adecuada descripción de los datos debe preceder y complementar al análisis estadístico formal. Las variables que sean importantes para la validéz e interpretación de subsecuentes análisis estadísticos es importante que se describan en mayor detalle. Esto se puede lograr mediante métodos gráficos, como los diagramas de dispersión y los histogramas, o bien usando

estadísticas de resumen. Las variables contínuas (como son el peso y la tensión arterial) pueden resumirse usando la media aritmética y la desviación estándar (DE) o bien la mediana y un rango percentilar, digamos el rango intercuartilar (percentiles 25 y 75). Las dos últimas son más adecuadas cuando las medidas contínuas tienen una distribución asimétrica. Para datos cualitativos ordinales (como los estados de enfermedad I al V) es incorrecto calcular la media aritmética y la desviación estándar; en su lugar se deben reportar proporciones.

Se deben describir las desviaciones que se presentaron con respecto a lo establecido en el diseño del estudio. Por ejemplo, en los ensayos clínicos es particularmente importante detallar las bajas del estudio con sus respectivas razones, si son conocidas, y la asignación a tratamientos. Para las encuestas, donde la tasa de respuesta es de fundamental importancia, es útil proporcionar información sobre las características de los que no respondieron en comparación con aquellas de las que si participaron. Si se tiene la intención de extrapolar los resultados a una población que sea pertinente, entonces se requerirá investigar la representatividad de la muestra estudiada.

Es útil comparar las características base de diferentes grupos, como lo son los grupos de tratamiento en un ensayo clínico. Si estas diferencias existen, aun cuando no sean estadísticamente significativas, son reales y deberán ser adecuadamente consideradas en el análisis (vea sección 3.12).

3.2 Supuestos (underlying assumptions)

Algunos métodos de análisis como la prueba de t, correlación, regresión y análisis de varianza, dependen de alguna manera de ciertos supuestos sobre la distribución de la(s) variable(s) analizada(s). Técnicamente, estos supuestos se refieren a que los datos provengan de una distribución Normal y a que si se comparan dos o más grupos, la variabilidad en cada uno de ellos sea la misma.

Es imposible decir hasta que grado estos supuestos pueden ser violados sin que se invalide el análisis. Sin embargo, cuando los datos tienen una distribución muy sesgada (asimétrica) o cuando la variabilidad es considerablemente diferente en los grupos, entonces se puede requerir hacer alguna transformación de los datos antes del análisis (ver sección 3.7) o usar los método alternativos "libres de distribución" que no dependen de supuestos acerca de la distribución (con frecuencias conocidos como métodos no-paramétricos). Por ejemplo, la prueba U de Mann-Whitney es el método libre de distribución equivalente a la prueba de t para dos muestras. Estos métodos libres de distribución también son apropiados para conjuntos de datos pequeños, para los que los supuestos no pueden ser validados adecuadamente.

Algunas veces el supuesto de Normalidad es especialmente importante, por ejemplo, cuando el rango de valores calculado como dos desviaciones estándar a ambos lados de la media se toma como el 95% "normal" o rango de referencia. En estos casos el supuesto sobre la distribución debe demostrarse para poder ser justificado.

3.3 Pruebas de significancia

El principal objetivo de las pruebas de significancia es evaluar un número limitado de hipótesis preformuladas. Otras pruebas de significancia que se llevan a cabo porque fueron sugeridas por una inspección preliminar de los datos, nos van a dar una falsa impresión ya que en estas circunstancias el valor de p calculado es muy pequeño. Por ejemplo, no es válido probar la diferencia entre el valor mínimo y el máximo de un conjunto de medias sin hacer referencia a las razones por las que se prueba esa diferncia en particular; existen técnicas especiales para hacer comparaciones pareadas entre varios grupos.

Es costumbre que se hagan pruebas de significancia de dos colas. Si se hace la prueba de una sola cola entonces hay que señalarlo y justificarlo con respecto al problema que se tiene a la mano.

La presentación e interpretación de resultados de pruebas de significancia se discuten en las secciones 4.3, 5.1 y 5.2.

3.4 Intervalos de confianza

En la mayor parte de los estudios existe la preocupación por estimar el valor de alguna cantidad, como por ejemplo la diferencia entre medias o un riesgo relativo. Es deseable calcular el intervalo de confianza alrededor de esa estimación. Este es un rango de valores entre el que nos encontramos, digamos, con la confianza del 95% de que este rango incluye al verdadero valor. Existe una estrecha relación entre el resultado de una prueba de significancia y su respectivo intervalo de confianza: si la diferencia entre tratamientos es significativa a un nivel del 5% entonces su respectivo intervalo de confianza del 95% excluye la diferencia igual a cero. El intervalo de confianza nos proporciona mayor información ya que nos indica el mínimo y máximo efecto verdadero que es compatible con las observaciones muestrales (vea también la sección 5.1).

Los intervalos de confianza ponen de manifiesto la precisión de una estimación. Cuando tenemos un intervalo de confianza amplio nos está señalando la falta de información, independientemente de que las diferencias sean estadísticamente significativas o no, y esto es una advertencia para no sobre-interpretar los resultados de estudios pequeños.

3.5 Observaciones pareadas

Es esencial distinguir el caso de observaciones no-pareadas, donde la comparación es entre medidas de dos grupos diferentes, por ejemplo, sujetos que reciben tratamientos alternativos; de las observaciones pareadas, en donde la comparación es entre dos medidas hechas en los mismos individuos en diferentes circunstancias (como sería antes y después del tratamiento). Por ejemplo, con datos no-pareados debemos usar la prueba de t para dos muestras, mientras que con datos pareados lo correcto es usar la prueba de t pareada. De forma similar, la prueba U de Mann-Whitney para datos no-pareados se sustituye por la prueba pareada de Wilcoxon, y la prueba usual de ji-cuadrada para tablas de 2 X 2 se sustituye por la prueba de McNemar. Siempre debemos aclarar que forma de la prueba se usó.

La misma distinción se debe hacer cuando hay tres o más grupos de observaciones. Todos los métodos estadísticos mencionados en esta sección pueden ser generalizados a más de dos grupos; en particular, la prueba de t pareada y para dos muestras se generaliza a diferentes formas de análisis de varianza.

3.6 Mediciones repetidas

Existe un diseño de estudio que es común y que incluye el registro seriado de mediciones de la misma variable(s) en los mismos individuos en diversos puntos en el tiempo. Este tipo de datos con frecuencia se analizan calculando las medias y las desviaciones estándar para cada tiempo y se presentan graficamente mediante una linea que une esas medias. La forma de esta curva promedio puede ser que no nos de una buena idea de la forma de las curvas independientes. A no ser que las respuestas individuales sean muy similares, es de mayor valor analizar alguna característica de los perfiles individuales, como sería el tiempo que toma alcanzar un pico o el período de tiempo por arriba de un determinado nivel. Lo anterior también ayuda a evitar el problema asociado con pruebas de significancia múltiples (vea sección 5.2).

Cuando se toman medidas repetidas de la misma variable en un mismo individuo bajo las mismas condiciones experimentales, conocidas como lecturas repetidas, no deben ser tratadas como observaciones independientes cuando se comparan grupos de individuos. Si el número de réplicas es el mismo para todos los individuos, el análisis no resulta difícil; en particular, se usa el análisis de varianza, mientras que si se hubiera tratado de datos no replicados hubieramos usado la prueba de t. Si el número de réplicas varía entre los individuos, intentar hacer un análisis general puede resultar muy complejo. Si usamos el valor máximo o mínimo de una serie de medidas (como el valor máximo de tensión arterial durante el embarazo) puede ser erroneamente interpretado si el número de observaciones varía en forma importante entre los individuos.

3.7 Transformación de datos

Muchas variables biomédicas están sesgadas positivamente, esto es que existen valores muy altos, y por tanto para poder hacer un análisis adecuado de ellas puede ser necesario hacer alguna transformación matemática de las mismas. En estas circunstancias con frecuencia se aplica la transformación logarítmica (log) aunque también pueden aplicarse otro tipo de transformaciones (como la raíz cuadrada o reciproca) por ser más adecuadas.

Después de efectuado el análisis se recomienda que los resultados se transformen nuevamente a la escala original antes de ser reportados. En el caso más común que es la transformación logarítmica, se deberá usar el antilogarítmo de la media (conocido como la media geométrica). Cuando se trata de la desviación o error estándar entonces no es correcto sacar el antilogarítmo sino que a los limites de confianza en escala logarítmica se les saca el antilogarítmo para obtener los intervalos estimados en la escala original. Se usa un procedimiento similar con otros tipos de transformaciones.

Si usamos una transformación es importante corroborar que se logró el objetivo deseado (una distribución Normal aproximada). No se debe asumir que la transformación logarítmica, por ejemplo, es necesariamente la más adecuada para todas las variables sesgadas positivamente.

3.8 Valores aberrantes (outliers)

Las observaciones que resulten fuertemente inconsistentes con la mayoría de los datos no deben ser excluidas del análisis a no ser que existan razones para dudar de su credibilidad. Si se excluyen estos valores aberrantes se debe reportar. Como los valores aberrantes pueden tener un efecto importante sobre el análisis estadístico, resulta útil analizar los datos con y sin dichos valores para valorar que tanto se modifican las conclusiones.

3.9 Correlación

De preferencia se debe incluir un diagrama de dispersión por cada coeficiente de correlación presentado, aunque esto podría no ser posible si hay muchas variables. Cuando se investigan varias variables es útil mostrar las correlaciones entre todos los pares de variables en una tabla (matríz de correlación), en lugar de solo mencionar los valores más altos o los que resulten significativos.

Cuando tenemos datos que se distribuyen en forma irregular podemos calcular la correlación de rangos en lugar del usual coeficiente de correlación de Pearson "producto de momentos" (r). La correlación de rangos también la podemos usar cuando se trata de variables limitadas a estar por arriba o por abajo de cierto

valor, por ejemplo, pesos al nacimiento por abajo de 2500 gramos, o si se trata de variables categóricas ordinales. También es útil cuando la relación entre las variables no es lineal, o cuando los valores de una variable fueron seleccionados por el experimentador en lugar de haber sido irrestringidos.

El coeficiente de correlación es útil para resumir el grado de asociación lineal que hay entre dos variables cuantitativas pero es uno de los métodos estadísticos que con mayor frecuencia se usan incorrectamente. Existen numerosas circunstancias en las que la correlación no debe ser usada. Es incorrecto calcular el coeficiente de correlación simple para datos que incluyan más de una observación en algunos o todos los sujetos en estudio ya que dichas observaciones no son independientes. La correlación también es inapropiada cuando se quieren comparar métodos alternativos de medición de la misma variable porque la correlación valora asociación y no concordancia. Cuando se usa la correlación para relacionar el cambio a través del tiempo con respecto al valor inicial puede llevarnos a una interpretación de resultados incorrecta.

También puede ser incorrecto calcular el coeficiente de correlación a partir de datos que comprenden subgrupos que se sabe difieren en su nivel promedio de una o de las dos variables, por ejemplo, combinar datos de hombres y mujeres cuando una de las variables es la altura.

La regresión y la correlación son dos técnicas separadas y con objetivos diferentes y no necesariamente una debe de ir acompañada de la otra. La interpretación de los coeficientes de correlación se discute en la sección 5.3.

3.10 Regresión

Es altamente recomendable presentar la linea de regresión ajustada junto con el diagrama de dispersion de los datos. Si presentamos unicamente la linea sin incluir los datos obtenemos muy poca información adicional a cuando únicamente nos hubieran presentado la ecuación de la regresión. Es útil proporcionar los valores de intersección y una medida de la dispersión de los puntos en torno a la linea ajustada (la desviación estándar residual). También se deben presentar los limites de confianza en torno a la linea de regresión para mostrar la incertidumbre de las predicciones basadas en la interrelación ajustada. Estos límites no son paralelos a la linea sino curvos, demuestran que hay mayor incertidumbre en la predicción de valores en el eje horizontal (x) que se encuentran alejados de la mayoría de las observaciones.

Cuando se efectúa un análisis de regresión con datos que incluyen distintos subgrupos puede llevarnos a resultados equivocados, en particular si los grupos difieren en el valor promedio de la variable dependiente (y). En estos casos se obtienen resultados más confiables usando el análisis de covarianza.

La regresión y la correlación son técnicas separadas que sirven a diferentes propósitos y no necesitan automáticamente acompañarse una de la otra. La interpretación del análisis de regresión se discute en la sección 5.4.

3.11 Datos de sobrevivencia

El reporte de datos de sobrevivencia debe incluir la presentación tabular o gráfica de tablas de vida, con detalles sobre cuantos pacientes se encontraban en riesgo (digamos, de morir) en diferentes tiempos del seguimiento. La tabla de vida maneja eficientemente los tiempos de sobrevivencia "censurados" que surgen cuando los pacientes se pierden para el seguimiento o continúan vivos; se sabe que su tiempo de sobrevivencia es de cuando menos "tantos dias". El cálculo del tiempo promedio de sobrevivencia es impredecible si existen censurados y porque la distribución de los tiempos de sobrevivencia por lo general está sesgada positivamente.

Si se hacen comparaciones entre las proporciones que sobreviven a un determinado tiempo, arbitrariamente fijado, entre grupos de tratamiento éstas pueden ser mal interpretadas, y por lo general estas comparaciones son menos eficientes que cuando se comparan tablas de vida por métodos como el de la prueba de los rangos logarítmicos.

Cuando existen suficientes defunciones se puede mostrar cómo el riesgo de morir varía con el tiempo graficando, de acuerdo a intervalos de tiempo iguales y convenientes, la proporción de los que continúan vivos al inicio de cada intervalo de tiempo y de los que mueren durante ese intervalo. Es posible hacer ajustes de acuerdo a ciertos factores de los pacientes que puedan tener influencia sobre el pronóstico, usando modelos de regresión apropiados para datos de sobrevivencia (ver sección 3.12).

3.12 Análisis complejos

En muchos estudios las variables de principal interés pueden estar influenciadas por muchas otras variables. Estas pueden ser cualquier cosa que varíe entre los sujetos y puedan haber afectado los resultados observados. Por ejemplo, en los ensayos clínicos se puede tratar de algunas características de los pacientes, o signos y síntomas. Algunas o todas las covariables pueden combinarse mediante técnicas apropiadas de regresión múltiple para explicar o predecir una variable de respuesta, sea una variable contínua (tensión arterial), una variable cualitativa (trombosis post-operatoria), o el tiempo de sobrevivencia. Aún en ensayos clínicos de asignación aleatoria los investigadores necesitan asegurarse de que el efecto del tratamiento se mantiene presente aún después de ajustar simultaneamente varios factores de riesgo.

Las técnicas multivariadas que manejan más de una variable de respuesta simultaneamente, requieren de la ayuda de un experto y van más allá del campo de acción de estos lineamintos.

Cualquier método estadístico complejo debe presentarse de una manera que sea comprensible para el lector.

(4) Sección de resultados: presentación de resultados

4.1 Presentación de estadísticas de resumen

Los valores de la media no deben de ser presentados sin acompañarse de alguna medida de dispersión o precisión. La desviación estándar (DE) se debe usar para mostrar la variabilidad entre los individuos y el error estándar de la media (EE) para mostrar la precisión de la media muestral. Se debe aclarar cuál medida se está presentando.

El uso del simbolo ± conjuntamente con el error estándar o la desviación estándar y la media (como en 14.2 ± 1.9) causa confusión y se debe evitar. Es preferible presentar la media como, por ejemplo, 14.2 (DE 1.9) o 14.2 (DE 7.4). Una manera adecuada de presentar la media conjuntamente con sus límites razonables de incertidumbre son los intervalos de confianza, un intervalo de confianza del 95% para el verdadero valor de la media va de aproximadamente dos errores estándar por abajo de la media observada a dos errores estándar por arriba de ella. Los intervalos de confianza se presentan con mayor claridad cuando se dan los límites, como en (10.4, 18.0), que cuando se usa el símbolo ±.

Cuando se hacen comparaciones pareadas, como cuando se usa la prueba de t pareada, es recomendable que se proporcionen los valores de la media y el error estándar (o desviación estándar) de las diferencias entre las observaciones.

Cuando se trata de datos analizados mediante métodos libres de distribución es más apropiado proporcionar la mediana y el rango cental, que cubre, por ejemplo, al 95% de las observaciones, que usar la media y la desviación estándar (ver sección 3.1). De igual manera, si el análisis se llevó a cabo con datos transformados entonces el valor de la media y de la desviación estándar de los datos crudos probablemente no sean buenas medidas de tendencia central y de dispersión y no deberán ser presentadas.

Cuando se presentan porcentajes el denominador debe quedar claramente especificado. Cuando se trata de muestras pequeñas no es muy útil el uso de porcentajes. Es importante que cuando se comparan porcentajes se distinga entre diferencias absolutas y diferencias relativas. Por ejemplo, una reducción del 25% al 20% puede ser expresada como 5% o como 20%.

4.2 Resultados individuales

El rango general no es un buen indicador de la variabilidad de un conjunto de observaciones ya que se puede ver fuertemente afectado por un solo valor extremo y además se va a ver incrementado con el tamaño de muestra. Si los datos se aproximan razonablemente a una distribución normal entonces el intervalo de dos desviaciones estándar hacia cualquier lado de la media cubre alrededor del 95% de las observaciones, pero con otro tipo de distribuciones el rango percentilar es más ampliamente aplicable (ver sección 3.1).

Aunque el análisis estadístico se concentra en los efectos promedio, en muchas circunstancias es importante que también se considere como respondieron los sujetos estudiados a nivel individual. De este modo, por ejemplo, con frecuencia es relevante clinicamente saber cuantos pacientes no mejoraron con el tratamiento además de conocer el beneficio promedio. No se debe interpretar un efecto promedio como aplicable a todos los individuos (vea también la sección 3.6).

4.3 Presentación de resultados de pruebas de significancia

Las pruebas de significancia generan valores observados de los estadísticos de prueba que se comparan con los valores tabulados de acuerdo a la distribución de que se trate (Normal, t, X^2, etc) para derivar los correspondientes valores de p. Se aconseja reportar los valores observados de los estadísticos de prueba y no únicamente los valores de p. Independientemente de que la prueba resulte significativa o no, se deben proporcionar los resultados cuantitativos que fueron probados, tal como valores de medias, proporciones, o coeficientes de correlación. Se debe precisar claramente cuales datos fueron analizados. Si se usan símbolos, como los asteriscos, para denotar niveles de probabilidad, estos deben ser definidos y de preferencia deben ser los mismos a lo largo de todo el artículo.

Los valores de p convencionalmente se presentan como <0.05, <0.01, o <0.001 pero no existe ninguna razón aparte de la familiaridad que justifique usar estos valores en particular. Los valores exactos de p (con no más de dos cifras significativas), como $p=0.18$ o 0.03, son de mayor utilidad. Es poco probable que sea necesario especificar valores de p por abajo de 0.0001. No es recomendable que a cualquier valor de $p > 0.05$ se le etiquete como "no significativo" ya que de esta manera se pueden obscurecer resultados que si bien no son estadísticamente significativos pueden sugerir efectos reales (vea la sección 5.1). Cuando se utilizan simbolos para expresar los valores de p, es importante distinguir $<$ (menor que) de $>$ (mayor que). Los valores de p que están entre dos límites deben ser expresados en orden lógico, por ejemplo, $0.01 < p < 0.05$ en donde p se encuentra entre 0.01 y 0.05. Los valores de p que se presentan en las tablas no necesitan repetirse en el texto.

La interpretación de las pruebas de significancia y de los valores de p se discuten en la sección 5.1.

4.4 Figuras (presentación gráfica)

La presentación gráfica de resultados ayuda al lector, se debe fomentar el uso de gráficas que presentan las observaciones individuales. Cuando en una gráfica se presentan puntos que corresponden a un mismo individuo en diferentes ocasiones deben unirse o bien usar símbolos para indicar que son puntos relacionados. Una alternativa útil es graficar para cada individuo la diferencia entre diferentes ocasiones.

Las acostumbradas "barras de error" de un error estándar por abajo y arriba de la media representan solamente un intervalo de confianza del 67%, y son por lo tanto sujetas a malas interpretaciones; es preferible usar el intervalo de confiaza del 95%. La presentación gráfica de este tipo de información está sujeta a las mismas consideraciones que se discutieron en la sección 4.1.

En los diagramas de dispersión en que se relacionan dos variables se deben mostrar todas las observaciones, aún cuando esto signifique pequeños ajustes para acomodar los puntos duplicados. Los puntos múltiples también se pueden reemplazar por simbolos que representen el número de observaciones que coinciden en un mismo punto.

4.5 Tablas

Es más fácil explorar resultados numéricos siguiendo columnas que siguiendo renglones, y también es mejor tener diferentes tipos de información (como medias y errores estándar) en columnas separadas. Se debe señalar el número de observaciones para cada resultado de la tabla. Resulta más sencillo leer tablas que proporcionan información sobre pacientes en forma individual, áreas geográficas, etc., si los renglones están ordenados de acuerdo a los diferentes niveles de una de las variables presentadas.

4.6 Precisión numérica

La falsa precisión no añade valor a un artículo y más aún le resta confiabilidad y credibilidad. Los resultados obtenidos con una calculadora o una computadora por lo general necesitan ser redondeados. Cuando se presentan medias, desviaciones estándar u otras estadísticas el autor debe tener en mente la precisión de los datos originales. Normalmente las medias deben expresarse con no más de un decimal adicional al de los datos crudos; las desviaciones estándar y los errores estándar pueden requerir de un decimal extra. Pocas veces es necesario que los porcentajes tengan más de una cifra decimal, y aún una cifra decimal con frecuencia no es necesaria. Con muestras mayores a 100 individuos el uso de cifras decimales implican una precisión no garantizada y por tanto deben evitarse. Es importante hacer notar que estos señalamientos se aplican solamente para la presentación de resultados, no se deben redondear cifras antes o durante el análisis de los datos. Es suficiente con que los valores de t, X^2, y r tengan dos cifras decimales.

4.7 Términos técnicos diversos

Es imposible que aquí se definan todos los términos estadísticos. Los siguientes comentarios se relacionan con términos que con frecuencia se usan en forma inadecuada o confusa.

El término **correlación** de preferencia no se debe usar como un término general que describe cualquier relación. Tiene un significado técnico específico como medida de asociación y como tal se debe reservar para el trabajo estadístico.

Incidencia debe usarse para describir la tasa de ocurrencia de casos nuevos de una determinada característica en una muestra o población en estudio, como es el número de nuevas notificaciones de casos de cancer en un año. La proporción de una muestra que ya tiene la característica es la **prevalencia**.

No-paramétrico se refiere a ciertos análisis estadísticos, como la prueba U de Mann-Whitney; no es una característica de las observaciones como tales.

Parámetro no debe usarse como sinónimo de "variable" para referirse a la medida o atributo de donde las observaciones fueron hechas. Los parámetros son características de las distribuciones o relaciones en las poblaciones, que se estiman mediante el análisis estadístico de una muestra de observaciones.

Percentiles. Cuando el rango de valores de una variable se divide en grupos iguales, los puntos de corte son la media, terciles, cuartiles, quintiles y así sucesivamente; nos referimos a estos grupos como mitades, tercios, cuartos, quintos, etc.

Sensibilidad es la habilidad de una prueba para identificar una enfermedad cuando en realidad ésta está presente, esto es, la proporción positiva de aquellos que tienen la enfermedad. **Especificidad** es la habilidad de una prueba para identificar la ausencia de una enfermedad cuando en realidad la enfermedad no está presente, esto es, la proporción negativa de aquellos que no tienen la enfermedad. Vea también la sección 5.4.

(5) Sección de discusión: interpretación

5.1 Interpretación de las pruebas de significancia

Una prueba de significancia valora, por medio de la probabilidad p, la factibilidad de los datos observados cuando alguna "hipótesis nula" (como la de no diferencia entre grupos) es verdadera. El valor de p se refiere a la probabilidad de que los datos observados, o un resultado extremo, haya ocurrido por azar, esto es, unicamente debido a variación muestral, cuando la hipótesis nula es verdadera. Si el valor de p es pequeño entonces dudaríamos de la hipótesis nula. Si el valor de p es grande entonces es factible que los datos sean consistentes con la hipótesis nula, y por esta razón dicha hipótesis no puede ser rechazada. Por lo tanto, el valor de p es la probabilidad de que haya un efecto real.

Aún si existe un efecto real muy importante es probable que encontremos un resultado no-significativo si el número de observaciones es pequeño. Por el contrario, si el tamaño de la muestra es muy grande entonces puede ocurrir que encontremos un resultado estadísticamente significativo aún cuando solamente exista un efecto real pequeño. De lo anterior se deriva que la significancia estadística no debe ser tomada como sinónimo de importancia clínica.

La interpretación de las pruebas de significancia en buena parte se deriva de lo antes comentado. Un resultado significativo no necesariamente indica un efecto real. Siempre hay el riesgo de obtener un hallazgo falso positivo; este riesgo disminuye para valores de p pequeños. Más aún, un resultado no-significativo (convencionalmente >0.05) no significa que no haya efecto sino unicamente que los datos son compatibles con que no haya efecto. Es deseable que exista cierta flexibilidad para interpretar los valores de p. El nivel 0.05 es un punto de corte conveniente, pero valores de p de 0.04 y 0.06, que no difieren grandemente, nos llevan a interpretaciones similares y no a interpretaciones radicalmente diferentes. El que se considere cualquier resultado con un valor de p >0.05 como no significativo puede confundir al lector (y a los autores); de aquí la sugerencia de la sección 4.3 de presentar el valor exacto de p.

Los intervalos de confianza son extremadamente útiles en la interpretación, particularmente en estudios pequeños, en tanto que muestran el grado de incertidumbre relacionado con un resultado, como es la diferencia entre dos medias, sea o no estadísticamente significativo. Pueden ser especialmente ilustrativos si se usan conjuntamente con resultados no-significativos.

5.2 Pruebas de significancia diversas

En muchos proyectos de investigación algunas pruebas de significancia se relacionan con comparaciones importantes que se planearon desde que la investigación se inició. Las pruebas de hipótesis que no se planearon desde la fase inicial de la investigación son secundarias, especialmente cuando son sugeridas por los resultados. Es importante distinguir entre estas dos situaciones y se deberá dar mucho mayor peso a las pruebas de las hipótesis que se plantearon desde el principio. Las otras pruebas deben considerarse únicamente como exploratorias, que permitirán formular nuevas hipótesis que serán probadas en estudios posteriores. Una de las razones que justifica lo anterior es que cuando se llevan a cabo multiples pruebas de significancia en el análisis de un mismo estudio, quizá comparando muchos subgrupos o explorando muchas variables, es de esperarse que surjan cierto número de resultados falsos positivos debidos únicamente al azar, mismos que van a ocasionar numerosos problemas para la interpretación. Sin embargo, es claro que mientras más pruebas se efectúen es más probable que se encuentren resultados significativos, pero el número de resultados falsos positivos esperados también va a aumentar. Una

manera de permitir el riesgo de resultados falsos positivos es establecer un nivel bajo de p como criterio para la significancia estadística.

Un problema más complejo surge cuando se llevan a cabo pruebas de significancia con datos dependientes (correlacionados). Un ejemplo de esta situación es el análisis de datos seriados (discutido en la sección 3.6), donde la misma prueba se efectua con datos para la misma variable recolectados en diferentes tiempos. Otro caso es cuando se hacen análisis por separado de dos o más variables correlacionadas como si estas fueran independientes; cualquier corroboración no incrementa grandemente el peso de la evidencia porque las pruebas se refieren a datos muy similares. Por ejemplo, las tensiones sistólica y diastólica se comportan de manera muy similar, al igual que las diferentes formas para valorar la respuesta de los pacientes. En estos casos se requiere hacer una interpretación muy cuidadosa de los resultados.

5.3 Asociación y causalidad

Una asociación estadísticamente significativa (obtenida por correlación o análisis de X^2) por sí misma no aporta evidencia directa de una relación causal entre las variables involucradas. En estudios observacionales la causalidad puede ser establecida solo en terrenos no-estadísticos; es más sencillo inferir causalidad en ensayos clínicos de asignación aleatoria. Se debe tener gran cuidado al comparar variables que varíen con el tiempo, porque en estos casos es fácil detectar una aparente asociación que es falsa.

5.4 Predicción y pruebas diagnósticas

Aún cuando el análisis de regresión haya indicado una relación estadísticamente significativa entre dos variables pudo haber habido una considerable imprecisión cuando se usó la ecuación de la regresión para predecir el nivel numérico de una variable (y) a partir de la otra (x) para casos individuales. La exactitud de dichas predicciones no puede ser valorada a partir de los coeficientes de correlación o regresión sino que se requiere calcular el intervalo de confianza para el valor de (y) que se predijo y que corresponde a un valor específico de (x) (ver sección 3.10). La recta de regresión debe usarse unicamente para predecir la variable (y) a partir de la variable (x), y no al contrario.

Una prueba diagnóstica con una alta sensibilidad y especificidad no necesariamente es una prueba útil para fines diagnósticos , especialmente cuando se aplica en una población en donde la prevalencia de la enfermedad es baja. En estos casos es útil calcular la proporción de sujetos con resultados de la prueba positivos y que en realidad tienen la enfermedad. Hay que hacer notar que no hay consenso en la definición de "tasa de falsos positivos" y "tasa de falsos negativos", por lo que siempre es necesario aclarar exactamente que es lo que se está calculando, y la mejor manera de ilustrarlo es mediante una tabla de 2 X 2 en que se relacione los resultados de la prueba con el verdadero estado de enfermedad de los pacientes.

Un problema diagnóstico similar surge con variables continuas. Es común clasificar como "anormal" a aquellos valores que salen del "rango normal" de esa variable, pero si la prevalencia de la verdadera anormalidad es baja entonces la mayoría de los valores que estén fuera del rango normal van a ser normales. La definición de anormalidad se debe basar tanto en el criterio clínico como en el estadístico.

5.5 Debilidades

Es mejor señalar los puntos débiles en el diseño y la ejecución de una investigación, si uno está conciente de ellos, y considerar los posibles efectos que estas debilidades pudieron tener en los resultados y en su interpretación, que ignorarlos con la esperanza de que no se noten.

(6) Notas finales

El propósito de los métodos estadísticos es el de proveer directamente de información verídica sobre la evidencia científica derivada de una pieza de investigación. La habilidad y experiencia que se necesita para diseñar estudios adecuados, efectuar análisis estadísticos razonables, y comunicar los hallazgos de una manera clara y objetiva no son fáciles de adquirir. Esperamos que estos lineamientos contribuyan a mejorar el estándar del trabajo estadístico que se reporta en las publicaciones médicas.

Referencias

1. Altman DG. Statistics in medical journals. Statistics in Medicine 1982; 1: 59-71.

2. O'Fallon JR, Dubey SB, Salsburg DS, et al. Should there be statistical guidelines for medical research papers? Biometrics 1978; 34: 687-95.

3. Armitage P. Statistical methods in medical research. Oxford: Blackwell, 1971.

4. Colton T. Statistics in medicine. Boston: Little, Brown, 1974.