Artículo de revisión

Common conceptual statistical mistakes in scientific literature

Errores conceptuales de estadística más comunes en publicaciones científicas

Equívocos comuns de estatísticas em publicações científicas

Roberto Antonio Matamoros Pinel[1*], MV, MSc, PhD, CvLAC; Alejandro Ceballos Márquez[2], MVZ, MSc, PhD, CvLAC

Fecha correspondencia:
Recibido: 11 de junio de 2017.
Aceptado: 9 de noviembre de 2017.

Forma de citar:
Matamoros Pinel RA, Ceballos Márquez A. Errores conceptuales de estadística más comunes en publicaciones científicas. Rev. CES Med. Vet. Zoot. Vol 12 (3): 211-229.

Open access
© Copyright
Creative commons
Éthics of publications
Peer review
Open Journal System
DOI: http://dx.doi.org/10.21615/cesmvz.12.3.4
ISSN 1900-9607

Filiación:

[1*] Profesor Titular, Universidad Santo Tomás, Av. Ejército 146, 6° piso, Torre A. Santiago, Chile.

[2] Profesor Asociado, Grupo de Investigación Calidad de Leche y Epidemiología Veterinaria, Universidad de Caldas, Manizales, Colombia.

Abstract

This work seeks to provide up-to-date information on commonly made statistical mistakes and statistical reports in scientific papers, and how these can be avoided. The main goal is to comprehensively review frequently observed statistical errors, flaws and pitfalls in medical and veterinary science in order to help researchers to produce statistically correct output in their future reports. At the same time, it can help readers to identify questionable statistical analysis, and estimate what the authors would have concluded when appropriate statistical methods have been used.

Keywords: *statistical mistakes, statistical methods, research statistics*.

Resumen

Este trabajo proporciona información actualizada sobre los errores estadísticos que comúnmente se cometen en la literatura científica y como estos podrían ser evitados. El objetivo principal es una revisión de manera conceptual de los errores estadísticos frecuentemente observados, así como defectos y trampas en la ciencia médica y veterinaria para ayudar a los investigadores a producir resultados estadísticamente correctos en sus futuras investigaciones. Al mismo tiempo, esta revisión podría ayudar a que los lectores de revistas científicas identifiquen análisis estadísticos o presentación de datos cuestionables y puedan estimar lo qué los autores habrían concluido si hubieran utilizado métodos estadísticos apropiados.

Palabras clave: *errores estadísticos, estadística en investigación, métodos estadísticos.*

Resumo

Este trabalho procura fornecer uma atualização sobre os erros estatísticos comumente cometidos na literatura científica e como eles podem ser evitados. O principal objetivo é analisar de uma maneira conceitual defeitos observados frequentemente, erros estatísticos e armadilhas na ciência médica e veterinária para ajudar os pesquisadores a produzir resultados estatisticamente corretos em futuras investigações. Ao mesmo tempo, esta revisão pode ajudar os leitores de revistas científicas a identificar análise estatística ou apresentação de dados questionáveis e estimar o que os autores teria concluído teve eles usaram métodos estatísticos adequados.

Palavras-chave: *erros estatísticos, métodos estatísticos, estatística de pesquisa*.

Introducción

El conocimiento de la bioestadística brinda múltiples ventajas al profesional del área de las ciencias de la salud, ciencias biológicas, medicina veterinaria, zootecnia y otras profesiones afines, ya sea porque busque actualizarse continuamente, se dedique a la investigación o simplemente se busque contar con una mirada más crítica a los ensayos clínicos que se publican diariamente. Este conocimiento ayuda al profesional a interpretar mejor los resultados de estudios originales, permite comprender muchos de los tecnicismos que se encuentran en ellos y enjuiciarlos críticamente. En este contexto, la estadística es considerada una herramienta poderosa en la investigación científica con un uso incrementado de métodos y paquetes estadísticos que han sido ampliamente documentados en la literatura científica en la última década, además de ser incorporados en la mayoría de los programas estadísticos [4, 58, 59]. Sin embargo, existe un consenso sobre el hecho que la calidad estadística de los trabajos es generalmente baja ya que una proporción relevante de publicaciones científicas, tanto en medicina humana como veterinaria, todavía contienen errores estadísticos en la aplicación de las técnicas y en la interpretación de los resultados, lo que no ha variado en los últimos 30 años [5, 8, 17, 18, 54].

Este problema compromete la calidad en el reporte de los resultados porque el uso inapropiado de los análisis estadísticos puede generar conclusiones incorrectas, un aumento de la probabilidad de cometer los errores Tipo I y II en las pruebas de hipótesis, resultados falsos en investigación y una pérdida considerable de recursos [19, 22, 32, 40]. Por lo anterior, el objetivo de esta revisión es describir de manera conceptual los errores más comunes, que se presentan en las publicaciones científicas.

Materiales y métodos

El presente trabajo se realizó considerando una revisión narrativa, no siguiendo la metodología para revisiones sistemáticas. Se seleccionaron, según la experiencia de los autores, los artículos y capítulos de libros que directa o indirectamente abordaban errores estadísticos que comúnmente se cometen en la literatura científica, principalmente en el área biomédica. El factor determinante para incluir las publicaciones referenciadas fue la revisión de una manera conceptual más que operacional, sobre los errores estadísticos frecuentemente observados.

Los criterios de inclusión de la literatura revisada contemplaron libros y artículos publicados entre 1960 y 2016 en inglés y español. Las palabras clave que se utilizaron para la búsqueda fueron: investigación biomédica, errores estadísticos y métodos estadísticos. Las bases de datos consultadas para esta selección fueron: Proquest, Medline, Pubmed, Elsevier, Ebsco y Pubmed.

Posteriormente, los trabajos fueron citados según la experiencia de los autores, quienes han sido profesores y asesores en bioestadística por más de 18 años en diferentes universidades. Las publicaciones fueron todas citadas según la necesidad del tema que se estuviera abordando en la revisión.

Conceptos generales

Antes de abordar los errores en la aplicación de la estadística, es necesario recordar ciertos principios fundamentales de la estadística inferencial.

Introducción a las pruebas de hipótesis

El mayor objetivo del análisis estadístico es obtener conclusiones o inferencias sobre la población examinando una muestra obtenida de una población objetivo[37]. Se comienza por realizar un enunciado sobre el promedio poblacional. Este enunciado corresponde al concepto de hipótesis nula (abreviada H_0) que expresa la "no diferencia" entre los grupos[59]. Por ejemplo, la H_0 acerca de un promedio poblacional (μ) puede decir que el μ no es diferente de cero, que podría expresarse como H_0: $\mu = 0$, lo que significa que la diferencia entre las medias de los dos grupos a comparar es igual a 0. Podríamos hipotetizar que el promedio poblacional no es diferente de (o es igual a) 3,5 cm o que no es diferente de 10,5 kg, que en cada caso se escribiría como H_0: $\mu = 3{,}5$ cm y H_0: $\mu = 10{,}5$ kg. Si se concluye que es probable que la hipótesis nula sea falsa, entonces se considera que hay suficiente evidencia para rechazarla[58]. Generalmente se enuncian una H_0 y una H_a por cada prueba estadística que se desea realizar. Para los ejemplos anteriores: H_0: $\mu = 0$, H_a: $\mu \neq 0$; H_0: $\mu = 3{,}5$ cm, H_a: $\mu \neq 3{,}5$ cm; y H_0: $\mu = 10{,}5$ kg, H_a: $\mu \neq 10{,}5$ kg.

Las hipótesis estadísticas deben ser enunciadas antes de la colección de los datos para ser analizados, así como el tipo de análisis debe ser seleccionado a priori una vez se tiene claridad sobre los datos que serán recolectados[59], complementario a ello debe considerarse también que los datos no deben direccionar el análisis estadístico a realizar. Se necesita un criterio objetivo para rechazar o no la H_0 por una prueba estadística. En teoría, una diferencia encontrada, por ejemplo, entre dos tratamientos farmacológicos, daría lugar a un valor de la prueba estadística que necesitamos para saber si la hipótesis se rechaza o no[3].

Cabría preguntarse: ¿Cuán alto es el valor de la prueba estadística o cuán baja debe ser la probabilidad para rechazar la hipótesis nula? Por convención internacional y según el criterio del investigador, una probabilidad de 5% o menos es comúnmente usada como criterio para rechazar la H_0. La probabilidad usada como criterio *a priori* de rechazo se denomina nivel de significancia denotada por alfa (α) y el valor de la prueba estadística (t, X^2, prueba Z, etc.) que corresponde a α se denomina valor crítico de la prueba estadística[59].

Pese a lo anterior, las pruebas de significancia han llevado a acentuar la divulgación de los resultados considerados positivos en la literatura médica en detrimento de la publicación de aquellos resultados nulos o negativos, lo que sin duda contribuye a incrementar el sesgo de publicación[53]. Estos autores mencionan que no solamente el sesgo de publicación se puede incrementar por la publicación de resultados positivos, sino por la falta de claridad en el entendimiento de la significancia estadística. En los últimos años se han publicado revisiones sobre la arbitrariedad en la escogencia del valor de P para declarar la significancia, que desde 1926 R. Fisher estableció como el valor "estándar" a partir del cual deberían repetirse experimentos bien diseñados para concluir finalmente que los resultados no son producto del azar[21, 47, 53].

Tal ha sido la generalización de la aceptación de un valor-P<0,05 como indicador de significancia, que todos aquellos resultados que no cumplen con esta premisa se han considerado como no significantes, aunque el valor de P sea ligeramente superior a 0,05. Como una forma de darle la oportunidad a la publicación de los resultados "no significantes", se ha iniciado desde 2002 la publicación seriada "Journal of Negative Results in Biomedicine" (http://www.jnrbm.com). Esta publicación tipo acceso abierto, con comité editor y arbitraje por parte de pares evaluadores provee una

plataforma para la discusión de todos aquellos resultados que "no se esperaban", que pueden ser controversiales o provocar diversas reacciones en la comunidad científica, justamente por ser resultados negativos. A su vez, en la actualidad, de acuerdo a las normas de presentación de manuscritos del International Committee of Medical Journal Editors (http://www.icmje.org), se recomienda describir los intervalos de confianza y no confiar únicamente en la prueba de hipótesis estadística, que incluye valores P, ya que no transmiten información importante sobre el tamaño del efecto y la precisión de las estimaciones.

Errores estadísticos en las pruebas de hipótesis

Es muy importante darse cuenta que una H_0 verdadera ocasionalmente será rechazada lo que significa que se habrá cometido un error en obtener una conclusión acerca de la población muestreada. Más aún, este error se cometerá con una frecuencia de α [59]. Esto es, si H_0 es de hecho un enunciado verdadero acerca de la población (por ejemplo, no hay diferencia entre tratamientos) se concluirá (erróneamente) que es falsa en un 5% de los casos (es decir, se van a encontrar diferencias donde realmente no las hay). El rechazo de H_0 cuando es de hecho verdadera es conocido como error tipo I o error de primera clase [37].

Por otro lado, si la H_0 es falsa, una prueba estadística, en ocasiones, no va a detectar este hecho y estaríamos cometiendo un error al no rechazarla (existen diferencias, pero no se fue capaz de detectarlas). La probabilidad de cometer este error (no rechazar la H_0 cuando es de hecho falsa) se denomina error tipo II o error de segunda clase[37]. Mientras que la probabilidad de cometer el error tipo I es α, el nivel de significancia especificado, la probabilidad de cometer el error tipo II es β, un valor que generalmente no se especifica porque se desconoce (generalmente no existe un valor único de β) [58]. Lo que se sabe es que para una muestra dada de tamaño "n" el valor de α esta inversamente relacionado con el valor de β. Es decir, una menor probabilidad de cometer el error tipo I está asociada con una mayor probabilidad de cometer el error tipo II y la única manera de reducir la probabilidad de cometer ambos tipos de error simultáneamente es incrementar el tamaño de la muestra [59] (esto lleva a la recomendación de tener el máximo tamaño de muestra posible, garantizando la disminución de β, pero dejando desconocido su valor). Así, para un α dado, muestras más grandes resultarán en una prueba estadística con mayor poder estadístico (concepto que se aclara más adelante). Para un "n" dado no se pueden minimizar ambos errores, por lo que es preciso saber cuál es la combinación aceptable para ambas probabilidades de error.

Por convención internacional, un α de 0,05 es usualmente considerado lo suficientemente bajo como probabilidad de cometer un error tipo I, pero no lo suficientemente bajo para que resulte en una probabilidad alta de cometer un error tipo II [59]. Sin embargo, no hay una regla general que indique la necesidad de fijar el valor de α en 0,05. A pesar de que es el nivel de significancia usado con más frecuencia, los investigadores pueden decidir si es más importante minimizar un tipo de error o el otro. En algunas situaciones, el 5% de probabilidad de rechazar incorrectamente la H_0 puede ser considerada inaceptablemente alta así que un nivel de significancia de 1% es usado.

Estos errores son análogos a los errores que se pueden cometer en un juicio criminal. Como resultado de una sentencia podemos declarar culpable a una persona que en realidad es inocente. Alternativamente, podemos declarar inocente a una persona que en realidad es culpable [38]. El sistema de justicia moderno considera que el primer error, declarar culpable a un inocente, es mucho más grave. Consecuentemente, el sistema se basa en la inocencia de toda persona a menos que se demuestre lo contrario.

En estadística se realiza una consideración similar. Se determina a priori cuál es el nivel aceptable de cometer un error Tipo I, (rechazar una H_0 que es verdadera = culpar a un inocente) y se deja variar la probabilidad de cometer un error Tipo II (aceptar una H_0 que es falsa = declarar inocente a un culpable).

Definición de poder estadístico

Una vez realizado el experimento, muchas veces se necesita saber si estos experimentos tienen el suficiente "poder" estadístico, comúnmente denominado *power* [58]. Si el análisis estadístico resulta en una conclusión estadísticamente significativa, es bastante fácil de interpretar. Pero la interpretación de resultados no estadísticamente significativos es más difícil [4,5]. Aun cuando el tratamiento realmente puede afectar el resultado de un experimento (es decir tiene un efecto), es posible que no se obtenga una diferencia estadísticamente significativa en la prueba estadística realizada. Sólo por casualidad o azar, los datos, podrían proporcionar un valor-P mayor que 0,05 (o cualquier valor usado como punto de corte, α). El concepto de valor-P y su interpretación correcta se verá más adelante, pero es básicamente la probabilidad de observar una diferencia tan grande como la que se observó incluso si las medias de población son idénticas (la hipótesis nula es verdadera o sea que no hay diferencia entre los grupos comparados).

Por otro lado, supongamos que estamos comparando dos medias con una prueba "*t* de Student". Asuma que los dos promedios realmente se diferencian por una cantidad determinada, y que se realizan muchos experimentos con el mismo tamaño de muestra. Cada experimento va a tener valores diferentes (por casualidad), y, por lo tanto, la prueba *t* de Student producirá resultados diferentes cada vez que se aplique a cada experimento [38].

En algunos experimentos, el valor-P será menor que α (por lo general se establece en 0,05), por lo que se dice que los resultados son estadísticamente significativos e indica la fortaleza de la evidencia. En otros experimentos, el valor-P será mayor que α por lo que se dirá que la evidencia sobre la diferencia no es estadísticamente significativa. Si realmente hay una diferencia (de un tamaño especificado) entre medias de los grupos, no se va a encontrar una evidencia estadísticamente significativa sobre la diferencia en cada uno de todos los posibles experimentos realizados en esos grupos [27]. El poder estadístico es la fracción de los experimentos que se esperan van a producir un valor-P estadísticamente significativo si es que hubiera un efecto del tratamiento. Si el diseño experimental tiene un alto poder o potencia, entonces hay una alta probabilidad de que el experimento encuentre un resultado estadísticamente significativo [58]; si el tratamiento realmente tiene el efecto esperado. Por otro lado, si no se encontraron diferencias estadísticamente significativas y la prueba estadística tiene un alto "poder" se puede confiar que los resultados son correctos [38].

Definición de "outlier" o datos inusuales

Cualquier valor que no sea común o un dato no esperado en los resultados de experimento para una variable medida es considerado un "*outlier*" (valor anómalo, atípico o inusual). Es decir, es un valor que está tan lejos del resto de los valores que pareciera provenir de otra población. Estos valores ocurren por diversas razones: ingreso no válido de datos, diversidad biológica, azar (en cualquier distribución, algunos valores por azar están alejados del resto), errores experimentales y, por último, un supuesto equivocado (si se asume que los datos provienen de una distribución gaussiana, se podría concluir que un valor extremo es un *outlier*, pero si la distribución es, por ejemplo, lognormal, los valores extremos son comunes y no *outliers* [29].

Errores comunes

A continuación, se presentan los errores de concepto más frecuentemente cometidos en el área de medicina veterinaria y humana que pueden ocurrir en diferentes etapas del proceso de investigación científica, desde la planificación de un estudio, a través de la realización de análisis apropiado de los datos hasta la presentación de los resultados e interpretación de los mismos.

Errores de concepto con las pruebas de hipótesis estadística

El primer error en esta área es creer que la prueba de hipótesis es una parte esencial de todos los análisis estadísticos. El punto principal de las pruebas de hipótesis es tomar decisiones. Se toma una decisión cuando los resultados son estadísticamente significativos y otra decisión cuando los resultados no son estadísticamente significativos. Esta situación es muy común en procesos de control de calidad, pero no así en investigación biomédica exploratoria.

En muchas situaciones de experimentos científicos no es necesario, y de hecho puede ser contraproducente, llegar a una conclusión de que un resultado es o no estadísticamente significativo. Los valores P y los intervalos de confianza pueden ayudar a evaluar y presentar evidencia científica sin siquiera usar la frase "estadísticamente significativo" [30]. La palabra significativo fue ligada a los valores bajos de P planteados por R. Fisher en forma totalmente deliberada, y quiso darle un significado cercano a la interpretación de la palabra en el lenguaje común [21]. Cuando se lee la literatura científica, no se debe permitir que la conclusión acerca de la significancia estadística impida entender lo que los resultados del experimento muestran [20, 55].

Para evitar el énfasis en la significancia estadística que se muestra con gráficos, tablas y asteriscos, cuando se lee un artículo científico, se deben ignorar inicialmente los enunciados sobre significancia estadística y enfocarse en el tamaño del efecto del tratamiento (diferencias, asociaciones, etc.) y su intervalo de confianza (IC). Aquí se debe enfatizar que los IC cuantifican precisión y que casi todos los resultados: proporciones, riesgos relativos, medias, diferencias entre medias, pendientes, constantes de tasas, etc., pueden y deben ser reportados con un intervalo de confianza [30, 59]. El segundo error es creer que, si un resultado es estadísticamente significativo, el efecto debe ser grande. La conclusión sobre la significancia de un resultado estadísticamente significativo debe aplicarse a la "fortaleza de la evidencia" y no al tamaño del efecto. Las pruebas de hipótesis estadística responden la pregunta ¿existe evidencia significativa sobre la diferencia encontrada? y no la pregunta ¿existe evidencia sobre la diferencia significativa? Solo basta recordar que, si el tamaño de muestra es grande, aún diferencias pequeñas e inconsecuentes serán estadísticamente significativas, pero no necesariamente científica, biológica o clínicamente significativas [52]. Por otro lado, debe tenerse presente que la ausencia de evidencia no es la evidencia de ausencia [1].

Desestimar la importancia de revisar previamente los supuestos estadísticos

Cada prueba estadística tiene sus propios supuestos. Por ejemplo, si consideramos una investigación cuyo interés es evaluar las diferencias entre cuatro grupos de tratamientos, se debiera escoger el análisis de varianza (Anova) de una vía y esta prueba estadística requiere cumplir con los siguientes tres supuestos: los datos de la muestra siguen una distribución normal o gaussiana, las observaciones son independientes entre sí y las varianzas entre los diferentes grupos son homogéneas. Para tal efecto, por ejemplo, existen pruebas estadísticas para evaluar la distribución normal de los datos como D'Agostino-Pearson (que detecta simultáneamente sesgo y curtosis), Shapiro-Wilk, etc. [38, 59].

Sin embargo, estos supuestos podrían ser ignorados o el investigador deliberadamente decide llevar a cabo esta prueba desestimando el no cumplimiento de los supuestos mencionados [49]. Este problema fue mencionado en un artículo a finales de los años 90 [26], donde se revisaron artículos de 17 revistas científicas indicando que la mayoría de los trabajos no verificaban el cumplimiento de los supuestos de las pruebas respectivas y desafortunadamente esta situación todavía continúa [25,38,42].

En el ejemplo anterior, el Anova compara los promedios de cada grupo, pero los promedios de los grupos estarán sesgados cuando la distribución de la variable no es gaussiana. Por lo tanto, los resultados no serían confiables, estarían sesgados o las conclusiones serían limitadas. En el caso de que no se cumplan estos supuestos, el investigador debe remediar la situación a través de la transformación de los datos para lo que la mayoría de paquetes estadísticos que existen actualmente poseen módulos de diferentes transformaciones como, por ejemplo, transformación logarítmica, transformación de raíz cuadrada, etc.[23]. Si la transformación no corrige esta desviación de los supuestos, se debiera aplicar pruebas no paramétricas o intentar otro tipo de distribuciones aplicando una regresión de Box-Cox para obtener el coeficiente más adecuado para la transformación [10]. Los análisis no paramétricos presentan generalmente un bajo poder estadístico y precisión (esto depende principalmente del tamaño de la muestra cómo se verá más adelante) si las comparamos con las pruebas paramétricas; pero no requieren los supuestos sobre la distribución de los valores en la población, basándose en algunos otros supuestos. Por ejemplo, al igual que las pruebas paramétricas, asumen que la muestra fue obtenida al azar de una población más grande y que cada valor fue recolectado en forma independiente [37].

Seleccionar arbitrariamente el tamaño de la muestra

Hay que recordar que existe una relación cercana entre error Tipo I, error Tipo II, tamaño de muestra y el poder de las pruebas de hipótesis [27,58]. Cuando el tamaño de muestra es pequeño, hay una mayor probabilidad de cometer el error Tipo II lo que resulta en un menor poder estadístico. En otras palabras, no se va a rechazar la hipótesis nula cuando se debiera rechazar (existen diferencias entre los grupos, pero no se pudieron detectar). Por lo anterior, tiene sentido realizar *a priori* un análisis de poder estadístico para asegurarse de que la prueba usada tiene suficiente poder para detectar significativamente un efecto o diferencia con un tamaño de muestra adecuado; no obstante, hay momentos en los cuales es necesario realizar un análisis *post-hoc* de la potencia estadística cuando no se han obtenido resultados similares a los valores usados *a priori* para determinar el tamaño de la muestra. Afortunadamente, el cálculo del tamaño de la muestra requerida y los análisis de poder están presentes en la mayoría de los paquetes estadísticos y de forma gratuita en Internet [45].

Eliminar datos perdidos sin justificación

En cualquier diseño experimental hay varias razones por las cuales se pueden perder datos. Debido a que los datos perdidos pueden potencialmente introducir sesgos en los resultados, el investigador debe decidir qué hacer con estos datos.

Es común ver que un(a) investigador(a) excluya las unidades experimentales que incluyen estos datos perdidos, pero esto también introduce sesgo en el análisis. Una pregunta importante antes de eliminar los datos perdidos sería ¿los individuos o unidades experimentales que participaron del experimento son diferentes de aquéllas que no participaron o que fallecieron? Si no son diferentes, la eliminación de datos perdidos no causaría una distorsión de los resultados. Sin embargo, esta distorsión puede ser seria cuando los datos eliminados son diferentes de los que se mantienen

en el experimento (se pierde información y esto introduce sesgo en los resultados). Otra pregunta relevante es establecer si la cantidad de datos perdidos es baja o alta al compararla con el tamaño total de la muestra. Si es una pequeña fracción del tamaño de la muestra, el eliminar los datos perdidos no tendrá mayores consecuencias, de lo contrario habrá un impacto en los resultados y conclusiones.

En ocasiones se puede intentar hacer un análisis de sensibilidad con una predicción de los datos faltantes a partir de los valores obtenidos en el estudio. Este análisis requiere la realización de una regresión inicial para predecir los valores faltantes. Posteriormente, se hace el análisis incluyendo los valores predichos y el análisis excluyendo estos valores. El investigador puede decidir sobre la conveniencia o no de incluir los datos faltantes una vez realizados ambos análisis. Este procedimiento debe ser adecuadamente descrito en la sección de materiales y métodos para que el lector esté informado sobre este procedimiento[38].

Presentación inapropiada de los datos

Una buena investigación merece ser bien presentada y esta presentación es tan importante como la colección y análisis de los datos [11, 15]. Por lo tanto, cuando se están describiendo estadísticamente o presentando los datos de investigación, se debería tener cuidado en usar medidas estadísticas adecuadas de tendencia central y dispersión. Por ejemplo, si se usan promedios aritméticos y desviaciones estándar, se asume que los datos presentan, al menos aproximadamente, una distribución normal y no están sesgados. De no ser así, estas medidas no pueden ser usadas para describir los datos. Para datos sesgados, comúnmente encontrados en experimentos biológicos y médicos, es más apropiado usar la mediana, cuartiles o rangos (aunque se debe tener en consideración que hay ciertas limitantes, como por ejemplo que el rango es sensible a los valores inusuales o "*outliers*"). Asimismo, si se están aplicando pruebas no paramétricas, se debe evitar presentar promedios y desviación estándar ya que estos parámetros, por definición, no son evaluados en pruebas no paramétricas y no tendría sentido describir los datos de esa manera [38]. Por lo tanto, aquí lo que corresponde es usar medianas, rangos y rangos intercuartílicos.

Por otro lado, no es suficiente presentar medidas de tendencia central sin acompañarlas por medidas de dispersión [12, 33, 57]. Asimismo, el error estándar del promedio (EEM), aunque comúnmente y erróneamente usado para descripción estadística (tal vez porque presenta los datos con menos variabilidad ya que por fórmula es la desviación estándar dividido por la raíz cuadrada del tamaño de la muestra), no es un estadístico descriptivo sino más bien un método inferencial usado para estimación estadística [14, 39]. La desviación estándar (DE) se usa para describir dispersión de datos y el EEM cuantifica cuán preciso es el promedio de la muestra con respecto al promedio de la población del cual se tomó la muestra. Con respecto al uso del EEM, algunos investigadores consideran que cuando se comparan promedios, mostrar el EEM da una idea acerca de si las diferencias son o no estadísticamente significativas. Sin embargo, al describir los datos, se necesita la DE para dar al lector una idea de la dispersión entre los sujetos experimentales, ya que cuando se comparan promedios las DE entregan información sobre la magnitud de la diferencia entre los promedios lo que se conoce como "tamaño del efecto" [38]. Asimismo, es importante visualizar la DE cuando hay varios grupos, porque si las DE difieren mucho entre los grupos, el investigador podría tener que usar la transformación logarítmica u otra antes de calcular el intervalo de confianza o los valores P. Finalmente, si se considera importante mostrar la significancia en los gráficos se debe mostrar el valor-P exacto del estadígrafo respectivo (esto es más exacto que mostrar el EEM).

En general, se puede decir que la aplicación apropiada de técnicas de estimación estadística puede entregar, si se presentan adecuadamente, más información a los lectores del estudio. Por ejemplo, si se usan estimaciones estadísticas para comparar grupos, los intervalos de confianza debieran ser presentados para las diferencias entre grupos en lugar de intervalo de confianza para cada grupo. Finalmente, los valores P debieran ser reportados de manera exacta en lugar de mencionarlo considerando el umbral ampliamente usado en la literatura científica (P = NS por no significativo; P < 0,05; P > 0,05). De hecho, indicar el valor-P específico o exacto obtenido en cada análisis es necesario ya que no es lo mismo, un valor-P < 0,049 que un valor-P < 0,023, ambos considerados P < 0,05. En la práctica actual en publicaciones científicas, esto es sugerido y aplicado más que sólo indicar que el valor-P correspondió al nivel de significancia preestablecido ya que se ha generado una peculiar prevalencia de valores ligeramente por debajo del P<0,05, lo que ha puesto en cuestionamiento la validez de los resultados en determinadas publicaciones científicas de diferentes áreas [31, 38].

En el siguiente ejemplo, los datos son presentados de diferente forma, variando de menor a mayor información, siendo la última la más apropiada: "*El efecto del fármaco fue significativo. El efecto del fármaco para disminuir la presión arterial fue significativo (P < 0,05). La presión diastólica en el grupo que recibió el fármaco disminuyó de 110 a 92 mmHg (P = 0,02). La presión diastólica en el grupo que recibió el fármaco disminuyó en promedio 18 mmHg (IC 95%: 2-34 mmHg), de 110 a 92 mmHg (P = 0,02)*". Es decir, lo ideal es presentar el valor exacto de P y el intervalo de confianza [9]. Es mejor cuantificar la asociación más que entregar solamente el valor-P, ya que un intervalo amplio indica considerable incerteza aunque el valor-P sea bajo.

Implicar relación de causalidad de una correlación

El coeficiente de correlación es una medida de relación entre dos o más variables. Sin embargo, esta relación no implica una relación de causa y efecto, aunque en algunas investigaciones se pretenda establecer este tipo de dependencia. Por ejemplo, Messerli en el año 2012 [34] se preguntó por qué algunos países producían más ganadores de premios Nobel que otros y para responder esta pregunta se graficaron los datos. En el eje vertical (Y) se colocaron el número total de premios Nobel por número de ciudadanos en los países. En el eje horizontal (X) se colocó el consumo de chocolate en un año reciente (diferentes años para diferentes países, basado en la disponibilidad de los datos). Ambos valores X e Y fueron estandarizados a la población actual de cada país. La correlación fue extremadamente fuerte (r = 0,79). Por supuesto que estos datos no prueban que el comer chocolate ayuda a que la gente gane los premios Nobel. Tampoco prueba que el incrementar las importaciones de chocolate en un país determinado va a aumentar las posibilidades de ganar premios Nobel en ese país.

Cuando dos variables están correlacionadas o asociadas, existe la posibilidad de que los cambios en una variable causen cambios en la otra variable, pero también es probable que ambas variables estén relacionadas con una tercera variable que las influencia a ambas. Hay que considerar la siguiente pregunta: ¿por qué dos variables pueden correlacionar tan bien? Al respecto hay 5 posibles explicaciones [38, 59]. Tomemos como ejemplo hipotético la relación entre sensibilidad de insulina (x_2) y el contenido de lípidos (x_1) en la membrana celular muscular. Las explicaciones, en este ejemplo, pueden ser: **a**. La variable x_1 determina la variable x_2; **b**. La variable x_2 de alguna manera afecta la variable x_1; **c**. Ambas variables están bajo el control de algún otro factor, tal vez una hormona; **d**. Tanto la variable x_1, x_2 u otros factores forman parte

de una compleja red bioquímica, molecular o fisiológica, y tal vez con componentes de retroalimentación positiva o negativa. En este caso la correlación observada es solo una consideración de una más compleja serie de relaciones; y, por último, **e**. Las dos variables no se correlacionan para nada en la población, y la correlación observada en la muestra es una coincidencia o producto del azar[37].

Confundir significancia estadística con importancia práctica

Tal como se mencionó al inicio de esta revisión, la significancia estadística por sí sola no indica la magnitud del efecto o cuán importante es el tamaño del efecto en relación con la importancia clínica o económica. En este contexto, la no significancia estadística puede tener una importancia práctica. Por ejemplo, un(a) investigador(a) puede encontrar que un nuevo tratamiento farmacológico produce un mejoramiento clínico significativo, pero no produce resultados estadísticamente significativos. Por otro lado, una significancia estadística puede no tener importancia práctica. Es decir, significancia estadística e importancia práctica no necesariamente están relacionadas [27]. El término significante o significativo tiene un significado específico en estadística. Significa que, por simple azar, una diferencia (o una asociación o correlación, etc.) tan grande o más grande que lo observado en el experimento ocurrirá en menos del 5% (u otro valor preseleccionado) de las veces que se repita un experimento. Es decir, un efecto pequeño puede ser estadísticamente significativo y sin embargo ser trivial desde el punto de vista clínico o científico. A continuación, un ejemplo [52]: El uso de un fármaco para el tratamiento de rinitis alérgicas reduce los síntomas y los resultados son estadísticamente significativos. Sin embargo, el fármaco reduce los síntomas de alergia solamente en un 7% por lo que no es clínicamente ni comercialmente útil.

Debe recordarse que en este contexto la ausencia de evidencia no es evidencia de ausencia, lo que no es nuevo en la interpretación de la significancia de los resultados en la literatura médica [2]. La verdad con respecto a los estudios que han probado efectos "negativos" de una intervención particular merecen la realización de otros estudios que ayuden a colectar evidencia de un real "no efecto"; pero, la forma de enunciar las conclusiones o los resultados puede dar la impresión de tener la respuesta definitiva con respecto a la intervención en cuestión [1].

Considerar que una diferencia no significativa significa que el efecto del tratamiento está ausente

Un valor-P elevado significa que una diferencia (o una asociación) tan grande como la que se observó en el experimento va a pasar frecuentemente como resultado del azar (y no por el efecto del tratamiento). Pero esto no significa que la hipótesis nula de no diferencia (H_0) sea verdadera [56]. Podemos cometer el error de tipo II (hay diferencias y fallamos en detectarlas). De hecho, en la literatura estadística en lugar de decir que se "acepta" la hipótesis nula, se habla de "no rechazar" la hipótesis nula (el experimento no es concluyente o no hay suficiente evidencia para rechazarla) porque decidir no rechazar la H_0 no es lo mismo que creer que la H_0 es definitivamente verdadera (recordar que "la ausencia de evidencia no significa evidencia de ausencia" [2]).

Como corolario a este error común en la interpretación de los resultados, podemos mencionar una frase acuñada por el cosmólogo inglés Martin Rees y ampliamente discutida por el astrónomo Carl Sagan cuando en una de sus publicaciones señala que cualquier cosa que no haya sido probada falsa es cierta y viceversa, apoyándose en la frase: "La ausencia de evidencia no es evidencia de ausencia".

Aplicar pruebas estadísticas inapropiadas

Aparte de considerar siempre los supuestos de cada prueba estadística, es necesario saber de antemano cuál es la prueba estadística apropiada para el análisis del tipo de datos que se tiene [51]. Para decidir cuál es la prueba adecuada, se deben considerar dos aspectos generales: **a**. Cuál es el objetivo del investigador(a) como, por ejemplo, comparar un grupo con un valor hipotético, comparar dos grupos independientes versus comparar dos grupos pareados, comparar tres o más grupos no relacionados versus comparar tres o más grupos relacionados, cuantificar la asociación entre dos variables, predecir el valor de una variable, etc.; y **b**. El tipo de variable que vamos a medir o analizar. Por ejemplo, si es cuantitativa o cualitativa (o categórica) y su escala de medida (las variables difieren en "cuán bien" pueden medirse, es decir, cuánta información medible se puede proporcionar: nominal, ordinal, de intervalo, de razón o proporción). Asimismo, si es una medida de una población gaussiana o no, si es binomial (dos resultados posibles), etc. Por ejemplo, si se está interesado(a) en examinar la relación entre dos variables categóricas como género y opinión (binominal, en contra o a favor) se debería aplicar la prueba de chi-cuadrado de independencia. Sin embargo, se debiera aplicar un Anova de una vía si la opinión es medida en un nivel continúo usando la escala de Likert).

Pese a las consideraciones anteriores, pruebas estadísticas simples como la prueba de *t* de Student y la de chi-cuadrado, son frecuentemente mal utilizadas en la investigación ya que no se toma en consideración la necesidad de seleccionar la versión correcta de la prueba ya que existen varias formas de estas [37, 59]. A su vez, si la frecuencia esperada presenta un valor menor que 5, la prueba de chi-cuadrado no debiera usarse ya que su aproximación en estas circunstancias no es confiable. Si el tamaño de la muestra es pequeño, la corrección por continuidad de Yates debiera usarse o mejor aún, se debe aplicar la prueba exacta de Fisher (37). Por otro lado, si un estudio requiere pruebas múltiples, es importante considerar la tasa de resultados positivos falsos y el aumento de la probabilidad de cometer el error Tipo I al no aplicar correcciones adecuadas de comparación múltiple [13, 24, 35]. Especialmente importante es el reconocimiento que al comparar más de dos grupos requiere el uso de Anova paramétrica o su equivalente no paramétrico (prueba de Kruskal-Wallis) y no una prueba *t* de Student ya que la aplicación repetida de la prueba para evaluar dos grupos aumenta el riesgo de generar resultados positivos falsos [38]. Si este es el caso, debe hacerse un ajuste del valor-P según el número de comparaciones que se hayan hecho en forma independiente (ajuste o corrección de Bonferroni [38]).

A manera de ejemplo, consideremos que tres muestras fueron tomadas de una misma población. Al realizar las tres posibles pruebas *t* de Student de dos muestras con un $\alpha = 0{,}05$, la probabilidad de concluir erróneamente que dos de los promedios son diferentes es 14% (considerablemente mayor que α). Entre más promedios estén involucrados mayor es la probabilidad de error. Esto es porque para cada prueba *t* de dos muestras ejecutada a un nivel de significancia de 5%, hay una probabilidad de 95% que correctamente concluyamos que no se rechaza la hipótesis estadística que enuncia que no hay diferencia entre los dos grupos. Para un grupo de tres hipótesis (H_0: $\mu 1 = \mu 2$, H_0: $\mu 1 = \mu 3$, H_0: $\mu 2 = \mu 3$; en lugar de H_0: $\mu 1 = \mu 2 = \mu 3$) la probabilidad de no rechazarlas en forma correcta es solo $0{,}95^3 = 0{,}86$. Esto significa que la probabilidad de rechazar incorrectamente al menos una de las hipótesis estadísticas es 1 - 0,86 = 0,14 aunque las muestras realmente no presenten diferencias estadísticamente significativas. En este mismo aspecto, Loannidis argumenta [24] que la mayoría de los resultados publicados como estadísticamente significativos son falsos (independientemente del prestigio de la revista científica). La investigación reciente

confirma esta afirmación [50,51]. Como ejemplo, dos compañías farmacéuticas han tabulado información sistemática para reproducir hallazgos preclínicos publicados y se ha descubierto que solo una pequeña fracción de los hallazgos publicados son reproducibles [7, 46].

Otro error común en este aspecto, es realizar un análisis paramétrico aun cuando los datos se desvían de los supuestos de esta prueba como la distribución normal de la variable. Todas las pruebas paramétricas asumen que la variable presenta una distribución normal por lo que los resultados no son confiables si es que no se cumple este supuesto. Por ejemplo, el uso de pruebas paramétricas con datos que no están normalmente distribuidos puede generar resultados estadísticamente significativos, pero es muy probable que se esté cometiendo el error Tipo I (encontrar diferencias donde no las hay).

Como las pruebas no paramétricas tienen un mayor poder estadístico cuando no se cumple este supuesto, es recomendable su aplicación en estos casos. Sin embargo, es necesario considerar lo siguiente. Los métodos no paramétricos no se basan en el supuesto de distribución de la población (también conocidos como métodos de distribución libre), su principio se basa en ordenar los datos en una escala jerárquica y se analiza solamente la posición o ubicación del dato con respecto a los demás, ignorando los valores reales. Esto asegura que el análisis no se vea afectado por medidas inusuales ("*outliers*") y no se asume ninguna distribución en particular. La ventaja de las pruebas no paramétricas es clara. No requieren el supuesto de una muestra proveniente de una población gaussiana por lo que puede ser usada cuando la validez de este supuesto está en duda [38]. Cuando el supuesto es falso estas pruebas tienen más poder estadístico que las pruebas paramétricas para detectar diferencias [58].

Por lo anterior, podría decidirse usar las pruebas no paramétricas como alternativa a las paramétricas al no cumplir con los supuestos de normalidad. Entonces, ¿por qué no siempre usar estas pruebas en estos casos? Las pruebas no paramétricas presentan menos poder cuando los datos siguen una distribución gaussiana. Por otro lado, cuando los datos provienen de poblaciones no gaussianas, las pruebas no paramétricas pueden ser tan poderosas como las pruebas paramétricas [28, 48]. La desventaja principal se debe a que estas pruebas solo consideran los rangos y no los datos reales, esencialmente se elimina parte de la información.

Si realmente existe una diferencia entre poblaciones gaussianas, el valor-P probablemente va a ser más elevado con una prueba no paramétrica (no encontrando diferencias entre grupos cuando efectivamente si existen diferencias). Ahora bien, ¿cuánto más elevado sería el valor-P en estas circunstancias? La respuesta es: depende del tamaño de la muestra. Con muestras grandes, las pruebas no paramétricas tienen tanto poder como las pruebas paramétricas cuando los datos son muestreados de una población gaussiana. Esto se evalúa a través de un valor denominado *eficiencia relativa asintótica* (ERA[41]).

Por ejemplo, con muestras grandes obtenidas de una población gaussiana, la ERA de la prueba de U de Mann-Whitney (el análogo no paramétrico de la prueba *t* de Student para muestras independientes) es 95%. Esto significa que el poder de esta prueba es igual al poder de una prueba *t* de Student para muestras independientes con el 95% de los datos. Las otras pruebas no paramétricas tienen similares eficiencias relativas asintóticas. Con muestras pequeñas provenientes de poblaciones gaussianas, las pruebas no paramétricas tienen mucho menos poder que las

pruebas paramétricas [28]. De hecho, con muestras muy pequeñas, las pruebas no paramétricas tienen poder estadístico igual a cero. Solo a manera de ejemplo, con 7 o menos valores, la prueba de Mann-Whitney usualmente reporta valores P mayor que 0,05 (no encuentra diferencias). A su vez, con 5 o menos pares de datos, la prueba de Wilcoxon (el equivalente no paramétrico de la prueba *t* de Student para muestras pareadas) usualmente reporta valores P mayor que 0,05. En este sentido, la gran incógnita es la decisión de cuándo aplicar pruebas no paramétricas.

Inicialmente es necesario considerar que la decisión sobre cuáles pruebas usar no es tan automática. Lo primero es realizar una prueba de normalidad (D'Agostino-Pearson, Shapiro-Wilk, u otra) y evaluar el potencial efecto de las observaciones inusuales, ya que muchas veces la prueba de normalidad está afectada por valores extremos [38]. Si los datos pasan esta prueba, se usa una prueba paramétrica. Si los datos fallan la prueba de normalidad, se tienen dos opciones: transformar los datos y aplicar sobre estos la prueba paramétrica o aplicar a los datos originales una prueba no paramétrica. Hay que considerar lo siguiente: Si el desvío del supuesto de normalidad es bajo (valores cercanos o ligeramente por debajo del P<0,05) y los datos no pasan la prueba de normalidad, se puede decidir no realizar ninguna modificación [37].

Las pruebas estadísticas paramétricas tienden a ser bastante "robustas" con violaciones moderadas del supuesto de distribución gaussiana [59]. A su vez, en ocasiones los datos fallan la prueba de normalidad porque efectivamente no presentan una distribución gaussiana y en estos casos, transformar los datos a logaritmos, o a sus valores recíprocos u otro tipo de transformación, convierte una distribución no gaussiana en gaussiana. En medicina humana[5] y veterinaria[44], como ejemplo esto ocurre comúnmente con mediciones relacionadas con potencias de fármacos (EC_{50}), concentración sérica de metabolitos, concentración sérica de sustancias tóxicas, período de incubación de una enfermedad, títulos de anticuerpo a un virus, el tiempo de supervivencia en pacientes con cáncer o después de una intervención quirúrgica u otras medidas (todos estos presentan una distribución denominada log-normal y se deben transformar los datos a su logaritmo [16, 29].

La decisión de cuándo usar una prueba paramétrica o no paramétrica cobra importancia con muestras pequeñas por el bajo poder de las pruebas no paramétricas. Pero, con este tipo de muestras las pruebas de normalidad igualmente tienen bajo poder [58]. Entonces, el problema principal es con muestras pequeñas por lo que siempre se recomienda obtener un tamaño de muestra aplicando el método apropiado de muestreo estadístico [23]. En resumen, un gran número de datos no presentan problemas. Usualmente es fácil detectar si los datos provienen de una distribución gaussiana (histograma, prueba de normalidad, etc.) pero no importa mucho porque en estas circunstancias las pruebas no paramétricas son "potentes" y las paramétricas son "robustas" (no se ven afectadas significativamente por la presencia de *outliers*). El inconveniente es cuán grande debe ser la muestra porque depende de la naturaleza de la distribución no gaussiana en particular. Una regla general es que a menos que la distribución de la población sea bastante atípica, se puede aplicar una prueba paramétrica si hay al menos 24 valores en cada grupo [37, 43, 48].

Eliminar outliers de manera subjetiva o influenciado por los resultados que se buscan
Para evitar este error, lo primero es revisar las posibles razones que generaron estos valores y que fueron mencionados al inicio de esta revisión. Después de haber abordado las razones anteriores, existen dos posibilidades [36]: **a**. El o los valor(es) extremo(s) provienen de la misma distribución que los otros valores y simplemente

son más elevados (o más bajos) que el resto. En este caso, el valor no debe ser tratado de manera especial ni debe ser eliminado; o **b**. El valor extremo fue el resultado de un error. Esto podría ser debido a un mal procesamiento de la muestra, un alza de voltaje, agujeros en los filtros, error en el registro de los datos, etc. Debido a que incluir un valor erróneo en el análisis dará resultados no válidos, estos se deben remover e indicarlo claramente en la sección que corresponda en el manuscrito. El problema es que nunca se está seguro de cuál de estas dos posibilidades es la correcta. ¿Error o azar? Ningún cálculo matemático puede decirnos con certeza si un "*outlier*" viene de una misma o diferente población. Sin embargo, una prueba estadística para observaciones inusuales (por ejemplo, el método de Grubbs o Rout) puede responder la siguiente pregunta [36]: Si los valores fueron realmente obtenidos de una distribución gaussiana, ¿cuál es la probabilidad de que se encuentre un valor tan alejado del resto como el que se observó en el experimento? Si el valor-P es bajo, se concluirá que el "*outlier*" no es de la misma distribución que los otros valores y se tiene justificación para eliminarlo del análisis. Si el valor-P es elevado, no se tiene evidencia de que el valor extremo provenga de una distribución diferente al resto de los valores.

Lo anterior no prueba que el valor extremo de hecho provino de la misma distribución que el resto de los valores. Lo único que se puede decir es que no hay evidencia que el valor proviene de una distribución diferente por lo que no puede ser eliminado de los datos. Sin embargo, varios autores [58,59] consideran que más que eliminar "*outliers*", una alternativa es usar métodos estadísticos diseñados para que los "*outliers*" tengan poco efecto en el resultado. Estos métodos se denominan "robustos" [59]. Son métodos existentes que soportan pequeñas desviaciones a los supuestos que lo fundamentan. Los métodos robustos en general son métodos paramétricos que no sufren grandes cambios en su poder estadístico cuando los supuestos varían en razón de múltiples causas, entre ellas, la presencia de "*outliers*" en la muestra. No se requiere decidir cuándo eliminar un "*outlier*" porque el método es diseñado para que el valor atípico tenga poca influencia en el resultado. Algunas aplicaciones en esta área son el uso de modelaje lineal, métodos robustos en el análisis de varianza y algunos estadígrafos. Por ejemplo, el estadígrafo robusto más simple es la mediana. Si un valor es muy alto o muy bajo, el valor de la mediana no va a cambiar, mientras que el valor del promedio se puede ver afectado por observaciones extremas.

Finalmente, varios científicos se hacen la pregunta: ¿Es legítimo remover un "*outlier*"? Es importante mencionar que no se engaña al sistema cuando la decisión de remover o no un "*outlier*" se basa en reglas y métodos establecidos antes de que los datos se colecten y cuando los métodos usados y el número y valor de los "*outliers*" removidos se reportan en la publicación respectiva. Cuando un "*outlier*" es eliminado, los análisis se ejecutan como si el valor no hubiera existido; pero, remover un "*outlier*" de un análisis no significa eliminarlo del registro de resultados de laboratorio. Todo lo contrario, el valor del "*outlier*", el criterio usado para identificarlo y la razón para excluirlo deben ser registrados [6].

Interpretar equivocadamente el concepto de valor-P

Además del error mencionado anteriormente (considerar que una diferencia no significativa – valor-P elevado- significa que el efecto del tratamiento está ausente), otro error es el conocido enunciado *el valor-P es la probabilidad que la hipótesis nula es verdadera*. El valor-P se calcula asumiendo que la hipótesis nula es verdadera. Por lo tanto, no puede ser la probabilidad que sea verdadera. Por ejemplo, si existe un valor-P de 0,03, existe una probabilidad del 3% de observar una diferencia tan

grande como la que se observó si las dos medias de población fueran idénticas (y la hipótesis nula es verdadera). Es decir, El valor-P mide la fuerza de la evidencia en contra de la hipótesis nula. Entre más bajo el valor-P, más fuerte la evidencia en contra de la H_0 [21].

El valor-P indica cuán raramente uno observaría una diferencia tan grande o más grande que la observada en el experimento si la hipótesis nula fuera verdadera [56]. La pregunta que el científico debe responder es si el resultado es tan inverosímil que H_0 debe ser desechada. Finalmente, no se debe interpretar el valor-P aisladamente. La conclusión depende del contexto del experimento ya que la interpretación de resultados requiere tanto sentido común como buen juicio.

Conclusión

La estadística es una herramienta muy importante. Sin embargo, errores estadísticos o interpretaciones erróneas pueden ocurrir en varias etapas del estudio o proyecto. Esto conlleva a que no se produzcan resultados confiables y que las conclusiones sean incorrectas cuando las diferentes fuentes de error no son cuidadosamente consideradas con anticipación. Algunos de los errores más comúnmente cometidos incluyen, pero no limitados a, desestimar la importancia de revisar previamente los supuestos estadísticos, seleccionar arbitrariamente el tamaño de la muestra, eliminar datos perdidos sin justificación, presentar equivocadamente los datos, implicar relación de causalidad de una correlación, confundir significancia estadística con importancia práctica, económica o clínica, considerar que una diferencia no significativa significa que el efecto está ausente o científicamente irrelevante, aplicar pruebas estadísticas inapropiadas, en el reporte de los resultados no enunciar el valor exacto de P ni colocar el intervalo de confianza respectivo, y finalmente, interpretar equivocadamente el concepto de valor-P. Por todo lo anterior, es muy importante comprender los posibles errores al aplicar la bioestadística de tal manera de poder abordar de manera crítica la lectura científica y la aplicación apropiada de técnicas estadísticas en la investigación biomédica.

Referencias

1. Alderson, P. Absence of evidence is not evidence of absence. *Brit Med J.* 2004; 328: 476-477.

2. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *Brit Med J* (Clinical Research Ed.) 1995; 311: 485.

3. Altman DG, Goodman SN, Schroter S. How statistical expertise is used in medical research. *JAMA.* 2002; 287: 2817-2820.

4. Altman DG. Statistics and ethics in medical research. Improving the quality of statistics in medical journals. *Brit Med J.* 1981; 282: 44-47.

5. Altman DG. Statistics in medical journals: some recent trends. *Stat Med.* 2000; 19: 3275–3289.

6. Barnett V, Lewis T. Outliers in Statistical Data. 3rd ed., John Wiley and sons, New York. 1994.

7. Begley CGC, Ellis LML. Drug development: Raise standards for preclinical cancer research. *Nature.* 2012; 483: 531-533.

8. Bland JMJ, Altman DGD. Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials.* 2011; 12: 264-266.

9. Boos DD, Stefanski LA. P-value precision and reproducibility. *Am Stat.* 2011; 65: 213-221.

10. Box, GEP, Cox, DGD. An analysis of transformations. *J Royal Stat Soc, Series B.* 1964; 26: 211-252.

11. Cooper RJ, Schriger DL, Close RJH. Graphical literacy: the quality of graphs in a large-circulation journal. *Ann Emerg Med.* 2002; 40: 317-322.

12. Cumming G, Fidler F, Vaux DL. Error bars in experimental biology. *J Cell Biol.* 2007; 177: 7-11.

13. Dar R, Serlin R, Omer H. Misuse of Statistical Tests in three Decades of Psychotherapy Research. *J Consult Clin Psychol.* 1994; 62: 75-82.

14. Davies HT. Describing and estimating: use and abuse of standard deviations and standard errors. *Hosp Med*. 1998; 59: 327-328.

15. Evans M. Presentation of manuscripts for publication in the British Journal of Surgery. *Br J Surg*. 1989; 76: 1311-1314.

16. Frazier E P, Schneider T, Michel M C. Effects of gender, age, and hypertension on beta-adrenergic receptor function in rat urinary bladder. *Arch Pharmacology.* 2006; 373: 300-309.

17. García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. *BMC Med Res Method.* 2004; 4: 13-17.

18. Gardenier JS, Resnik DB. The misuse of statistics: Concepts, tools, and a research agenda. *Account Res.* 2002; 9: 65-74.

19. Gardner MJ, Bond J. An exploratory study of statistical assessment of papers published in the British Medical Journal. *JAMA.* 1990; 263: 1355-1357.

20. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. *Am Stat.* 2006; 60: 328-331.

21. Goodman S. A dirty dozen: Twelve p-value misconceptions. *Sem Hematol.* 2008; 45: 135-140.

22. Gore SM, Jones G, Thompson SG. The Lancet's statistical review process: Areas for improvement by authors. *Lancet.* 1992; 340:100-102.

23. Hawkes J, Marsh W. *Discovering Statistics*. 2nd ed., Hawkes Publishing Inc., Charleston, SC. 2004.

24. Ioannidis JPA. Why most published research findings are false. *PLoS Med.* 2005; 2(8): 696-701.

25. Kanter MH, Taylor JR. Accuracy of statistical methods in Transfusion: a review of articles from July/August 1992 through June 1993. *Transfusion.* 1994; 34: 697-701.

26. Keselman HJ, Huberty CJ, Lix LM, Olejnik S, Cribbie RA, Donahue B, *et al*. Statistical practices of educational researchers: An analysis of their Anova, Manova and Ancova analyses. *Rev Educ Res.* 1998. 68: 350-386.

27. Kim A, Mallory C. *Statistics for Evidence-based Practice in Nursing*. Jones & Bartlett Learning, Burlington, MA. 2014.

28. Lehman E. *Nonparametrics; Statistical Methods Based on Ranks*. Springer, New York. 2007.

29. Limpert E, Stahel WA, Abbt M. Log-normal distribution across the sciences: Keys and clues. *Biosciences.* 2001; 51: 341-352.

30. Marshall SW. Testing with confidence: The use (and misuse) of confidence intervals in biomedical research. J. Sci Med Sport. 2004; 7: 135-137.

31. Masicampo EJ, Lalande DR. A peculiar prevalence of P values just below 0.05. *Q J Exp Psychol.* 2012; 65: 2271-2279.

32. McCance I. Assessment of statistical procedures used in papers in the Australian Veterinary Journal. *Aust Vet J.* 1995: 72: 322-328.

33. McGuigan SM. The use of statistics in the British Journal of Psychiatry. *Br J Psych*. 1995; 167: 683–688.

34. Messerli FH. Chocolate consumption, cognitive function, and Nobel laureates. *New Engl J Med*. 2012; 367: 1562-1564.

35. Moreira ED, Stein Z, Susser E. Reporting on methods of sub-group analysis in clinical trials: a survey of four scientific journals. *Brazilian J Med Biol Res.* 2001; 34:1441-1446.

36. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression-A new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics.* 2006; 7: 123-143.

37. Motulsky, HJ. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. 2nd ed., Oxford University Press, New York. 2010.

38. Motulsky, HJ. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. 3rd ed., Oxford University Press, New York. 2014.

39. Nagele P. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *Br J Anaesth.* 2001; 90: 514-516.

40. Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. Erroneous analyses of interactions in neuroscience: A problem of significance. Nature Neuroscience. 2011; 14: 1105-1107.

41. Nikitin Y. Asymptotic relative efficiency in testing. *International Encyclopedia of Statistical Sciences*. Springer, New York. 2010.

42. Olsen CH. Review of the use of statistics in infection and immunity. *Infect Immun.* 2003; 71: 6689-6692.

43. Parker RA, Berman, NG. Sample size: More than calculations. *Am Stat.* 2003; 57: 166-170.

44. Petrie A, Watson, P. *Statistics for Veterinary and Animal Science*. 2nd ed., Blackwell Publishing, Oxford. 2006.

45. Power, Sample Size and Experimental Design Calculations. Washington, DC. Actualizado 18 julio 2015; citado 30 julio 2015. Disponible en: http://statpages.org.

46. Prinz FF, Schlange TT, Asadullah KK. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Rev Drug Discovery.* 2011; 10: 712.

47. Schervish MJ. P values: What they are and what they are not? *Am Stat.* 1996; 50: 203-206.

48. Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures.* 4th ed., Chapman & Hall/CRC, New York. 2007.

49. Shott S. Statistics simplified. Detecting statistical errors in veterinary research. *JAVMA*. 2011; 237: 305-308.

50. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci.* 2011; 22: 1359-1366.

51. Academy of Medical Sciences. Symposium report "Reproducibility and reliability of biomedical research: improving research practice". BBSRC, MRC & Wellcome Trust 2015.

52. Spector R, Vesell ES. Pharmacology and statistics: Recommendations to strengthen a productive partnership. *Pharmacology.* 2006; 77: 85-92.

53. Sterne JAC, Smith GD. Sifting the evidence-what's wrong with significance tests? *Brit Med J.* 2001; 322: 226-231.

54. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer, H. Statistical errors in medical research: A review of common pitfalls. *Swiss Med Wkly.* 2007; 137: 44-49.

55. Svensson S, Menkes DB, Lexchin J. Surrogate outcomes in clinical trials: A cautionary tale. *JAMA Intern Med.* 2013; 173: 611-612.

56. Vickers A. *What is a P-value Anyway?* Addison-Wesley, Pearson Education Inc. Boston. 2010.

57. White SJ. Statistical errors in papers in the British Journal of Psychiatry. *Br J Psych.* 1979; 135: 336-342.

58. Wilcox RR. Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy. 2nd ed., Springer, New York. 2010.

59. Zar JH. *Biostatistical Analysis*. 5th ed., Prentice Hall, New Jersey. 2010.