ISSN 2311-9659

CRITERIOS PARA LA ELECCIÓN DE UNA PRUEBA ESTADÍSTICA INFERENCIAL UNIVARIANTE

A. Chocó
Químico Farmacéutico - Epidemiólogo
Clínica de Enfermedades Infecciosas, Hospital Roosevelt
"If you only have a hammer, you tend to see every problem as a nail"
The Psychology of Science, Abraham Maslow, 1966.

Resumen

Es indiscutible la importancia de la Estadística en la investigación científica, como el medio que permite obtener información válida a partir de los datos; sin embargo, ante la disponibilidad de una gran cantidad de enfoques y pruebas estadísticas, resulta prioritario conocer los criterios para elegir las herramientas idóneas; entre ellos pueden enumerarse: el nivel de medida de cada variable incluida en el análisis, la pregunta de investigación o la hipótesis, el diseño de la investigación, la distribución de probabilidad de la variable independiente, el tipo de muestras utilizadas y el tamaño y diseño de muestra utilizados. En esta revisión bibliográfica se exponen los criterios para elegir una prueba estadística inferencial univariante, entre las alternativas paramétricas y no paramétricas, agrupándolos estos en tres categorías: la pregunta de investigación, la estructura de los datos y el diseño de la investigación. Respecto a la pregunta de investigación hay que distinguir entre las descriptivas, analíticas y experimentales. La estructura de los datos se refiere a las características de las variables como lo son, la escala de medición, la distribución de probabilidad, la independencia en la recolección de datos y la homocedasticidad. En cuanto al diseño de investigación, se debe considerar el carácter apareado o independiente de los grupos de comparación, el tipo de hipótesis planteada y los aspectos relacionados al muestreo. Finalmente, se presentan algoritmos concisos para la elección de las pruebas partiendo de la escala de medición de la variable respuesta y distinguiendo entre procedimientos paramétricos y no paramétricos.

Palabras clave:
Estadística paramétrica, Estadística no paramétrica, Estadística univariante, elección de pruebas estadísticas, Bioestadística.

Abstract

The importance of the Statistics in scientific research is unquestionable, as the medium to obtain valid information from the data; however, given the availability of a large number of approaches and statistical tests, it is a priority to know the criteria for choosing the right tools; among them, it can be enumerated: the measurement level of each variable included in the analysis, the research question and the hypotesis, the research designs, the probability distribution of the outcome variable, the sample size and sample designd used. This review expose the criteria to choosing an univariate inferential statistical test, between the parametric and non-parametrica alternatives, grouping them in three categories:research question, data structure and research design. Regarding the research question, it is neccesary distinguish between descriptive, analytical and experimental ones. The data structure refers to the characteristics of the variables as they are, the scale of measurement, the probability distribution, independence in data collection and homoscedasticity. In terms of research design, should be considered the carácter paired or independent of the comparision groups, the type of hypotesis raised and issues related to sampling process. Finally, concise algorithms for the choice of statistica tests are presented; they are based on the identification of the scale of measurement of the outcome variable and distinguishing between parametric and nonparametric procedures.

Key words:
Parametric tests, Non-parametric tests, univariate statistics, choosing statistical tests, Biostatistics.

Introducción

La inferencia estadística pretende tomar información de una muestra para generalizarla a la población de donde esta proviene, con un nivel de certeza predeterminado y haciendo uso de procedimientos probabilísticos. Cuando la información que se extrae de la muestra hace referencia a comparaciones entre grupos o consiste en la evaluación de hipótesis, se define, además, un nivel de probabilidad con el que se desea encontrar diferencias respecto a una variable de desenlace entre tales grupos de comparación. Los estadísticos estudian muestras que se suponen son representativas y adecuadas en tamaño de la población de la que se extrajeron. Sin embargo, existe una gran cantidad de recursos estadísticos y pruebas alternativas disponibles, lo que hace difícil el poder determinar cuál es el procedimiento estadístico que se aplicará a un problema en particular. El objetivo de este artículo es exponer los criterios para seleccionar un análisis estadístico inferencial univariante entre las alternativas paramétricas y no paramétricas.

Desarrollo

¿Cuándo utilizar una prueba estadística?

El objeto de analizar datos a través de la Estadística se basa en la necesidad de presentar la conclusión más fuerte posible a partir de cantidades limitadas de datos. La validez de esta conclusión puede verse afectada por los dos aspectos siguientes: a) La variabilidad biológica y la imprecisión experimental dificultan la identificación de diferencias reales; b) el cerebro humano tiende a buscar patrones, incluso en datos aleatorios. Cuando las diferencias observadas son pequeñas comparado a la imprecisión experimental y la variabilidad biológica deben usarse pruebas estadísticas (Krzywinski & Altman, 2013; Motulsky, 1995).

La lógica de las pruebas estadísticas asume que todos los eventos ocurren por azar y calculan cual es la probabilidad de que un determinando evento se haya producido por azar para confirmar este supuesto (Henquin, 2013); es decir la estadística contribuye a estimar la probabilidad de que el azar justifique resultados clínicos observados (Fletcher, Fletcher, & Wagner, 1998).

¿Cómo elegir una prueba estadística?

Cuando se dispone de pruebas estadísticas alternativas y válidas para evaluar una hipótesis de investigación, deben emplearse algunas racionalizaciones para elegir entre ellas (Siegel & Castellan, 1995); en general, la elección de una prueba estadística dependerá de: a) la pregunta de investigación, b) de la estructura de los datos y c) del diseño del estudio (Baptist, Röhrig, Hommel, & Blettner, 2010). Cada uno de estos tópicos se tratará a continuación.

Pregunta de investigación

Existe una clasificación general para todas las posibles preguntas de investigación, que se relaciona con los alcances de una investigación; a) las investigaciones descriptivas indican como es o como está la situación de las variables que se estudian en una población; la presencia o ausencia de algo; la frecuencia con que un fenómeno ocurre; y en quienes donde y cuando; b) las investigaciones analíticas indican porque suceden determinados fenómenos, cual es la causa o factores asociados a ese fenómeno o cual es el efecto de la presencia de las determinadas variables; c) las investigaciones experimentales son aquellas en las cuales el investigador manipula las variables y los niveles de exposición de un factor para evaluar su efecto en los diferentes grupos de investigación. (Hernández Sampieri, Fernández, & Baptista, 2003). Cada una de las modalidades anteriormente descritas orientara a un diferente enfoque de análisis estadístico. Además, en la investigación biomédica las preguntas pueden clasificarse de acuerdo a los diferentes tipos de encuentros médico-pacientes (Fletcher et al., 1998). En el cuadro número uno se muestra tanto el tipo de preguntas clínicas más frecuentes, así como los resultados de la enfermedad en dichos estudios.

Cuadro I. Tipos de eventos estudiados en la investigación biomédica

Tipo de evento	Eventos clínicos	Definición
Preguntas clínicas frecuentes	Anomalía	Identificación de individuos sanos y enfermos
	Diagnóstico	Grado de precisión de las pruebas utilizadas para diagnosticar la enfermedad.
	Frecuencia	Frecuencia con que se presenta una enfermedad.
	Riesgo	Factores se asocian con mayor o menor riesgo de enfermedad.
	Pronóstico	Consecuencias de ser portador de la enfermedad.
	Tratamiento	Forma en la que el tratamiento cambia en el curso de la enfermedad.
	Prevención	Evaluación de la mejora del curso de la enfermedad cuando la detección y abordaje son precoces.
	Causa	Condiciones que conducen a la enfermedad.
	Coste	Coste de la asistencia de la enfermedad.
	Muerte	Se considera mal resultado si se presenta prematuramente.
Desenlaces en salud	Enfermedad	Conjunto de signos, síntomas físicos y anomalías de laboratorio.
	Malestar	Síntomas como dolor, náuseas, disnea, prurito y otros.
	Discapacidad	Deterioro de la capacidad para realizar las actividades habituales.
	Falta de satisfacción	Reacción emocional a la enfermedad y a su asistencia.

Fuente: adaptado de Fletcher y otros, 1998.

Estructura de los datos: el nivel de medición de las variables

El proceso de medición implica la asignación de números a observaciones de manera que los números sean factibles de análisis a través de la manipulación u operación de acuerdo a ciertas reglas, con el propósito de revelar nueva información acerca de los objetos y las propiedades medidas. Las operaciones interpretables en un conjunto dado de datos dependen del nivel de medición alcanzado (Siegel & Castellan, 1995).

La escala más básica es la nominal que consiste en variables categóricas cuyas características no presentan ningún orden, y por tanto no provee información alguna de la relación entre las categorías; esta puede ser dicotómica si consiste en solo dos valores posibles o politómica si hay posibilidad de la existencia de más de dos categorías. A continuación se presenta la escala ordinal que se refiere a una variable categórica cuyas categorías si pueden ordenarse, aunque estos no son considerados números reales; continuando con la jerarquía, la escala de intervalo se refiere a números verdaderos, susceptibles de medición que informan que tan grande o tan pequeña resulta una medida respecto de otra, es decir las distancias entre puntos sucesivos en la escala son equivalentes; la escala de intervalo no posee un cero absoluto mientras que la escala más compleja llamada de razón posee todas las características de las escalas anteriores y un cero absoluto. Además, ya sea la escala de intervalo o de razón pueden clasificarse como discretas o continuas; las escalas discretas son números que no permiten divisiones decimales mientras que las continuas si lo permiten (Argimon Pallas & Jiménez Villa, 2000; Blair & Taylor, 2008; Gunawardena, 2011; Henquin, 2013; Jaykaran, 2010; Siegel & Castellan, 1995). En el cuadro II se resumen los tipos de relaciones y operaciones posibles para cada escala de medición.

Cuadro II. Escala de medición de las variables

Escala	Tipo de relaciones	Tipos de operaciones
Nominal	Equivalencia	Transformación de categorías
	Equivalencia	Transformación monotónica
Ordinal	Orden	Cálculo de mediana y cuartiles
	Equivalencia	Operaciones aritméticas
	Orden	Cálculo de media y desviación estándar
De intervalo	Razón conocida entre dos intervalos	Correlación
	Equivalencia	Operaciones aritméticas
	Orden	Cálculo de media y desviación estándar
	Razón conocida entre cualesquier de dos intervalos	Correlación
De razón	Razón conocida entre cualesquiera de dos valores	Modelos de regresión

Fuente: adaptado de Siegel y Castellan, 1995.

El análisis de la escala de medición de las variables implicadas debe realizarse tanto para la variable independiente como para la dependiente; es importante tomar en cuenta que cuando las variables dependientes tiene una escala de intervalo o razón, las pruebas estadísticas que se pueden aplicar poseen más potencia (Argimon Pallas & Jiménez Villa, 2000). Por otro lado si solo existe una variable independiente se habla de análisis bivariados, mientras que si existe más de una variable dependiente se habla de análisis multivariados; si, por otro lado existen varias variables dependientes se habla de métodos de respuesta múltiple (Watt & van den Berg, 2002).

Conviene terminar el análisis de la escala de medición hablando de variables duras y blancas. Variables duras son aquellas en las cuales existe una medición objetiva, al repetirse se obtienen resultados similares, generalmente tienen poca posibilidad de influencia externa y para las cuales hay instrumentos precisos; variables blandas son aquellas que no son objetivas o que tienen un grado alto de subjetividad, no hay instrumentos de medición estandarizados, no hay necesariamente repetibilidad cuando se hacen varias mediciones en condiciones similares y puede haber influencia externa (Ruiz & Morillo, 2004). Más adelante veremos que la distribución de probabilidad de estas variables no suele ser paramétrica.

Estructura de los datos: distribución de probabilidad

Cuando la variable independiente tiene una escala de intervalo o razón, hay que preguntarse cuál es la distribución de probabilidad poblacional de la misma. Si los datos presentan una distribución de probabilidad similar a la normal (gaussiana), deberían utilizarse pruebas paramétricas para evaluar las hipótesis propuestas; de lo contrario, deberán utilizarse métodos no paramétricos, los cuales pueden aplicarse también a variables ordinales (Gunawardena, 2011; Jaykaran, 2010; Siegel & Castellan, 1995).

Existen varios métodos para evaluar la distribución de probabilidad de los datos, como lo son el uso de histogramas, diagramas de cajas (boxplots), gráficos de normalidad Q-Q (Q-Q plots), bean plots, cálculo del coeficiente de asimetría y la curtosis, la evaluación de la regla empírica, los test de bondad de ajuste como la prueba de Shapiro-Wilk, la prueba de Kolmogorov-Smirnof, la prueba de Agostino-Pearson; éstas últimas evalúan la hipótesis que los datos se distribuyen poblacionalmente de forma normal (hipótesis nula) (Albert & Rizzo, 2012;

Barton & Peat, 2014; Jaykaran, 2010; Kanji, 2006).
Estructura de los datos: homocedasticidad, independencia.

Además de la normalidad, los métodos paramétricos deben satisfacer otros supuestos como la independencia, que implica que la selección de un caso no debe sesgar la oportunidad de seleccionar a cualquier otro caso; la homocedasticidad, que implica que los grupos de comparación posean varianzas (dispersión) similares; y que las variables respuesta sean medidas por lo menos en una escala de intervalo (Glantz, 2006; Siegel & Castellan, 1995). Sin embargo, los métodos paramétricos también son conocidos como robustos, pues aun cuando no se satisfagan completamente estas asunciones, pueden aplicarse, sobre todo si se trata de muestras lo suficientemente grandes. Cuando las muestras son pequeñas, generalmente, es difícil determinar si se cumplen dichas asunciones (Argimon Pallas & Jiménez Villa, 2000).

Los métodos no paramétricos, típicamente, hacen menos suposiciones sobre los datos; esto tiene dos implicaciones importantes: a) que las conclusiones derivadas de la aplicación de una prueba no paramétrica son menos específicas que las derivadas de la aplicación de una prueba paramétrica equivalente; b) que cuando no se cumplen esas suposiciones o no se puede comprobar que existan, por ejemplo en caso de muestras pequeñas, es recomendable usar pruebas no paramétricas, pues en estas condiciones, éstas resultarán más potentes que sus equivalentes paramétricos; la potencia es la capacidad de una prueba para detectar como estadísticamente significativa una determinada asociación o diferencia que existe en la realidad. Otro aspecto importante a considerar en la aplicación de los métodos no paramétricos es la consideración de la fuerza de las variables; si la escala es blanda, aplicar una prueba paramétrica podría causar distorsiones que comprometan la validez de los resultados; por ello, cuando se trate de aplicar métodos estadísticos a datos categóricos, rangos, variables con distribución no paramétrica y variables blandas, se utilizarán los no paramétricos (Argimon Pallas & Jiménez Villa, 2000; Siegel & Castellan, 1995).
Otras consideraciones para considerar la elección entre pruebas paramétricas o no paramétricas son las siguientes: a) cuando los datos muestran distribuciones sesgadas o valores atípicos o atípicos extremos los métodos no paramétricos podrían ser los métodos más adecuados; b) cuando se disponen de pocos datos experimentales pero sí se disponen de datos previos de la misma población, pueden examinarse éstos para evaluar la normalidad de los mismos; c) pueden aplicarse a datos no paramétricos transformaciones, como la transformación logarítmica, aunque las transformaciones convierten las variables originales en otras de difícil interpretación y comparación (Barton & Peat, 2014; Motulsky, 1995).
Los procedimientos modernos de Estadística robusta, pueden resolver problemas inherentes al uso de métodos clásicos paramétricos cuando las suposiciones han sido violadas (Erceg-Hurn & Mirosevich, 2008), sin embargo, no se hablará más de estos métodos en este artículo.

Diseño del estudio: la cantidad de grupos de comparación

La cantidad de grupos de comparación, en efecto influyen sobre la prueba estadística a utilizar. Mientras más grupos se comparen entre sí, necesitará utilizarse un mayor tamaño de muestra, para garantizar la validez de las conclusiones obtenidas a través de la aplicación de las pruebas estadísticas. (Argimon Pallas & Jiménez Villa, 2000; Barton & Peat, 2014; Glantz, 2006) De esto se hablará con más detalle cuando se desarrolle el tópico correspondiente al tamaño de la muestra.

Diseño del estudio: el carácter apareado o independiente de los grupos

Las medidas repetidas o apareadas son aquellas que se han realizado sobre los mismos sujetos, por ejemplo cuando se evalúa el efecto de una intervención se toma una medición basal (anterior a la aplicación de la intervención) que se comparará con una medición final (posterior a la aplicación de la intervención), o cuando se comparan individuos con características similares o equivalentes que difieren únicamente en que unos están expuestos y los otros no a un factor modulador de la variable respuesta o a una intervención, como ocurre en los estudios de casos y controles apareados, donde a cada caso le corresponde al menos un control que tiene características clínicas o demográficas similares. Se habla de grupos independientes cuando se comparan individuos diferentes (Argimon Pallas & Jiménez Villa, 2000; Gunawardena, 2011).
La ventaja de diseñar estudios apareados consiste en que, dado que los sujetos son los mismos, existen menos fuentes de variación, y, por ende, en estos casos se cuenta con pruebas más potentes. En los estudios apareados se cuantifica la magnitud de las diferencias entre diferentes mediciones realizadas en los mismos individuos; en los estudios independientes se comparan los valores resumidos entre los diferentes grupos

independientes. Además, el tamaño de muestra utilizado difiere entre uno y otro diseño; en particular, los diseños pareados tienen la ventaja de usar muestras más pequeñas (Argimon Pallas & Jiménez Villa, 2000; Baptist et al., 2010; Gunawardena, 2011; Ruiz & Morillo, 2004). Cuando se comparan tres o más grupos el término que se utiliza es el de medidas repetidas (Motulsky, 1995).

Diseño del estudio: pruebas de una o dos colas

En el apartado anterior se indicó que en los estudios apareados se evalúa la diferencia entre las mediciones realizadas, es decir, se usan hipótesis de diferencias, y por tanto, en estudios con grupos independientes se utilizan hipótesis de comparación de parámetros entre grupos (Jaykaran, 2010). Otras hipótesis que pueden evaluarse son las de una o dos colas; la hipótesis de una cola se aplican cuando existe evidencia clara que la intervención podría actuar en una dirección; las de dos colas cuando la dirección de las diferencias esperadas no está clara (Argimon Pallas & Jiménez Villa, 2000; Henquin, 2013; Motulsky, 1995). En la mayoría de los estudios clínicos, el uso de pruebas de una cola son raras, debido que a menudo se busca un efecto en ambas direcciones y porque los estudios de dos colas reducen la oportunidad de que una diferencia entre grupos sea declarada estadísticamente significativa en error, y en consecuencia, que un nuevo tratamiento sea incorrectamente aceptado como ser más efectivo que un tratamiento existente (Barton & Peat, 2014).

Diseño del estudio: aspectos referentes al muestreo

En las secciones anteriores se definió el concepto de potencia y se mencionaron algunos tópicos relacionado con el muestreo. Ahora se discutirá el concepto de potencia-eficacia, que implica que un incremento particular en el tamaño de la muestra hará equivalentes en potencia a dos pruebas estadísticas. Una prueba estadística no paramétrica poseerá entre el 90 a 95% de la potencia de su equivalente paramétrico, y según el concepto de potencia-eficacia, aumentando de forma suficiente el tamaño de muestra una prueba no paramétrica será tan potente como su equivalente paramétrico (Siegel & Castellan, 1995). También hay que recordar que, según el teorema del límite central, las distribuciones muestrales de ciertas clases de estadísticos se aproximarán a la normalidad, a medida que el tamaño de la muestra se incrementa, e independientemente de la forma de la población muestreada (Blair & Taylor, 2008).

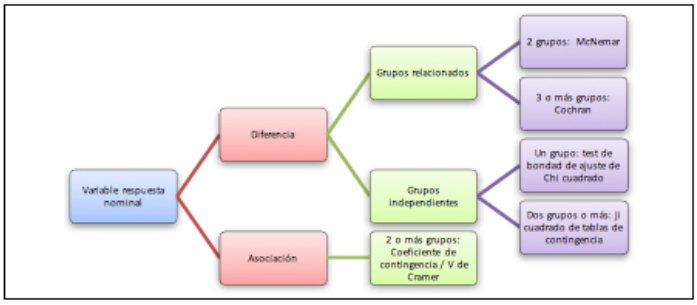
Otro aspecto que se debe considerar es el radio de los grupos de comparación; la potencia se ve comprometida cuando el tamaño de los grupos de comparación se encuentra desequilibrado; pruebas como la T de Student y el análisis de varianza requieren radios de comparación que no excedan 1:4 (Barton & Peat, 2014).

En síntesis, se dirá que en casos que por, tamaño reducido de muestra, no se pueda conocer la distribución de probabilidad de la variable respuesta, y no se disponga de más evidencia, se usarán pruebas no paramétricas. En algunos casos particulares, la elección entra una y otra prueba será basada únicamente en no contar con una muestra más grande como lo es el caso de la prueba exacta de Fisher que se utiliza como una alternativa de la prueba de ji cuadrado, al no contarse con la suficiente muestra para satisfacer las demandas de la última prueba (Siegel & Castellan, 1995).

Algoritmos para elegir una prueba estadística

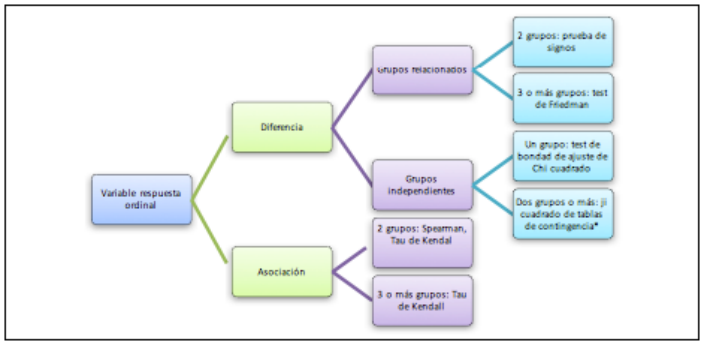
En esta última sección se presentarán algunos algoritmos para decidir cuál es la prueba idónea, tomando en cuenta todo lo expuesto con anterioridad. Es necesario, sin embargo, hacer hincapié en algunas advertencias y recomendaciones. Estos algoritmos han sido elaborados con base a la literatura citada en este artículo, en otros libros no citados, y en el software UNStat4 elaborado por Marta García-Granero y Diego Calavia Gil, de la Facultad de Ciencias de la Universidad de Navarra. Para conocer detalles de la aplicación y supuestos de las pruebas puede consultarse el Handbook of Parametric and Nonparametric Statistical Procedures de David Sheskin.

Algoritmo 1: variable respuesta nominal



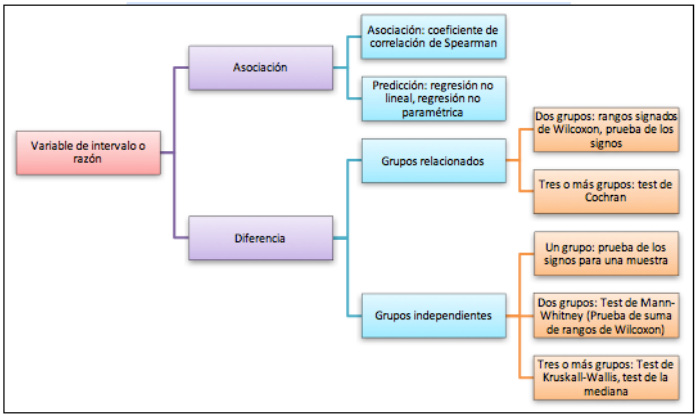
ISSN 2311-9659

Algoritmo 2: variable respuesta ordinal

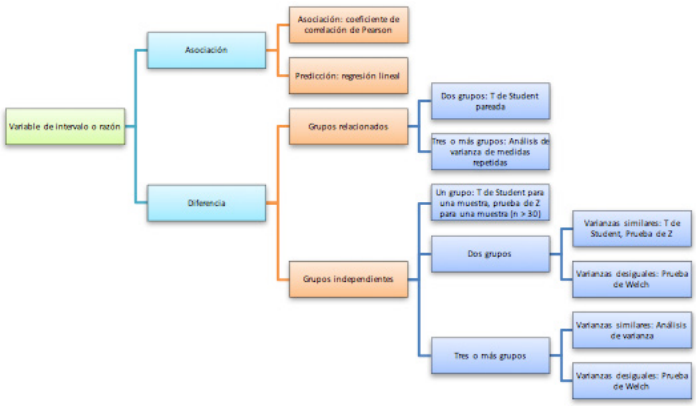


* Cuando el rango es amplio, por ejemplo > 6 puede usarse la prueba de la suma de rangos de Wilcoxon para comparar dos grupos independientes o la prueba de Kruskal-Wallis para comparer tres grupos independientes.

Algoritmo 3. Variable respuesta cuantitativa, pruebas no paramétricas



Algoritmo 4. Variable respuesta cuantitativa, pruebas paramétricas



Referencias

Albert, J., & Rizzo, M. (2012). R by Example. New York, NY: Springer New York. http://doi.org/10.1007/978-1-4614-1365-3

Argimon Pallas, J., & Jiménez Villa, J. (2000). Métodos de investigación clínica y epidemiológica (3rd ed.). Madrid: Elsevier.

Baptist, J., Röhrig, B., Hommel, G., & Blettner, M. (2010). Choosing Statistical Tests. Deutsches Ärzteblatt International, 107(19), 343–348. http://doi.org/10.3238/arztebl.2010.0343

Barton, B., & Peat, J. (2014). Medical Statistics: A Guide to SPSS, data analysis and critical appraisal (2nd ed.). United Kingdom: BMJ Books.

Blair, C., & Taylor, R. (2008). Bioestadística. México: Pearson Educación, S.A.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. American Psychologist, 63(7), 591–601. http://doi.org/10.1037/0003-066X.63.7.591

Fletcher, R., Fletcher, S., & Wagner, E. (1998). Epidemiología Clínica: Aspectos fundamentales (2nd ed.). Barcelona: Masson.
Glantz, S. (2006). Bioestadística (6th ed.). México, D.F.: Mc Graw Hill.

Gunawardena, N. (2011). Choosing the correct statistical test in research. Sri Lanka Journal of Child Health, (40), 149–153.
Henquin, R. (2013). Epidemiología y Estadística para principiantes. Buenos Aires: Corpus.

Hernández Sampieri, R., Fernández, C., & Baptista, P. (2003). Metodologia de la Investigación (3rd ed.). México, D.F.: Mc Graw Hill.
Jaykaran. (2010). How to select appropriate statistical test? Journal of Pharmaceutical Negative Results, 1(2), 61.

Kanji, G. (2006). 100 Statistical Tests (3rd ed.) (3rd ed.). Londres: SAGE Publications Ltd. http://doi.org/10.4135/9781849208499

Krzywinski, M., & Altman, N. (2013). Points of significance: Importance of being uncertain. Nature Methods, 10(9), 809–810. http://doi.org/10.1038/nmeth.2613

Motulsky, H. (1995). Intuitive Biostatistics. New York: Oxford University Press.

Ruiz, A., & Morillo, L. (2004). Epidemiología Clínica: Investigación clínica aplicada. Bogotá: Editorial Médica Panamericana.

Siegel, S., & Castellan, J. (1995). Estadística no paramétrica aplicada a las ciencias de la conducta (4th ed.). México, D.F.: Trillas.
Watt, J., & van den Berg, S. (2002). Research Methods for Communication Science.