Rev Med Chile 2018; 146: 1184-1189

La incertidumbre de la "significación" estadística

RAÚL J. DOMENECH[1]

[1]Programa de Fisiopatología, Instituto de Ciencias Biomédicas, Facultad de Medicina, Universidad de Chile. Santiago, Chile.

El autor declara no tener conflictos de interés. Trabajo no recibió financiamiento.

Recibido el 25 de agosto de 2017, aceptado el 16 de agosto de 2018.

Correspondencia a:
Raúl J. Domenech
Programa de Fisiopatología, Instituto de Ciencias Biomédicas, Facultad de Medicina, Universidad de Chile.
Av Salvador 486, Santiago (Providencia) Chile.
rdomenec@med.uhile.cl

The uncertainties of statistical "significance"

Statistical inference was introduced by Fisher and Neyman-Pearson more than 90 years ago to define the probability that the difference in results between several groups is due to randomness or is a real, "significant" difference. The usual procedure is to test the probability (P) against the null hypothesis that there is no real difference except because of the inevitable sampling variability. If this probability is high we accept the null hypothesis and infer that there is no real difference, but if P is low (P < 0.05) we reject the null hypothesis and infer that there is, a "significant" difference. However, a large amount of discoveries using this method are not reproducible. Statisticians have defined the deficiencies of the method and warned the researchers that P is a very unreliable measure. Two uncertainties of the "significance" concept are described in this review: a) The inefficacy of a P value to discard the null hypothesis; b) The low probability to reproduce a P value after an exact replication of the experiment. Due to the discredit of "significance" the American Statistical Association recently stated that P values do not provide a good measure of evidence for a hypothesis. Statisticians recommend to never use the word "significant" because it is misleading. Instead, the exact P value should be stated along with the effect size and confidence intervals. Nothing greater than P = 0.001 should be considered as a demonstration that something was discovered. Currently, several alternatives are being studied to replace the classical concepts.

(Rev Med Chile 2018; 146: 1184-1189)

Key words: *Biostatistics; Confidence Intervals; Reproducibility of Results.*

"The function of significance tests is to prevent you from making a fool of yourself, and not to make unpublishable results publishable".

D. Colquhoun[2]

Es nuestra costumbre durante una investigación biomédica, determinar si nuestros resultados son "significativos". Por ejemplo, si una droga muestra un efecto realmente diferente comparada con un placebo u otra droga. "Realmente diferente" significa que la diferencia encontrada, no se debe al azar por la inevitable variabilidad que presenta una diferencia aunque la droga no tenga efecto. Existen muchos "test de significación" pero en los últimos años los expertos en estadística han advertido del alto error cometido en la determinación de "significación" con estos tests. Basta usar para el análisis el test de Student ("t-test") no pareado porque es uno de los más usados[1] y el principio del problema es igual para todos los estadísticos.

Supongamos que deseamos comparar el efecto de una nueva droga (droga A) contra una droga conocida (droga B) sobre la presión arterial media en mujeres con hipertensión arterial esencial entre 20 y 30 años de edad de una determinada población. Se toma un grupo de 40 mujeres y se las distribuye en forma aleatoria a las 2 drogas obteniendo 2 grupos en los cuales la distribución de la magnitud de presión arterial es aproximadamente normal y las varianzas son similares. Después de transcurrido

un determinado tiempo se analiza el cambio de presión arterial producido por ambas drogas, se calcula la diferencia de sus promedios, $\bar{x}_A$-$\bar{x}_B$, y se la divide por el error estándar de la diferencia de los promedios.($s_{\bar{x}A-\bar{x}B}$). Esto produce el valor del estadístico "t":

$$t = \bar{x}_A - \bar{x}_B / s_{\bar{x}A-\bar{x}B}$$

Para determinar si este valor de t indica una diferencia real, "significativa", establecemos la **hipótesis de nulidad (H0),** es decir, suponemos que no hay diferencia real entre los promedios poblacionales de ambos grupos, que ambos pertenecen a una misma población. En este caso los valores de t de un infinito número de diferencias de promedios de muestras de igual tamaño deberían distribuirse en una curva como se muestra en la Figura 1. Diferentes valores de t en la abscisa determinan diferentes probabilidades (áreas bajo la curva), de obtener estos valores. Los valores de t están agrupados alrededor de t = 0 dado que la mayor parte de las diferencias entre los promedios sería 0 o muy cercano a 0, pero la inevitable variabilidad entre las muestras producirá también valores de t que se disponen, por azar, simétricamente a ambos lados de t = 0. En nuestro caso, debido al tamaño muestral, rara vez se encontrarán valores de t mayores de + 2,1 o menores de − 2,1 (ambas colas de la curva). El área bajo la curva, distal a estos valores de t comprenden, cada una de ellas, la probabilidad P ≤ 0,025 (2,5%), en total P ≤ 0,05 (5,0%), es decir una baja probabilidad de encontrar por azar una diferencia de promedios alejadas de 0 cuando la hipótesis de nulidad cierta. Así, el valor de P queda definido como la probabilidad de obtener un valor de t (u otro estadístico), o uno más distal que el computado de los datos, cuando la hipótesis de nulidad es cierta. Por esta razón, si en nuestro estudio encontráramos un valor de t igual o superior a + 2,1 o igual o inferior a -2,1, es decir P ≤ 0,05, podríamos *considerar razonable, o estar dispuestos a aceptar que:* siendo tan baja la probabilidad (≤ 5,0%) de encontrar estos valores por azar, la hipótesis de nulidad no es cierta, la rechazamos y decimos que la diferencia encontrada muy probablemente es una diferencia real entre dos muestras provenientes de poblaciones diferentes (una población con la droga A y otra con la droga B). Decimos que el resultado es "*significativo*", que una de las drogas tiene mayor efecto que la otra y que solo existe 5% de probabilidad (1 en 20) de estar equivocados y que el resultado se explique por azar es decir que sea un **falso positivo** (un falso resultado). Este es el Test de Análisis de la Hipótesis de nulidad. Nótese que en esta búsqueda de la felicidad (la significación) aceptamos la posibilidad de estar cometiendo un error. Este error se denomina Error tipo I y la probabilidad de cometerlo, P, se denomina nivel de significación α (en este caso P ≤ 0,05 o 5%) y es el máximo Error Tipo I que estamos dispuestosa cometer. El

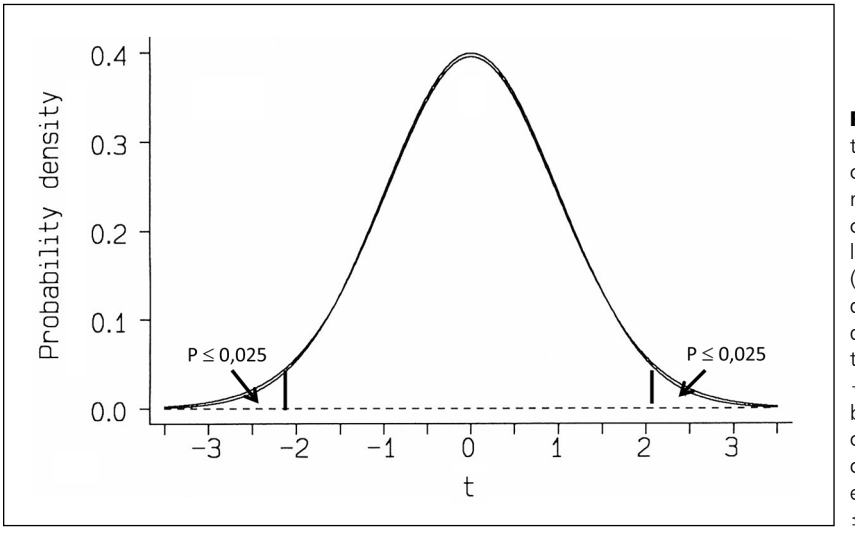

Figura 1. Distribución de t.-Distribución de los valores de t al comparar los promedios de dos muestras de una misma población (hipótesis nula). El área bajo la curva define la probabilidad (p) de obtener una diferencia de promedios ($\bar{x}_1 - \bar{x}_2$) para los diferentes valores de t. Cuando t es mayor a + 2,1 o menor de −2,1 (ambas colas de la distribución distales a la líneas verticales cortas), la probabilidad de obtener estos valores es ≤ 0,025 en cada uno de esos extremos, o ≤ 0,050 para ambos.

valor de t, (2,1), correspondiente a $P \leq 0,05$ que elegimos en forma convencional para rechazar la hipótesis de nulidad, se denomina valor crítico de t. Nótese que esta elección es arbitraria y *a priori*; si queremos cometer menos error podemos fijar un valor crítico de t correspondiente a $P \leq 0,01$ (1,0%) o $P \leq 0,001$ (1,0 %0) pero siempre arbitrarios. Sin embargo, desde hace varios años los expertos en estadística, vienen advirtiendo que el error cometido al elegir un valor de t equivalente a $P \leq 0,05$ es muy superior a 5% por lo que se ha puesto en duda los resultados de muchas publicaciones considerándolos falsos descubrimientos como se explica a continuación.

Un análisis muy demostrativo de este problema lo realizó recientemente Colquhoun, D[2]. Para comenzar debe notarse que el valor de $P \leq 0,05$ nos entrega, en base a un solo experimento, una baja probabilidad de encontrar falsos positivos cuando suponemos que no existe una diferencia real (la hipótesis de nulidad es cierta) pero no nos entrega la probabilidad de encontrar esos resultados cuando existe una diferencia real, con verdaderos positivos (la hipótesis de nulidad es falsa) si se repite varias veces el experimento en el largo plazo[3]. Para determinar la probabilidad de resultados falsamente positivos es necesario conocer también la probabilidad de obtener resultados verdaderos positivos y calcular la proporción de resultados falsos positivos del total de resultados positivos (verdaderos + falsos). Supongamos que se realizan 1.000 t tests y que conocemos la prevalencia del efecto real de la droga, por ejemplo 10%, es decir $1.000 \times 0,10 = 100$ test debieran ser verdaderos positivos, y con un poder (capacidad del test de detectar verdaderos positivos) de 0,80, obtendremos una población de 80 test verdaderos positivos en la cual la hipótesis de nulidad es falsa (existe una diferencia real entre las drogas). Si la prevalencia del efecto real de la droga es 10% quiere decir que la droga no tiene efecto en 90% de los experimentos, es decir $1.000 \times 0,90 = 900$ tests serán verdaderos negativos constituyendo una población en que la hipótesis de nulidad es cierta (no existe diferencia entre las drogas) y si elegimos un α de 5% para rechazar la hipótesis nula, 5% de ellos 900 x 0,05 = 45 serán falsamente positivos (por azar). En total los exámenes positivos suman 80 + 45 = 125, de los cuales 45 son falsos positivos es decir tenemos 45/125 = 36% de falsos positivos en lugar de 5% como pensábamos a partir de nuestro valor crítico de t correspondiente a $P \leq 0,05$. El error es muy superior a lo esperado. Este efecto se comprueba simulando computacionalmente un gran número de t test con variables aleatorias y comparando t tests en que la hipótesis nula es cierta, (para obtener los falsos positivos), con t tests en que la hipótesis nula no es cierta, (para obtener los verdaderos positivos). y un poder = $0,80^2$.

Como se puede apreciar del análisis arriba descrito la diferencia entre el error presunto y el real disminuye si suponemos una mayor prevalencia del efecto real de la droga para un determinado poder del experimento, porque aumenta el número de verdaderos positivos. Pero solo asumiendo que el efecto real ocurre en 50% de los test, y conservando un poder de 0,80 el error disminuye a 6%, no muy diferente de 5%. Por otra parte, el error aumenta si disminuye el poder del experimento porque disminuye la probabilidad de detectar verdaderos positivos. Muchas publicaciones no describen el poder de su estudio o describen un poder muy bajo, frecuentemente 0,50 o menos, generalmente debido a un tamaño muestral pequeño. Incluso en los ensayos clínicos randomizados con alto número de pacientes y un poder cercano a 0,80 presentan un Error Tipo I de 36% al reportar una significancia al valor de $P \leq 0,05$ como el ejemplo arriba analizado.

Es obvio que este análisis no se puede realizar si no se conoce la prevalencia del efecto real de la droga (generalmente no se conoce) que permita obtener los verdaderos positivos, pero revela la incertidumbre de trabajar con solo un valor de P asumiendo que la hipótesis de nulidad es cierta. Esto puede explicar en gran parte los falsos resultados publicados con tests de significación usando $P \leq 0,05$ para descartar la hipótesis de nulidad como lo explica Ioannidis en su artículo "Why most published research findings are false"[4]. En otras palabras ¿estamos haciendo el ridículo publicando descubrimientos que no se reproducen porque son falsos?

El problema recién analizado revela la incertidumbre del error cometido para descartar la hipótesis de nulidad al usar un valor de P. Sin embargo, el problema va más allá, a la incertidumbre en la replicación de los valores de P. Al respecto existen diferencias de opinión entre los expertos y el tema es muy bien analizado por J. E. Hoffman en su libro "Biostatistics for Medical and Biomedical

Practitioners"[5], Hubbard and Bayarrri[6] y Cumming G[7]. El test de significancia fue introducido por Fisher en 1925[8] y basado en la hipótesis de nulidad, como se expuso más arriba, que supone que las diferencia entre dos grupos no es real sino debida al azar por la inevitable variabilidad de los datos alrededor de una medida central. Por lo tanto, había que definir la probabilidad, P, de que la diferencia observada se deba al azar. Si la probabilidad de azar es alta (Ej. $P > 0,05$) la hipótesis de nulidad no se puede descartar y si la probabilidad es muy baja ($P < 0,05$) la hipótesis de nulidad es difícil de aceptar. Fisher usaba el valor de P obtenido en un solo experimento (lo que usualmente hacemos) como el peso de la evidencia en contra de la hipótesis de nulidad y no como un criterio exacto de error como el que se obtendría por un análisis repetitivo del experimento (raramente factible) o su repetición en el largo plazo y aconsejaba otros factores como la magnitud de la diferencia encontrada para decidir el rechazo de la hipótesis de nulidad. Esto es la inferencia inductiva de Fisher que involucra cierta subjetividad. Posteriormente Neyman y Pearson en 1928[9,10] agregaron un concepto más rígido con la intención de superar esta subjetividad. Establecieron el concepto de fijar *a priori* un máximo valor de P (generalmente 0,05 o 0,01) para rechazar la hipótesis de nulidad y lo denominaron nivel de significación α, y denominaron como Error Tipo I el riesgo de rechazar falsamente la hipótesis de nulidad con esta determinación como se explicó más arriba. Postularon entonces una hipótesis alternativa (HA), es decir, la droga tiene real efecto (concepto imposible de no plantear cuando se rechaza la hipótesis de nulidad). La hipótesis alternativa tiene entonces su propia curva de distribución de t (Figura 2) con valores de t diferentes a los de la hipótesis de nulidad y centrados alrededor del valor de t correspondiente a la diferencia de promedios producido por la droga. De la comparación de esta curva con la curva de la hipótesis de nulidad nace el error Tipo II, la probabilidad de obtener valores verdaderos positivos, bajo la curva de la hipótesis alternativa, pero no detectables por quedar incluidos también bajo el área de la curva de la hipótesis de nulidad sobre $P = 0,05$ (Figura 2). Se lo denomina también error Tipo β. Por lo tanto nos queda $1-\beta$ = probabilidad de detectar verdaderos positivos y rechazar correctamente la hipótesis de nulidad, lo que se denomina Poder (Figura 2). El Poder aumenta al incrementar el tamaño del efecto y/o el tamaño de la muestra y disminuye al disminuir el valor crítico de P seleccionado para descartar la hipótesis de nulidad (al intentar disminuir el error Tipo I aumenta el error Tipo II). Así, en contraste con la terminología "Test de Significancia" de Fisher, Neyman y Pearson introdujeron la terminología "Test de Hipótesis" la cual a diferencia con la inferencia inductiva de Fisher corresponde a una conducta inductiva que establece reglas para tomar decisiones entre dos hipótesis. A diferencia del test de significancia de Fisher quien sostenía que la hipótesis de nulidad

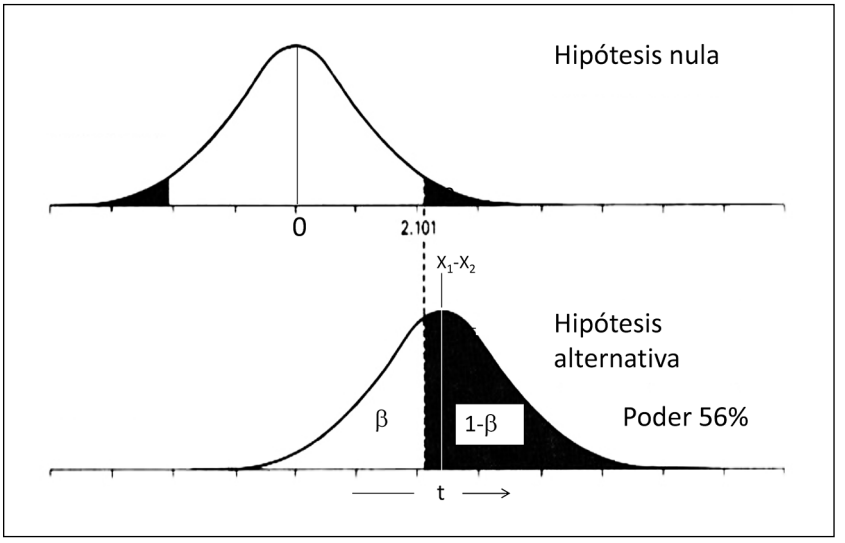

Figura 2. Hipótesis nula e hipótesis alternativa. La imagen inferior muestra la distribución de t de la hipótesis alternativa desplazada con respecto a la hipótesis nula en la imagen superior. β es la probabilidad de verdaderos positivos no detectables por quedar bajo el valor de $t = 2,01$ de la hipótesis nula correspondiente a $P \leq 0,05$ (Error tipo II). $1-\beta$ es la probabilidad de detectar verdaderos positivos o Poder del estudio.

nunca puede ser aceptada sino difícil de rechazar, el test de hipótesis de Neyman y Pearson sostiene que la hipótesis de nulidad se puede aceptar o rechazar. La mezcla de estas 2 concepciones es lo que hoy se nos enseña.

Sin embargo, esta mezcla de los conceptos de Fisher y de Neyman y Pearson ha sido cuestionada por varios expertos en estadística[3,11-14] fundametalmente porque Neyman y Pearson consideran el valor de P como la probabilidad de error que se obtendría (teóricamente) en el largo plazo con la repetición de muestras (probabilidad frecuencial) y no como una inferencia inductiva de un simple experimento como lo concibió Fisher. Esta suposición ha sido desafiada con un análisis computacional de tests repetidos de situaciones simuladas que permite determinar la distribución de los valores de P y que revela que suponiendo que la mitad de los tests corresponden a una hipótesis de nulidad cierta (no hay diferencias entre los grupos) y la otra mitad a tests en que la hipótesis alternativa es cierta (hay diferencias entre los grupos), 20-50% de las veces el valor de P = 0,05 proviene de la hipótesis de nulidad y el resto de las veces de la hipótesis alternativa. Es decir el valor de P = 0,05 provee solo una leve evidencia en contra de la hipótesis de nulidad[6]. Los valores de P no equivalen entonces a un error en el sentido de probabilidad frecuencial como suponen Neyman y Pearson, lo cual tiene consecuencias prácticas como por ejemplo:

a) Para un determinado tamaño del efecto, P es función de la variabilidad y tamaño de las muestras y dos experimentos con igual resultado en cuanto a tamaño del efecto pero diferentes valores de P no son necesariamente contradictorios[14].
b) Un determinado valor de P no garantiza que se reproducirá al repetir el experimento. Al respecto Cumming[7] demostró que un valor de P de 0,05 tiene 80% de probabilidad de variar entre 0,00008 y 0,44 en subsecuentes repeticiones del experimento. En otras palabras, si en un experimento se obtiene un valor de P = 0,05 ¿qué valor de P se espera obtener al repetir exactamente el experimento con otra muestra?, prácticamente cualesquier valor de P simplemente por variabilidad del muestreo. (conceptos de "intervalos de P" y "la ruleta de los valores de P" de Cumming). Más aún, este resultado es función del poder pero independiente del tamaño muestral.

En conclusión, existen al menos dos fuentes de incertidumbre del valor de P en la "significación" de un resultado. Por un lado, la ineficacia del valor de P en determinar el error tipo I (cuantificar probabilidad de falsos positivos) para descartar la hipótesis n de nulidad y, por otro, la incertidumbre de la replicación de un valor de P al repetir el experimento.

Las críticas sobre "significación" estadística son cada vez más abundantes (ver Simposium en "Biostatistics", vol 14, número 1, 2014) y recientemente la "American Statistical Association" expuso sus puntos de vista al respecto[15] de los cuales sobresalen los siguientes:

1. El valor de P no mide la probabilidad de que la hipótesis estudiada sea cierta ni la probabilidad de que los resultados se deban solo al azar.
2. Conclusiones científicas y decisiones comerciales no deben basarse solamente en si el valor de P sobrepasa un valor umbral.
3. Un valor de P, o significación estadística, no mide el tamaño de un efecto ni la importancia de un resultado.
4. El valor de P no provee una buena medida de evidencia para un modelo o hipótesis.

La pregunta es ¿qué hacer nosotros los investigadores y los lectores de publicaciones de investigaciones clínicas? Al respecto los expertos en estadística aconsejan:

1. Nunca usar la, palabra "significativo" ni las expresiones "casi significativo" o "tendencia a la "significación" porque son engañosas.
2. Si se realiza un test de significación, comunicar el valor de P exacto (no el área distal al valor crítico de t), el tamaño del efecto obtenido y los límites de confianza.
3. Considerar que la obtención de un valor de P ≤ 0,05 solo indica que vale la pena repetir el experimento mejorando el poder con un mayor tamaño muestral y asegurar la aleatorización.
4. Para disminuir la probabilidad de un falso resultado en primera instancia o al repetir un experimento, considerar como muy probable verdadero positivo solo un valor de P < 0,001.

Existen varias proposiciones de los expertos para evitar estas dificultades como el uso de valores umbrales de P menores de 0,05 para la significaión estadística[16] o el uso del Factor de Bayes[17,18] que

busca una evidencia más directa de la relación entre la hipótesis alternativa y la hipótesis de nulidad; pero la discusión de estas proposiciones está fuera del objetivo de este artículo.

Sin embargo, cualesquiera sea la evolución de los intentos por mejorar nuestras conclusiones sobre los resultados de una investigación y evitar la presión de publicar a todo trance (vanidad, figuración, ascenso, renovar el proyecto, mantener o ascender en el cargo) parece cierto que, muchas veces, como dijo Vin Scully, un periodista deportivo sagaz:

"*La Estadística se usa más bien como un borracho usa un farol de alumbrado: para sujetarse, no para iluminarse*".

Referencias

1. Glantz S. Primer of Biostatistics. Seventh edition. McGraw-Hill Companies, Inc. 2012.
2. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of P values. R Soc open sci 2014; 1: 140216, http://dx.doi.org/10.1098/rsos.140216.
3. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypothesis. Am Stat 2001; 55: 62-71.
4. Ioannidis JP. Why most published research findings are false. PLoS Med 2005. 2, e124. (doi:10.1371/journal.pmed.0020124).
5. Hoffman JIE. Biostatistics for Medical and Biomedical Practitioners. USA Academic Press, Elseviere. 2015, p. 151-70.
6. Hubbard R, Bayarri MJ. P values are not error probabilities (Online). Available: http://ftp.stat.duke.edu/WorkingPapers/03-26.pdf [Accessed 2003].
7. Cumming G. Replication and P Intervals : p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. Perspect Psychol Sci 2008; 3: 286-300.
8. Fisher RA. Statistical Methods for Research Workers. 1925, Edinburg: Oliver and Boyd.
9. Neyman J, Pearson ES. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I, Biometrika 1928; 20A: 175-240.
10. Neyman J, Pearson ES. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part II, Biometrika 1928; 20A: 263-94.
11. Rozeboom WW. The fallacy of the null-hypothesis significant test. Psychol bull 1960; 57: 416-28.
12. Cohen J. The earth is round ($p \leq .05$). American Psychologist 1994; 49: 997-1003.
13. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008; 45: 135-40.
14. Motulsky HJ. Common misconceptions about data analysis and statistics . Br J Pharmacol 2015; 172: 2126-32.
15. Wassertein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. Amer Stat 2016; 70: 129-33.
16. Benjamin DJ, et al. Redefine statistical significance. Credit Psy Ar Xiv reprint service, 22 July 2017. Web dx. Doi. Org/10.17605/OSF.10/ MKY9J.
17. Goodman SN. Toward evidence-based medical statistics. 1 The P value fallacy. Ann Intern Med 1999a; 130: 995-1004.
18. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 1999b; 130: 1005-13.