Statistical Significance, Power of the Test, and Effect Size Measures in Two-independent-samples t-test case

Svetlana Todorova - Chief Assist. Prof. Dr.
University of Economics - Varna, Varna, Bulgaria
svetlana.todorova@ue-varna.bg

Abstract

Quantitative research is focused on determining the occurrence of certain population phenomena by analyzing data from a sample. Statistics is a tool that is used to check hypotheses and make decisions to reject or fail to reject such hypotheses. In this paper, the various statistical tools are reviewed: the limitations of null hypothesis significance testing and the advantages of using effect size as two measurements, which can provide important information about the results of a study. These measurements also can help interpretation of data results and easily detect trivial effects. Further, it is recommended to establish an appropriate sample size by using the optimum statistical power of the test that the research is designed. We discuss statistical significance, sample size, power of the test, and effect size, all of which have an enormous impact on how we interpret the results.

Keywords: statistical significance, power of the test, effect size, sample size, independent samples t-test

JEL Code: C12, C18, C88

Въведение

Роналд Фишър е бащата на съвременните статистически заключения. От тогава, по правило е прието, че ако p-value е по-малка от 0.05, то например, наблюдаваната разлика между средните от две независими извадки (групи) е статистически значима. Казано с други думи, има по-малко от 5% вероятност за възникване на случайна грешка и нулевата хипотеза може да бъде отхвърлена, защото 95 пъти от 100 пъти наблюдаваната разлика между средните няма да се дължи на случайност. Въпреки, че някои изследователи използват по-строги равнища на значимост, като $\alpha = 0.01$ или $\alpha = 0.005$, общоприетата гранична стойност на $\alpha = 0.05$ се приема по поразбиране и е заложена така в почти всички статистически софтуери. Всъщност няма научен аргумент, който да поддържа идеята, че $\alpha = 0.05$ и съответно търсеното p-value да е по-малко от 0.05. Това обстоятелство кара някои статистически експерти, като Rosnow и Rosenthal (1989, стр. 1277), да се изразят саркастично, че „със сигурност Бог обича 0.06 почти толкова, колкото 0.05“.

Проверката за статистическа значимост е дълбоко залегнала в съзнанието на изследователите. Това е типа мислене, което възприема света само в черен или бял цвят, без междинни нюанси (Kirk, 2001). Този подход кара изследователите да не докладват резултати от работата си ако се получи така, че те не са статистически значими и според тях „няма важни резултати“ или „хипотезата не е доказана“. От друга страна, публикуването само на статистически значими резултати в научните списания изкривява съответните знания и дава погрешна представа за изследванията (Cumming, 2014). Поради тази причина Cumming (2014) предлага да се разглежда така наречените „нови статистики“, като се използва размер на ефекта, мощност на критерия и мета-анализите.

Обикновено проверката за статистическата значимост е съсредоточена върху отхвърлянето на нулева хипотеза при равнище на значимост, което е избрано преди това, най-често 0.05. По този начин изследователите получат информация за това дали „има или няма промяна, която е различна от нула“. Но вероятността за отхвърляне на нулевата хипотеза нараства с увеличаване на обема на извадката или извадките. Следователно проверката за статистическата значимост ни дава повече информация за обема на извадката, отколкото за хипотезата. Тълкуването на статистическата значимост става безсмислено, когато размерът на извадката е толкова голям, че всяка открита разлика, колкото и малка или дори тривиална да е, ще позволи отхвърлянето на нулевата хипотеза. Някои статистици

смятат, че статистическата значимост е не само ненужен инструмент, но и е навредила на научното развитие. Както е посочено от Schmidt и Hunter (2002, стр. 65): „Тестовете за значимост са катастрофален метод за тестване на хипотези“. А Kirk (2001, стр. 214) твърди, че науката би имала по-големи ползи, ако изследователят се фокусира върху размера на ефекта и неговото практическо значение, вярвайки също, че „никой не е в по-добра позиция от изследователя, който е събрал и анализирал данните, за да реши дали ефектите са тривиални или не”.

Така през 1999 г., след дълъг период на работа, Wilkinson и Американска психологическа асоциация (APA) подготвят доклад, в който се заявява: „изследователите трябва винаги да публикуват размера на ефекта в основните резултати“ (Wilkinson, 1999, стр. 599). А главният автор на Revista Panamericana de Salud Publica, критикувайки използването на статистическата значимост в изследванията, заявява, че „изследователите, които получават (статистически) значими резултати, са били удовлетворени, когато са постигнали целта си, без да осъзнават, че са не са постигнали никакво подобрение в разбирането на изучавания феномен” (Clark, 2004, стр. 293).

Целта на това изследване е да покаже, че при тестването на статистически хипотези, освен проверката за статистическа значимост, от изключително значение е да се изследва размера на ефекта и да се отчете мощността на критерия.

1. Статистическа значимост

Статистическата проверка на хипотези при независими извадки се прилага, за да се определи дали разликата между средните на две групи е наистина значима или се дължи на случайни фактори. Статистическата значимост е крайъгълния камък на статистическите заключения. За да обсъдим нейното значението в контекста на истинските експерименти ще стъпим на един пример. Например проверката на тази статистическа хипотеза може да определи дали средната успеваемост на студентите е различна, по-висока или по-ниска при приложението на два различни метода на преподаване. Първата стъпка е определяне на нулевата и алтернативната хипотези. Нулевата хипотеза е хипотеза за нулев ефект, а алтернативната показва, че тази разлика е различна, по-голяма или по-малка от дадена стойност като това се указва от знака в алтернативната хипотеза ($\neq$; $>$; $<$):

Нулева хипотеза: Предполагаема хипотеза, която твърди, че няма значими разлики между групите, т.е. нулевата хипотеза не предвижда разлики между резултатите на студентите обучавани по двата метода.

Алтернативна или изследователска хипотеза: Единият метод на преподаване ще доведе до по-високите средни резултати на студентите.

За проверка на хипотезата се излъчват две случайни извадки от двете генерални съвкупности (групи). В края на семестъра се пуска еднакъв тест, за да се сравнят получените резултати. Да приемем, че средният резултат от теста на едната група студенти е 80%, а от другата е 90%. Каква е вероятността този ефект, породен от различни методи на обучение, върху резултатите от тестовете на студентите да е възникнал случайно или поради друга латентна (т.е. неизмерена) променлива? И не на последно място, дали 10% по-високи резултати са „достатъчно високи“, за да бъдат толкова различни от 80%?

Получаването на статистически значим резултат означава, че p-value, получена от t-теста при двете независими извадки е равна или по-малка от определеното равнище на значимост, което в повечето случаи е 0.05. Но p-value по-малка от 0.05 би означавала, че има по-малко от 5% шанс нулева хипотеза да е вярна. В този случай се отхвърля нулевата хипотеза в полза на алтернативната и се взема решение, че има статистически значима разлика в тестовите резултати на студентите, обучавани по двата метода. Забележете, в

случая не се казва, че различните методи на обучение са причина за значими разлики в резултатите от тестовете на студентите. P-value дава информация само за това, че двете групи са различни една от друга или не, а не че методът на обучение е повлиял в това различие. Ако зададем по-малко равнище на значимост, например 0.01, то за да се счита, че получената разлика е статистически значима, трябва получената p-value да е равна или по-малка от 0.01 и тогава може да се заключи, че има по-малко от 1% шанс нулевата хипотеза да е вярна.

От друга страна, обемът на извадката или броят на участниците в проучването има огромно влияние върху това дали получените резултати са значими или не. Колкото по-голяма е действителната разлика между средните резултатите от тестовете на студентите, толкова по-малка част от извадката ще ни е необходима, за да намерим значима разлика (т.е. p-value $\leq$ 0,05). Теоретично обаче, може да се намери значима разлика в почти всички експерименти с достатъчно голям обем на извадката, въпреки че изключително големите извадки изискват голям финансов ресурс.

2. Мощност на критерия

При статистическата проверка на хипотези, поради извадковия характер на изследването, съществува риск от допускане на грешки. Илюстративно това може да се представи, чрез следната схема:

Таблица 1. Статистическа проверка на хипотези и грешки от I и II род

Действително състояние → и вземане на решение ↓		**Действително състояние**	
		Но е вярна	**Но не е вярна**
Решение	**Не се отхвърля Но**	**Правилно решение (1 – α)**	**Грешка от II род (β)**
	Отхвърля се Но	**Грешка от I род (α)**	**Правилно решение (1 - β)**

Грешка от I род (α) или фалшив положителен резултат е вероятността, че средните в двете групи са значими, когато в действителност не е така. При равнище на значимост 0.05 има 5% шанс за неправилно отхвърляне на нулевата хипотеза.

Грешка от II род (β) или фалшив отрицателен резултат е вероятността да се заключи, че средните в двете групи не се различават значително, когато в действителност са различни.

Грешка от I род може да възникне само ако нулевата хипотеза е вярна, докато грешка от II род възниква само ако нулевата хипотеза не е вярна. С други думи грешка от I род и грешка от II род не могат да възникнат едновременно. Въпреки това, при един и същ обем на извадката ако се намали α, β се увеличава и ако се намали β, α се увеличава. Ако обаче, се увеличи обема на извадката, то и α и β намаляват.

За да се намали вероятността от извършване на грешка от II род, трябва да се увеличи мощността на критерия. Мощността на критерия се определя като 1-β. Това е вероятността за откриване на разлика в средните между групите, когато разликата действително съществува (т.е. вероятността за правилно отхвърляне на нулевата хипотеза). Следователно, когато се увеличава мощността на критерия се увеличава и възможността за откриване на

статистически значима разлика (т.е. p-value ≤ 0.05) между групите.

Общоприето е, че трябва да се стремим към мощност от 0.80 или по-голяма. В този случай, шансът да се открие статистически значима разлика е 80%. Въпреки това, все още съществуват 20% вероятност, при които няма да се открие действително значима разлика между средните на групите.

Размерът на ефекта е полезен статистически инструмент, чиято употреба се препоръчва широко от Американската психологическа асоциация (1994, 2001, 2010), Американската асоциация за образователни изследвания (2006) в областта на образованието и Международния комитет на редакторите на медицински списания в областта на медицинските изследвания. Подходящо избрана мощност на критерия намалява вероятността от неправилни заключения.

3. Размер на ефекта

Когато се правят промени в начина, по който се преподава в един курс, често искаме да измерим въздействието на тези промени върху обучението на студентите. Има няколко различни обобщаващи характеристики, които могат да се използват, за да се сравни обучението в различните курсове при различни подходи на преподаването. В изследванията на социалните науки тази обобщаваща характеристика се нарича *размер на ефекта*. Размерът на ефекта е мярка за това колко важна е разликата: големите размери на ефекта означават, че разликата е важна, малките размери на ефекта означават, че разликата е маловажна. Ефектът на размера се получава като се нормира разликата между двете средни и дава мярка за това колко съществено се различават резултатите от различното преподаване в двата курса.

Размерът на ефекта обаче, не е статистическата значимост. Докато значимостта показва колко вероятно е резултатът да се дължи на случайни фактори, то размерът на ефекта разкрива колко важен е резултатът. В статията „Изявление относно статистическата значимост и p-values“, Американската статистическа асоциация обяснява, че статистическата значимост не е еквивалентна на научна, човешка или икономическа значимост. По-малките p-values не предполагат непременно наличието на по-големи или по-важни ефекти и по-големите p-values не означават липса на важност или дори липса на ефект. Всеки ефект, независимо колко малък е, може да доведе до малка p-value, ако размерът на извадката или прецизността на измерване е достатъчно висока, а големите ефекти могат да доведат до по-високи p-values, ако размерът на извадката е малък или измерванията са неточни (Wasserstein and Lazar, 2016).

Размерът на ефекта се отнася до концепцията за „минимална важна разлика“, която гласи, че в определен момент значимата разлика (т.е. p-value ≤ 0.05) е толкова малка, че не би допринесла никакви ползи в реалния свят. По-малки p-values (0.05 и по-ниски) не предполагат доказателства за големи или важни ефекти, нито пък високите p-values (0.05+) предполагат незначителна важност или малки ефекти. При достатъчно голям размер на извадката, дори много малки размери на ефекта могат да доведат до значителни p-values (0.05 и по-ниски). С други думи, статистическата значимост изследва вероятността дали резултатите се дължат на случайност, а размерът на ефекта обяснява важността на резултати.

Най-разпространеният показател за размера на ефекта е *Cohen's d,* който използва за измерител стандартизираната разлика между две средни. Той може да се приложи и за сравняване на ефектите между отделни проучвания, дори когато се използват данни в различни мерни единици, защото е нормиран чрез стандартното отклонение (S_{pooled}). *Cohen's d* се тълкува по следния начин: ако стойностите му са по-малки или равни на 0.2 - ефектът е малък, ако са около 0.5 - ефектът е среден и ако са 0.8 и по-големи -

ефектът е голям.

4. Статистическа значимост, мощност на критерия и размер на ефекта при определяне на обема на извадката

Ако се обедини всичко казано до тук, статистическа значимост, мощност на критерия и размер на ефекта, може да се изчисли минимален необходим обем на извадката за дадено проучване или експеримент като се постигне необходимата мощност на критерия и размер на ефекта. Това трябва да се извърши предварително, преди действителното провеждане на експеримента, по време на фазата на проектиране на изследването. По този начин се постига по-добра представа за обема на извадката (броя на участниците, необходими за всяка от групите) и може да се намери статистически значима разлика, ако има такава.

Тези четири аспекта - статистическа значимост, мощност на критерия, размер на ефекта и обем на извадката са толкова взаимосвързани, че когато три от тях са известни, четвъртият лесно може да бъде определен. Задавайки равнище на значимост, мощност на критерия и размер на ефекта, е възможно да се определи най-малкия обем на извадката, който е необходим за получаване на желания значим статистически резултат. По този начин, ако резултатите от изследването не са статистически значими, но имат практическа значимост, което означава, че е на лице среден или голям ефект на размера, тогава, за да се получи статистически значим резултат, ще е необходимо само да се увеличи обема на извадката. По същия начин, колкото по-висок е размерът на ефекта, толкова по-малък обем на извадката е ще е необходим за откриване на статистически значими резултати.

G*Power (Faul, Erdfelder, Lang, Buchner, 2007) е безплатен статистически софтуер, който се използва в тази разработка при намирането на обема на две независими извадки при тестването на разликата между средни с t-тест, въз основа на зададени параметри: равнище на значимост 0.05; мощност на критерия – 0.80 и 0.90 и размер на ефекта – 0.2, 0.5 и 0.8. При процедурата е заложено, че двете извадки са с еднакъв обем, а при тестването на хипотезите се използва двустранна критична област. Получени са следните резултати в таблица 2.

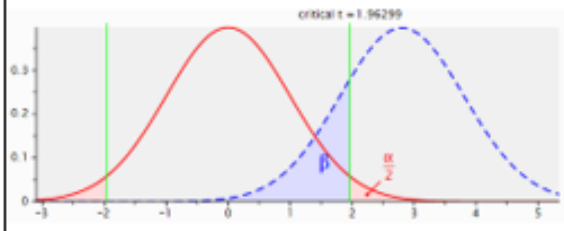
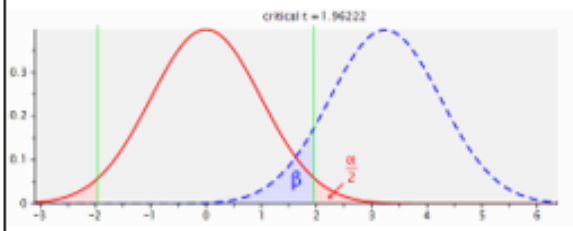
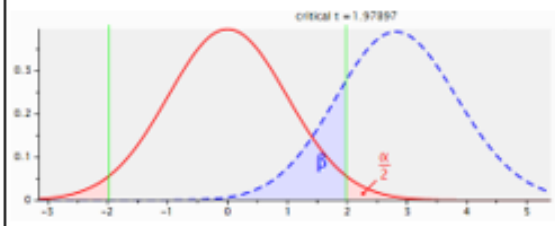
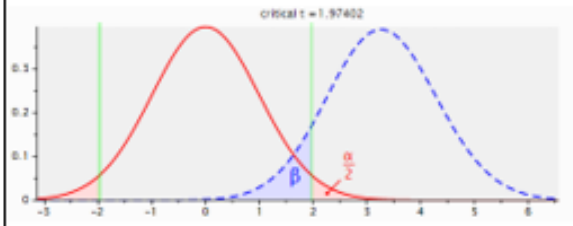
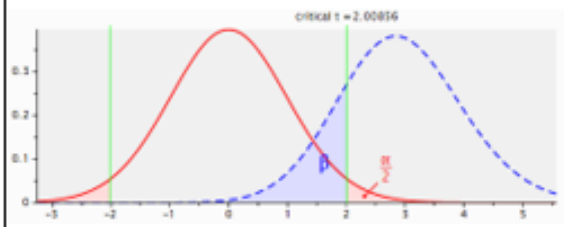
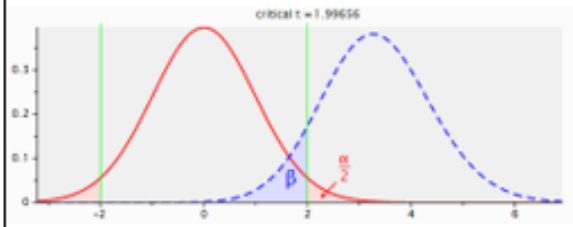
Съгласно описаните параметри ако се избере мощност на критерия 0.80 и се търси среден размер на ефекта (0.5), то минималният необходим обем за всяка извадка ще е 64. Ако обаче се зададе голям размер на ефекта (0.8), то тогава ще се изисква минимален необходим обем на всяка извадка – 26. Ако вместо 0.80 се зададе мощност на критерия 0.90 и отново се търси среден размер на ефекта, то минималният необходим обем на всяка извадка ще е 86. Следователно увеличаването на обема на извадката и/или размера на ефекта подобрява мощността на критерия и прецизността на анализа.

Мощност на критерия от 0.80 и статистическа значимост 0.05 е четири пъти по-вероятно да доведе до грешка от II род, отколкото грешка от I род. Според Cohen (1992a) обаче, за предпочитане е да се направи грешка от II род, отколкото грешка I род. Той отчита, че повечето от статиите, публикували в научни списания, показват средна мощност от 0.47, дори от 0.25, което показва, че вероятността за вземане на грешни решения е много висока. Cashen и Geiger (2004) откриват, че само 9.3% от 43 проучвания анализират мощността на критерия при тестването на хипотези. Така много изследователи не са наясно с мощността на критерия и неговото значение за изследванията. В случай, че не се използва подходяща мощност на критерия може да се увеличи риска от получаване на неверни заключения, като например да се приеме, че няма разлика между групите, когато различие наистина съществува (грешка II род).

И докато доскоро, при класическата проверка на хипотези, стоеше въпросът каква ще е мощността на критерия при първоначално зададения обем на извадката, то днес съществува друг по-приемлив подход. Този подход изисква първо да се определи желаната мощност на

критерия и размер на ефекта (малък, среден или голям), а след това да се търси минималния обем на извадката, при който това ще се получи.

Таблица 2. Обем на извадките при α = 0.05 и при различна мощност на критерия и различен размер на ефекта

1-β / d	0.8	0.9
0.2	 $n_1 = 394$ $n_2 = 394$ $n_1 + n_2 = 788$	 $n_1 = 526$ $n_2 = 526$ $n_1 + n_2 = 1052$
0.5	 $n_1 = 64$ $n_2 = 64$ $n_1 + n_2 = 128$	 $n_1 = 86$ $n_2 = 86$ $n_1 + n_2 = 172$
0.8	 $n_1 = 26$ $n_2 = 26$ $n_1 + n_2 = 52$	 $n_1 = 34$ $n_2 = 34$ $n_1 + n_2 = 68$

Заключение

В резултат на проведеното проучване се достига до извода, че при тестването на статистически хипотези (например за разликата между средни от две независими извадки), освен избраното равнище на значимост и докладването на статистическата значимост, от изключително значение е да се отчете и ефекта на размера и мощността на критерия. Използването на поне, оптимална мощност на критерия от 0.80 е важна за генерирането на прецизни научни знания. По-високата мощност на критерия намалява риска да не се отхвърли нулевата хипотеза, когато тя не е вярна. От друга страна при достатъчно голям размер на извадката, дори и много малък размер на ефекта може да доведе до статистически значими разлики (p-value $\leq$ 0.05), а резултатът на практика да е нищожен и маловажен. Затова се препоръчва при публикуване и тълкуване на резултати в научни статии, освен статистическата значимост, да се отчита и ефекта на размера и мощността на критерия. Това ще даде възможност на бъдещите изследователите да внесат друга перспектива в анализа и тълкуването на научни резултати.

References

1. **American Educational Research Association**. (2006). *Standards for Reporting on Empirical Social Science Research.* Educational Researcher, 35 (6), 33-40. doi: 10.3102/0013189X035006033
2. **American Psychological Association**. (1994). Publication Manual of the American Psychological Association (4th ed.). Washington, DC: Author.
3. **American Psychological Association**. (2001). Publication Manual of the American Psychological Association (5th ed.). Washington, DC: Author.
4. **American Psychological Association.** (2010). *Publication Manual of the American Psychological Association*. Washington, DC: Author.
5. **Anderson S. Kelley K, & Maxwell S.** (2017). *Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty*. Psychological Science, 28: 1547–1562.
6. **Camacho-Sandoval, J.** (2007). *Potencia estadistica en el diseno de estudios chnicos*. Acta Medica Costarricense, 49 (4), 203-204.
7. **Cashen, L. H., & Geiger, S. W.** (2004). *Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies*. Organizational Research Methods, 7(2), 151-167. doi: 10.1177/1094428104263676
8. **Clark, M. L.** (2004). *Los valores P y los intervalos de confianza: ^en que confiar?*. Revista Pan-americana de Salud Pdblica/Pan American Journal of Public Health 15(5), 293-296. doi: 10.1590/S1020-49892004000500001
9. **Cohen, J.** (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum Associates.
10. **Cohen, J.** (1994). *The Earth is Round (p < .05)*. American Psychologist. 49(12), 997-1003. doi: 10.1037/0003-066X.49.12.997
11. **Cohen, J.** (1992a). *A power primer*. Psychological Bulletin, 112(1), 155-159. doi:10.1037/0033-2909. 112.1.155
12. **Cohen, J.** (1992b). *Statistical power analysis*. Current Directions in Psychological Science, 1(3), 98-101. Retrieved from http://www.jstor.org/stable/20182143. doi:10.1111/1467-8721.epl0768783

13. **Cumming, G.** (2014). *The new statistics: Why and how*. Psychological Science, 25. doi: 10.1177/0956797613504966
14. **Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A.** (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.
15. **Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G.** (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.
16. **Ferguson, C. J.** (2016). *An effect size primer: A guide for clinicians and researchers*. In A. E. Kazdin (Ed.), Methodological issues and strategies in clinical research, pp. 301–310.
17. **Kirk, R. E.** (2001). *Promoting good statistical practices: Some suggestions*. Educational and Psychological Measurement, 61(2), 213-218. doi: 10.1177/00131640121971185
18. **Lakens, D.** (2013). *Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs.* Frontiers Psychology 4:863.doi: 10.3389/fpsyg.2013.00863
19. **McLeod, S. A.** (2019, July 10). What does effect size tell you? Simply psychology: Available at: https://www.simplypsychology.org/effect-size.html (Accessed: 10 September 2022)
20. **Norcross J., Hogan T., Koocher G., & Maggio L.** (2017). *Clinician's guide to evidence-based practices: Behavioral health and addictions* (2nd ed.). New York: Oxford.
21. **Rosnow, R. L., & Rosenthal, R.** (1989). *Statistical procedures and the justification of knowledge in psychological science*. American Psychologist, 44(10), 1276. doi: 10.1037/0003. 066X.44.10.1276
22. **Schmidt, R, & Hunter, J.** (2002). A*re there Benefits From NHST?* American Psychologist, 57. doi: 10.1037/0003-066X.57.1.65
23. **Téllez A., García C.H., Corral-Verdugo V.** (2015). *Effect size, confidence intervals and statistical power in psychological research*. Psychology in Russia: State of the Art, 8(3), 27-47.
24. **Wasserstein R., Lazar, N.** (2016) The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, pp. 129-133, DOI: 10.1080/00031305.2016.1154108
25. **Wilkinson, L. & APA Task Force on Statistical Inference.** (1999). *Statistical methods in psychology journals: Guidelines and explanations*, American Psychologist, 54(8), 594. doi:10.1037/0003-066X.54.8.594