АНАЛИЗ НА СТАТИСТИЧЕСКА МОЩНОСТ: ЗА И ПРОТИВ

Стефан Матеев

Нов български университет

Разработването на статистически методи за третиране на данни вече има почти стогодишна история. На пръв поглед нещата изглеждат ясни – изследователят подбира подходящ дизайн, и третира данните от изследването си със съответния стандартен метод за статистически анализ. Остава въз основа на резултатите от анализа да се формулират твърдения и становища относно обекта на изследване. Но развитието на математическия апарат за статистически анализ в последните години показва, че нещата не са толкова прости. Особено след публикуването на фундаменталния труд на Cohen (1988) става ясно, че целите и намеренията на изследователя следва да се анализират и *преди* да е предприел изследването. Именно в това се състои т.н. анализ на статистическата мощност. Целта на настоящата работа е да разгледа най-общо анализа на мощността, като разкрие неговите силни и слаби страни. В текста е използвано, като работен пример, съвсем просто изследване в областта на психологията и поведенческите науки, но разглежданията лесно могат да се обобщят и за други клонове на науката.

Нека си представим, че имаме за задача да определим дали средната стойност на характеристиката X на лицата от дадена популация се различава от дадена фиксирана норма C. Популацията може да се състои от лицата от даден етнос, или от тези, които живеят в дадена географска област, или тези, които имат определено ниво на образование. Предполагаме, че характеристиката X може да бъде измерена в рамките на интервална скала или скала на отношения за всеки член на популацията с достатъчно висока точност. X може да е горна граница на кръвно налягане, измерена в mmHg, ръст, измерен в cm, или ниво на агресивност, измерено с подходящ психологически тест.

Най-добрият начин да изпълним задачата е да измерим X за *всички* лица от популацията, и да пресметнем средната стойност на измерванията. Тази средна стойност се обозначава със символа $\mu(X)$ и се нарича популационна средна. За съжаление, обикновено този начин е неосъществим, поради големия брой на лицата в популацията и поради ограничените ни ресурси. Принудени сме да изследваме ограничена по обем група лица, която наричаме извадка (sample), и от данните за

извадката да направим заключение, колко е популационната средна стойност μ, и дали тя се различава от фиксираната норма C.

Методът, чрез който се формулира заключението, изисква основни познания относно този дял от статистиката, който се нарича статистически извод. Предполагаме, че читателят има такива познания и може да продължи да чете нататък. Добър литературен източник на български език по тези въпроси е Калинов (2013).

Изключително популярен е следният начин на разсъждение. Предполагаме, че има две възможности: $\mu = C$ или $\mu \neq C$. Първата възможност се нарича нулева хипотеза, втората – алтернативна хипотеза. Питаме се, коя от двете хипотези е вярна? Точен отговор на този въпрос не може да се даде. Разработен е метод, с който може да се пресметне вероятността да грешим, ако твърдим, че $\mu \neq C$, но в действителност $\mu = C$. Тази вероятност се обозначава със символа p и се нарича вероятност за грешка от първи вид. Това е вероятност за грешка, подобна на „лъжлива тревога“ (виж приказката за лъжливото овчарче) – твърдим, че популацията се различава средно взето от нормата C, а пък в действителност тя не се различава.

Разсъждението продължава по следния начин. Предварително си избираме една число, една вероятност, наричана ниво на значимост (significance level) и обозначавана със символа α. В огромното многообразие от изследвания най-популярното число е $\alpha = 0.05$. Изследователите са се споразумели, че ако пресметнатата стойност на p е по-малка от 0.05, вероятността за грешка от първи род (т.е. за лъжлива тревога) е достатъчно малка, и алтернативната хипотеза може да се приеме. Тогава е прието за се казва, че разликата между средната стойност на характеристиката X в извадката, m(X), и нормата C е *значима*. Получаването на „значимост“ обикновено се приема като подкрепа на твърдението, че разлика между μ и C има, и че тя представлява практически интерес. Както ще покажем по-надолу, тези твърдения може да се окажат най-малкото необосновани, и даже грешни.

Но пресмятанията може да покажат, че вероятността за грешка от първи род е по-голяма от приетото число $\alpha = 0.05$. Вербално този резултат се изразява с думите „разликата между m(X) и C не е значима, $p > 0.05$“. Строгото разсъждение е, че няма основания да се отхвърли нулевата хипотеза $\mu = C$. За съжаление, липсата на значимост масово се интерпретира като „няма разлика между μ и C“. Това е логически невярно.

Липсата на основание за наличие на разлика не е наличие на основание за липса на разлика. Статистическият извод, поне във вида, в който се описва в много учебници, не дава решение на този проблем. Преподавателите обикновено се задоволяват с твърдението, че нулевата хипотеза не може да се потвърди.[1]

Когато се планира изследване, обичайно е да се формулират научни хипотези. В преобладаващия брой случаи, научните хипотези са предсказания за наличието на някакъв ефект. Подразбира се, или даже директно се формулира, че при определени манипулации на независимите променливи, зависимата променлива се очаква да претърпи *значими* промени. С други думи, очаква се прилагането на статистически тест да доведе до стойности на грешката от първи вид по-малки от 0.05, или от друго предварително избрано ниво на значимост α. Получаването на „значимост“ обикновено се приема като подкрепа на научната хипотеза. По този начин, успешното отхвърляне на нулевата хипотеза се превръща в залог за успешна защита на тезис или потенциално успешно публикуване на статия. Значимите резултати създават впечатлението, че изследователят правилно е анализирал изследователския проблем и е предсказал интересни (т.е. „значими“) резултати. „Негативният“, т.е. незначим, резултат от изследването често може да се интерпретира като плод на неадекватен анализ на научния проблем или на недостатъци на изследването. Не е чудно, че негативните резултати се възприемат като неуспех, или като безинтересни. Ако става въпрос за статия, тя може да остане непубликувана (или „остава в чекмеджето“, file drawer problem), въпреки иначе добре проведеното изследване. Този факт е известен отдавна (Rosenthal, 1979) и е дискутиран в литературата.

Опитният изследовател или научен ръководител добре знае, че за получаването на „значими“ резултати са необходими извадки с голям обем, или брой n лица. Това твърдение не се нуждае от математически доказателства. Както отбелязахме по-горе, ако характеристиката X се измери при *всичките* N лица от популацията, тогава средната стойност μ ще бъде определена точно, и на въпроса, дали тя се различава или не от нормата C, ще се даде точен отговор. Опасност от лъжлива тревога няма, нейната вероятност p е нула. Вероятността за грешка от първи род нараства прогресивно,

[1] Разработени са специални методи за статистическа подкрепа на нулевата хипотеза. Това са т.н. тестове за еквивалентност (например Schuirman, 1987, Cribbie et al., 2004), които не са предмет на настоящия текст.

когато обемът на извадката n значително намалява спрямо броя на лицата N в популацията. При недопустимо малки извадки, p може значително да надвиши избраното ниво от 0.05. Именно статистическият извод е необходим за точното пресмятане на вероятността p. Но какво е „недопустимо малък“ обем на извадката? Отговорът на този въпрос е свързан с понятието *статистическа мощност*. За да изясним понятието мощност, и начините за нейното пресмятане, трябва да въведем някои понятия.

Преди всичко грешката от първи вид не е единствената, която можем да допуснем, когато анализираме данните от изследване. Възможен е следният изход от статистическия анализ. Възможно е популационната средна да е различна от нормата, т.е. $\mu \neq C$, но в резултат от пресмятанията да се получи $p > 0.05$, т.е. резултатът да не е „значим“. Без допълнителни разглеждания, този резултат може погрешно да се интерпретира като „няма разлика между μ и нормата C“. Този вид грешка се нарича грешка от втори вид. Тя е вид пропуск - има ефект в популацията, но ние не можем да го установим. В литературата може да се срещне понятието „неуспех да се открие ефект“ (failure to detect effect). Всъщност, „откриването“ на ефекта се състои в това, да се получи „значимост“, т.е. да получим $p < 0.05$. Грешка от втори вид може да възникне с определена вероятност, която се означава със символа β. Ако е известна тази вероятност, можем лесно да пресметнем вероятността да получим „значимост“, когато действително $\mu \neq C$ в популацията. Това е вид успех - в популацията има разлика между μ и C, и ние демонстрираме, че тя е значима. Вероятността за този успех е просто $1- \beta$, и именно тази вероятност се нарича мощност (power). С други думи, *мощност е вероятността да отхвърлим нулевата хипотеза, при условие, че алтернативната е вярна*. Когато се дефинира понятието мощност, в литературата се говори за мощност на статистическия тест или за мощност на изследването. Смятаме, че второто наименование е по-правилно, тъй като мощността се определя не само от прилагания статистически тест, но и от начина на провеждане на изследването. Все още няма утвърден символ за мощност, използва се или думата power, или $1- \beta$. Кои фактори влияят върху мощността?

Първият фактор е доста тривиален. Твърдението, че една разлика е „значима“ се изказва, когато при пресмятанията се получи стойност на вероятността за грешка от първи вид p, която е по-малка от избраното ниво на значимост α. Стойността на p не

зависи от α, но становището за значимост зависи от α. Например, ако пресметнатата стойност на p е 0.04, разликата се обявява за значима при $\alpha = 0.05$, но тя не е значима при $\alpha = 0.01$. Следователно, ако предварително сме избрали да работим с ниска стойност на α, обикновено 0.01 вместо 0.05, трябва да се примирим с по-ниската вероятност да декларираме, че „разликата е значима“. С други думи, мощността намалява с намаляването на нивото на значимост α.

Тук има една тънкост, която е свързана с начина, по който е формулирана алтернативната хипотеза. Когато тя е $\mu \neq C$, пресмятаме т.н. двустранен (two-tailed) статистически тест. Ако тя е формулирана едностранно, като $\mu > C$ или $\mu < C$, пресмятаме т.н. едностранен (one-tailed) тест. При повечето статистически тестове алтернативната хипотеза може да се формулира както едностранно, така и двустранно. Формулировката зависи от съображения на изследователя, които тук няма да обсъждаме. Факт е, че един и същ масив от данни може да се анализира както с едностранен, така и с двустранен тест, и винаги при едностранния тест вероятността за грешка p се получава два пъти по-ниска, отколкото при двустранния. Но нивото на значимост е фиксирано, следователно при едностранен тест нулевата хипотеза се отхвърля по-лесно и съответно мощността е по-висока[2].

Следващият фактор, който влияе върху мощността, е обемът n на извадката. При изследване на една и съща популация, и при една и съща алтернативна хипотеза, например $\mu \neq C$, по-големият обем на извадката неизбежно води до по-ниски стойности на вероятността p, и при достатъчно голям обем n може да се надяваме, че ще получим $p < \alpha$ и основания да декларираме, че разликата е значима.

---

[2] Пример: приели сме $\alpha = 0.05$, пресмятането на p с двустранен тест дава 0.07, имаме $p > 0.05$, т.е. разликата между m(X) и C не е значима. Не сме доволни от този извод, и прилагаме едностранен тест. Получаваме $p = 0.035$, което е по-малко от 0.05 и обявяваме разликата за значима. Glenberg (1996, стр. 172) пише, че подобен „трик“ е „забранен“, и че алтернативната хипотеза трябва да се формулира предварително, преди да започне изследването. Това е слаб аргумент. Проблемите при статистическия анализ на данни са не толкова математически, колкото логически. А в логиката последователността, в която се формулират предпоставките, не влияе върху крайния извод.

Последният фактор, който влияе върху мощността, е т.н. *сила на ефекта в популацията* (population effect size). Този фактор изисква по-внимателно разглеждане. В предишна наша работа (Матеев, 2014) разгледахме понятието сила на ефект (effect size, ES) като вид описателна статистика, като индекс, който ни позволява да оценим големината и практическата важност на една емпирично определена разлика. За примера, който все още използваме, силата на ефекта се пресмята като

$$d = |m(X) - C|/s \qquad (1)$$

където m(X) е средната стойност на измерванията X на отделните лица в извадката, C е фиксираната норма, s е стандартното отклонение на разпределението на X. Символът d използваме тук вместо ES (effect size) (виж Матеев, 2014). В много текстове силата на ефекта се обозначава с Cohen's d, или само с d.

Ако m(X) и C съвпадат, d = 0, и казваме, че средно взето лицата в извадката не се отличават от нормата C. Ако s остава едно и също, прогресивното нарастване (или намаляване) на m(X) води до нарастване на стойността на d, което може да бъде (теоретически) до безкрайност. Забележете, че намаляването на s също води до нарастване на d. Величината |m(X) – C| се нарича суров ефект (raw effect size), тя се измерва в единиците на измерване на X – сантиметри, секунди или точки от психологически тест. Стандартното отклонение s се измерва в същите единици, така че d е безразмерна величина. Тя се нарича още стандартизиран ефект (standardized effect size). Чрез равенство (1) се пресмята броят на стандартните отклонения, с които m(X) и C се различават една от друга. Пресмятането е валидно, ако разпределението на стойностите X е поне близо до нормалното.

Ролята на индекса d при пресмятането на значимостта на разликата m(X) – C може лесно да се демонстрира. Използваме t-тест във вида

$$t = |m(X) - C|/(s/\sqrt{n}) \qquad (2)$$

където n е броят на лицата в извадката, а останалите символи в дясната страна на (2) са същите, както и в равенство (1)

Ако полученото t надхвърли една критична стойност, определена от нивото на значимост α и от степените на свобода, то декларираме, че разликата |m(X) – C| е значима (или по-строго, значимо различна от нула). Съвременните статистически пакети директно пресмятат стойността на вероятността за грешка от първи вид p.

Колкото по-висока е стойността на t, толкова по-малка е вероятността p. С проста аритметика равенство (2) може да се преобразува като

$$t = (|m(X) - C|/s)*\sqrt{n} = d*\sqrt{n} \quad (3)$$

Получаваме, че t, или „значимостта“ на разликата, е функция на произведението на силата на ефекта d и на √n. При по-сложни дизайни d се пресмята по по-друг начин, и величината под корена е по-сложна, но *винаги* грешката от първи вид p намалява при по-силни ефекти d и при по-голям обем на извадката (или извадките).

Дефиницията на силата на ефекта в популацията е:

$$\delta = |\mu(X) - C|/\sigma \quad (4)$$

където μ(X) е средната стойност на измерванията X на отделните лица в популацията, C отново е фиксираната норма, σ е стандартното отклонение на разпределението на X в популацията. Следвайки традицията популационните параметри да се обозначават с гръцки букви, в равенство (4) използваме символа δ вместо d. Равенство (1) се пресмята от емпиричните данни, то е описание на това, което се получава от изследването. За стойността на δ може само да предполагаме въз основа на лични впечатления и на литературни данни. Но ефектът d, който наблюдаваме в извадката, отразява силата на ефекта в популацията δ. Ако предполагаемата стойност на δ е висока, можем да се надяваме на висок ефект d и в извадката, и съответно на успешно отхвърляне на нулевата хипотеза. С други думи, мощността е по-висока, когато силата на ефекта в популацията δ също е висока.

От горните разглеждания се вижда, че мощността, като вероятност да отхвърлим нулевата хипотеза, се определя от: нивото на значимост α, обемът на извадката n, и силата на ефекта в популацията δ. Към тях може да добавим и намерението ни да приложим едностранен или двустранен статистически тест. Този фактор не зависи от статистически съображения, и няма да го разглеждаме нататък.

Мощност, α, n и δ са величини, които са математически свързани. Ако знаем три от тях, можем да изчислим четвъртата. Нека разгледаме приложението на тези изчисления за примера, който използваме.

Изчисляване на α няма особен смисъл. Обичайната стойност е $\alpha = 0.05$ и становището за значимост се основава на сравнението на p с 0.05.

Нека си представим, че възнамеряваме да проведем изследване, като разполагаме с извадка от n лица. Знаем, че мощността нараства с нарастването на n, но не можем да си позволим по-голям брой лица поради недостиг на ресурси, поради етични причини, и др. Задаваме си въпроса, може ли с този брой лица да разчитаме, че ще отхвърлим нулевата хипотеза $\mu = C$ (при $\alpha = 0.05$) ако е вярна алтернативната? Думите „да разчитаме“ следва да се прецизират, именно като „каква е вероятността да получа $p < 0.05$ с този брой лица“ или най-точно, „каква ще е мощността на това изследване“? Ако пресмятането покаже, че мощността е например 0.9, усилието да проведем изследването си струва, почти сигурно е, че ще отхвърлим нулевата хипотеза. Но ако мощността се окаже 0.3, може би въобще не си струва да започваме! За пресмятането е нужно още нещо. Трябва да определим една минимална сила на ефекта в популацията, която представлява интерес. Тук използваме термина целеви ефект (target effect, Cumming, 2012), и го обозначаваме с δ_C. С прости думи, предполагаме, че в популацията съществува някакъв ефект, но ако неговата сила е по-ниска от дадена стойност δ_C, то той не представлява интерес, и не си струва да се занимаваме с неговото „откриване“.

Колко да бъде този минимален ефект δ_C, който трябва да заложим в изчисленията? Cohen (1988) е решил въпроса, като е предложил репери за слаб, среден и силен ефект. Тези репери са определени от него въз основа на преглед на статии в областта на психологията на консултирането и образованието. Слаб ефект е 0.2, среден 0.5 и силен е 0.8. Тези репери не са абсолютни, може да се приемат други стойности от литературата. Но при липса на други сведения те вършат работа. Да допуснем, че броят на лицата в извадката ни е 50, и че целевият ефект, който представлява интерес за нас е среден, $\delta_C = 0.5$. Вече имаме $\alpha = 0.05$, $n = 50$ и $\delta_C = 0.5$, и мощността може да се пресметне. Faul et al. (2007, 2009) създадоха програмата G*Power 3, която е свободна за изтегляне от www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/, и е сравнително проста за обслужване. В Допълнение 1 е представен начинът за пресмятане на мощността с G*Power 3 за настоящия пример. Препоръчваме на читателя и програмата ESCI (Cumming, 2012). С нея може да се пресмятат както мощност, така и други важни величини, като доверителни интервали. Тя е свободна за изтегляне от www.thenewstatistics.com.

Пресмятането на обем на извадка е от особена важност при изследването. Нека отново сме заложили $\alpha = 0.05$ и $\delta_C = 0.5$. За да определим обема на извадката, трябва да заложим още един параметър, това е желаната мощност. В неговия фундаментален труд Cohen (1988) препоръчва тя да е 0.8. Имаме три параметъра на изследването, можем да пресметнем четвъртия. В Допълнение 2 е показано как става това.

Накрая, използвайки същата програма, можем да определим минималния популационен ефект δ_C, който можем да „открием“ с вероятност, зададена от мощността (например 0.8), наличния брой от 50 лица и $\alpha = 0.05$. Това е показано в Допълнение 3.

Тези анализи се провеждат (или по-скоро следва да се провеждат*) преди* изследването. Те се наричат проспективен анализ на мощността (prospective power analysis, a priori power analysis). От тях най-важният може би е определянето на обема на извадката. Той е от съществено значение при планиране на предстоящото изследване и при изготвяне на проекти за финансиране.

Анализ на мощността може да се провежда и след като изследването е приключило и данните са налице. Такъв анализ се нарича ретроспективен (retrospective power analysis, или post hoc power analysis). Налице са стойностите на α и обемът на извадката n. Вместо обаче, да се допуска определена целева (наричана още независима) стойност на популационната сила на ефекта δ_C, при изчисленията се залага силата на ефекта d, пресметната от данните чрез равенство (1). До такъв анализ обикновено се прибягва, когато нулевата хипотеза не се отхвърля. Някои автори смятат, че ако ретроспективната мощност се получи висока, това е довод в подкрепа на нулевата хипотеза. В последните години беше показано, че това не е вярно (Nakagawa & Foster, 2004, Baguley, 2004). Най-солидният довод против ретроспективния анализ на мощността представиха Hoenig & Heisley (2001). Те доказаха строго, че ретроспективната мощност е функционално, едно към едно, свързана с вероятността за грешка от първи вид p. С други думи, ретроспективната мощност носи точно толкова информация, колкото и p, и както едната, така и другата величина не може да послужи за потвърждаване на нулевата хипотеза. Общото мнение в литературата понастоящем е, че ретроспективен анализ на мощността *не следва* да се провежда и представя. Съветваме читателя да се съобразява с това мнение и да не дава повод за критики, че некомпетентно анализира данните си. Обръщаме внимание, че в статистическия пакет

SPSS има опция за пресмятане на “observed power”. Това е именно ретроспективната мощност, чието споменаване в статия би направило лошо впечатление.

Читателят може да остане с впечатление, че анализът на мощността решава проблема с постигането на значим резултат при провеждането на изследването. За съжаление, това не е така. Прието е, и това дължим отново на Cohen (1988), при пресмятането на обема на извадката да се залага предварително power = 0.8. Това число е конвенция, подобна на конвенцията $\alpha = 0.05$ и общо взето, е приета от повечето автори. За да се пресметне обемът на извадката n, необходимо е да се приеме, или допусне, сила на ефекта δ_C в популацията, която представлява някакъв практически интерес. Тогава, в резултат на анализа, може да се твърди, че ако *точно* такъв ефект δ_C действително съществува в популацията, с извадка от n лица вероятността да се отхвърли нулевата хипотеза е равна на 0.8. При тези обстоятелства, вероятността за грешка от втори вид е $\beta = 1 – 0.8 = 0.2$. Толкова е вероятността да *не* отхвърлим нулевата хипотеза, когато алтернативната е вярна. Вероятността 0.2 не е малка, и не следва да се учудваме, ако при изследването не отхвърлим нулевата хипотеза, въпреки грижливото предварително пресмятане на обема на извадката. Освен това, при фиксиран обем на извадката, мощността зависи от избрания целеви ефект δ_C. Но реалният ефект δ в популацията може да е по-силен или по-слаб от избрания δ_C. Ако имаме късмет реалният ефект δ да е по-силен от δ_C, шансовете ни да отхвърлим нулевата хипотеза нарастват. Но ако реалният ефект се окаже по-слаб, тогава тези шансове намаляват. И ако статистическият тест покаже, че $p > 0.05$, не е лесно да се установи, дали това се дължи на сравнително високата вероятност за грешка от втори вид $\beta = 0.2$, заложена в проспективния анализ на мощността, или на това, че реалният ефект в популацията се е оказал по-слаб от δ_C. Някой би си помислил, че е възможно да се застрахова от подобна ситуация. Просто в анализа на мощността трябва да заложим power = 0.9 или даже 0.95, и целеви ефект достатъчно нисък, например $\delta_C = 0.1$. Но тогава ще се окаже, че броят на лицата в извадката трябва да е n = 1084 (използвайте Допълнение 2), което може значително да надхвърли ресурсите ни. Така изследователят се оказва в ситуация, в която е принуден да прави компромиси при планирането на изследването. Самото изследване се оказва вид облог, който се сключва с природата. Анализът на мощността представлява метод, с който могат да се оценяват шансове за успех или неуспех, и дали планираните средства и ресурси за бъдещото изследване са оправдани.

Директни доводи против провеждането на проспективен анализ на мощността е трудно да се намерят в литературата. Но трябва да се има пред вид, че самото понятие мощност е тясно свързано с анализа на значимостта на наблюдаваните ефекти. Помним, че мощност е вероятността да отхвърлим нулевата хипотеза при условие, че алтернативната е вярна. Всеки анализ има някаква цел, и целта на анализа на мощността е именно да осигури предпоставки за отхвърляне на нулевата хипотеза. И тук възникват възраженията.

В литературата е възприет термина null hypothesis significance testing, съкратено NHST. Този термин служи за обозначаване на всички методи, с които се пресмятат p-стойности. През последните две десетилетия NHST е критикуван все повече. Подробни текстове по този въпрос са Kline (2004) и Cumming (2012, 2014). За някои логически проблеми на NHST насочваме читателя към Матеев (2014). В критиките към NHST не става въпрос за дефекти или грешки при пресмятането на p-стойностите. Работата е там, че с течение на десетилетията у изследователите са се натрупали схващания и практики, които водят до неправилна интерпретация на данни и необосновани изводи въз основа на „значимости“. По-долу ще засегнем някои популярни случаи на злоупотреба с p-стойности. С курсив са отбелязани становища, които горещо съветваме читателя да избягва.

Високо значимите ефекти, т.е. тези, при които се получават стойности на p, значително по-ниски от 0.05 (например p = 0.00003), са интересни и важни. Това може да е вярно, но може и да не е вярно. Както се вижда от равенство (3), значимостта на ефекта зависи и от броя на изследваните лица в извадката. Възможно е ефектът да е нищожен по сила, но благодарение на големия обем на извадката той да е „високо значим“. Високата значимост не означава обезателно същественост.

Ако нулевата хипотеза не се отхвърли, това означава отсъствие на ефект. Отново това може да е вярно, но и да не е вярно. Ефектът d може и да е силен, но поради ограничения брой на лицата в извадката пресмятането на p да даде висока стойност, значително над 0.05 (равенство (3)).

Стойността на p носи информация за силата на наблюдавания ефект. Това може да е вярно само ако се сравняват резултати от две (или повече) изследвания, проведени с един и същ дизайн и с еднакъв обем на извадките. Тогава по-ниската стойност на p наистина показва, че в даденото изследване ефектът е по-силен.

Стойността на p сама по себе си не дава количествена оценка на силата на ефекта, каквато например, дава стойността на d.

Стойността на p дава вероятността за намеса на случайни, странични фактори в изследването. Не е вярно. Стойността на p може да варира от едно към друго изследване, в зависимост от това, на каква извадка от популацията сме попаднали. Този въпрос е великолепно илюстриран от Cumming (2012).

Kline (2004) посвещава 34 страници за изреждане на всички възможни погрешни изводи и становища, които могат да се направят въз основа на NHST. На стр. 43 той заключава, че „ако статистически тестове изобщо не се използват, то и анализът на мощност няма смисъл“.

В раздела „Статистика“ на „Напътствия към авторите“ на авторитетното списание Psychological Science можем да прочетем, че при анализа на данните се препоръчва използването на сила на ефекта и доверителни интервали „за да се избегнат проблемите, свързани с отхвърлянето на нулеви хипотези“. Но от друга страна, редакторите на списанието изискват от авторите да обяснят защо вярват, че използваните обеми на извадките са адекватни за целите на изследванията. Точни напътствия не се дават. Авторите биха могли да изтъкват различни доводи; от тях проспективният анализ на мощността (например, както е описан в Допълнение 2) определено би изглеждал неоспорим. Но понятието мощност е тясно свързано именно с отхвърлянето на нулеви хипотези, което редакторите съветват да се избягва…..

В заключение, въпреки критиките и възраженията срещу проверката на статистически хипотези като инструмент за анализ на данни, може би е твърде рано да се откажем от p-стойностите и съответно, от анализа на мощността. Най-малкото, тяхната теория следва да се преподава, за да могат студенти, преподаватели и изследователи да разбират текстовете на хилядите вече публикувани статии, в които резултатите се описват и интерпретират въз основа на p-стойности. Освен това, същата теория служи и при пресмятането на силата на ефекта и доверителните интервали, за които говорят редакторите на Psychological Science. Статистическият извод, във вида, в който всички ние сме го учили навремето, представлява инструмент за изказване на становище относно популация въз основа на данни от извадка. В много случаи такова становище има смисъл и е полезно. Това, което бихме посъветвали читателя, е да си

пресмята мощност и p-стойности, но да не разчита много на тях при *психологическата* интерпретация на данните.

ЛИТЕРАТУРА

Калинов, К. (2013). *Статистически методи в поведенческите и социалните науки*. Нов български университет, София (3-то издание)

Матеев, С. (2014). Същественост и значимост на резултати от научни изследвания в психологията. *Научен електронен архив на НБУ*, http://eprints.nbu.bg/2277/

Baguley, T. (2004). Understanding statistical power in the context of applied research, *Applied Ergonomics* **35**, 73–80

Cribbie, R.A., Gruman, J.A. & Arpin-Cribbie, C.A. (2004). Recommendations for Applying Tests of Equivalence. *Journal of Clinical Psychology*, **60**, 1–10

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Ass., Hillsdale, New Jersey

Cumming, G. (2012). *Understanding the new statistics. Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, **25**, 7-29

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

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses, *Behavior Research Methods*, **41**, 1149-1160

Glenberg, A. (1996). *Learning from data. An introduction to statistical reasoning*. Lawrence Erlbaum Ass., Mahwah, New Jersey

Hoenig JM, Heisley DM (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, **55,** 19–24

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Nakagawa, S. & Foster, T.M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica,* **7,** 103–108

O'Keefe, D.J. (2007). Post hoc power, observed power, a priori power, retrospective power, achieved power: sorting out appropriate uses of statistical power analyses. *Communications Methods and Measures*, **1**, 291-299

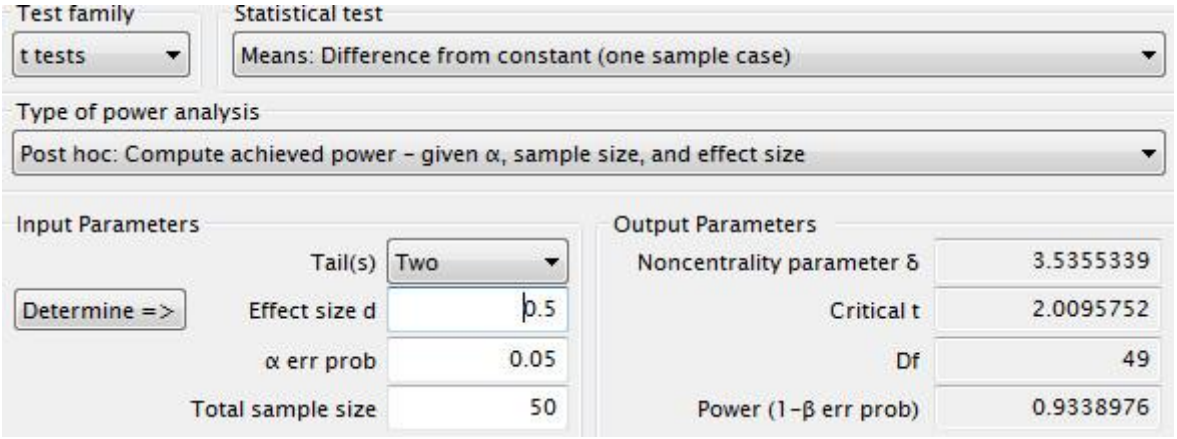
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, **86**, 638-641

Schuirman, D.J. (1987). A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657-679

ДОПЪЛНЕНИЯ: работа с програмата G*Power 3

Допълненията са посветени на предварителен (prospective) анализ на възможностите за проверка на алтернативна хипотеза от вида $|\mu(X) \neq C|$ при ниво на значимост $\alpha = 0.05$ и прилагане на двустранен t-тест (равенство 3).

ДОПЪЛНЕНИЕ 1: Определяне на мощност 1- β при дадени обем на извадка n и желана целева сила на ефекта δ_C



Фиг. 1.

На фиг 1 са показани по-важните елементи от дисплея на G*Power 3. В падащия прозорец “Test family” се появява “t tests” по подразбиране. Внимание, падащият прозорец “Tail(s)” по подразбиране е “one”, сменяме го с “two”. От прозореца “Statistical test” избираме “Means: difference from constant (one sample case)”. От прозореца “Type of power analysis” избираме “Post hoc: Compute achieved power – given α, sample size and effect size”. Не се смущаваме от думите “Post hoc” в прозореца. Пресмятането на мощността е математически едно и също, независимо от това, дали го правим преди или след изследването.

В клетката “Effect size d” записваме целевия ефект δ_C. За примера в текста сме избрали той да е той е 0.5. Записваме стойността $\alpha = 0.05$ и наличния обем на

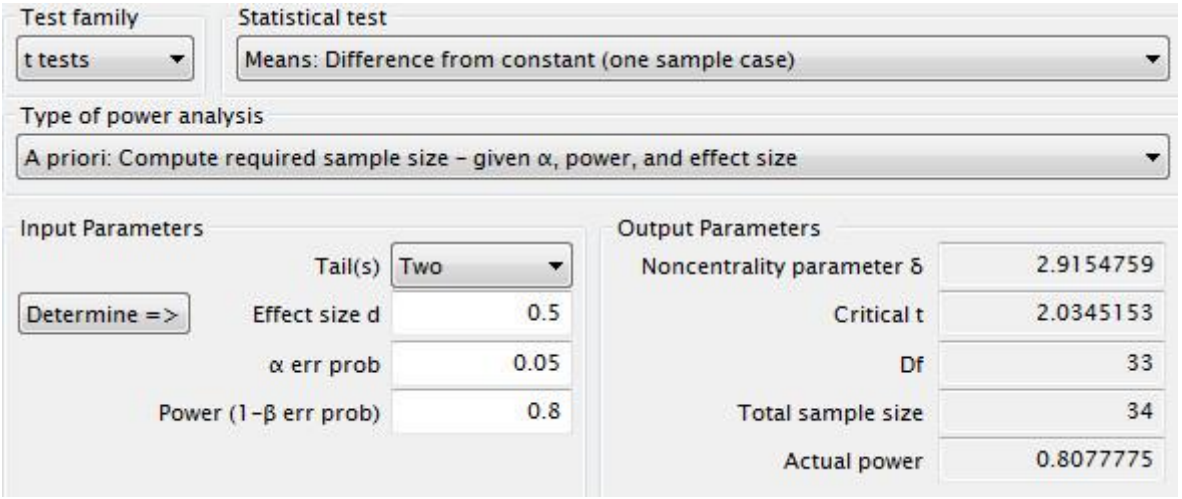
извадката n = 50, с който разполагаме, в клетката “Total sample size”. Следва Enter (или бутон Calculate).

Получаваме резултатите. Мощността 1 - β = 0.933. Толкова е вероятността да отхвърлим нулевата хипотеза при условие, че ефектът в популацията е със сила 0.5. Това е висока вероятност, и си струва да проведем изследването. Но ако ще публикуваме изследването в списание от нивото на Psychological Science, ще се наложи да обясняваме защо избираме толкова силен целеви ефект. За упражнение съветваме читателя да променя стойностите на δ_C, α и n, и да се убеди, че когато всяка от тях нараства, мощността също нараства.

За сведение на читателя: т.н. нецентрален параметър не е нищо друго освен стойността на статистиката t, която се получава от равенство (3). Ако имаме късмета в извадката да се получи сила на ефекта d точно 0.5, то t = 0.5*√50 = 3.53, колкото го дава програмата. Критичното t = 2.01 е същото, което може да се определи от таблица при α = 0.05 и степени на свобода df = 50 - 1 = 49

ДОПЪЛНЕНИЕ 2: Определяне на обем на извадка n при дадени желана целева сила на ефекта δ_C и желана мощност 1 - β

Приемаме желана сила на целевия ефект отново 0.5 и желана мощност от 0.8. От падащия прозорец “Type of power analysis” избираме “A priori: Compute required sample size – given α, power and effect size”. Записваме данните за желаните от нас стойности на δ_C = 0.5, α = 0.05 и 1 - β = 0.8, следва Enter и получаваме (фиг.2), че са ни необходими 34 лица в извадката.



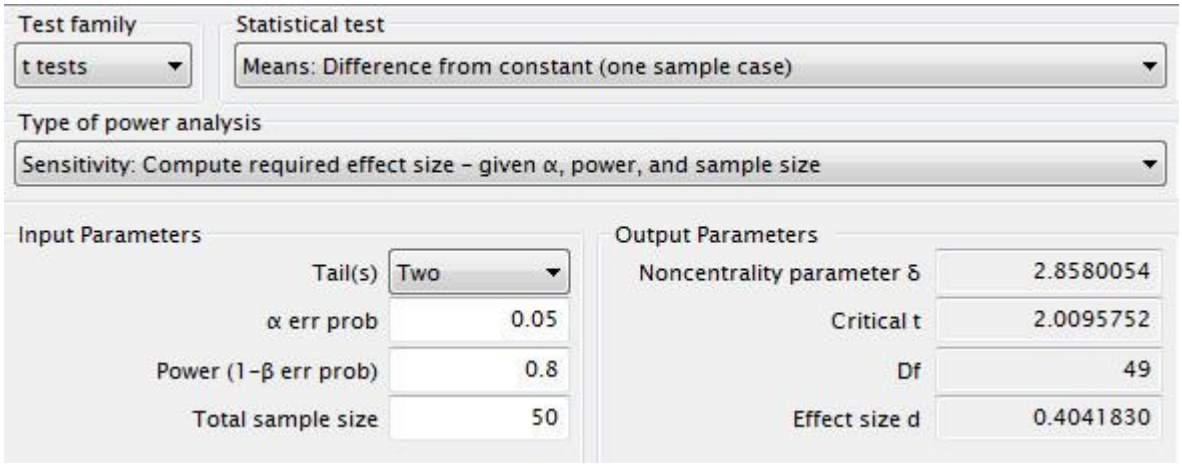
Фиг. 2

Програмата дава и стойност на “actual power”, тя не е винаги съвсем същата както желаната от 0.8. Така е защото авторите на програмата предпочитат да дават броя на лицата с цяло число, и това налага „напасване“ на мощността. Actual power е и меродавната стойност на мощността при 34 лица в извадката.

ДОПЪЛНЕНИЕ 3: Очаквана целева сила на ефекта δ_C при даден обем на извадка n и желана мощност 1- β

Тук става въпрос за т.н. „чувствителност“ на изследването. Ако желаната мощност е например 0.8, и броят на лицата в извадката е например 50, то колко ще е силата на популационния ефект δ_C, който ще се „открие“ с вероятност 0.8?

От падащия прозорец “Type of power analysis” избираме “Sensitivity: Compute required effect size – given α, power, and sample size”. Записваме $\alpha = 0.05$, $1 - \beta = 0.8$ и $n = 50$, и програмата пресмята $\delta_C = 0.404$ (фиг. 3)



Фиг. 3.

Следователно, можем да твърдим, че ако популационният ефект е със сила точно 0.4, вероятността да установим (с извадка от 50 лица), че той е значим, е 0.8. Стойността от 0.4 не е кой знае колко впечатляваща, тя е близо до среден ефект. Увеличете обема на извадката на 100 души, и ще получите ефект от 0.28. Като минимален ефект, който представлява интерес, това число изглежда малко по-добре.