СЪЩЕСТВЕНОСТ И ЗНАЧИМОСТ НА РЕЗУЛТАТИ ОТ НАУЧНИ ИЗСЛЕДВАНИЯ В ПСИХОЛОГИЯТА

Стефан Матеев

§ 1. Статистическият извод: логически проблеми

Ще започнем изложението с една типична ситуация, с която психологът се сблъсква в своята практика. Нека даден психологически инструмент, тест или въпросник, е приложен на две групи лица, група 1 и група 2. Инструментът измерва определена характеристика X на лицата. Групите 1 и 2 са подбрани така, че те да принадлежат към различни популации, например различни етноси, професии, възрасти, и т.н. Задачата на психолога е да определи, дали популациите се различават по отношение на характеристиката X.

В течение на десетилетия тази задача е решавана по начин, известен на всеки първокурсник по психология като *статистически извод*. Формулираме две хипотези: нулева и алтернативна. Нулевата е, че популационните средни стойности μ_1 и μ_2 са еднакви, а алтернативната – че μ_1 и μ_2 са различни. След това, провеждаме (вече с компютър) подходящ статистически тест. Програмата пресмята едно магическо число, обозначавано с *p*, което ни дава вероятността, с която грешим, ако приемем алтернативната хипотеза. Ако това число е по-малко (обикновено) от 0.05, то приемаме алтернативната хипотеза. Казваме, че разликата между средните стойности на група 1 и група 2 е *значима*, и това ни дава повод за обширни разсъждения относно популациите, към които принадлежат групите. Ако се получи $p > 0.05$, то казваме, че разликата между двете средни *не е значима*, и изпадаме в малко неудобно положение. Възможно е, популациите наистина да не се различават по отношение на характеристиката X. Но вече и първокурсниците знаят, че $p > 0.05$ не е повод за такова твърдение. С други думи, липсата на доказателство за наличие на разлика не е наличие на доказателство за липса на разлика. Но от друга страна, разглеждането на данните понякога показва, че груповите средни стойности m_1 и m_2 съвсем не са еднакви, и че разликата между тях може и да представлява някакъв интерес....

Тук следва да се замислим върху въпроса, защо въобще се налага да провеждаме статистически тестове, като например t-теста. „Най-чист" отговор на въпроса, дали μ_1 и μ_2 са различни или не, може да се даде, ако приложим въпросника на *всички* лица от двете популации и пресметнем средните стойности на получените данни. Тогава няма нужда от никакъв статистически тест, за да сравним μ_1 и μ_2. Проблемът е, че нямаме възможността - нито времето, нито ресурсите, да изследваме целите популации. Налага се да извлечем две ограничени на брой групи (наричани извадки) от всяка популация, да разгледаме средните им стойности m_1 и m_2, и да направим извод от типа: „наблюдавам, че m_1 и m_2 се различават, следователно μ_1 и μ_2 *може би* също се различават". Изводът е малко особен - в логиката, завещана ни от Аристотел, „може би" няма. В случая, при едни и същи предпоставки изводът може да се окаже както верен, така и неверен. Именно за това е необходим статистическият тест – с него

пресмятаме вероятността, с която грешим, когато изказваме определено твърдение относно популациите. По-специално, стойността на *p* ни дава вероятността за нещо като „лъжлива тревога“ – да твърдим, че μ_1 и μ_2 са различни, когато те всъщност са еднакви. Психолозите (а и не само те) по света са се споразумяли, че ако тази вероятност е по-голяма от 0.05 (стойност, предложена на времето от големия статистик сър Роналд Фишер), то твърдението, че μ_1 и μ_2 са различни, вече не е обосновано и не следва да се изказва.

Следвайки традицията данните да се анализират *само* чрез статистически тестове, стигаме да едно разделение на видовете резултати, които могат да се получат от едно изследване: разликите, които наблюдаваме, са или значими ($p<0.05$) или незначими ($p>0.05$). На пръв поглед, тази дихотомия е полезна; тя би означавала, че разсъжденията може да се провеждат според правилата на формалната логика, в която твърденията са или верни, или неверни. За съжаление не е така. Не трябва да забравяме, че „(не)значимостта“ на един резултат се определя от стойността на някаква ненулева вероятност. Това може сериозно да подкопае изводите. Pollard & Richardson (1987), също и Cohen (1994) дават великолепен пример за това. Те разглеждат следната парафраза на един силогизъм, известен в логиката като modus tollens:

(1) Ако едно лице е американец, то вероятно не е член на Конгреса (вярно)
(2) Дадено лице е член на Конгреса (вярно)
Следователно, лицето вероятно не е американец (грешно!)

Така от верните предпоставки (1) и (2) стигаме до грешен извод. Същият грешен извод се получава и когато се отхвърля нулевата хипотеза:

(1) Ако нулевата хипотеза е вярна, то вероятно резултатът ($m_1 \neq m_2$, $p < 0.05$) нямаше да се получи
(2) Резултатът се получава
Следователно, нулевата хипотеза вероятно не е вярна и следва да бъде отхвърлена

Подобен пример привеждат Beck-Bornholdt & Dubben (1996). Двамата автори провеждат разсъждение, аналогично на горните две, с което „показват“, че папа Йоан-Павел II е извънземно същество. Нямаше да споменаваме примера, ако той не бе публикуван в авторитетното списание *Nature*. Това най-малкото показва, че редакторите на списанието са се отнесли сериозно към логическите проблеми, които създава статистическият извод. Въпросната публикация предизвика поток от гневни писма на различни автори, които се стремяха да посочат грешката – все пак покойният папа не беше извънземно същество. Но в писмата няма две еднакви становища; авторите значително се разминават в мненията си относно това, как следва да се проведе разсъждението.

Тези примери не са единствените против използването на статистическия извод като инструмент за формулиране на психологически извод. Мнозина автори критикуват проверката за „значимост“ в психологическите изследвания (например Bakan, 1966,

Cohen, 1990, 1994, Loftus, 1996, Nickerson, 2000, Kirk, 2003, Kline, 2004). Предлагат се различни подходи за решаване на проблема. Един от тях ще разгледаме в този текст.

§ 2. Значимост и същественост на ефекта

Нека разгледаме понятието „значима разлика“. На английски се използува терминът significant, което в буквален превод е точно „значим“. И в двата езика, английски и български, „значим“ означава също и „значителен“, „съществен“, изобщо нещо, което представлява интерес. Излиза, че ако резултатът е „значим“, т.е. ако $p < 0.05$, то той е съществен и представлява интерес. В общия случай, това просто не е вярно. „Значимостта“ на един резултат означава, че при пресмятането на съответния статистически тест се е получила вероятност за „лъжлива тревога“, по-малка от 0.05, *и нищо повече*. „Значимостта“ все още не носи никаква информация относно *съществеността* на резултата, или ползувайки примера, с който започнахме, относно *големината* на разликата между двете изследвани извадки. Ако тази разлика е „голяма“, то можем да се надяваме, че и разликата между популациите ще е голяма и съществена. Но как да определим „големината“ на разликата между групите ?

Опитният изследовател ще каже веднага: „знам как да определя големина на разлика“. Ако става въпрос за разлика от две точки по скала на въпросник, в който резултатите могат да варират например от 5 до 50 точки, то разликата е малка. Но ако става въпрос за оценки по шестобалната система, разликата от две точки е много голяма. Това е така. Съществеността на една разлика може да се определи въз основа на познаване на инструмента, на минал опит, на данни от литературата. Но това не е достатъчно.

В последните десетилетия много учени положиха усилия да преодолеят мисленето в термини на „значимо“ – „незначимо“. Предложени бяха методи за описание на различни типове данни, които да дават количествена представа за съществеността на даден резултат от изследване. Психологическата гилдия по света робува на традиции и навици – някои от тези методи са вече възприети и се прилагат в психологическата литература, други се споменават по-скоро за пълнота в статистическите учебници и наръчници. Не са малко и авторите, които въобще пренебрегват тези методи и продължават да описват данните си със „значимости“. Както отбелязва Cohen (1988), минаха над 30 години, докато психолозите започнат да прилагат t-теста при анализа на данните си. Изглежда ще трябва да почакаме, докато методите за оценка на големина на разлики станат стандартна практика.

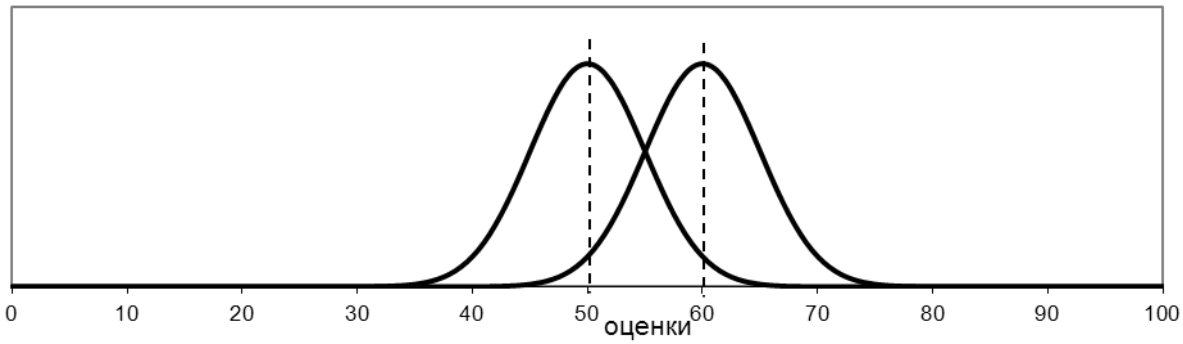
§ 3. Силата на ефекта – индекс за същественост

Ще започнем с един основен принцип, чрез който психологът може да информира за съществеността на резултатите, които е получил. За простота, отново разглеждаме случая, в който с въпросник са изследвани две групи, 1 и 2, от различни лица, като оценките им X са нормално разпределени и с еднакви стандартни отклонения.

На фиг. 1 са представени криви на нормални разпределения, които илюстрират хипотетични данни, получени от двете групи. Средните им стойности са отбелязани с

вертикални прекъснати линии. При описанието на разликата между групите, следваме основния принцип:

Разликата е толкова по-голяма, колкото е по-малко се припокриват двете разпределения.



Фиг. 1. Илюстрация на две разпределения, които се припокриват. Средните им стойности са означени с прекъснати линии.

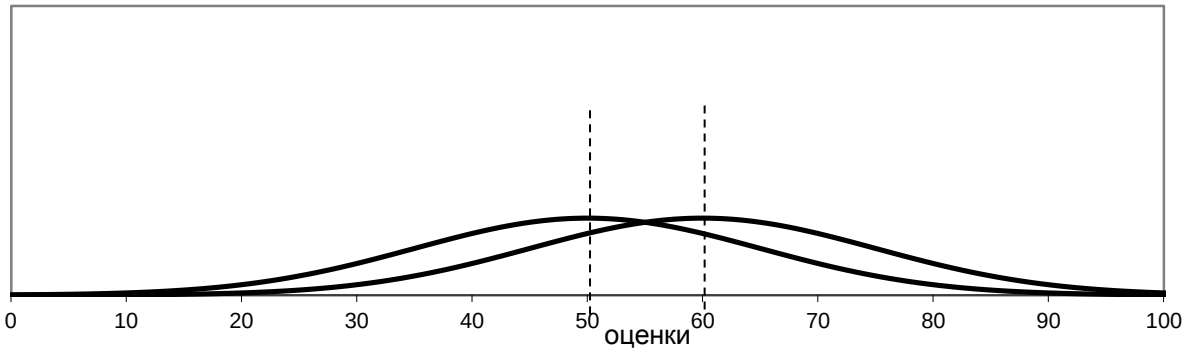
Ако разпределенията се припокриват напълно, групите не се различават по отношение на характеристиката, която измерва въпросникът. Разделянето на двете разпределения води до прогресивно нарастване на разликата, като поне на теория, тя може да расте до безкрайност. Следователно, нужна ни е мярка, която да описва степента на припокриване на разпределенията. Такава мярка е вдъхновена от Cohen (1988), и тя е

$$ES = \frac{m_2 - m_1}{s} \qquad (1)$$

т.е. това е разликата между двете средни стойности, нормирана с общото стандартно отклонение. Мярката се нарича *сила на ефекта, effect size,* или *Cohen's d.* Използват се абревиатурите ES или само d, като е добре в текст да се упоменава какво точно се има пред вид.

По-честа практика е в числителя на (1) да се пресмята абсолютната стойност $|m_1 - m_2|$, а в придружаващия текст да се даде информация, в коя група средната стойност е по-висока. ES може да се пресмята според (1) и с неговия знак, но отново в текста се дава информация, коя средна стойност от коя се изважда.

Ясно е, че когато определяме големината на една разлика в термини на сила на ефекта, трябва да взимаме пред вид не само разликата между двете средни стойности, но и разсейването на данните, дадено от стандартното отклонение. На фиг. 2 са показани две разпределения, чиито разлики между средните стойности са еднакви с тези на фиг 1, но разсейванията на данните са по-големи. „На око" се вижда, че разпределенията на фиг. 2 се припокриват повече, отколкото тези на фиг. 1. Съответно и разликата между групите, илюстрирана на фиг 1 е по-голяма, отколкото разликата на фиг 2.



Фиг. 2. Илюстрирация на две разпределения със същите средни стойности, като тези на фиг. 1, но с по-големи стандартни отклонения. Съответно, те се припокриват в по-голяма степен.

§ 4. Сила на ефекта и статистически извод

Изложеното пресмятане на силата на ефекта е операция за *описание* на данните, т.е., ES е вид описателна статистика. Индексът обаче, е тясно свързан със статистическия извод. Ще илюстрираме тази връзка със следния опростен пример. Да разгледаме t-теста за две независими извадки с еднакъв обем n, във вида

$$t = \frac{|m_2 - m_1|}{\sqrt{\frac{2.s^2}{n}}} \qquad (2)$$

където t е коефициентът на Стюдент и s е еднаквото стандартно отклонение на двете извадки. С малко аритметика изразът (2) може да се преобразува по следния начин:

$$t = \frac{|m_2 - m_1|}{\sqrt{\frac{2.s^2}{n}}} = \frac{|m_2 - m_1|}{s}\sqrt{\frac{n}{2}} = ES.\sqrt{n/2} \qquad (3)$$

Както добре знаем, високата стойност на t означава „висока значимост“ на разликата, т.е., ниска стойност на вероятността *p*. Следователно, значимостта на една разлика се определя от произведението на две неща – силата на ефекта и обема на извадките, или

$$\text{значимост} = ES * f(n) \qquad (4)$$

където f(n) е функция от обема на извадките. За този опростен случай функцията е $\sqrt{n/2}$.

Равенствата (3) са изведени за опростения случай за t-тест при две нормално разпределени извадки с еднакъв обем и еднакви стандартни отклонения. При този случай се допуска също, че оценките на лицата от въпросника са в рамките на интервална скала. Тези допускания не винаги са верни в реалната психологическа практика. Видът на данните поставя и изисквания за различни статистически тестове - F-тест, $\chi 2$, и т.н. за определяне на значимост. Нека читателят повярва и приеме без строго доказателство (то такова и няма), че равенство (4), или по-скоро твърдението,

което то изразява, е валидно за всички възможни статистически тестове и видове на данните. Единствено начинът на пресмятане на ES и функцията f(n) зависят от вида на данните и съответно от статистическия тест, който се прилага.

Равенство (4) има огромно значение при анализа на данните. Именно то показва, че за да определим една разлика между две групи като съществена, и представляваща практически интерес, не е достатъчно да определим дали тя е значима. Възможен е резултат, при който разликата между групите, определена с ES, да е мизерно малка и несъществена, но благодарение на големия брой лица в извадките, тя да се окаже „високо значима“ и това да даде повод за изводи и твърдения, които са направо неверни. Обратно, разликата може да не е значима (поради недостатъчния обем на извадките), но ES да е съществен, което да даде импулс за анализ и размишления, и провеждане на по-нататъшни изследвания, ако е нужно. Не трябва да се забравя, че ES, определен от данните в извадките, или sample effect size, представлява оценка на силата на ефекта в популацията. *Не съществува статистически трик, с който слаб ефект в извадките да се превърне в силен ефект в популацията!*

Да разгледаме по-подробно ползата от пресмятането на силата на ефекта. Преди всичко ES дава възможност за сравняване на данни, получени от различни изследвания, с различни въпросници, с различни обеми на извадките. Твърдения като „факторът етнос оказва по-силен ефект върху променливата X, отколкото върху Y” или „възрастта оказва по-силен ефект върху X, отколкото етносът“, са смислени, когато се изказват в термини на ES, а не на разлики между средни стойности или на стойности на *p*. ES е безразмерен индекс. Това му свойство се използува в мета-изследванията, при които авторите им се стремят да направят изводи за определени популационни ефекти въз основа на третиране на данни от множество различни изследвания в литературата. За тази цел те трябва да уеднаквят „единиците“ на измерванията в различните статии. Пресмятането на ES (поне засега) е единственият възможен начин за това.

В множество статии, дисертации и дипломни работи може да срещнем следната формулировка на „психологическа“ хипотеза: „очакваме променливата (или факторът) A да окаже значим ефект върху променливата B“. При проверката на така формулираната хипотеза няма наука, в частност психология. Ефектът съществува в природата, него Бог го дава и изследователят по принцип няма контрол върху него. Равенство (4) показва, че тази хипотеза ще се подкрепи, ако броят на изследваните лица е достатъчно голям. Ако не е голям, хипотезата няма да се потвърди. Къде е науката, ако психологическият извод се определя от обема на извадката? Тук оставяме настрана въпроса, дали извадката е представителна, или не е.

Препоръката ни е, в такива текстове думата „значим ефект“ да се заменя със „съществен“ или „представляващ практически (или теоретически) интерес“. Лошо няма, ако се провежда статистически тест, той се изисква от комисии при защити и при рецензиране на статии. Но не трябва да се забравя, че стойността на *p*, която се получава в резултат от провеждането на статистическия тест, не дава информация за съществеността на наблюдаваната разлика. Когато се сравняват резултатите от две

изследвания, проведени с еднакъв брой лица и анализирани с един и същ статистически тест, по-високата „значимост“ (т.е. по-ниска стойност на p) според (4) наистина показва по-силен ефект. Но дори и тогава, p не ни казва нищо за това, дали въпросният ефект представлява интерес и доколко заслужава да му се обърне внимание.

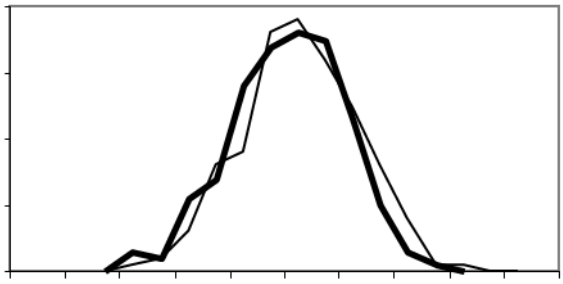
§ 5. Класификацията слаб, среден и силен ефект

Представянето на описания на данни със сила на ефекта определено води до ментално натоварване на изследователя. Тук не става въпрос за усилията, необходими за разглеждане на разпределения и за допълнителни пресмятания. Откакто има компютри, това не е особен проблем. Работата е там, че когато се прави статистически извод, психологът може лесно да вземе решение. Ако $p < 0.05$, решава едно, ако $p > 0.05$, решава друго. Понякога се изказват твърдения като „разликата е маргинално значима, $p = 0.06$“ или „наблюдава се тенденция за значима разлика“. Те демонстрират усилията на изследователите да се измъкнат от дихотомията „значимо-незначимо“. В литературата се цитира саркастичната забележка на Rosnow & Rosenthal (1989), че „Бог обича 0.05 толкова, колкото и 0.06“. При разглеждането на ES нещата изглеждат още по-зле. Какво да твърдим, ако получим, например ES = 0.65, много или малко е това? Представлява ли този ефект някакъв интерес, или не?

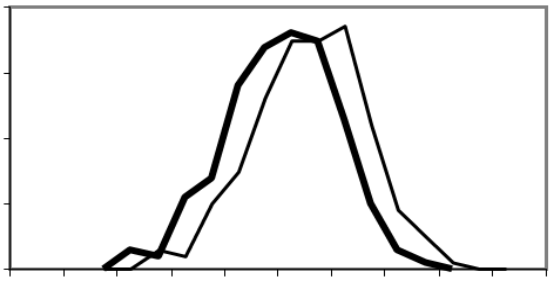
Въпросът е много основателен, и за съжаление отговорът му не е особено еднозначен. Cohen (1988) го е осъзнал и го е решил по следния начин. Той въвежда понятията слаб (small), среден (medium) и силен (large) ефект. Ефектът е слаб при ES = 0.2, среден при ES = 0.5 и силен при ES = 0.8. Авторът отбелязва, че това разделение е направено след преглед на резултати от изследвания в областта на социалната психология и психологията на образованието . Трябва веднага да отбележим, че при тази класификация Cohen не изказва становища относно това, кой от тези ефекти представлява практически интерес и кой не. Целта му е да облекчи пресмятането на т.н. мощност на определени статистически тестове, също и предварителното определяне на обема на извадките.

Авторитетът на Cohen е огромен, и мнозина автори след него цитират неговата класификация. Всеки от тях прави уговорката, че тя е ориентировъчна, и не бива да се възприема съвсем буквално при оценката на наблюдавания ефект. За да ориентираме читателя за какво става въпрос, на следващите фигури са илюстрирани трите ефекта. Илюстрациите са направени, като в Excel са генерирани двойки нормални разпределения с еднакви разсейвания, но разместени помежду си съответно с 0.2, 0.5 и 0.8 стандартни отклонения.

Фиг. 3. илюстрира „слабия“ ефект, ES = 0.2. Разликата между двете разпределения според нас не може да се види „на око“. Пресметнатият ефект може и да не е нулев, но разликата между двете групи явно не представлява практически интерес. Друга работа е, ако изследователят предварително знае, че се очаква именно такъв слаб ефект; тогава ES = 0.2 може да придобие „теоретическо“ значение. От равенствата (3) не е трудно да се пресметне, че този ефект ще е „значим“, ако са изследвани две групи от поне 200 лица всяка.



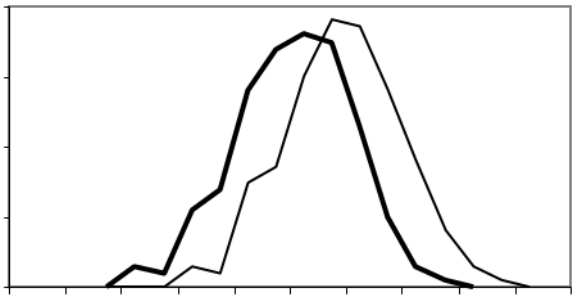
Фиг.3. „Слаб“ ефект, ES = 0.2. Средната стойност на разпределението с тънка линия е отместена с 0.2 стандартни отклонения надясно.



Фиг.4. „Среден“ ефект, ES = 0.5. Средната стойност на разпределението с тънка линия е отместена с 0.5 стандартни отклонения надясно.

При „средния“ ефект, ES = 0.5, разликата между двете разпределения вече може да се забележи и „на око“. Преценката, дали ефектът представлява практически интерес, е проблем на изследователя. Най-добре е той да се позове на подобни изследвания в литературата, и тогава да реши дали ефектът е съществен. Ако обемът на всяка от извадките е поне 32 лица, то ES = 0.5 и по-висок, t-тестът ще покаже „значимост“

Фиг. 5. илюстрира „силния“ ефект, ES = 0.8. Разликата между групите вече е отчетлива. Много изследователи биха казали, че този ефект е съществен и представлява практически интерес. Cohen (1988) отбелязва, че такъв ефект се получава при сравнение на IQ между първокурсници и докторанти. Ако обемът на всяка от извадките е около 13 лица, то ES = 0.8 и по-висок, t-тестът ще покаже „значимост“.



Фиг.5. „Силен“ ефект, ES = 0.8. Средната стойност на разпределението с тънка линия е отместена с 0.8 стандартни отклонения надясно.

В крайна сметка, опитният изследовател трябва да успее да прецени, дали разликата, която наблюдава, представлява интерес, практически или теоретически. Henson (2006) предупреждава за опастността, реперите за слаб, среден и силен ефект да се превърнат в догма, подобна на тази, която представлява нивото $p < 0.05$. Все пак, класификацията на Cohen може да подскаже на дипломанти и докторанти, как приблизително да класифицират ефектите, които се получават в техните изследвания. Това е по-добре от нищо.

При оценката на ефекта, дали е съществен и дали представлява представлява практически интерес, следва да се вземе пред вид не само припокриването на разпределенията на данните. Добре е да се прецени, какво предизвиква този ефект, т.е. какви манипулации на независимата променлива водят до появата му. Може да се окаже, че ефектът не е кой знае колко силен, определен с индекса ES, но се предизвиква от незначителна, минимална манипулация. Например, на студенти в читалнята се подарява кексче. След това, те показват по-висока степен на готовност да помагат на колегите си, отколкото студентите, които не са получили подарък (Prentice & Miller, 1992). Третирането на лицата (с кексче) е минимално; дори и да доведе само до слаб ефект, резултатът може да представлява интерес. Abelson, (1995, стр 47) разглежда подробно подобни случаи и въвежда терминологията cause effect size, в свободен превод "сила на ефекта на причината". Количествено охарактеризиране на силата на ефекта на причината може да се получи при регресионния анализ – ъгловият коефициент на регресионната права показва с колко единици се променя критерийната променлива при промяна на предикторната с една единица. Но извън този случай, няма разработени процедури за пресмятането му, също и репери за слаб, среден и силен ефект. Аргументът за същественост остава въпрос на творчество.

По-горе разглеждахме случая на разлика между две извадки с еднакъв обем, чиито разпределения на оценките са нормални, и стандартните им отклонения са еднакви. Ако обемите на извадките, n_1 и n_2, са различни, и стандартните отклонения s_1 и s_2 не са еднакви, но разпределенията са близки до нормалните, то пресмятанията се усложняват, но логиката им остава същата. В този случай знаменателят в равенство (1) е обединеното стандартно отклонение, s_{pooled}, което се пресмята като

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_2^{\,2} + (n_2 - 1)s_1^{\,2}}{n_1 + n_2 - 2}} \qquad (5)$$

В числителя под корена се пресмята сборът на квадратите на отклоненията на данните от m_1 и m_2, а в знаменателя е сборът на съответните степени на свобода, $(n_1 - 1)+(n_2 - 1) = n_1 + n_2 - 2$.

Glass (1976, цитиран по Henson, 2006) предлага индекс за пресмятане на силата на ефекта при случая, при който едната от групите се приема за контролна, а другата – за експериментална. Индексът се обозначава с Glass's Δ и се пресмята като

$$\Delta = \frac{m_E - m_C}{s_C} \qquad (6)$$

където m_E и m_C са средните стойности на данните от експерименталната и контролна групи, а s_C е стандартното отклонение на контролната група. Впечатлението ни е, че този индекс се среща основно в учебниците по статистика, а не в изследователските трудове.

§ 6. Сила на ефекта при вътрегрупов дизайн

Разглежданията на силата ефекта в параграфи 4 и 5 се отнасят за случая, когато имаме две групи с различни лица, т.е. става въпрос за т.н. междугрупов дизайн (between-subjects) на изследването. По-долу ще разгледаме случая на вътрегрупов дизайн (within-subjects), именно, когато данните се събират от една и съща група лица, но в две различни условия. В този случай пресмятането на силата на ефекта, ES, става по следния начин.

Нека имаме група от n лица, които са изследвани в две условия, 1 и 2. Така всяко от лицата получава по две оценки от измерването. Обозначаваме ги с X_1 и X_2. Разглеждането на средните стойности m_1 и m_2, получени при двете условия, и особено на стандартните отклонения s_1 и s_2, не носи кой знае каква информация за ефекта на условията върху оценките X. Следва да се направи следното.
Пресмята се разликата $D = X_2 - X_1$ с нейния знак за всяко от n-те лица поотделно
Пресмята се средната стойност на всички разлики D, m_D
Пресмята се стандартното отклонение s_D на n-те разлики.
Силата на ефекта ES се определя като $ES = |m_D|/s_D$

Зад този израз всъщност се крие сметката $ES = |m_D - 0|/s_D$, т.е. ES показва с колко стандартни отклонения средната стойност $m_D = m_2 - m_1$ се отклонява от стойността нула. Читателят може да използва същите репери, както и при междугруповия дизайн, 0.2, 0.5 и 0.8 за квалифициране на ефекта като слаб, среден и силен.

Връзката между така пресметнатия ES и статистическия извод, който се провежда чрез t-теста за зависими извадки, е проста. Имаме

$$t = \frac{m_D}{s_D / \sqrt{n}} = ES.\sqrt{n} \qquad (6)$$

Отново виждаме, че значимостта на разликата между двете условия се определя от произведението на ES и $\sqrt{n}$, както описахме с равенство (4).

При вътрегруповия дизайн, когато едни и същи лица се изследват при две условия, данните обикновено са положително корелирани. Лицата, които имат високи стойности на X при едното условие имат и високи стойности в другото, и обратно. В този случай

стандартното отклонение на разликите s_D е по-малко от отделните стандартни отклонения s_1и s_2 на данните от условията. (При перфектна положителна корелация, r = +1, стандартното отклонение на разликите даже става нула.) Въпреки че стандартните отклонение при условията 1 и 2 може да показват силно разсеяни данни, разликите m_D обикновено не са толкова разсеяни, което води до по-високи стойности на ES. В резултат, при по-висока корелация между данните, нулевата хипотеза се отхвърля с по-малък брой лица. Това е и една от причините, поради която вътрегруповият дизайн се предпочита от изследователите.

Много препоръчително е, данните от вътрегрупов дизайн да се представят с хистограма на разпределението на разликите $X_2 - X_1$, (или с таблица, ако лицата са по-малко). Например, ако едното условие е „контролно“, а другото включва някакво третиране, е полезно да се знае при какъв процент от лицата третирането е предизвикало ефект, обратен на очаквания. Хистограмата на разликите ще илюстрира този процент.

§ 7. Описание на разлика между групи с помощта на относителни дялове

Изброените досега индекси за описание на разликата между две групи лица „работят“ добре при допускането за нормални разпределения на данните. Ако разпределенията не са поне близки до нормалното, прилагането на тези индекси става проблематично, също и връзката на индексите със параметричните статистически тестове започва да се губи. Ако разпределенията не са нормални, например са с изявена асиметрия, прилагат се други методи за описание на разликата.

Един съвсем прост метод е следният. На око се определя в коя от двете групи оценките от прилагането на въпросника или теста са по-високи. Пресмята се медианата на данните от тази група. След това се определя процентът на лицата от „по-слабата“ група, които имат оценки по-ниски от медианата. По този начин се пресмята индексът: „относителният дял на лицата от по-слабата група, които имат оценки, по-ниски от тези на по-добрата половина на по-силната група“ (Cohen, 1988). Ако този дял е 50%, разлика между групите няма. С нарастването на разликата, този процент нараства, докато стане 100%. За съжаление, този индекс не информира достатъчно за това, в каква степен двете разпределения се припокриват. Дори и когато той е 100%, т.е. всички лица от по-слабата група имат оценки, по-ниски от медианата на по-добрата група, разпределенията пак може да се припокриват в някаква степен. Но ако разпределенията са припокрити в по-голяма степен, тогава индексът варира между 50% и 100%. Като „слаб“, среден“ и „силен“ ефекти могат да се използуват ориентировъчните стойности 58%, 69% и 79%. Тези стойности съответстват на сила на ефекта съответно 0.2, 0.5 и 0.8 при условие, че разпределенията са нормални.

McGraw & Wong (1992) предлагат една мярка за силата на ефекта, която започва да „набира скорост“ в публикационната дейност. Това е common language (CL) effect size, , или в по-свободен превод, „сила на ефекта с прости думи“. Авторите изтъкват, че индексите ES, които се пресмятат като стандартизираната разлика между две средни стойности, са трудно разбираеми за хора, които нямат статистически познания. Те

механично четат стойностите на ES, без да разбират техния психологически смисъл. McGraw & Wong предлагат следното.

Нека изберем случайно едно лице от по-силната група 2. То има оценка X_2. Въпреки, че лицето е от по-силната група, това съвсем не означава, че X_2 ще е по-висока от *всички* оценки X_1 на по-слабата група 1. Тъй като разпределенията се припокриват, много вероятно е в по-слабата група да има лица с оценки X_1, които да са по-високи от оценката на това лице. Но все пак, следва да се очаква, че оценките на повечето лица от група 1 ще са по-ниски от X_2. А какво ще се получи, ако изберем друго лице от по-силната група? Математически въпросът се поставя по следния начин.

Нека изберем случайно две лица, по едно от всяка група, и се запитаме, каква е вероятността лицето от по-силната група да има по-висока оценка от лицето от по-слабата група. По принцип, тази вероятност може да се оцени, ако изготвим списък на всички възможни двойки лица, които са изследвани, и преброим при колко от тях лицата от по-силната група имат по-високи оценки от лицата от по-слабата.

Ако процентът на тези двойки е 50%, разлика между групите няма. Това означава, че колкото лица от слабата група имат по-ниски оценки от лицата в силната група, точно толкова имат и по-високи оценки. От двете оценки не можем да определим, кое от лицата в двойката към коя група принадлежи. Групите не се различават по отношение на характеристиката X.

Ако процентът на тези двойки е 100%, това означава, че разпределенията на данните от групите въобще не се припокриват. *Всички* лица от по-слабата група имат оценки, които са по-ниски от всяка от оценките в по-силната група. Ако знаем оценките на лицата в случайно избраната двойка, можем с пълна сигурност да определим, кое лице към коя група принадлежи. Групите напълно се различават по отношение на характеристиката X.

Изложените два случая са две крайности – на липса на разлика между групите и на групи, които напълно се различават по отношение на X. Междинните случаи се описват от индекса „сила на ефекта с прости думи", или common language ES, който не е нищо друго, освен оценка на вероятността $Pr(X_1 < X_2)$, т.е. вероятността случайно избрано лице от едната група да има по-висока оценка от случайно избрано лице от другата група. Или индексът „сила на ефекта с прости думи" просто информира читателя или слушателя за процентът на лицата от едната група, които имат по-високи оценки от лицата в другата група. Не са нужни специални познания по статистика, за да се разбере, какво означава този индекс, всеки сам може да си направи изводи за това, доколко съществена е разликата между групите. Този индекс все още няма утвърден етикет, както ES или d, по-нататък в текста го означаваме с $Pr(X_1 < X_2)$, както го дефинирахме по-горе.

Пресмятането на индекса е лесно, ако данните от групите са нормално разпределени и стандартните отклонения са еднакви. Тогава $Pr(X_1 < X_2)$ може да се преизчисли от израза за $ES = (m_2 – m_1)/s$ като

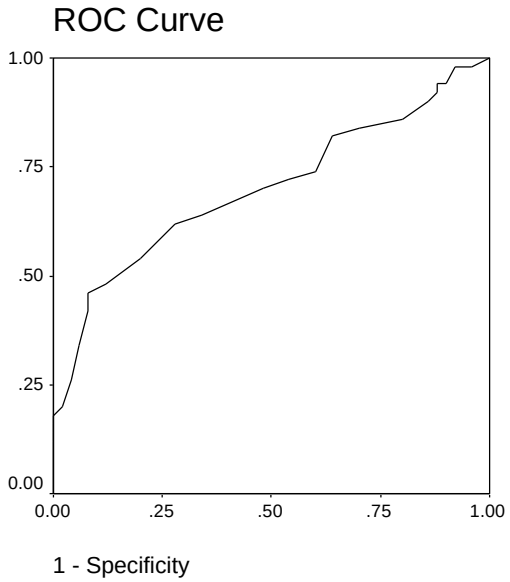
$$Pr(X_1 < X_2) = z'(ES/\sqrt{2}), \quad (7)$$

където z’ е обратната z-трансформация (виж Матеев, 2011, за подробно описание на z и z’-трансформациите). В Excel тя се осъществява с командата =normsdist(value). Реперите 0.2, 0.5 и 0.8, които могат ориентировъчно да служат за класификация на ефекта като слаб, среден и силен, след преизчисляването им във вероятности или в проценти стават 56%, 64% и 71%. Информацията, която носят тези числа е същата, както и стойностите на ES, но данните за $Pr(X_1 < X_2)$ са определено по-ясни за хора, които нямат специализирани статистически познания.

Изчисляването на $Pr(X_1 < X_2)$ става по-комплицирано, ако разпределенията на данните от двете групи значително се отклоняват от нормалните. На тези читатели, които са по-наясно с непараметричните статистически тестове, препоръчваме работата на Ruscio (2008), в която се разглежда връзката между $Pr(X_1 < X_2)$ и статистиките на Mann-Whitney и Wilcoxon. Но при изчислението съществено може да помогне построяването на т.н. работна характеристика, receiver-operating characteristic, или ROC-крива.

Подробно разглеждане на въпроса за работната характеристика е представено в Матеев (2011). Тук ще изясним понятието, без да навлизаме в математически подробности. Отново разглеждаме две групи лица, които са тествани с въпросник, който измерва характеристиката X. Работната характеристика е свързана със задачата, от оценката X на дадено лице да „познаем“, към коя от двете групи то принадлежи. Може да предприемем следното.

Избираме дадена стойност от зависимата променлива X като критерийна, X_c. Решаваме, че лицата, които имат оценки от въпросника, по-високи от X_c, принадлежат към по-силната група 2, а лицата, при които $X < X_c$, принадлежат към по-слабата група 1. Част от лицата, при които $X > X_c$, се оказва, че наистина принадлежат към по-силната група. Техният относителен дял се нарича *сензитивност* (sensitivity). Но друга част от лицата, при които X също е по-висока от X_c, се оказва, че принадлежат към по-слабата група. Техният относителен дял се обозначава с *1-специфичност*, (1-specificity). Двата относителни дяла, сензитивност и 1-специфичност, могат да се пресметнат за *всяка* стойност на критерия X_c. Те се нанасят на графика, на която по абсцисата е 1-специфичност, а по ординатата – сензитивност. Двойките относителни дялове се свързват с гладка (или начупена линия). Тази линия, илюстрирана на фиг.6., представлява ROC-кривата, или работната характеристика. Ако тя съвпада с диагонала, двете групи от данни се припокриват напълно. Ако кривата съвпадне с раменете на горния ляв ъгъл, имаме пълно разделяне на двете групи от данни. Като индекс за степента, в която данните се разделят, се използува площта под ROC-кривата, area under the curve, AUC. Когато площта е 0.5, т.е. половината от графиката, кривата



Фиг 6. Илюстрация на хипотетични данни от две разпределения, които се отклоняват от нормалното и са с различни стандартни отклонения. Със SPSS се пресмята площ под кривата, равна на 0.7, т.е. 70% от лицата на „по-добрата“ група имат оценки, по-високи от тези на лицата от „по-слабата група“

.

съвпада с диагонала. Когато кривата се съвпада с горния ляв ъгъл, площта под нея е единица.

Площта под кривата представлява индекс за това, доколко двете разпределения са разделени едно от друго. Предимствата на индекса са, че той „работи“ независимо от формата на двете разпределения, от съотношението на броя на лицата във всяка от групите, и не изисква интервална скала на зависимата променлива X. Неговият смисъл става ясен от една теорема, доказана навремето от Green & Swets (1966), именно, че площта под кривата на работната характеристика дава вероятността $Pr(X_1 < X_2)$. Този резултат облекчава пресмятането на вероятността с използването на опцията ROC-curve в пакета SPSS. Програмата не само пресмята площта под кривата, но също и стандартната грешка и 95%-ния доверителен интервал.

Значението на теоремата на Green & Swets извън сферата на психофизиката започна да се осъзнава в последните години. Анализът на работната характеристика започна да се прилага в сферата на диагностиката, психологическа и медицинска (например Калчев, 2008) , но става ясно, че той може съществено да подпомогне описанието и анализа на значително по-широк клас от данни. В това отношение много полезна е работата на Ruscio (2008), в която се разглеждат индекси за силата на ефекта в сравнителен план. Освен лесното възприемане на индекса $Pr(X_1 < X_2)$ като средство за описание на данни и за комуникация между изследователи, авторът изтъква и други негови преимущества, като устойчивост при различни обеми на извадките, устойчивост срещу „бегълци“ (outliers) в данните, нечувствителност към трансформации на данните (например

логаритмуване), и др. Много вероятно е в близко бъдеще този индекс да се превърне в стандарт при описание на данни от научни изследвания.

§ 7. Заключение

Настоящата статия не изчерпва в никакъв случай въпроса за описанието на данни с помощта на сила на ефекта, също и приложенията на този тип описателна статистика, например, при определяне на мощност на статистически тест или на обем на извадка. Не са разгледани различните видове индекси, които се основават на пресмятане на корелационни коефициенти, или индекси, които съпътстват прилагането на дисперсионен анализ. Целта ни бе да покажем философията на този вид описателна статистика и на ползите от нея, както и да представим пресмятането на силата на ефекта в някои по-прости и по-често срещани случаи..

Аргументираме, че статистическият извод, т.е. изчисляването на значимост, не е удачен инструмент при количественото установяване на съществеността на един резултат. Статистическият извод не трябва да се използува за цели, за които той не е пригоден. Въпреки логическите проблеми, които изложихма в § 1, статистическият извод ни дава идея за това, с каква увереност данни от извадки може да се обобщят за цели популации. Това е безспорно важно и полезно, но то не е психологически извод. Тази истина е осъзната от редакционните колегии на по-авторитетните списания (виж например упътванията за авторите, които се дават в списанието Psychological Science), които препоръчват на авторите да не правят психологически изводи въз основа на резултати от статистически тестове. Вместо това, авторите следва да представят и интерпретират данни за силата на ефекта, който се получава от изследването.

Литература

Калчев, П. (2008). *Индиректна скала за употреба на психоактивни вещества*. София: изд. Изток-Запад

Матеев, С. (2001). *Начала на психофизиката*. София: изд. НБУ

Abelson, R.P. (1995*). Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum,

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, **66**, 423-437

Beck-Bornholdt, H.P. & Dubben, H.H. (1996). Is the Pope an alien? *Nature*, **381**, 730

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum (2^{nd} edition)

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304-1312

Cohen J (1994). The earth is round ($p < 0.05$). *American Psychologist*, **49**, 997-1003

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, ***5***, 3-8.

Green, D.M. & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: Willey

Henson, R.K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, **34**, 601-629

Kirk, R. (2003). The importance of effect magnitude. In: *Handbook of Research Methods in Experimental Psychology* (Edited by Davis, S.F.) Chapter 5, pp. 83 – 105, Blackwell Publishing Ltd

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

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current directions in psychological science*, **5**, 161-171

McGraw, K.O & Wong, S.P. (1992). A common language effect size statistics. *Psychological Bulletin*, **111**, 361-365

Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, **5**, 241-301

Pollard, P. & Richardson, J. T. E. (1987). On the probability of making type I errors. *Psychological Bulletin*, **102**, 159 – 163

Prentice, D.A. & Miller, D.T. (1992). When small effects are impressive. *Psychological Bulletin*, **112**, 160-164

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, **44**, 1276-1284