

СТАТИСТИЧЕСКАТА ЗНАЧИМОСТ – ПАНАЦЕЯ ИЛИ ПРЕПЪНИКАМЪК?

Маргарита Ламбова

Икономически университет – Варна

Резюме. Представени са разсъждения относно понятието „статистическа значимост“, популярно средство за „доказване“ на съвместимостта с практиката на конструирани на базата на предположения теоретични модели.

В съответствие с поставената цел въз основа на теоретичните особености и логиката на понятието са разкрити проблемни моменти и възможности за злоупотреби с него при проверката на статистически хипотези. Представената аргументация позволява твърдението, че „рецептурната“ проверка на статистически хипотези създава условия за развитие на псевдонаука, подпомагана от погрешни изводи, които са направени на базата на „статистически значими“ резултати.

Ключови думи: статистическа значимост; тестване на статистическа хипотеза; p – стойност; ниво на значимост; псевдонаука

Въведение

Проверката на статистически хипотези е един от „най-удобните“ статистически инструменти, използвани при научни изследвания с цел доказване на съвместимостта с практиката на конструирани на базата на предположения теоретични модели. Използва се както при работа с извадки (случайни и недотам случайни), така и при изследването на динамични редове, като в много случаи въпросът относно изискуемото от гледна точка на статистическата методология наличие на условия за провеждане на случаен експеримент остава извън ползрението. Проблемът за качеството на информацията, използвана за проверка на направените предположения, тук няма да бъде засегнат, въпреки че той е свързан с допълнително неконтролируемо увеличаване степента на ненадеждност на получените резултати. Вниманието ще бъде насочено единствено към категорията „статистическа значимост“, проблемните моменти, свързани с нейната същност и логика, както и с опасностите от създаване на псевдонаука при злоупотреби с нея.

При проверката на статистически предположения предварителните очаквания преобладаващо са свързани с търсенето на доказателства за наличие на предполагаемо различие, залегнало в теоретичния модел на дадено изследва-

не. Това означава, че теоретичните тези биха се потвърдили само в случай че се стигне до заключение за статистическа значимост на резултатите, получени въз основа на емпиричната информация. Не е за подценяване изкушението да се тръгне „на лов“ за статистически значими резултати, тъй като без тях проведеното изследване би могло да се окаже безсмислено. По този начин статистическата значимост се използва в ролята на панацея, която може да „излекува“ всяко изследване, дори и такова, осъществено на базата на нелогична концепция, стига резултатите да се окажат „статистически значими“. Този „лек“ се използва все по-често с навлизането на потребителския статистически софтуер, който не изисква осмисляне на логиката и „подводните камъни“ на статистическите тестове и позволява не само чисто „рецептурното“ им приложение, но и целенасочена манипулация на резултатите с помощта на вездесъщата p -величина (p -value), която се приема от ползвателите като единствен критерий за дихотомна категоризация на резултатите – статистически значими и статистически незначими.

Целта е въз основа на теоретичните особености и логиката на понятието „статистическа значимост“ да се разкрият проблемни моменти и възможности за злоупотреби с него при проверката на статистически хипотези.

1. Теоретични основи на понятието „статистическа значимост“

Понятието „статистическа значимост“ се използва изключително често в научни изследвания, като в повечето случаи неговата логическа същност остава на заден план и се гласува необосновано високо доверие на получените резултати. Обикновено се „забравя“ или не се взема под внимание, че проверката на статистически хипотези е способ, който все още предизвиква ожесточени дискусии относно правилността на използваните логически подходи, стоящи в основата на различните концепции.

До заключението „статистически значим резултат“ може да се стигне само в контекста на даден логически подход и предварително определени критерии за вземане на решение, зависещи от рисковете за допускане на невярно заключение, които изследователят е склонен да поеме.

За изясняване на същността и логиката на статистическата значимост могат да помогнат различните, отчасти противоречащи си теоретични концепции относно статистическите заключения, върху които стъпва съвременната класическа проверка на статистически хипотези.

Като предтеча на теорията за проверката на статистически хипотези се възприема шотландецът John Arbuthnot¹⁾ (Bewersdorff, 2011), който още през 1710 г. прави опити за доказване на предположение относно превеса на броя на ражданията на покръстени момчета над броя на ражданията на покръстени момичета, след като му е направило впечатление, че за всяка от предходните 82 години е имало такъв превес. При доказването на всеобщата валидност

на това според него „божие провидение“, Arbuthnot първоначално предполага еднакъв шанс за двата пола при раждане, т.е. случайност на пола подобна на тази при хвърляне на монета. По този начин той косвено е формулирал работна хипотеза, която според съвременните схващания би трябвало да гласи: вероятността новородено дете да бъде момче, респ. момиче, възлиза на 50%. В случай че емпиричната информация влиза в противоречие със съотношението 50:50, според неговите виждания е налице доказателство, че по правило се раждат повече момчета, отколкото момичета. Според неговите изчисления вероятността броят на новородените момчета да превишава този на новородените момичета за всяка от тези 82 години при общовалидно съотношение 50:50, е изключително малка и възлиза на $1/2^{82}$, поради което той възприема като доказана тезата, че превесът не се дължи на случайността, а е налице „божие провидение“ за него. Съобразно съвременните понятия той би заключил, че резултатът е статистически значим.

В логическия подход на Arbuthnot се открояват два съществени момента, които до днес формират основата, върху която стъпва проверката на статистически хипотези.

1. Проверява се предположение за наличие на определена закономерност, валидна за съвкупност, на базата на наблюдение на относително малка извадка от нея.

2. За доказване на тезата с помощта на вероятностни изчисления се използва обратното предположение, формулирано като работна хипотеза.

Още на този етап от развитието на статистическите тестове се приема, че когато вероятността за определен емпиричен резултат при валидни в зададената работна хипотеза закономерности е твърде малка, тогава наблюдаваното различие не може да бъде обяснено със случайността, а се дължи на това, че резултатът е проявление на други закономерности. Колко малка трябва да бъде тази вероятност, все още не се поставя като въпрос, т.е. не е определена конкретна величина, на базата на която да се стигне до заключение за статистическа значимост на различието. Точно тази величина се оказва препъника-мък в развитието на теорията на статистическите заключения и предизвиква ожесточени научни дискусии, дори вражди, между представителите на двете основни школи, започнали в началото на XX век развитието на различни по своята логика концепции, върху които стъпва съвременната класическа проверка на статистически хипотези.

Дълго време концепцията на Нейман – Пирсън (Jerzy Neyman²⁾ и Egon Sharpe Pearson³⁾ се конкурира с тази на Фишер (Ronald Aymler Fisher⁴⁾), като противоречието между тях е свързано в голяма степен с критериите, въз основа на които се стига до заключение за статистическа значимост на резултата.

В началото на процеса на развитие на класическата проверка на хипотези стоят научните постижения на Карл Пирсън (Karl Pearson⁵⁾) – професор

по евгеника и основоположник на биометрията. Научен сътрудник в неговия институт е Роналд Фишер, един от най-изтъкнатите биолози, генетици и статистици на XX век, с когото не споделят единно мнение относно теорията на статистическите заключения, в частност относно необходимия обем на извадките, като постепенно научната дискусия прераства в открита вражда. Синът на Карл Пирсън – Egon Sharpe Pearson също работи в института на баща си и е опонент на Фишер. Известно време той си сътрудничи с полския математик Jerzy Neyman, с когото имат редица съвместни публикации в областта на теорията за проверката на хипотези, като Нейман също се намира в научна „вражда“ с Фишер.

В крайна сметка, научният спор между двата враждуващи лагера води до качествен скок в развитието на теорията на статистическите заключения, като съвременната класическа проверка на хипотези е своеобразен хибрид от две концепции, основаващи се на различна логика. Ядрото на класическата теория за проверка на хипотези съставлява концепцията на Нейман – Пирсън (Rüger, 2002), но тя съдържа и известни елементи от концепцията на Фишер. Основните характеристики на съвременната класическа теория за проверка на хипотези според Rüger (2002) са следните: изхожда се от обективистична гледна точка, преследва се когнитивистична цел и се осъществява честотна оценка.

Концепцията на Фишер предхожда по развитие тази на Нейман – Пирсън и се основава на тест за значимост, включващ само една хипотеза, наречена нулева, предполагаща липсата на статистически значимо различие, чрез която трябва да се провери дали емпиричното разпределение в извадката се различава от определено хипотетично разпределение. Тестовите по Фишер имат за задача единствено да верифицират наличието или отсъствието на статистическо значимо отклонение в наблюдаваната извадка (Rüger, 2002). Действителната цел на такъв тест е чрез отхвърлянето на хипотезата да се намери доказателство за статистическата значимост на получения резултат. За такова доказателство е необходима мярка, която да оправдава отхвърлянето на хипотезата. През 20-те години на XX век като такава мярка Фишер въвежда p -величината. Тази величина е обвързана с резултата от теста, т.е. с конкретната реализация на статистическия критерий, получена въз основа на данните от случайната извадка. Това е вероятността, с която при валидна нулева хипотеза статистическият критерий може да приеме значение, което е равно на получената емпирична характеристика или е още по-екстремно (Rüger, 2002). Според Nuzzo (2014) p -величината е вероятността, с която даден емпиричен резултат е възникнал без наличие на предполагаемия ефект на базата на случайността.

P -величината показва колко „екстремна“ е получената на базата на данните от случайната извадка емпирична характеристика на теста, в случай че е вярна нулевата хипотеза.

Като критерий за оценка на степента на екстремност на получения резултат и оттам вземане на решение за приемане или отхвърляне на хипотезата Фишер първоначално (1935) предлага равнището на значимост в смисъла на априорна норма (Gigerenzer, Krauss, 2000). То е фиксирано в таблиците му, разработени за t -теста на Стюдънт, като стандартни величини, възлизащи на 1% и на 5%, без обосновка на техния избор. По-късно Фишер (Fisher, 1956) променя вижданията си относно равнището на значимост и предлага да бъде определено апостериорно равнище на значимост на самите емпирични данни, т.е. p -величината (Gigerenzer, Krauss, 2000). По този начин той създава един удобен за потребителите подход за проверка на статистическата значимост на получения резултат (Lehmann, 1993). Въпреки че тестът за значимост е удобен за ползване, той е обременен със съществени недостатъци. Не е възможна смислена интерпретация на мощността на критерия и оттам на риска за приемане на невярна нулева хипотеза, като при него интуитивно се приема, че алтернативната хипотеза обхваща всичко, което не е включено в нулевата. Равнището на значимост първоначално е стандартна величина, а по-късно се отъждествява с p -величината, като не се разсъждава върху проблемите, свързани с риска за отхвърляне на вярна нулева хипотеза, респ. приемане на невярна такава, както и върху последствията при допускане на грешно заключение.

Развитието на концепцията на Нейман – Пирсън е провокирано от несъвършенствата на тестовата теория на Фишер, и най-вече от несъгласието им с интерпретациите му на равнището на значимост, като двамата предлагат алтернативен подход, включващ конкретизация на алтернативната хипотеза (Saint-Mont, 2011). Концепцията на Нейман – Пирсън се основава на тест на двойка хипотези, т.е. на работа с две вероятностни разпределения, като е възможно, но не е задължително алтернативната хипотеза да допълва нулевата. Това означава, че невинаги тя обхваща остатъка от множеството на всички възможности, който не е включен в нулевата хипотеза. Ако с Θ бъде означено цялото множество от възможни величини на даден параметър, а с Θ_0 и Θ_1 – не-пресичащи се негови подмножества, тогава общият вид на двойката хипотези според концепцията на Пирсън – Нейман може да бъде представен по следния начин:

$$H_0: \theta \in \Theta_0 \text{ срещу } H_1: \theta \in \Theta_1,$$

където е възможно, но не е необходимо Θ_0 и Θ_1 да се допълват. В повечето тестови ситуации $H_0 \cup H_1 = \Theta$ (Rüger, 2002).

Допустимият риск за грешка от първи род α , т.е. допустимата вероятност, с която вярна нулева хипотеза може да бъде отхвърлена на базата на случайните резултати от теста, се задава предварително, като за тази величина Нейман и Пирсън използват въведеното от Фишер понятие равнище на значимост. Резултатите, получени въз основа на случайна извадка, по своята същност са реализации на случайна величина. При валидност на нулевата хипотеза тази

случайна величина е с определено разпределение, което задава вероятностите за възможните ѝ реализации, респ. вероятността за реализация, непревишаваща дадена величина. Реализациите от краищата на разпределението са малко вероятни и се приема, че резултат, който е с нищожна вероятност за сбъдване, не принадлежи към това разпределение, а към друго, зададено чрез алтернативната хипотеза. Въпреки всичко, макар и нищожна, различна от 0 вероятност за екстреман резултат при работа със случайни извадки винаги съществува и в случаите, когато тестът доведе до такъв резултат заради малката вероятност, с която той е възможен при вярна нулева хипотеза, последната се отхвърля и ако тя в действителност е вярна, се допуска грешка от първи род. Чрез равнището на значимост се отрязват малко вероятните краища на разпределението, като резултатите, попадащи в областта с голяма вероятностна маса, се приемат за съвместими с нулевата хипотеза, а тези, които попаднат в отрязаните краища, за несъвместими с нея. По такъв начин вероятностната маса в тези краища формира толерирания риск за допускане на грешка от първи род.

При подхода на Нейман – Пирсън заключението се прави въз основа на сравнението на емпиричната характеристика, представляваща конкретна реализация на статистическия критерий, чието разпределение при валидна нулева хипотеза е известно предварително, и границите на областта на приемане, които по своята същност са квантили на съответното разпределение от порядък, който се определя от равнището на значимост. Нейман и Пирсън въвеждат още една вероятност и това е рискът за допускане на грешка от втори род β , т.е. за приемане на невярна нулева хипотеза, който е обвързан с разпределението, зададено чрез алтернативната хипотеза. Този риск в повечето случаи може да се контролира само косвено с помощта на равнището на значимост и обема на извадката и зависи от мощността на критерия – функция, задаваща вероятността за отхвърляне на H_0 в зависимост от действителната величина на проверявания параметър, чрез която е възможно установяването на вероятността $1 - \beta$ за отхвърляне на невярна H_0 при различни величини на проверявания параметър (Ламбова, 2012).

Концепцията на Нейман – Пирсън се различава съществено по своята логика от тази на Фишер, като това различие предопределя редица спорни моменти. Според Нейман тестът на Фишер от гледна точка на математиката е „по-лош, отколкото безполезен“ в случаите, когато $1 - \beta < \alpha$, където α е вероятността за допускане на грешка от първи род, а β – вероятността за допускане на грешка от втори род (Gigerenzer, Krauss, 2000). Един от спорните въпроси на двете конкуриращи се концепции е познавателната същност на статистически значимите резултати. Фишер е застъпвал мнението, че статистически значимият резултат дава конкретно познание за осъществения единичен експеримент, докато Пирсън и Нейман го интерпретират бихевиористично:

според тях статистически значимият резултат само подсказва, че поведението трябва да бъдат такова, като че нулевата хипотеза е невярна. Това означава, че концепцията на Пирсън – Нейман съдържа трета възможна интерпретация на равнището на значимост. Вероятностите α и β следва да бъдат зададени преди осъществяване на теста, но не съобразно общовалидна норма, а на базата на точен анализ на вредите и ползите при евентуална грешка от първи, респ. от втори род (Gigerenzer, Krauss, 2000).

„Ранен“ Фишер	„Късен“ Фишер	Нейман и Пирсън
Определяне на равнище на значимост преди осъществяване на теста в смисъл на установена норма (напр. 5%). Равнището на значимост е естествено свойство на теста.	Определяне на екзактно равнище на значимост на базата на емпиричните данни (p -величината). Равнището на значимост е свойство на данните и вече не е необходима произволно установена норма.	Вероятностите α и β се определят преди осъществяване на теста. Те представляват относителните честоти за събдяване на грешка от първи, респ. втори род, следователно са свойства на теста. За тяхното определяне не е необходима установена норма, а оценка на ползите и вредите от възможните грешни заключения.

Интерпретации на равнището на значимост (по Gigerenzer, Krauss, 2000)

Различието в интерпретацията на равнището на значимост на представителите на двете концепции обуславя и различната познавателна същност на понятието „статистическа значимост“.

Статистическата значимост според концепцията на Фишер е свойство на емпиричните данни, т.е. се отнася за конкретната извадка и показва, че разпределението в извадката се различава от определено хипотетично разпределение, зададено чрез нулевата хипотеза. Прави се заключение за статистически значим резултат, когато вероятността за събдяването му при валидна нулева хипотеза е много малка. Статистическата значимост по Фишер е свързана с т.нар. индуктивно заключение (Lehmann, 1993), чрез което се верифицира наличието на различие между разпределението в извадката и зададеното чрез нулевата хипотеза разпределение. Привържениците на концепцията на Фишер приемат $p = 0.05$ като критична граница на статистическата значимост, т.е. вероятност, по-малка или равна на 5% за получаването на подобен резултат, в случай че е валидна нулевата хипотеза. Според Bortz (2006) при $p \leq 0.05$ резултатът се приема за статистически значим, при $p \leq 0.01$ е много значим, а при $p \leq 0.001$ е изключително значим. Граничните величини за съответната интерпретация са произволно установени, като имат за своя основа първоначално въведените от Фишер стандартни величини за равнището на значимост, за които също липсва аргументация. Извън поле-

зрението остава втората възможност за допускане на невярно заключение: приемането на невярна нулева хипотеза, т.е. вероятността за допускане на грешка от втори род (β).

Според концепцията на Нейман – Пирсън статистическата значимост е свойство на теста, а не на емпиричните данни, като няма установени стандартни величини за равнището на значимост, което варира в зависимост от преценката на ползите и вредите от приемането на невярна нулева или алтернативна хипотеза. Когато вероятността за даден резултат или още по-екстремна от него при валидна нулева хипотеза не надвишава възприетото равнище на значимост, е налице статистически значим резултат, при който е желателно поведението да бъде такова, като че нулевата хипотеза не е вярна. Статистическата значимост по Нейман – Пирсън е свързана с т.нар. индуктивно поведение (Neuman and Pearson, 1933), като не е възможно и логично нейното степенуване. За разлика от теста за значимост по Фишер тя не се обвързва с величината на равнището на значимост. По-малкият допустим риск за отхвърляне на вярна нулева хипотеза не означава автоматично по-голяма статистическа значимост на резултата, водещ до приемане на алтернативната хипотеза. Докато при концепцията на Фишер по правило стремежът е „доказването“ на значимостта на някакво различие, свързано с разпределението, заложено в единствената хипотеза, и емпиричното разпределение в извадката, при концепцията на Нейман – Пирсън двете конкуриращи се хипотези са равностойни и предварителната нагласа невинаги изисква отхвърляне на нулевата хипотеза. Когато целта е приемане на нулевата хипотеза, подсигуриятването на резултата зависи от по правило неизвестната вероятност β за допускане на грешка от втори род, която е в обратна зависимост с равнището на значимост. Колкото е по-малка вероятността α , толкова по-предпазливо се подхожда към отхвърляне на H_0 , т.е. толкова по-голям е рискът тя да бъде приета, въпреки че не е вярна, при равни други условия. Подсигуриятването на приемането на H_0 може да се осъществи чрез избор на сравнително високо равнище на значимост, т.е. толериране на по-голям риск за неправилно отхвърляне на вярна нулева хипотеза. Hartung (2005) препоръчва при тестове за съответствие между емпирично и дадено теоретично разпределение избор на много високо равнище на значимост – например $\alpha = 0.9$, тъй като при тях целта е потвърждаване на предположението, че извадката произхожда от съвкупност с разпределение, съответстващо на зададеното в нулевата хипотеза. Ако въпреки високото равнище на значимост нулевата хипотеза бъде приета, тогава рискът, действителното състояние да не съответства на зададеното чрез H_0 , ще бъде много малък, но не може да се заключи, че тя е вярна, тъй като величината на β е неизвестна.

На базата на сравнението на двете концепции за проверка на хипотези, на които стъпва съвременната класическа теория, могат да бъдат изведени следните основни различия.

1. Концепцията на Фишер се основава на тест за значимост, осъществяван с помощта на една хипотеза, докато тази на Нейман – Пирсън включва тест на две конкуриращи се хипотези.
2. Концепцията на Фишер цели индуктивно заключение, а тази на Нейман – Пирсън – индуктивно поведение.
3. За равнището на значимост по Фишер няма еднозначна дефиниция, като в началото то се възприема като свойство на теста, по-късно – като свойство на данните, и липсва аргументация за установените стандартни величини. Според Нейман – Пирсън равнището на значимост е допустимият риск за грешка от първи род, който се задава предварително на базата на оценка на последствията при допускане на грешно заключение и представлява свойство на теста.
4. Поради липса на алтернативна хипотеза при концепцията на Фишер вероятността за допускане на грешка от втори род и мощността на критерия остават извън полезрението.
5. Според Фишер статистическата значимост е свойство на емпиричните данни, което подлежи на степенуване в зависимост от равнището на значимост и се отнася за конкретната извадка, а според Нейман и Пирсън – свойство на теста, което не е обвързано с величината на равнището на значимост и не е логично неговото степенуване.

2. Статистическата значимост – инструмент за създаване на псевдо-наука

„Резултатът е статистически значим“ е магически израз, който придава нужната тежест на защитаваната научна теза и се възприема като основен аргумент в нейна полза, като по правило не се разсъждава върху смисъла на понятието „статистическа значимост“ и логическия подход, довел до този резултат. Както беше посочено, съвременната класическа проверка на хипотези е своеобразен хибрид между концепцията на Фишер и тази на Нейман и Пирсън, като водеща роля има втората, която с помощта на алтернативна хипотеза и мощност на критерия позволява по-комплексен и логически издържан подход при осъществяване на проверката. Логическата несъвместимост на двете концепции е причината, поради която на тяхна основа е създаден комбиниран подход, който не е лишен от вътрешни противоречия и създава възможности както за недоразумения и заблуди, така и за целенасочена манипулация при интерпретацията на получените резултати. Този подход в учебниците по статистика и при преподаването на статистически тестове обикновено се представя чисто алгоритмично под формата на нетърпящ възражения „ста-

тистически ритуал“ или „статистически рецептурник“ (Gigerenzer, Krauss, 2000), като проблемните моменти услужливо биват подминавани. „Една готварска рецепта за тестване на хипотези в края на краищата не би следвало да поощрява разсъждения относно цената“ (Gigerenzer, Krauss, 2000). Така се симулира логическа състоятелност и се създават предпоставки за съзнателна и несъзнателна неправилна интерпретация.

Докато в учебната и специализираната литература теорията за проверката на хипотези в повечето случаи все още стъпва върху концепцията на Нейман – Пирсън, като се разглежда логиката на статистическите критерии като случайни величини с определено вероятностно разпределение, емпиричните характеристики – като техни реализации, получени на базата на конкретни случайни извадки, и критичните величини – като квантили на разпределението на статистическия критерий при валидна нулева хипотеза от порядък, зависещ от възприетото равнище на значимост, с навлизането на статистическия софтуер традиционният начин за вземане на решение отстъпва място на подход, при който до заключение се стига с помощта на p -величината на Фишер. В някои софтуерни пакети (напр. SPSS) тя е означена със „Sig.“, т.е. значимост (*Significance*). Съкращението „Sig.“ интуитивно се обвързва с равнището на значимост (*significance level*), като понякога величината бива наричана гранично равнище на значимост (Хаджиев, 2002), точна вероятност за допускане на грешка от I тип (Калинов, 2013), емпирично равнище на значимост или критично равнище на значимост (Ross, 2006) и по този начин донякъде съответства на т. нар. екзактно равнище на значимост на късния Фишер, представляващо свойство на данните, за което не е необходимо да бъде сравнявано с предварително установена норма.

Както равнището на значимост α , така и p -величината представляват вероятности, свързани с разпределението на статистическия критерий, но логиката на двете величини е различна. Докато равнището на значимост се задава предварително и представлява допустимият риск за отхвърляне на вярна нулева хипотеза, p -величината е обвързана с резултата от теста, т.е. с конкретната реализация на статистическия критерий, получена въз основа на данните от случайната извадка. Колкото е по-малка тази вероятност, толкова по-екстремна е резултатът от теста за разпределението на статистическия критерий при валидна нулева хипотеза. Ако p -величината е по-малка от зададеното предварително равнище на значимост, тогава емпиричната характеристика попада в отрязаните краища на разпределението и се налага отхвърляне на нулевата хипотеза. В този случай се говори за статистическа значимост на резултата при зададеното равнище на значимост (Lambova, 2016).

Тук ще бъде обърнато внимание на няколко проблемни момента, които поставят под въпрос приписването на прекалено голяма познавателна тежест на

резултати, които на базата на „ритуална“ проверка на хипотези са категоризирани като статистически значими.

1. Използване на p -величината извън контекста на комплексната теория на Нейман – Пирсън като единствен критерий за дихотомна категоризация на резултатите – статистически значими и статистически незначими.
2. На базата на непознаване логическата същност на величините, използвани като критерии за вземане на решение, неправилна интерпретация на статистическата значимост.
3. Използване на статистическата значимост като механично доказателство за спорни тези.
4. Степенуване на статистическата значимост.
5. Целенасочено търсене на статистически значим резултат чрез т.нар. *p-Hacking*.

С навлизането на потребителския статистически софтуер p -величината постепенно се превръща в единствен критерий, на базата на който се преценява дали резултатът да бъде квалифициран като статистически значим, или незначим. Потребителският подход все повече се отклонява от комплексната теория на Нейман – Пирсън, формираща ядрото на съвременната класическа проверка на хипотези.

Обикновено при работа с p -величината извън полезрението остава втората възможност за допускане на невярно заключение: приемането на невярна нулева хипотеза, т.е. вероятността за допускане на грешка от втори род (β). Единственият стремеж се състои в получаването на възможно най-малка p -величина, чрез която да бъде „доказана“ значимостта на някакво различие. В много случаи при използване на статистически софтуер дори не се определя предварително равнище на значимост, като допустимият риск за грешка от първи род се нагласява впоследствие, след като вече е ясен резултатът от теста. Според Hartung (2005) всяка нулева хипотеза може да бъде отхвърлена, когато равнището на значимост последващо бъде зададено малко по-голямо от p -величината. Ясно е, че подобен подход влиза в противоречие с изискванията на класическата теория за проверка на хипотези, като получените резултати могат да бъдат определени като псевдонаучни и в много случаи заслужават да бъдат обозначени с наложилото се в разговорния език понятие „стъкмистика“. Изискването за предварително задаване на равнището на значимост според Hartung (2005) е изискване, свързано с почтеността на статистиците.

Не се взема под внимание, че при равни други условия колкото е по-малко равнището на значимост, толкова по-голяма е вероятността β за приемане на невярна нулева хипотеза, т.е. расте рискът да бъде подценен действително съществуващ ефект. При категоризация на резултата като статистически незначим, въпреки че в действителност е вярна алтернативната хипотеза, е възможно например да бъде констатирана липсата на ефект от действието на нов

лекарствен препарат, който в действителност е по-добър от вече допуснатите за употреба, и по този начин пациентите да бъдат лишени от възможност за ефективно лечение.

Като критична граница за „статистическата значимост“ на резултатите се е наложила величината $p = 0.05$, заимствана от Фишер, въвел я като стандарт без никаква обосновка. Това означава, че като критерий за преценка на резултата се използва една напълно произволно избрана величина, която не е свързана с анализ на последствията при приемане на невярна или отхвърляне на вярна нулева хипотеза.

Въпреки че p -величината е удобен инструмент при осъществяването на статистически тестове, от теоретична гледна точка използването ѝ не е напълно обосновано. Rüger (2002) посочва следните възражения срещу приложението на този подход.

1. *Изолираната p -величина все още не представлява статистически тест, тя е само резултат от наблюдението на конкретна извадка, като действителното заключение относно H_0 и H_1 се отлага за по-късен етап и се предоставя на следващ наблюдател, който въз основа на въведено от него равнище на значимост да прецени дали да обяви, че е налице статистически значимо различие, или да приеме нулевата хипотеза.*
2. *Чрез използването на p -величини възниква следната опасност от злоупотреба със статистическите тестове: наблюдаваната p -величина се интерпретира като равнище на значимост, по-конкретно като равнището на значимост, при което въпросният тест е довел до отхвърляне на нулевата хипотеза. Тази интерпретация подвежда към подход, при който първо се определя p и ако величината не е прекалено голяма, се задава равнище на значимост, при което алтернативната хипотеза да бъде обявена за статистически значима.*
3. *Наблюдаваната p -величина, стриктно погледнато, изобищо не представлява честотно интерпретируема вероятност, камо ли това е вероятността, с която при въпросния тест може да се допусне грешка от първи род. Поради тази причина статистически заключения, базирани се само на определянето на p , не са съвместими с честотните принципи на класическата теория за проверка на статистически хипотези. Само събитие, получено с помощта на предварително зададено равнище на значимост ($p \leq \alpha$), притежава честотно интерпретируема вероятност и това е α .*

Според Rüger (2002) посочените възражения стават безпредметни единствено когато при осъществяването на даден статистически тест наблюдаваната p -величина се обвърже с предварително зададена допустима вероятност

за грешка от първи род α , при което α се възприема като равнището на значимост на теста, а p – като неговата стандартизирана емпирична характеристика. Въпреки тази възможност за коректно прилагане на p -величината класическата теория за проверка на статистически хипотези не използва този инструмент, като освен споменатата в много случаи тенденциозна злоупотреба чрез последващо нагласяване на равнището на значимост според Rüger (2002) се откроява още следната основна причина: при използването на p -величината вече не е непосредствено видимо кой е първоначалният статистически критерий, залегал в основата на теста и отговорен за степента му на надеждност. Теоретичните постановки относно статистическите тестове са обвързани тясно със статистическия критерий. Само чрез неговото познаване е възможно да се разбере логическата същност на теста. При осъществяването на тестове с помощта на p -величина тази логика става невидима за ползвателя и последният не е в състояние да осъзнае напълно това, което прави.

Проблеми възникват и при интерпретацията на статистическата значимост, като при непознаване на логическата същност на величините, използвани като критерии за вземане на решение, тя често се обвързва със сигурността, с която се приема алтернативната хипотеза, и по този начин чрез нея се симулира необосновано голяма надеждност на получените резултати.

По какъв начин следва да бъде интерпретирана статистическата значимост на резултата, при положение че предварително е зададено определено равнище на значимост? Обикновено при „рецептурно“ преподаване и осъществяване на статистически тестове на този въпрос не се обръща достатъчно внимание, като това води до грешки и заблуди при тълкуването на получените резултати. Проблеми с интерпретацията имат не само ползвателите, но дори и преподавателите.

Осъществено е емпирично изследване с цел установяване на възприятието за статистическа значимост (Krauss, Wassner, 2001). Като респонденти в проучването са участвали 44 студенти по психология и 69 преподаватели от берлински университети, които е трябвало да отговорят на следния въпрос:

Представете си, че с помощта на t -тест проверявате дали откритото различие между две извадки е статистически значимо. Оказва се, че при равнище на значимост, възлизащо на 1%, разликата е статистически значима. Кое от следните твърдения следва да бъде изведено на базата на тази фактология? (възможно е да бъдат верни повече от едно твърдения или нито едно от тях).

1. Еднозначно е доказано, че нулевата хипотеза е невярна.
2. Откритата е вероятността за валидност на нулевата хипотеза.
3. Еднозначно е доказано, че алтернативната хипотеза е вярна.
4. Може да бъде изведена вероятността, с която е вярна алтернативната хипотеза.

5. Ако бъде направено заключение за отхвърляне на нулевата хипотеза, тогава е известна вероятността това заключение да бъде неправилно.
6. Ако експериментът бъде повторен много пъти, то в 99% от случаите би се стигнало до статистически значим резултат.

Всяко от тези твърдения е невярно, въпреки това 100% от студентите и 86% от преподавателите са предположили, че едно или няколко са верни. Твърденията са грешни, тъй като понятието „статистическа значимост“ имплицира винаги „свят“, в който H_0 е вярна (Krauss, Wassner, 2001). Следователно вероятността на конкретните резултати би трябвало да се разглежда на базата на предположението за вярна нулева хипотеза. Ако например $p = 0.02$, правилно ще бъде да се твърди, че в случай на вярна нулева хипотеза вероятността за сбъждане на получения или по-екстремален от него резултат възлиза на 2%. Обратното твърдение, според което при сбъждане на получения резултат вероятността нулевата хипотеза да е вярна, възлиза на 2%, няма да бъде вярно. Също така няма да е правилно твърдението, че нулевата хипотеза е отхвърлена при риск за допускане на грешка от първи род, възлизащ на 2%. Ако предварително е зададено равнище на значимост $\alpha \leq 0.05$, полученият резултат $p = 0.02$ ще доведе до квалифициране на резултата като статистически значим. Тогава може да възникне въпросът: каква е вероятността действително да е налице статистически значимо различие? Удобният, но грешен отговор е обратното твърдение: $(1 - p) 100 = 98\%$.

Подобни неверни твърдения относно вероятности за сбъждане симулират надеждност на резултатите, превръщат статистическата значимост в еднозначно и необоримо доказателство за верността на защитавани тези, с което може да се придаде наукообразен характер дори на напълно нелогични изследвания. Несъстоятелността на изводи от този род Dubben и Beck-Bornholdt (2006) онагледяват с помощта на примери, един от които е особено показателен: *Ако дадено същество е човек, то тогава с вероятност 50% то е мъж. Обратно твърдение: ако някой е мъж, то тогава той с 50% вероятност е човек.*

При $\alpha = 0.05$ и действително вярна нулева хипотеза тестът със сигурност 95% ще покаже отсъствие на статистически значими различия. При $\alpha = 0.05$ и $p = 0.02$ тестът води до заключение за статистическа значимост на различието, но не е възможно да се правят заключения относно вероятността, с която резултатът е статистически значим.

Погрешно тълкуване се допуска и в случаите, когато тестът доведе до приемане на нулевата хипотеза. Въз основа на осъществен анализ на 791 статии, публикувани в пет научни списания, е установено, че 51% от тях включват неправилна интерпретация на статистически незначим резултат, като се твърди, че проверяваният ефект не е налице (Amrhein, Greenland, McShane, 2019).

Неправилната интерпретация на статистическата значимост, нейното откъсване от логическата същност на величините, използвани като критерии за формиране на заключение, води до използването ѝ като чисто механично доказателство за съвместимостта с практиката на конструирани на базата на предположения теоретични модели. От една страна, не се взема под внимание, че формиран на базата на случайността резултат не може да бъде доказателство за съществуването на каквато и да е закономерност, а от друга – извън полезрението остава произволният характер на използваните стандартни гранични стойности на a и/или p , чрез които се издава „присъдата“ относно наличието или отсъствието на статистическа значимост на различието. Както вече беше посочено, при „рецептурното“ приложение на статистически тестове обикновено вероятността за допускане на грешка от втори род β остава на заден план, като предварително не се прави анализ на последствията при приемането на невярна или отхвърлянето на вярна нулева хипотеза, целящ определяне на подходяща величина на равнището на значимост, чрез която косвено да се контролира β . Използва се преди всичко всеобщо възприетата стандартна величина $a = 0.05$, която според Dubben и Beck-Bornholdt (2010) придава на науката хазартен характер, тъй като примерно се толерира в действителност неефективен лекарствен препарат с вероятност 5% случайно да покаже статистически значимо действие. Те сравняват научните изследвания с игра с двадесетостранен зар, като всяка хвърлена двадесетица съответства на една научна публикация, в която статистическата значимост е резултат на случайността (т.е. в действителност не е налице съответният ефект – бел. на автора).

Не би трябвало да се подценява и обратният ефект от налагането на механично заключение относно статистическа значимост или незначимост на резултата на базата на наложения стандарт от 5%. Също така случайно може за обещаващо успех ново лекарствено средство при даден експеримент да се стигне до $p = 0.06$ и на тази основа то да бъде квалифицирано като неефективно (Gigerenzer, Krauss, 2000).

Не на последно място стои проблемът с логическата състоятелност на степенуването на статистическата значимост, наложило се в случаите, когато като мерило се използва единствено p -величината. Както вече беше посочено в т. 1, при $a \leq 0.05$ се приема, че резултатът е статистически значим, при $p \leq 0.01$ – много значим, а при $p \leq 0.001$ – изключително значим, като произволно установените гранични величини са заимствани от стандартните величини за равнището на значимост на ранния Фишер, без за се разсъждава върху последствията при приемане на невярна нулева хипотеза.

При степенуването на статистическата значимост на резултата не се взема под внимание, че „ p -величината не е показателна за размера на действителния ефект, нито е мерило за вероятността за допускане на грешка от първи род“ (Nuzzo, 2014).

В крайна сметка, при „рецептурната“ проверка на хипотези всичко е подвластно на постигането на „статистически значимо“ заключение, което потвърждава предварителната теза, следователно на получаването на p -величина, която е по-малка от 0.05. За манипулирането на данни и целенасоченото търсене на „статистически значим“ резултат се използват термините *p-Hacking*, *data dredging* и *significance chasing* (Nuzzo, 2014). Според изследване на Simonsohn (2011) в публикувани студии по психология има струпване на p -величини в близост до 0.05, което може да се очаква, когато изследователите толкова дълго са били на лов за значими p -величини, докато подходящата им е попаднала в мрежата.

Предговорът на статията „Пратете статистическата значимост в пенсия“ (Amrhein, Greenland, McShane, 2019) започва по следния начин:

„Фенове на книгата „Пътеводител на галактическия стопаджия“ познават отговора за живота, вселената и всичко останало: 42. Закачката тук естествено е свързана с това, че истината не може да се разкрие чрез едно-единствено число. Но точно с такава цел често се използва p -величината: като мярка за това колко изненадващ е даден резултат при определени предположения – например, че не е налице ефект. Дали p -величината лежи под или над произволна граница за статистическа значимост – например 0.05, определя дали хипотези ще бъдат потвърдени, студии ще бъдат публикувани или продукти ще бъдат пуснати на пазара. Привличането на p -величината като единствен арбитър обаче може да доведе до небалансирани преценки, превъзнасяне на лъжеположителни резултати и незабелязване на действителни ефекти“.

Самата статия представлява своеобразен призив за отказ от понятието „статистическа значимост“, подписан освен от тримата автори от още над 800 статистици и изследователи, според които p -величината от статистически инструмент се е превърнала в тиранин, който налага нетърпяща възражения дихотомна категоризация на резултатите – статистически значими и статистически незначими, без възможност за комплексен анализ на конкретната ситуация. Според тримата автори заблудата, че прекрачването на границата към статистическата значимост означава „действителен“ резултат, за нещастие води до това, учени и редактори на научни издания да предпочитат подобни резултати. По този начин те изкривяват отразяването на изследователската дейност в специализираната литература. Статистически значими резултати са в голяма степен преекспонирани, докато статистически незначими се срещат несъразмерно рядко.

Сляпото фокусиране върху статистическата значимост обуславя целенасоченото търсене на данни и методи, които да са в състояние да осигурят такава, достатъчна за публикуването на даден резултат, независимо от логическата състоятелност на защитаваната теза.

В крайна сметка, може да се твърди, че проблемните моменти, предизвикани от съвременната „рецептурна“ проверка на хипотези, подпомагат развитието на псевдонаука, стъпваща върху „плаващите пясъци“ на статистическата значимост.

Заклучение

В съответствие с поставената цел въз основа на теоретичните особености и логиката на понятието „статистическа значимост“ са разкрити проблемни моменти и възможности за злоупотреби с него при проверката на статистически хипотези.

Изследването включва две основни направления:

- разкриване на теоретичните основи и логиката на понятието „статистическа значимост“ на базата на сравнителен анализ на двете концепции, на които стъпва съвременната класическа теория за проверка на хипотези;
- разсъждения относно проблемни моменти, свързани с понятието „статистическа значимост“, които поставят под въпрос приписването на прекалено голяма познавателна тежест на резултати, които на базата на „ритуална“ проверка на хипотези са категоризирани като статистически значими.

Фактологията и разсъжденията, свързани с първото направление, позволяват следните обобщения.

1. Налице са съществени логически противоречия между двете концепции, от които е произлязла съвременната класическа теория за проверка на хипотези.
2. Концепцията на Фишер се основава на тест за значимост, осъществяван с помощта на една хипотеза и целящ индуктивно заключение, докато тази на Нейман – Пирсън включва тест на две конкуриращи се хипотези, като целта е индуктивно поведение.
3. За равнището на значимост по Фишер няма еднозначна дефиниция, като в началото то се възприема като свойство на теста, по-късно като свойство на данните и липсва аргументация за установените стандартни величини. Според Нейман – Пирсън равнището на значимост е допустимият риск за грешка от първи род, който се задава предварително на базата на оценка на последствията при допускане на грешно заключение и представлява свойство на теста.
4. Според Фишер статистическата значимост е свойство на емпиричните данни, което подлежи на степенуване в зависимост от равнището на значимост и се отнася за конкретната извадка, а според Нейман и Пирсън – свойство на теста, което не е обвързано с величината на равнището на значимост и не е логично неговото степенуване.

Въз основа на разсъжденията и аргументите, свързани с второто направление на изследването, могат да бъдат направени следните основни изводи.

1. Статистическата значимост, чрез която резултатите се групират в две взаимноизключващи се категории, създава предпоставки за неправилна интерпретация.
2. При „рецептурното“ приложение на статистически тестове статистическата значимост позволява механично „доказване“ дори на логически несъстоятелни тези.
3. При „лова“ за колкото се може „по-значими“ резултати в много случаи извън ползването остава рискът за допускане на грешка от втори род, като статистическата значимост не се обвързва с предварителен анализ на тежестта на последствията при евентуално приемане на невярна или отхвърляне на вярна нулева хипотеза.
4. На базата на една-единствена произволно избрана величина ($p = 0,005$) се издава „присъда“ на получения резултат, която решава дали той (оправдано или неоправдано) да бъде превъзнесен до небесата, или да потъне в научното забвение.
5. Проблемите, които са предизвикани от съвременната „рецептурна“ проверка на хипотези, превръщат статистическата значимост по-скоро в препъникамък, отколкото в панацея, като може да се твърди, че подпомагат развитието на псевдонаука, основаваща се на изолирани от цветната палитра черно-бели резултати.

NOTES

1. <http://mathshistory.st-andrews.ac.uk/Biographies/Arbuthnot.html>.
2. <http://mathshistory.st-andrews.ac.uk/Biographies/Neyman.html>.
3. http://mathshistory.st-andrews.ac.uk/Biographies/Pearson_Egon.html.
4. <http://mathshistory.st-andrews.ac.uk/Biographies/Fisher.html>.
5. <http://mathshistory.st-andrews.ac.uk/Biographies/Pearson.html>.

ЛИТЕРАТУРА

- Калинов, Кр. (2013). *Статистически методи в поведенческите и социалните науки*. София: Нов български университет.
- Ламбова, М. (2016). Популярни заблуди при проверката на статистически хипотези, *Статистика*, НСИ, 2016, 3, 59 – 74.
- Ламбова, М., Русев, Ч., Косева, Д., Стоянова, В. (2012). *Въведение в статистиката*. Варна: СТЕНО.
- Хаджиев, В. (2002). *Статистически и иконометричен софтуер*. Варна: Университетско издателство ИУ – Варна.

REFERENCES

- Kalinov, Kr. (2013). *Statisticheski metodi v povedencheskite i sozialnite nauki*. Sofia: Nov bulgarski universitet.
- Lambova, M. (2016). Populyarni zabludi pri proverkata na statisticheski hipotezi. *Statistika*, NSI, 3, 59 – 74.
- Lambova, M., Rusev, Ch., Koseva, D., Stoyanova, V.(2012). *Vavedenie v statistikata*. Varna: STENO.
- Hadzhiev, V. (2002). *Statisticheski i ikonometrichen softuer*. Varna: Universitetsko izdatelstvo IU – Varna.
- Amrhein, V., Greenland, S. & McShane, B. (2019). Schickt die statistische Signifikanz in den Ruhestand, *Spektrum – die Woche*, 52/2019, S. 42 – 50.
- Bewersdorff, J. (2011). *Statistik – wie und warum sie funktioniert: Ein mathematisches Lesebuch*. Wiesbaden: Vieweg+Teubner Verlag.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer Verlag.
- Dubben, H.-H. & Beck-Bornholdt, H.-P. (2010). *Der Hund, der Eier legt. Erkennen von Fehlinformation durch Querdenken*. Reinbek bei Hamburg: Rowohlt Verlag.
- Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd.
- Fisher, R. (1935). *The Design of Experiments (2 ed.)*. Edinburgh: Oliver & Boyd.
- Gigerenzer, G. & Krauss, S. (2000). Statistisches Denken oder statistische Rituale? Was sollte man unterrichten, Anregungen zum Stochastikunterricht. *Die NCTMStandards*, S. 53 – 62.
- Hartung, J., Elpert, B. & Klösner, K.-H. (2005). *Statistik. Lehr- und Handbuch der angewandten Statistik*. München: Oldenbourg Verlag.
- Krauss, S. & Wassner, C. (2001). *Wie man das Testen von Hypothesen einführen sollte*. *Stochastik in der Schule*, 21 (1), S. 29 – 34.
- Lehmann, E. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?, *Journal of the American Statistical Assotiation*, 88 (424), pp. 1242 – 1249.
- Neyman, J. & Pearson, E. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231, pp. 289 – 337.
- Nuzzo, R. (2014). Der Fluch des p-Werts. *Spektrum der Wissenschaft*, 9/2014, S. 52 – 56.
- Rüger, B. (2002). *Test- und Schätztheorie. Band II: Statistische Tests*. München, Wien: Oldenbourg Verlag.

- Saint-Mont, U. (2011). *Statistik im Forschungsprozess: Eine Philosophie der Statistik als Baustein einer integrativen Wissenschaftstheorie*. Heidelberg: Physica-Verlag HD.
- Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22/2011, pp 1359 – 1366.

STATISTICAL SIGNIFICANCE – IS IT A PANACEA OR A PITFALL?

Abstract. Thoughts and considerations are presented about the term “statistical significance” – a common tool for “proving” the compatibility with practice of theoretical models designed on the ground of assumptions. In accordance with the set goal and based on the theoretical features and logics of the concept, there are revealed problematic issues and possibilities for abusing it when verifying statistical hypotheses. The suggested argumentation enables one to claim that the “prescriptive” verification of statistical hypotheses creates conditions for developing a pseudoscience supported by incorrect conclusions which are made on the ground of “statistically significant” results.

Keywords: statistical significance; statistical hypothesis testing; *p-value*; significance level; pseudoscience

✉ **Dr. Margarita Lambova, Assoc. Prof.**

University of Economics – Varna
77, Knyaz Boris I Blvd.
9002 Varna, Bulgaria
E-mail: lambowa@ue-varna.bg