ПОПУЛЯРНИ ЗАБЛУДИ ПРИ ПРОВЕРКАТА НА СТАТИСТИЧЕСКИ ХИПОТЕЗИ

Маргарита Ламбова[*]

Въведение

Статистическите тестове са един от най-популярните статистически инструменти, използвани както при работа със случайни извадки, така и при изследването на динамични редове, при които обикновено се приема, че са налице условията за провеждане на случаен експеримент. При проверката на статистически предположения предварителната нагласа в повечето случаи е свързана с отхвърляне на нулевата хипотеза, което е равнозначно на доказване на статистическата значимост на някакво различие, като само при подобен резултат определени последващи действия биха имали смисъл. Това означава, че в повечето случаи теоретичният модел, който сме построили на етапа на качествения анализ на ситуацията, би се потвърдил единствено, ако на базата на емпирични данни се стигне до заключение, според което има основания за отхвърляне на нулевата хипотеза и приемане на алтернативната. Тъй като за статистическите тестове има създаден потребителски софтуер, който не изисква осмисляне на тяхната логика и позволява чисто алгоритмичното им приложение, много често се допускат съществени логически грешки при интерпретацията на резултатите, като по този начин се придава несъществуваща реално вероятностна тежест на потвърждението на състоянието, което е предвидено от теоретичния модел. Вниманието тук ще бъде насочено основно към величините, на базата на които се стига

[*] Доц. д-р, катедра „Статистика и приложна математика“, Икономически университет - Варна; e-mail: lambowa@yahoo.de.

до заключение, и към сигурността, с която се приема, респ. отхвърля, нулевата хипотеза.

Целта се състои в разкриване на заблудите по отношение на т.нар. *p*-величина (*p-value*), зададена като показател за вземане на решение в статистическия софтуер, както и по отношение на вероятността, с която приетата хипотеза е вярна или невярна.

I. За същността на *p*-величината и заблудите при нейното използване

При традиционния подход на проверка на статистически хипотези, основаващ се на концепцията на Нейман - Пирсън, предварително се задава допустимият риск за грешка от първи род α, т.е. допустимата вероятност, с която вярна нулева хипотеза ще бъде отхвърлена на базата на случайните резултати от теста, като тази величина се нарича „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 екстремен 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 риска за допускане на грешка от първи род, който сме склонни да толерираме. Ако напр. $\alpha = 0.05$, вероятностната маса в отрязаните краища на разпределението възлиза на 5%, което означава, че вероятността вярна нулева хипотеза да бъде отхвърлена въз основа на резултата от теста, не бива да превишава 5%. Съответно вероятността вярна нулева хипотеза да не бъде отхвърлена, трябва да възлиза на минимум $1-\alpha = 0.95$. При традиционния подход заключението се прави въз основа на сравнението на емпиричната характеристика, представляваща конкретна реализация на статистическия критерий, чието разпределение при вярна 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. Например при двустранен параметричен z-тест границите ще бъдат квантили от порядък $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$, т.е. $z_{\frac{\alpha}{2}}$ и $z_{1-\frac{\alpha}{2}}$. При $\alpha = 0.05$ това означава, че при вярна нулева хипотеза вероятността за реализация на стандартно нормално разпределения статистически критерий, която не превишава лявата гранична стойност $z_{\frac{\alpha}{2}} = z_{0.025} = -1.96$, е равна на 2.5%, т.е. вероятностната маса на левия отрязан край на разпределението (лявата критична област) възлиза на 2.5%. Вероятността за реализация на статистическия критерий, която не превишава дясната гранична стойност $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$, възлиза на 97.5%, т.е. вероятностната маса на десния отрязан край на разпределението също е 2.5%. Вероятността статистическият критерий да приеме значение между двете граници при вярна нулева хипотеза, възлиза съответно на $1-\alpha = 0.95$, като тя определя сигурността (95%), с която **вярна** нулева хипотеза ще бъде приета на базата на резултатите от теста.

С навлизането на статистическия софтуер традиционният начин за вземане на решение постепенно отстъпва място на подход, при който до заключение се стига с помощта на т.нар. p-величина (p-*value*), която за първи път е въведена от Роналд Фишер през 20-те години на 20. век, но при друга концепция за осъществяване на статистически тестове (работа само с нулева хипотеза, без възможност за смислена интерпретация на β, съответно на мощността на критерия). Общата логика е почти същата, но в известна степен е по-неразбираема, особено за ползвателите на статистически софтуер, които не са добре запознати със статистическата методология, като това води в определени случаи до неправилна интерпретация на резултатите.

Какво всъщност представлява p-величината, която в някои софтуерни пакети (напр. SPSS) 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). Логиката на двете величини обаче е различна, въпреки че както равнището на значимост α, така и p-величината представляват вероятности, свързани с разпределението на статистическия критерий. 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. **Това е вероятността, с която при вярна нулева хипотеза статистическият критерий може да приеме значение, което е равно на получената емпирична характеристика или е още по-екстремно в посока на алтернативната хипотеза** (Rüger, 2002). Според Nuzzo (2014) p-величината е вероятността, с която даден емпиричен резултат е възникнал без наличие на предполагаемия ефект (различие, заложено в алтернативната хипотеза - бел. авт.) на базата на случайността.

Ако се осъществява например едностранен z-тест с дясна критична област, p-величината може да се изрази по следния начин:

$$p = W(Z \geq z / H_0),$$

където:

W е вероятност;

Z - стандартно нормално разпределения статистически критерий на теста;

z - емпиричната характеристика на теста като конкретна реализация на статистическия критерий;

т.е. вероятност стандартно нормално разпределената случайна величина Z, използвана като статистически критерий на теста, при вярна нулева хипотеза да приеме значение, което е равно на или по-голямо от получената емпирична характеристика z.

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. Колкото по-малка е тази вероятност, толкова по-екстремен е резултатът от теста за разпределението на статистическия критерий при вярна нулева хипотеза. Ако p-величината е по-малка от предварително зададеното 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.

Проблеми възникват при интерпретацията на p-величината, като при непознаване на логическата ѝ същност тази вероятност се обвързва със сигурността, с която се приема алтернативната хипотеза и се интерпретира като равнище на значимост, при което е отхвърлена нулевата хипотеза. Ако например $p = 0.03$, правилното твърдение гласи, че в случай на вярна нулева хипотеза вероятността за сбъдване на получения или по-екстремен от него резултат възлиза на 3%. **Заблуда обаче представлява обратното твърдение, според което при сбъдване на**

получения резултат вероятността нулевата хипотеза да е вярна, възлиза на 3%. Също така няма да е правилно твърдението, че нулевата хипотеза е отхвърлена при риск за допускане на грешка от първи род, възлизащ на 3%.

Ако предварително е зададено равнище на значимост $\alpha = 0.05$, полученият резултат $p = 0.03$ ще доведе до отхвърляне на нулевата хипотеза и приемане на статистическата значимост на различието. В подобни случаи обикновено се задава въпросът: Колко вероятно е действително да е налице статистически значимо различие, т.е. да е вярна приетата алтернативна хипотеза? Популярният отговор гласи: Вероятността заключението да не е вярно, възлиза на 3%, следователно вероятността правилно да сме приели алтернативната хипотеза е $(1-p)100 = 97\%$. За съжаление, този отговор не е верен.

Подобно обръщане на твърденията относно вероятности за сбъдване придава привидна тежест на резултата, към който се стремим. Несъстоятелността на изводи от този род Dubben и Beck-Bornholdt (2006) онагледяват с помощта на следните примери:

- 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.

- Ако дадено лице е заболяло, тогава тестовият резултат с 99% вероятност е положителен. Обратно твърдение: ако тестовият резултат е положителен, тогава лицето е заболяло с 99% вероятност.

- Когато участваме в спортния тотализатор, е много вероятно да не познаем всички числа. Обратно твърдение: ако познаем всички числа, е много вероятно да не сме участвали в спортния тотализатор.

- Ако нулевата хипотеза е вярна, тогава полученият резултат ($p = 0.0404$) е почти невероятен. Обратно твърдение: ако се е сбъднал резултатът, тогава нулевата хипотеза е почти невероятна ($p = 0.0404$).

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). Освен това не се обръща достатъчно внимание на емпиричната характеристика на теста, която фигурира като резултат в популярните статистически софтуерни продукти и също би могла да се използва при вземането на решение. Единственият стремеж се състои в получаването на възможно най-малка p-величина, чрез която да бъде „доказана“ значимостта на някакво различие. 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 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. Като критична граница за „статистическата значимост“ на резултатите се е наложила величината $p = 0.05$. 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. 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 съществуват термините *„p-Hacking“*, *„data dredging“*, и „*significance chasing“* (Nuzzo, 2014). Според 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. 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 статистическата методология, като получените резултати са 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.

P-величината е удобен инструмент при осъществяването на статистически тестове, даващ възможност с помощта на една-единствена величина да се стигне до заключение, без да е необходимо предварително задаване на допустимите рискове, както и осмислянето на логическата същност на процедурата. От теоретична гледна точка използването ѝ обаче 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, изобщо 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, базиращи 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, залегнал 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 той не е в състояние да осъзнае напълно това, което прави. Тестът по този начин се превръща в черна кутия с вход и изход, като съществува риск от неправилна интерпретация на изходящите данни заради непознаването на вътрешната ѝ структура (логиката на теста).

II. За измеримостта на сигурността, с която приетата статистическа хипотеза е вярна

Риск от неправилна интерпретация на резултатите от статистическите тестове съществува не само при използване на p-величина като критерий за вземане на решение, но и при класическия подход за формиране на заключение. Популярните заблуди при интерпретацията са свързани основно с двата вида риск от допускане на неправилно заключение - α и β, и обвързването им със сигурността, с която е вярна или невярна нулевата, респ. алтернативната, хипотеза.

Както вече беше посочено, α е допустимият риск за грешка от първи род, т.е. максимално допустимата вероятност, с която вярна нулева хипотеза може да бъде отхвърлена на базата на случайните резултати от теста. Ако $\alpha = 0.05$, тогава в поне 95% от случаите вярна нулева хипотеза ще бъде приета като такава, а в останалите случаи тя неправилно ще бъде отхвърлена. β е допустимият риск за грешка от втори род, т.е. максимално допустимата вероятност, с която невярна нулева хипотеза може да бъде приета за вярна въз основа на случайните резултати от теста. Ако $\beta = 0.02$, тогава в поне 98% от случаите действително съществуващото различие (вярна алтернативна хипотеза) ще бъде разпознато като такова, а в останалите случаи то няма да бъде разпознато и ще бъде приета неправилно нулевата хипотеза. С други думи, ако в действителност е вярна нулевата хипотеза, тестът със сигурност 95% ще покаже отсъствие на статистически значими различия, а ако в действителност тя не е вярна, тестът ще регистрира статистически значимите отклонения със сигурност 98%. Често при приложението на статистически тестове интерпретацията включва огледалния образ на тези твърдения, като се разсъждава по следния начин: след като в 98% от случаите тестът разпознава статистически значимите отклонения, когато в действителност е вярна алтернативната хипотеза, то тогава при резултат, показващ наличието на статистически значимо различие, със сигурност 98% е вярна алтернативната хипотеза. По същата логика, след като тестът със сигурност 95% не регистрира статистически значими отклонения, когато в действителност е вярна нулевата хипотеза, то при резултат, показващ отсъствие на такива отклонения, със сигурност 95% е вярна нулевата хипотеза. Проблемът в случая е, че при това обратно твърдение вероятността вече не се обвързва с резултатите от теста, а с направените предположения. Алогизмът тук не е съвсем явен и рискът за неволна грешка при интерпретацията или за манипулирането ѝ с цел обличането в наукообразна форма на незадоволителни резултати е много голям. 95%, респ. 98%, сигурност, че нулевата,

респ. алтернативната, хипотеза е вярна при приемане, респ. отхвърляне, на H_0 звучи много сериозно и на пръв поглед това твърдение е необоримо, но само на пръв поглед. Да се върнем на посочения вече пример, даден от Dubben и Beck-Bornholdt (2006), построен върху същата логика, при който алогизмът е явен: вероятността даден човек да е мъж е 50%. Огледалният образ на твърдението е съответно: вероятността даден мъж да е човек е 50%. Тук не трябва да се доказва, че второто твърдение не е вярно, тъй като е ясно, че останалите 50% мъже също са от вида хомо сапиенс. По друг начин стои въпросът при грешната интерпретация на вероятностите, свързани с проверката на статистически хипотези, тъй като само чрез допълнителни изчисления на базата на априорна информация, и то при ограничен брой тестове, е възможно обвързването на двете вероятности за допускане на невярно заключение и изчисляването на действителната сигурност, с която е вярна нулевата или алтернативната хипотеза, когато резултатите от теста доведат до приемането на съответната хипотеза. По правило при повечето тестове предварително се задава само допустимата величина на риска за грешка от първи род, т.е. равнището на значимост α, докато вероятността β за допускане на грешка от втори род не подлежи на директен контрол и варира в зависимост от възприетото равнище на значимост, от обема на извадката и от размера на отклонението между зададеното в нулевата хипотеза и действителното състояние. Единствено при параметрични тестове с конкретизирана алтернативна хипотеза е възможно директното контролиране на двата вида риск за допускане на невярно заключение, като за целта се определя минимално необходим обем на извадката, позволяващ спазването на предварително зададените вероятности α и β (Ламбова, 2012). Само при подобни тестове и наличие на априорна информация за вероятността състоянието на дадена съвкупност да съответства на една от хипотезите е възможно определянето на действителната сигурност, с която е вярна нулевата, респ. алтернативната, хипотеза, когато резултатите от теста доведат до приемането на съответната хипотеза.

С помощта на хипотетичен пример ще демонстрираме несъстоятелността на популярните твърдения относно сигурността, с която се приема нулевата, респ. алтернативната, хипотеза.

Нека приемем, че от 10 000 партиди тухли 1%, т.е. 100, са некачествени. Всяка партида формира генерална съвкупност, като оценката на годността ѝ се осъществява с помощта на параметричен тест с конкретизирана алтернативна хипотеза, напр. за проверка на предположение за величината на относителен дял, при който нулевата хипотеза предполага качественост, а алтернативната -

некачественост на партидата. Величините на двата вида риск са съответно $\alpha = 0.05$ и $\beta = 0.02$.

Въпросът е, каква е вероятността при тестови резултат, който показва статистически значими отклонения от нормите, съответната партида наистина да е некачествена, съответно вероятността при резултат, който води до приемане на нулевата хипотеза, партидата да е качествена?

Вече беше посочено, че популярните, но неверни отговори при подобна ситуация биха били следните: нулевата хипотеза се приема със сигурност 95%, когато тестовият резултат не показва статистически значими отклонения от нормите; алтернативната се приема със сигурност 98%, когато тестът регистрира такива отклонения.

Въз основа на зададените рискове за неправилно заключение и априорната информация за дела на некачествените партиди може да се установи посоченото в табл. 1 хипотетично двумерно разпределение на партидите според тяхното действително състояние и според резултатите от съответния статистически тест, при което се приема, че съотношенията съответстват напълно на зададените предварително вероятности за допускане на грешка от първи и втори род, което би било валидно при безкраен брой случаи (партиди). В конкретната ситуация, когато 100 от 10 000 партиди са некачествени и се толерират $\alpha = 0.05$ и $\beta = 0.02$, се очаква в 95% от случаите качествена партида да бъде оценена като такава на базата на резултатите от теста, т.е. при 9 405 партиди ще бъде приета правилно нулевата хипотеза. В 5% от случаите, т.е. при 495 качествени партиди, тестът ще регистрира статистически значими отклонения от нормите и ще бъде приета неправилно алтернативната хипотеза. Очаква се също така в 98% от случаите статистически значимите отклонения при некачествена партида да бъдат регистрирани от теста, т.е. при 98 от общо 100 некачествени партиди ще бъде приета правилно алтернативната хипотеза. В 2% от случаите, т.е. при две некачествени партиди, се очаква тестът да не открие статистически значимо отклонение, което ще доведе до неправилно приемане на нулевата хипотеза.

1. Разпределение на 10 000 партиди според тяхното действително състояние и според резултатите от съответния статистически тест при $\alpha = 0.05$, $\beta = 0.02$ и 1% вероятност за некачественост на партидата

Действително състояние	Резултат от теста		Общ брой
	статистически значимо отклонение - приема се H_1	статистически незначимо отклонение - приема се H_0	
Некачествена партида	98	2	100
Качествена партида	495	9405	9900
Общ брой	593	9407	10000

С помощта на тази информация може да бъде определена хипотетичната сигурност, с която се приема нулевата, респ. алтернативната, хипотеза в конкретната ситуация, т.е. може да бъде намерен отговор на следните два въпроса:

1. С каква сигурност се приема нулевата хипотеза, когато резултатите от теста не показват наличието на статистически значими отклонения?

2. С каква сигурност се отхвърля нулевата хипотеза и се приема алтернативната, когато резултатите от теста показват наличието на статистически значими отклонения?

След като сме задали близки по стойност вероятности за допускане на грешка от първи и втори род, ние очакваме, че сигурността на приемане на двете хипотези също няма да се различава съществено, но при подобно съотношение на качествени и некачествени партиди се получават озадачаващи величини, които поставят под съмнение заключения от рода: тестът регистрира наличието на статистически значими отклонения, следователно е вярна алтернативната хипотеза. В конкретната ситуация очакваме тестът да не отчете статистически значими отклонения при 9 407 партиди, като само при две от тях резултатът няма да съответства на действителното състояние. Това означава, че приемането на нулевата хипотеза е с много висока степен на сигурност:

$$W_{(H_0)} = \frac{9405}{9407} = 0.99979,$$

т.е. в 99.98% от случаите приемането на нулевата хипотеза ще е вярно заключение (съответната партида в действителност е качествена).

За съжаление, същото не се отнася за приемането на алтернативната хипотеза. Очаква се общият брой на тестовете, при които резултатът показва наличието на статистически значимо отклонение, да възлиза на 593, но само при 98 от тях заключението ще е вярно. При останалите 495 ще се допусне грешка от първи род, т.е.

ще бъде отхвърлена вярна нулева хипотеза. Вероятността, с която приемането на алтернативната хипотеза е вярно заключение, в случая е следната:

$$W_{(H_1)} = \frac{98}{593} = 0.16526.$$

Следователно със сигурност 16.53% от партидата действително е некачествена, когато резултатът от теста показва наличието на статистически значимо отклонение. В 83.5% от случаите заключението ще е неправилно, т.е. партида, която в действителност съответства на нормите, ще бъде оценена като некачествена.

Отговорите на поставените два въпроса в конкретната ситуация са следните:

1. 99.98% е сигурността за това, че партидата наистина е качествена, след като резултатите от теста водят до приемане на нулевата хипотеза.

2. 16.53% е сигурността за това, че партидата наистина е некачествена, след като резултатите от теста показват наличие на статистически значими отклонения и водят до приемане на алтернативната хипотеза.

Ясно е, че сигурността, с която при приемане на нулевата, респ. алтернативната, хипотеза е направено вярно заключение, зависи в много голяма степен от априорна информация, която по правило не е известна при тестването на статистически хипотези, в случая от относителния дял на качествените, респ. некачествените, партиди (съвкупности), следователно тази сигурност не подлежи на директен контрол и измерване. При промяна на съотношенията се променя и сигурността.

Нека приемем, че само половината от 10 000 партиди са качествени, т.е. 5 000. Тогава съобразно зададените рискове за допускане на грешка от първи и втори род се получава посоченото в табл. 2 хипотетично двумерно разпределение на партидите според тяхното действително състояние и според резултатите от съответния статистически тест.

2. Разпределение на 10 000 партиди според тяхното действително състояние и според резултатите от съответния статистически тест при $\alpha = 0{,}05$, $\beta = 0{,}02$ и 50% вероятност за некачественост на партидата

Действително състояние	Резултат от теста		Общ брой
	статистически значимо отклонение - приема се H_1	статистически незначимо отклонение - приема се H_0	
Некачествена партида	4900	100	5000
Качествена партида	250	4750	5000
Общ брой	5150	4850	10000

В случая вероятностите за правилно приемане на съответната хипотеза ще бъдат съответно:

$$W_{(H_0)} = \frac{4750}{4850} = 0.97938;$$

$$W_{(H_1)} = \frac{4900}{5150} = 0.95146.$$

Сигурността за приемане на нулева, респ. алтернативна, хипотеза е много висока, но това се дължи изцяло на предположението за равномерно разпределение на партидите според тяхната качественост, което в практиката може да се приеме за почти невероятно.

Следва да бъде подчертано, че статистическите тестове не позволяват твърдения относно вероятността приетата хипотеза да е вярна или невярна и не могат да се приемат като еднозначно доказателство в полза на или против дадена хипотеза. Твърдения относно вероятността на хипотези са възможни единствено в рамките на Бейсовската статистика (Gigerenzer, Krauss, 2000).

Заключение

Без претенции за изчерпателност по отношение на разпространените в практиката заблуди при проверката на статистически хипотези, в съответствие с набелязаната цел са засегнати проблемни моменти при вземането на решение за приемане или отхвърляне на дадено предположение, като акцентът се поставя, от една страна, върху логиката на *p*-величината, използвана като показател за вземане на решение, а от друга, върху вероятността, с която приетата хипотеза е вярна или невярна.

Направените разсъждения позволяват следните основни изводи:

1. Въпреки че *p*-величината е удобен инструмент за вземане на решение относно приемането или отхвърлянето на нулевата хипотеза, използването ѝ е съпроводено от значителен риск за неправилна интерпретация на резултатите или дори за целенасочената им манипулация чрез последващо нагласяване на равнището на значимост α.

2. *P*-величината не е мерило за вероятността за допускане на грешка от първи род, следователно тя не представлява нито точна вероятност за допускане на грешка от I тип, нито гранично или критично равнище на значимост.

3. Заблуда представляват твърденията, според които:

- при сбъдване на получения резултат вероятността нулевата хипотеза да е вярна, възлиза на p;

- нулевата хипотеза е отхвърлена при риск за допускане на грешка от първи род, възлизащ на p.

4. Статистическите тестове не позволяват твърдения относно вероятността приетата хипотеза да е вярна и не могат да служат като еднозначно доказателство в полза на или против дадена хипотеза. Изключение прави проверката на хипотези в рамките на Бейсовската статистика, където се използва априорна информация за вероятностите за различни състояния на съвкупността.

5. Заблуда представляват твърденията, според които:

- след като в $(1-\beta).100\,\%$ от случаите тестът разпознава статистически значимите отклонения, когато в действителност е вярна алтернативната хипотеза, то при резултат, показващ наличието на статистически значимо различие, със сигурност $(1-\beta).100\,\%$ е вярна алтернативната хипотеза;

- след като тестът със сигурност $(1-\alpha).100\,\%$ не регистрира статистически значими отклонения, когато в действителност е вярна нулевата хипотеза, то при резултат, показващ отсъствие на такива отклонения, със сигурност $(1-\alpha).100\,\%$ е вярна нулевата хипотеза.

ЦИТИРАНА ЛИТЕРАТУРА:

Калинов, Кр. (2013). Статистически методи в поведенческите и социалните науки. София, Нов български университет.

Ламбова, М., Ч. Русев, Д. Косева, В. Стоянова (2012). Въведение в статистиката. Варна, ИК „СТЕНО“.

Хаджиев, В. (2002). Статистически и иконометричен софтуер. Варна, Университетско издателство ИУ - Варна.

Bortz, J., N. Döring (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Springer Verlag, Heidelberg.

Bourier, G. (2002). Wahrscheinlichkeitsrechnung und schließende Statistik, 3. Auflage.Wiesbaden, Gabler Verlag.

Dubben, H.-H., H.-P. Beck-Bornholdt (2010). Der Hund, der Eier legt. Erkennen von Fehlinformation durch Querdenken. Reinbek bei Hamburg, Rowohlt Verlag.

Gigerenzer, G., S. Krauss (2000). Statistisches Denken oder statistische Rituale? Was sollte man unterrichten, Anregungen zum Stochastikunterricht. Die NCTMStandards, pp. 53 - 62.

Hartung, J., B. Elpert, K.-H. Klösner (2005). Statistik. Lehr- und Handbuch der angewandten Statistik. München, Oldenbourg Verlag.

Mosler, K., Fr. Schmid (2006). Wahrscheinlichkeitsrechnung und schließende Statistik, 2. Auflage. Berlin, Heidelberg, New York, Springer Verlag.

Nuzzo, R. (2014). Der Fluch des p-Werts. Spektrum der Wissenschaft, 9/2014, pp. 52 - 56.

Polasek, W. (1997). Schließende Statistik. Berlin, Heidelberg, New York, Springer Verlag.

Ross, Sh. M. (2006). Statistik für Ingenieure und Naturwissenschaftler, 3. Auflage. München, Spektrum Akademischer Verlag.

Rüger, B. (2012). Test- und Schätztheorie. Band I: Grundlagen. München, Wien, Oldenbourg Verlag.

Rüger, B. (2002). Test- und Schätztheorie. Band II: Statistische Tests. München, Wien, Oldenbourg Verlag.

Simmons, J. P., L. D. Nelson, U. Simonsohn (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychol. Sci. 22/2011, pp. 1359 - 1366.