Testiranje statističkih hipoteza i neke zamke

Statistical hypothesis testing and some pitfalls

Vesna Ilakovac

Katedra za biofiziku, medicinsku statistiku i medicinsku informatiku, Medicinski fakultet Sveučilišta J.J. Strossmayer, Osijek
Department of Biophysics, Medical Statistics and Medical Informatics, J.J.Strossmayer University of Osijek, School of Medicine, Osijek, Croatia

Sažetak

Analiza podataka za potrebe istraživanja obično teži rabljenju informacija dobivenih iz uzorka ispitanika, kako bi se moglo zaključiti o relevantnoj populaciji. Testiranje statističkih hipoteza je široko rasprostranjena metoda statističkog zaključivanja. Za čitatelja znanstvenih i stručnih časopisa, kao i za istraživača, bitno je da razumije opće koncepte postupka testiranja, kako bi donio ispravnu odluku i stvorio mišljenje o predočenim rezultatima.

Ovaj nam članak daje pregled osnovnih koraka općeg postupka testiranja statističkih hipoteza i ističe neke poznate zamke i zablude. Članak obrađuje i pitanja koja se posebno odnose na *P* vrijednost, određivanje razine značajnosti, dokazivanje nulte hipoteze i problem višestrukog testiranja.

Ključne riječi: testiranje statističke hipoteze; *P* vrijednost; razina statističke značajnosti; problem višestrukog testiranja hipoteze

Abstract

Data analysis for research purposes usually aims to use the information gained from a sample of individuals in order to make inferences about the relevant population.

Statistical hypothesis testing is a widely used method of statistical inference. It is important to a reader of scientific or expert journals, as well as to a researcher, to understand the basic concepts of the testing procedure, in order to make sound decision and opinion on presented results.

This article gives an overview of basic steps in the general procedure for statistical hypothesis testing and points out some common pitfalls and misconceptions. Questions with particular regard to *P* value, determining significance level, proving null hypothesis and multiplicity problem were addressed.

Key words: statistical hypothesis testing; *P* value; significance level; multiplicity problem

Pristiglo: 21. studenog 2008. *Received: November 21, 2008*

Prihvaćeno: 15. prosinca 2008. *Accepted: December 15, 2008*

Uvod

Analiza podataka za potrebe istraživanja teži obično k upotrebi informacija dobivenih iz uzorka ispitanika, kako bi se moglo zaključiti o relevantnoj populaciji. Testiranje statističkih hipoteza je široko rasprostranjena metoda statističkog zaključivanja. Na primjer, ako nas zanima ima li razlike između muškaraca i žena po pitanju njihove koncentracije kolesterola u serumu, ili hoće li rast bakterija pratiti neku poznatu raspodjelu, ili hoće li naš koeficijent korelacije biti različit od 0, rabiti ćemo testiranje hipoteze. Računala i specijalizirani statistički programi, sa svo-

Introduction

Data analysis for research purposes usually aims to use the information gained from a sample of individuals in order to make inferences about the relevant population. Statistical hypothesis testing is a widely used method of statistical inference. For example, if we are interested in whether there is a difference between men and women with respect to their serum cholesterol levels, or if the growth of bacteria follows some known distribution, or if our correlation coefficient is different from 0, we will use a hypothesis test. Computers and specialized statistical

jim opširnim uputama i objašnjenjima, prilično nam olakšavaju izvođenje statističkih testova. Statistički programi izračunavaju točnu *P* vrijednost i urednici časopisa danas zahtijevaju od autora da navedu tu dobivenu vrijednost te tako omoguće čitateljima vlastito u tumačenje (kao primjer, vidi Upute za autore na internetskoj stranici časopisa *Biochemia Medica*).

S druge strane, tumačenje rezultata testa nije puko navođenje „statističke značajnosti" kada je *P* vrijednost niža od 0,05 ili bilo koje druge proizvoljne granične vrijednosti. Stoga je jednako važno kako za istraživača, tako i za čitatelja znanstvenih i stručnih časopisa razumjeti postupak testiranja statističkih hipoteza i kako ga koristiti kada želimo predstaviti ili procijeniti rezultate istraživanja u objavljenom članku. Postoji jedno, donekle zapostavljeno pitanje koje se tiče statističkih testova. Mnogi objavljeni radovi navode prilično velik broj *P* vrijednosti, što može otežati tumačenje (1). Svrha je ovog rada dati kratak pregled osnovnih koraka u općenitom postupku testiranja statističkih hipoteza i istaknuti neke uobičajene zamke i zablude.

Testiranje statističkih hipoteza

Testiranje statističkih hipoteza je postupak koji uključuje formuliranje statističke hipoteze i upotrebu podataka iz uzorka, kako bi se moglo odlučiti o ispravnosti formulirane statističke hipoteze. Iako detalji testiranja mogu varirati od testa do testa, za svako testiranje statističkih hipoteza možemo koristiti ovaj postupak u četiri koraka:

1. Postaviti nultu hipotezu i alternativne hipoteze.
2. Definirati postupak testiranja uključujući odabir razine statističke značajnosti i snage testa.
3. Izračunati test statistiku i pripadajuću *P* vrijednost.
4. Zaključiti jesu li podatci u skladu s nultom hipotezom, odnosno donijeti odluku o nultoj hipotezi.

Dvije se moguće pogreške mogu potkrasti u odlučivanju o nultoj hipotezi (2).

Pogreška tipa I događa se u slučaju kada „vidimo" učinak kojeg zapravo nema. Vjerojatnost da će se napraviti pogreška tipa I obično se naziva alfa (α) i njena se vrijednost određuje prije testiranja statističke hipoteze. Alfa je ono što nazivamo „razinom značajnosti" i njena je vrijednost najčešće postavljena na 0,05 ili 0,01. Kada je *P* vrijednost, dobivena u trećem koraku općih uputa o postupku testiranja statističke hipoteze niža od vrijednosti α, tada se rezultat naziva „statistički značajnim na razini α".

Pogreška tipa II događa se kada ne „vidimo" razliku, a ona je zapravo prisutna. Vjerojatnost da će se napraviti pogreška tipa II naziva se beta (β) i njena vrijednost uvelike ovisi o veličini učinka koji nas zanima, veličini uzorka i odabranoj razini statističke značajnosti. Beta se povezuje sa snagom testa u otkrivanju učinka navedene veličine. Više o analizi snage testa može se pročitati u jednom

software, with their extensive help guides, make carrying out statistical tests rather easy. Statistical computer programs give the exact *P* value, and journal editors today demand that researchers quote actual *P* values, and let readers make their own interpretation (for example, see Instructions to Authors on *Biochemia Medica* web site).

On the other hand, there is more to interpretation of a test result than just stating "statistical significance" when *P* value is less then 0.05 or any other arbitrary cut-off value. So it is equally important for both, researchers and readers of scientific or expert journals, to understand statistical hypothesis testing procedures and how to use them when presenting or evaluating research results in the published article. There is one somewhat disregarded issue concerning statistical tests. Many published papers today quote rather a large number of *P* values which may be difficult to interpret (1). The purpose of this paper is to give a brief overview of basic steps in the general procedure for a statistical hypothesis testing, and to point out some common pitfalls and misconceptions.

Statistical hypothesis testing

A statistical hypothesis testing is a procedure that involves formulating a statistical hypothesis and using a sample data to decide on the validity of the formulated statistical hypothesis. Although details of the test might change from one test to another, we can use this four-step procedure to do any hypothesis testing:

1. Set up the null and alternative hypotheses.
2. Define the test procedure, including selection of significance level and power.
3. Calculate test statistics and associated *P* value.
4. Conclude that data are consistent or inconsistent with the null hypothesis, i.e. make a decision about null hypothesis.

Two possible errors can be made when deciding upon the null hypothesis (2).

Type I error occurs when we "see" the effect when actually there is none. The probability of making a Type I error is usually called alpha (α), and that value is determined in advance for any hypothesis test. Alpha is what we call "significance level" and its value is most commonly set at 0.05 or 0.01. When the *P* value, obtained in the third step of the general hypothesis test procedure is below the value of α, the result is called "statistically significant at the α level".

Type II error occurs when we fail to see the difference when it is actually present. The probability of making the Type II error is called beta (β) and its value depends greatly upon the size of the effect we are interested in, sample size and the chosen significance level. Beta is associated with the power of the test to detect an effect of a speci-

prethodno objavljenom članku iz serije *Odabrane teme iz biostatistike* (3).

Što je *P* vrijednost?

P vrijednost se često pogrešno interpretira kao vjerojatnost da je nulta hipoteza istinita. Nulta hipoteza nije nasumična te ona nema vjerojatnosti. Ona je ili istinita ili nije. Pravo značenje *P* vrijednosti jest, da je to vjerojatnost opažanja podataka kakvi su na promatranom uzorku (ili ekstremnijih podataka) kada je nulta hipoteza **istinita**. Na primjer, kada promatramo razliku u srednjim vrijednostima koncentracije kolesterola u serumu mjerenu u dva uzorka, želimo znati koliko je vjerojatno da ćemo dobiti takvu ili još ekstremniju razliku kad ne bi bilo stvarne razlike između ispitane populacije. To je ono što nam kazuje *P* vrijednost i ako je ona mala, recimo 0,003, smatramo da je opažena razlika malo vjerojatna u slučaju kada je nulta hipoteza istinita.
To nas dovodi do pitanja koliko malo je malo, odnosno do pitanja odabira razine statističke značajnosti.

Koju razinu statističke značajnosti odabrati?

Važno je naglasiti da je razina statističke značajnosti proizvoljna vrijednost koju odabiremo kao graničnu vrijednost u odlučivanju o nultoj hipotezi te da je treba odrediti prije analize. Čak i ako znamo točnu *P* vrijednost, potrebna nam je pomoć pri odlučivanju na temelju promatrane *P* vrijednosti.
Prihvatljivo i jednostavno rješenje je ustanoviti posljedice pogrešnih odluka, odnosno pogrešaka tipa I i II. Ako (pogrešno) uočavanje razlike koje zapravo nema može naštetiti populaciji koju ispitujemo (ili općenito, svoj populaciji), tada trebamo odabrati nižu razinu statističke značajnosti, odnosno pokušati smanjiti vjerojatnost pogreške tipa I.
Zamislite sljedeći scenarij: Prospektivno kliničko ispitivanje pokazalo je da bolesnici na liječenju A imaju jake štetne nuspojave. Liječenje A je otkazano i sada se ispituju učinci novog liječenja B. Primijećeno je smanjenje štetnih nuspojava kod novog liječenja B u odnosu na staro liječenje A.
Pitanje je: koju razinu statističke vjerojatnosti trebamo odabrati kako bi procijenili značajnost promatrane razlike, odnosno je li novo liječenje zaista bolje od starog?
Možemo donijeti dva pogrešna zaključka, od kojih svaki nosi posljedice za bolesnike.
Pogrešni zaključak 1: liječenje B je bolje, iako je zapravo jednako liječenju A.
Posljedica 1: prihvaćamo novo liječenje (liječenje B) i njemu izlažemo bolesnike, zajedno s štetnim nuspojavama koje će ono donijeti.
Pogrešni zaključak 2: oba su liječenja identična, iako je zapravo liječenje B bolje od liječenja A.

fied size. More about power analysis in research can be found in one of the previous articles in *Lessons in Biostatistics* series (3).

What is the *P* value?

The *P* value is often misinterpreted as probability that the null hypothesis is true. The null hypothesis is not random and has no probability. It is either true or not. The actual meaning of the *P* value is the probability of having observed our data (or more extreme data) when the null hypothesis **is** true. For example, when we observe the difference in means of serum cholesterol levels measured in two samples, we want to know how likely it is to get such or more extreme difference when there is no actual difference between underlying populations. This is what *P* value tells us, and if we find that the *P* value is low, say 0.003, we consider the observed difference quite unlikely under the terms of the null hypothesis.
This leads us to the question of how low is low, i.e. the question of choosing the significance level.

Which significance level should we choose?

It is important to emphasize that significance level is an arbitrary value we choose as a cut-off value for deciding upon the null hypothesis and that it should be determined prior to analysis. Even when we do know the exact *P* value, we need some guidance about reaching a decision from the observed *P* value.
The plausible and simple solution is to identify the consequences of wrong decisions, i.e. of making the Type I or Type II error. If seeing (wrongly) a difference when actually there is none can be harmful for the population under study (or in general), we should choose a lower significance level, thus trying to minimize the probability of making the Type I error.
Picture the following scenario: A prospective clinical trial has shown that patients under treatment A experience considerable adverse effects. Treatment A was called off and the effects of a new treatment B were investigated. The decrease in adverse effects was observed for the new treatment B comparing to the old treatment A.
The question is: what significance level should we choose to estimate the significance of the observed difference, i.e. is the new treatment really better than the old one?
We can make two wrong conclusions, and each of them with consequence regarding patients.
Wrong conclusion #1: Treatment B is better, when actually it is the same as treatment A.
Consequence #1: We adopt the new treatment exposing patients to the adverse effects of Treatment B.
Wrong conclusion #2: Both treatments are the same, when actually treatment B is better than the treatment A.

Posljedica 2: ne primjenjujemo novo liječenje u praksi, već nastavljamo tražiti bolje rješenje.
Dakle, prilično je jasno da pogreška tipa I (primjećivanje razlike koje zapravo nema) u ovom scenariju nanosi više štete bolesniku i da ju trebamo pokušati izbjeći.
Kod odabira razine statističke značajnosti α, moramo imati na umu da će, ako smanjimo vrijednost α, narasti vrijednost β, smanjujući istodobno snagu statističkog testa (3).

Možemo li „dokazati" nultu hipotezu?

Možda će Vas odgovor iznenaditi, no on jednostavno glasi: **ne**. Dobivanje neznačajnog rezultata o učinku nekog liječenja ne implicira da taj učinak ne postoji. Najviše što možemo reći, jest da nismo uspjeli naći dovoljno dokaza o njegovom postojanju. Da citiram naslov jednog članka u renomiranom britanskom medicinskom časopisu *British Medical Journal*: „Nepostojanje dokaza nije dokaz nepostojanja" (4). U odnosu na nultu hipotezu trebamo se izraziti da je „nismo odbacili" ili je „nismo uspjeli odbaciti" (5). Statistička hipoteza ne „dokazuje" ništa.

Višestruko testiranje hipoteza

Ako odaberemo 0,05 kao razinu statističke značajnosti i zatim napravimo 20 nezavisnih testova na istim podatcima, vjerojatnost pogreške tipa I, odnosno odbacivanja nulte hipoteze, iako je ona istinita, iznosi 0,64. To znači da je vjerojatnije da ćemo dobiti jedan statistički značajan rezultat nego niti jedan. Nadalje, među 20 takvih nezavisnih testova hipoteze, očekujemo da ćemo pukim slučajem dobiti 20 x 0,05 = 1 značajan rezultat. Kako je to moguće?
Vjerojatnost pogreške tipa I (razina statističke značajnosti) α može se opisati kao vjerojatnost odbacivanja nulte hipoteze kada je ona zapravo istinita. To možemo izraziti kao:

$$\alpha = 1 - (1 - \alpha).$$

U toj jednadžbi (1-α) predstavlja zapravo vjerojatnost suprotnog događaja, odnosno **ne** odbacivanje nulte hipoteze kada je ona zapravo istinita.
Ako testiramo nekoliko nezavisnih nultih hipoteza, kada su one zapravo sve istinite, vjerojatnost barem jedne pogreške tipa I jednaka je 1-(vjerojatnost da neće biti niti jedne pogreške tipa 1). U slučaju dva testa to bi bilo:

$$\text{vjerojatnost barem jedne pogreške tipa I} = \alpha_2 = 1 - [(1- \alpha) \times (1- \alpha)] = 1-(1- \alpha)^2.$$

Kada je α = 0,05, tada je vjerojatnost barem jedne pogreške tipa I kod testiranja dvije nezavisne nulte hipoteze:

$$\alpha_2 = 1 - 0{,}95^2 = 1 - 0{,}90 = 0{,}10.$$

Consequence #2: We do not adopt the new treatment in practice, but continue to search for a better solution.
Now it is quite obvious that making a Type I error (seeing the difference when there is none) in this scenario brings more harm to patients and that we should try to avoid making it.
When choosing the α significance level, we must bear in mind that if we decrease the value of α, the value of β will increase, thus decreasing the power of the test (3).

Can we "prove" the null hypothesis?

It may come as a surprise, but the answer to this question is very simple: **No**.
Getting an insignificant result about some treatment effect does not imply that there is none. The most we can say is that we failed to find sufficient evidence for its existence. To quote the title of an article in a highly respected medical journal - *British Medical Journal*: "Absence of evidence is not evidence of absence" (4). In terms of the null hypothesis we should say that "we have not rejected" or "have failed to reject" the null hypothesis (5). Statistical hypothesis test does not "prove" anything.

Multiple hypothesis testing

If we choose 0.05 as a significance level and then carry out 20 independent tests on the same data, the probability of making a Type I error when all underlying null hypotheses are in fact true is 0.64. This means that we are more likely to get one significant result than not. Furthermore, among 20 of such independent hypothesis tests, we expect to get 20 x 0.05 = 1 significant result purely by chance. How can this be?
The probability of making the Type I error (significance level) α can be described as probability of rejecting null hypothesis when the null hypothesis is actually true. We can also express it as:

$$\alpha = 1 - (1 - \alpha).$$

In this equation (1-α) is actually the probability of the complementary event, i.e. **not** rejecting the null hypothesis when the null hypothesis is actually true.
If we test several independent null hypotheses when all of them are actually true, the probability of making at least one Type I error equals 1- (probability of making none). In case of two tests that would be:

$$\text{probability of making at least one Type I error} = \alpha_2 = 1 - [(1-\alpha) \times (1-\alpha)] = 1-(1-\alpha)^2.$$

When α=0.05, the probability of making at least one Type I error when testing two independent null hypotheses is:

$$\alpha_2 = 1 - 0.95^2 = 1 - 0.90 = 0.10.$$

Za tri testa je vjerojatnost barem jedne pogreške tipa I:

$$\alpha_3 = 1 - 0{,}95^3 = 1 - 0{,}86 = 0{,}14.$$

Općenito, vjerojatnost barem jedne pogreške tipa I u seriji k nezavisnih nultih hipoteza kada su **sve** nulte hipoteze zapravo istinite jest:

$$\alpha_k = 1-(1-\alpha)^k.$$

Sada je zapravo jednostavno vidjeti da je za 20 testova vjerojatnost barem jedne pogreške tipa I kada su **sve** nulte hipoteze istinite jednaka 0,64. Slika 1. pokazuje da je potrebno 60 testova kako bi se dosegla vjerojatnost od 0,95 za dobivanje statistički značajnog rezultata o nekom učinku pukim slučajem i kada učinak zapravo ne postoji. Očekivani broj statistički značajnih rezultata u seriji k nezavisnih testiranja hipoteza kada su **sve** nulte hipoteze istinite izračunava se jednostavno kao:

$$k \times \alpha.$$

Problem te vrste pojavljuje se kada izvršimo testiranje hipoteze na višestrukim podskupinama ispitanih uzoraka. U istraživanju o vezi bolova u leđima i riziku od ishemijske bolesti srca sa smrtnim ishodom (6), autor je prikazao tablicu sa smrtnošću povezanom sa starosnom dobi kod muškaraca sa i bez boli u leđima podijeljenu u dvije dobne skupine i tri skupine prema uzroku smrti.
Rezultati testa usporedbe skupina sa i bez boli u leđima iskazani bili su iskazani pomoću *P* vrijednosti za svaku od

For 3 tests, the probability of making at least one Type I error is:

$$\alpha_3 = 1 - 0.95^3 = 1 - 0.86 = 0.14.$$

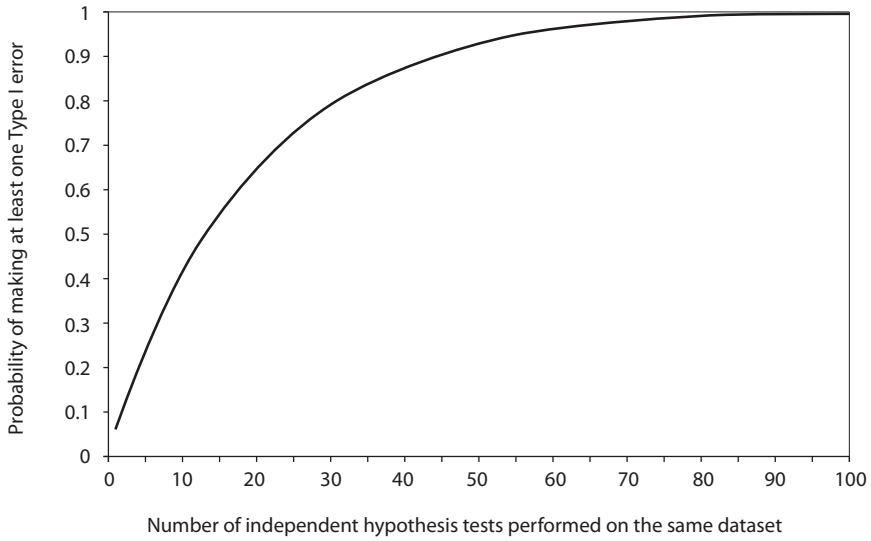
In general, the probability of making at least one Type I error in the series of k independent null hypothesis tests when **all** null hypotheses are actually true is:

$$\alpha_k = 1-(1-\alpha)^k.$$

Now it is easy to see that for 20 tests the probability of making at least one Type I error when **all** null hypotheses are actually true equals 0.64. From Figure 1, we can see that it takes about 60 tests to reach the probability of 0.95 to get a significant result about some effect purely by chance, when no effect actually exists.
The expected number of significant results in a series of k independent hypothesis tests when **all** null hypotheses are actually true is simply calculated as:

$$k \times \alpha.$$

A problem of this kind arises when we perform hypothesis tests on multiple subsets of studied samples. In a study of association of back pain and risk of fatal ischemic heart disease (6) the author presented a table with age-specific mortality for men with and without back pain divided into two age groups and three groups by cause of death.
The results of the test comparing groups with and without back pain by means of *P* values were presented for



Slika 1. Vjerojatnost barem jedne pogreške tipa I, kao funkcija broja nezavisnih testova hipoteza izvedenih na istom skupu podataka, u slučaju kada su sve nulte hipoteze istinite, a razina statističke značajnosti α postavljena na 0,05.

Figure 1. Probability of making at least one Type I error as a function of the number of independent hypothesis tests performed on the same dataset when all null hypotheses are true and significance level α is set to 0.05.

šest podskupina, plus dodatne dvije za sve uzroke u obje dobne skupine. Samo je jedna od objavljenih *P* vrijednosti bila niža od 0,05 (0,02), dok su ostale bile u rasponu od 0,10 do 0,99. Istaknuto je, dakle, da nije pronađena veza između bolova u leđima i bilo koje vaskularne bolesti kod žena, što navodi na zaključak da je autor izveo isti broj testova u podskupini žena. To bi bilo sveukupno barem 16 testova, među kojima je samo jedan smatran „značajnim", točno onoliko koliko bismo očekivali da će se dogoditi pukim slučajem. Problem višestrukog testiranja često se pojavljuje kod pokusa na mikropostrojima (7). Uzmimo kao primjer određivanje nekoliko alela u skupini nasumično odabranih bolesnika s određenom bolešću i skupinu nasumično odabranih kontrolnih ispitanika. Broj testova trebao bi odgovarati broju alela koji se određuju, ako je učestalost alela jednaka u obje populacije. Za 30 alela vjerojatnost lažnog značajnog rezultata iznosi 0,76, dok se za 50 alela povećava na 0,92.

Jedan način rješavanja tih problema jest prilagoditi ili minimalnu prihvaćenu razinu statističke značajnosti ili *P* vrijednosti dobivene iz serije nezavisnih testova, kako bi se očuvala ukupna razina značajnosti. Ako prilagodimo minimalnu prihvaćenu razinu statističke značajnosti, uspoređujemo „originalne" *P* vrijednosti s prilagođenom razinom statističke značajnosti. Ako prilagodimo *P* vrijednosti, uspoređujemo prilagođene *P* vrijednosti s originalno postavljenom razinom statističke značajnosti.

Uobičajen način prilagođavanja prvotne *P* vrijednosti (ponekad je nazivamo „nominalnom" *P* vrijednosti) za višestruko testiranje jest upotreba Bonferronijeve metode (1). Prema toj se metodi prilagodba radi množenjem nominalnih *P* vrijednosti s brojem izvedenih testova. Dakle, ako smo napravili tri nezavisna testa, koji su kao rezultat imali *P* vrijednosti 0,020, 0,030 i 0,040, onda su prilagođene *P* vrijednosti po Bonferroniju 0,060 (za 0,020), 0,090 (za 0,030) i 0,120 (za 0,040). Dok su „originalni" rezultati za sva tri testa bili smatrani statistički značajnima na razini 0,05, nakon prilagodbe niti jedan od njih nije ostao statistički značajan.

Bliska Bonferronijevoj metodi je Šidakova metoda (Sidak) (7). Korigirane *P* vrijednosti po Šidaku računaju se kao:

$$p_k = 1 - (1 - p)^k.$$

Drugi se problem pojavljuje ako imamo višestruke mjere ishoda, u tom slučaju testovi općenito neće biti nezavisni. Ostali problemi kod višestrukog testiranja nastaju kada imamo više od dvije skupine ispitanika te želimo usporediti svaki par skupina ili kad imamo niz promatranja unutar duljeg vremena te želimo testirati svako razdoblje posebno. Za probleme kod visoko koreliranih višestrukih testova spomenute metode nisu prikladne, budući da bi bile prekonzervativne te možda ne bi prepoznale stvaran učinak (1).

each of the six subgroups plus two more for all causes in both age groups. Only one of those eight reported *P* values was less than 0.05 (0.02) while other ranged from 0.10 to 0.99. It was also pointed out that no association between back pain and any vascular disease was found in women, which leads to the notion that the author performed the same number of tests in the women subgroup. That would make the total of at least 16 tests among which only one was found to be "significant", just about as many as we would expect to occur purely by chance.

The problem of multiple testing is common to microarray experiments (7). Let us consider the experiment of typing a series of random patients with a particular disease and a series of random controls without disease for a certain number of alleles. The number of tests would equal the number of alleles, testing whether each allele frequency was the same in the two populations. For 30 alleles, the probability of a false significant result when there is no association whatsoever is 0.76 while it increases to 0.92 for 50 alleles.

One way of dealing with these problems is to adjust either the minimum accepted significance level or to adjust *P* values obtained from the series of independent tests in order to preserve the overall significance level. If we adjust the minimum accepted significance level, we compare the "original" *P* values with the adjusted significance level. If we adjust *P* values, then we compare adjusted *P* values with the originally stated significance level.

A common way to adjust the original *P* values (sometimes called "nominal" P-values) for multiple testing is to use the Bonferroni method (1). By this method the adjustment is made by multiplying the nominal *P* values with the number of tests performed. So, if we made three independent tests which resulted in *P* values of 0.020, 0.030 and 0.040, the Bonferroni-adjusted *P* values would be 0.060, 0.090 and 0.120, respectively. While "original" results for all three tests would be considered significant at 0.05 level, after adjustment none of them remained significant.

Closely related to the Bonferroni method is the Šidak (Sidak) method (7). Šidak-adjusted P-values for k independent tests are calculated as:

$$p_k = 1 - (1 - p)^k.$$

Another problem occurs if we have multiple outcome measurements, in which case the tests will not be independent in general. Other multiple testing problems arise when we have more than two groups of study participants and wish to compare each pair of groups, or when we have a series of observations over time and wish to test each time point separately. For problems where the multiple tests are highly correlated, the aforementioned methods are not appropriate as they will be highly conservative and may miss the real effect (1).

Problem višestrukog testiranja je ozbiljan problem u znanstvenom istraživanju. Izostanak prilagodbe zbog višestrukog testiranja stvara ozbiljnu sumnju u dobivene rezultate, zbog činjenice da višestrukost povećava razinu značajnosti i smanjuje snagu testa u istraživanju, tim više ako usporedbe nisu prethodno planirane niti specificirane. Stoga se problem višestrukosti ne smije olako shvatiti. Danas postoje brojne metode za prilagođavanje višestrukosti koje bi se trebale upotrijebiti u susretu s tim problemom. Ako dovoljno dugo i uporno ispitujemo podatke, oni će naposljetku „priznati" da ipak negdje postoji statistički značajna razlika. No ta će značajnost biti nepouzdana te je vjerojatno da će i zaključci utemeljeni na takvim statistički značajnim rezultatima biti lažni.

Zaključak

Testiranje statističke hipoteze česta je metoda statističkog zaključivanja. Postoji nekoliko jednako važnih pitanja o kojima nismo raspravljali u ovom članku, neka od kojih su odabir pravog testa, upotreba jednosmjernih ili dvosmjernih testova, razlikovanje statističke značajnosti i praktične važnosti. Jednako je važno čitatelju znanstvenih i stručnih časopisa, kao i samom istraživaču, razumjeti osnovnu ideju postupka testiranja statističke hipoteze, kako bi mogao donijeti ispravne odluke i zaključke o rezultatima istraživanja.
Još jedna stvar zaslužuje pažnju, kada se govori o rezultatima istraživanja. Izvještavati o značajnoj razlici, a ne govoriti pritom i o veličini te promatrane razlike (odnosno o veličini učinka) i odgovarajućim intervalima pouzdanosti, ne daje potpunu sliku dobivenog rezultata. Stoga se u statističkom zaključivanju preporuča korištenje i statističkog testiranja i metoda za procjenu veličine učinka.

The multiple testing problem is a serious problem in scientific research. Failure to adjust for multiplicity raises a serious doubt in obtained results because multiplicity inflates the significance level and diminishes the power of the research, and even more so if comparisons were not planned and prespecified. Thus, the multiplicity problem should not be taken lightly. Today there are numerous methods for adjusting multiplicity and an appropriate method should be implemented when facing the multiplicity problem. If we torture the data long enough they will finally produce something which is "significant". But the significance will be unreliable, and conclusions based on those significant results are likely to be spurious.

Conclusion

Statistical hypothesis testing is a common method of statistical inference. There are several equally important issues not addressed in this article such as choosing the right test, performing one-tailed or two-tailed test, distinction of statistical significance and practical importance, just to name a few. It is important to a reader of scientific or expert journals as well as is to a researcher to understand the basic concepts of the testing procedure in order to make sound decisions about results and to draw accurate conclusions.
One more thing deserves attention when dealing with research results. Reporting a significant difference without reporting the size of the difference observed (i.e., the effect size) and associated confidence intervals is just half of the story. Thus, it is recommended to use both hypothesis testing and effect size estimation methods for statistical inference.

Adresa za dopisivanje:

Vesna Ilakovac
Katedra za biofiziku, medicinsku statistiku i medicinsku informatiku
Medicinski fakultet Sveučilišta J.J. Strossmayer
J. Huttlera 4
31000 Osijek
e-pošta: vilakov@mefos.hr

Corresponding author:

Vesna Ilakovac
Department of Biophysics, Medical Statistics and Medical Informatics
School of Medicine
J.J.Strossmayer University of Osijek
J. Huttlera 4
31000 Osijek
Croatia
e-mail: vilakov@mefos.hr

Literatura/References

1. *Bland JM, Altman DG. Statistics notes: Multiple significance tests: the Bonferroni method. BMJ. 1995;310:170.*
2. *Altman DG. Practical statistics for medical research. 1st ed. London: Chapman&Hall/CRC, 1991.*
3. *McHugh ML. Power analysis in research. Biochemia Medica. 2008;18:263-74.*
4. *Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. BMJ. 1995;311:485.*
5. *Bland M. An Introduction to Medical Statistics. 3rd ed. New York: Oxford University Press, 2000.*
6. *Penttinen J. Back pain and risk of fatal ischaemic heart disease: 13 yeas follow up of Finnish farmers. BMJ. 1994;309:1267-8.*
7. *Dudoit S, Popper Shaffer J, Boldrick JC. Multiple Hypothesis Testing in Microarray Experiments. Stat Sci. 2003;18:71-103.*