Arh. hig. rada toksikol., 38 (1987) 251—260.

Opažanje
UDK 001.3

KOLIKI POSTOTAK ZNANSTVENIH OTKRIĆA NISU OTKRIĆA?

B. Sorić i B. Petz

Dom zdravlja »Pešćenica«, Zagreb i Odsjek za psihologiju, Filozofski fakultet Sveučilišta u Zagrebu, Zagreb

(Primljeno 26. XII. 1986)

Pri statističkom provjeravanju rezultatâ znanstvenih istraživanja određuje se rizik pogrešnog odbacivanja istinite nul-hipoteze. Rezoniranje o tom riziku pogrešno se primjenjuje i na rizik (postotak) dobivanja »lažnih otkrića«, npr. na postotak neefikasnih lijekova među prihvaćenim lijekovima. Postotak pogrešno prihvaćenih rezultata u skupu statistički značajnih rezultata (»otkrićâ«) ne ovisi samo o razini statističke značajnosti, već i o nepoznatom stanju u skupu populacija koje istražujemo, odnosno o omjeru značajnih i neznačajnih rezultata, koji bi se mogao saznati, ali o kojem se danas ne vodi računa.

Ako u nekoj igri slučaja postoji 5% vjerojatnosti dobitka, onda obično rezoniramo ovako: ako tu igru igramo, recimo, 1000 puta, onda je najvjerojatnije da ćemo oko 50 puta dobiti, pa prema tome oko 950 puta izgubiti.

Većina ljudi, koji eksperimentalno rade na različitim istraživanjima, na jednaki način rezonira i kod znanstvenih eksperimenata, u kojima je ustanovljeno da je, npr., neka razlika između dviju aritmetičkih sredina »statistički značajna na nivou značajnosti« (ili »na nivou rizika«) od recimo 5% ($p < 0{,}05$). U tom slučaju, naime, rezonira se ovako: budući da u svim onim eksperimentima u kojima smo našli »statistički značajnu razliku«, šansa da smo pogriješili (tj. šansa da smo razliku proglasili statistički značajnom, a među populacijama razlike ustvari nema), iznosi oko 5%, to možemo smatrati da (1) kod aproksimativno 5% naših rezultata, kod kojih smo našli statistički značajne razlike, ta razlika zapravo ne postoji, i (2) kod približno 95% ostalih naših prihvaćenih rezultata ta razlika postoji i među populacijama. Prema tome, možemo smatrati da smo u oko 5% svih naših »znanstvenih otkrića« pogriješili, tj.: ako broj naših »otkrića« (statistički značajnih rezultata) iznosi 1000, oko 50 njih je promašeno i dovelo je do krivog zaključka, a ostalih 950 je ispravno. Primjer iz područja medicine, točnije rečeno iz područja farmakologije, mogao bi glasiti ovako: ako su lijekovi testirani na nivou značajnosti od oko 5%, to znači da među njima imamo oko 5% neefikasnih, a oko 95% korisnih i efikasnih lijekova.

Kao dokaz da istraživači, i ostali, koji poznaju osnovnu statističku metodologiju, tako rezoniraju, mogu poslužiti rezultati jednog ispitivanja što smo ga izveli na 77 ispitanika, koji se svi u svom radu služe statističkom metodologijom. Ispitivanje je provedeno u prosincu 1986. godine, a sastojalo se u tome da je ispitanicima pismeno predočen ovaj zadatak:

Kada za neku razliku između dvije aritmetičke sredine t-testom ustanovimo da je statistički značajna na nivou $p < 0,05$, to obično protumačimo ovako: postoji najviše 5% šanse da smo pogriješili, tj. da smo proglasili postojećom neku razliku koja ustvari među populacijama ne postoji. — Zamislite da, recimo, u 1000 različitih eksperimenata, koji su svi dali razlike između aritmetičkih sredina, koje su bile statistički značajne ($p < 0,05$), želimo prosuditi koliko po prilici od tih 1000 rezultata predstavlja pogrešan zaključak (tj. da ustvari nema razlike među aritmetičkim sredinama populacijâ), iako smo mi za sve slučajeve ustanovili da razlika postoji; onda je ispravno zaključiti ovo:

a) u najviše oko 500 eksperimenata pogrešno smo razliku proglasili statistički značajnom
b) u najviše 100 eksperimenata pogrešno smo razliku proglasili statistički značajnom
c) u najviše oko 50 eksperimenata pogrešno smo razliku proglasili statistički značajnom
d) u najviše oko 5 eksperimenata pogrešno smo razliku proglasili statistički značajnom
e) ništa ne možemo zaključiti o tome u koliko smo po prilici eksperimenata pogrešno zaključili da je razlika statistički značajna.

Ukupno 41 ispitanik odgovorio je (očekivanim) odgovorom c), 13 ispitanika odgovorima d) ili a), a 23 odgovorom e).

Budući da je — kao što ćemo vidjeti — odgovor e) jedini ispravan, traženo je od tih ispitanika da obrazlože zašto su odgovorili tim odgovorom. Niti jedan od ta 23 ispitanika nije dao ispravno obrazloženje; najčešće obrazloženje bilo je po prilici ovako formulirano: »Činjenica da u svakoj pojedinoj situaciji vjerojatnost pogreške iznosi 5%, ništa ne govori o vjerojatnosti pogreške među podacima, koji pripadaju različitim uzorcima.«

Ostali predloženi odgovori imaju još i manje smisla.

No, takvo zaključivanje nije ispravno, a da je tome tako, pokušat ćemo dokazati ovim člankom.

Da bi taj dokaz bio razumljiviji, potrebno je da se najprije podsjetimo načina rezoniranja uz takozvanu »nul-hipotezu« u statistici.

Kao što je poznato, »nul-hipoteza« u statističkom rezoniranju pretpostavlja da neka nađena razlika (npr. između dvije aritmetičke sredine) ustvari u populaciji ne postoji. No, iako ona ne postoji, mi ćemo — ako vršimo velik broj mjerenja na uzorcima — ipak neprestano nailaziti na manje ili veće razlike među aritmetičkim sredinama uzoraka. Jasno je međutim da su sve te nađene razlike slučajne.

Jedan jednostavan zamišljeni primjer to će lako demonstrirati: pretpostavimo da imamo dvije identične kutije u kojima se nalaze žetoni s rezultatima mjerenja visine svih odraslih muškaraca Zagreba. Te kutije sadrže dakle po-

pulaciju (u duplikatu). Iako je sadržaj obih kutija identičan, ipak, kada bismo iz svake od njih uzimali uzorke od po, recimo, 50 žetona, i kada bismo svaki put izračunali aritmetičke sredine, praktički nikad ne bismo dobili jednake aritmetičke sredine u oba paralelna uzorka. No naravno da te »razlike« moramo pripisati slučaju. Kod vrlo velikog broja tako uzimanih paralelnih uzoraka i bilježenja svih dobivenih razlika među aritmetičkim sredinama ustanovili bismo da se te razlike distribuiraju po normalnoj raspodjeli, kojoj je aritmetička sredina 0 (nula) (što odgovara pravoj razlici između aritmetičkih sredina obih populacija, tj. sadržaji kutija su identični, te nema razlike među aritmetičkim sredinama). Standardna devijacija te raspodjele odgovara drugom korijenu iz zbroja varijanci podijeljenih s N u svakom uzorku (N je opseg uzorka), i ovu standardnu devijaciju (tj. raspršenje razlikâ aritmetičkih sredina uzoraka oko prave razlike aritmetičkih sredina) nazivamo u statistici standardnom pogreškom razlike između dviju aritmetičkih sredina.

Pretpostavimo li sada da smo, na primjer, jednim mjerenjem našli da razlika između dviju aritmetičkih sredina dvaju uzoraka iznosi 3, a da standardna pogreška te razlike iznosi 2, to — ako se služimo terminologijom nul-hipoteze — znači ovo: Kad ustvari među populacijama, kojima pripadaju ti uzorci, ne bi bilo razlike, onda bi se — pri uzimanju velikog broja uzoraka koji su veličine kao i uzorci našeg mjerenja — dobila normalna distribucija razlikâ među aritmetičkim sredinama, kojoj je aritmetička sredina nula, a standardna devijacija 2. Razlika između oba uzorka mogla bi i slučajno biti čak i veličine 6 (tj. 0 ± 3 standardne devijacije), a budući da smo mi dobili razliku veličine 3, ta se razlika nalazi samo 1,5 standardnih devijacija udaljena od prosjeka, pa prema tome posve dobro »pristaje« u ovu krivulju populacije razlika, kojima je aritmetička sredina nula. Takvu razliku ne smijemo zato smatrati realnom, tj. statistički značajnom, jer se ona — kako vidimo — dosta lako može i slučajno dogoditi, pa makar među populacijama nema razlike.

Da smo u svom pokusu međutim dobili između aritmetičkih sredina obaju uzoraka razliku od 7, onda bi naše rezoniranje glasilo ovako: kada među populacijama ne bi bilo razlike, praktički bi najveća slučajna razlika mogla dostići veličinu 6 (tj. 0 ± 3 standardne devijacije); no budući da smo mi dobili razliku veličine 7, to je gotovo isključeno da se tako velika razlika mogla pojaviti slučajno. Iz toga slijedi da se ta razlika od 7 pojavila zato, što i među populacijama postoji razlika. Stoga za tu razliku od 7 kažemo da je statistički značajna, što znači da smo gotovo posve sigurni da nije slučajna.

Iz toga slijedi da smo gotovo potpuno sigurni da neka nađena razlika među aritmetičkim sredinama nije slučajna tek onda ako je ta razlika barem 3 puta veća od svoje vlastite pogreške, jer u tom slučaju ona se nalazi izvan ± 3 standardne devijacije naše distribucije razlikâ, kojima je aritmetička sredina nula (Posve sigurni ne možemo biti nikada, jer — kako znamo — aritmetička sredina ± 3 standardne devijacije obuhvaća 99,73% cijele površine normalne distribucije, pa na 10 000 rezultata još uvijek ima 27 rezultata koji su izvan tog raspona od 6 standardnih devijacija.)

No svojedobno su statističari (Fisher i drugi) ispravno upozorili da ne valja biti suviše strog prilikom ocjene je li neka razlika statistički značajna ili nije, jer ako tražimo da ona bude barem 3 puta veća od svoje pogreške, onda će se dogoditi da mnoge razlike, koje ustvari među populacijama postoje, nećemo

prihvatiti kao »statistički značajne«. Konkretno, ako stvarna razlika među aritmetičkim sredinama dviju populacija iznosi 2,5, a mi u uzorcima iz tih dviju populacija dobijemo, recimo, razliku veličine 3, i njezinu standardnu pogrešku veličine 2, onda tu razliku ne bismo proglasili statistički značajnom, jer je ona samo 1,5 puta veća od svoje pogreške (3/2 = 1,5), a ipak realna razlika među populacijama postoji.

Stoga je predložen kompromis koji bi se riječima mogao obrazložiti i formulirati ovako: pod pretpostavkom da nema razlike među aritmetičkim sredinama dviju populacija, mogli bismo i slučajno dobiti razlike veličine tri standardne pogreške, no slučajno dobivanje razlika, koje su veće od oko 2 standardne pogreške (točnije 1,96 stand. pogreške) već je vrlo rijetko: samo oko 5% slučajeva pada izvan granica aritmet. sredina ± 2 standardne devijacije. Stoga — predložili su ti statističari —uzmimo kao granicu upravo tih 5% (ili, ako želimo biti stroži, uzmimo kao granicu 1%), pa kažimo da sve razlike koje padaju izvan tih granica — naravno uz rizik od 5% (ili 1%) da smo pogriješili u odluci — stvarno postoje, tj. da su statistički značajne. U takvim se dakle slučajevima smatra da postoji 95% (ili 99%) vjerojatnosti da razlika među populacijama postoji, a samo 5% (ili 1%) vjerojatnosti da smo pogriješili u zaključku, jer razlike ustvari nema.

Kad smo ovako osvježili naše znanje i sjećanje o tome što ustvari znači »nul-hipoteza«, i kakve zaključke iz takvih računa možemo povlačiti, vratimo se našem problemu! — Podsjetimo se, najčešći način rezoniranja je ovaj:

Ako eksperimente radimo na nivou značajnosti (rizika) od 5%, onda smo po prilici kod 5% naših znanstvenih rezultata izveli pogrešan zaključak, tj. zaključak da je razlika između aritmetičkih sredina statistički značajna (jer se zapravo aritmetičke sredine papulacijâ međusobno ne razlikuju), a u preostalih 95% slučajeva naši su zaključci ispravni. — Kao što smo već rekli, to rezoniranje nije točno!

Ako se podsjetimo doslovnog značenja nul-hipoteze, ono glasi: Ako nema razlike među populacijama, onda je vjerojatnost 5% da ćemo u našem pokusu razliku smatrati značajnom. — Dakle, primijenjeno na naš primjer, to znači ovo:

Kada među populacijama, na koje se odnose naši eksperimenti, ne bi bilo razlike, u oko 5% eksperimenata mi bismo razliku ipak (pogrešno) »našli«. No iz toga naravno slijedi da niti u jednom od tih 5% eksperimenata nismo pogodili istinu, tj. istinu da razlika ne postoji!

Uzmimo jedan potpuno konkretan primjer, i to iz područja eksperimentiranja s preparatima, kod kojih se ispituje da li su efikasni kao lijekovi.

Pretpostavimo da želimo testirati 1000 novih farmaceutskih preparata, tj. želimo ih usporediti s dosadašnjim lijekovima i ustanoviti jesu li oni »statistički značajno« bolji od tih dosadašnjih lijekova.

Pretpostavimo nadalje (što nam u stvarnosti naravno ne može biti poznato) da se tih 1000 novih preparata ne razlikuje, po svom djelovanju, od dosadašnjih lijekova, te da su, dakle, neefikasni kao nova metoda, odnosno jednako su efikasni kao i prethodna klasična terapija. (To i nije tako fantastično nerealna pretpostavka, jer — prema mišljenju nekih autora (1, 2) — samo oko 0,1 promil (!) testiranih tvari dospije do toga da uđu u službenu distribuciju lijekova.) Iako, dakle, kod testiranja tih tvari mi za svaki novi preparat i do-

sadašnji lijek uzimamo uzorke iz identičnih populacija, ipak ćemo (budući da ispitivanje radimo na nivou značajnosti od 5% kod jednosmjernog testiranja) u oko 5% slučajeva, dakle kod 50 preparata, »ustanoviti« da je razlika »statistički značajna«, tj. da je taj novi preparat »bolji od dosadašnjeg«.

U ovom slučaju mi smo dakle odbacili (na temelju rezultata testiranja) oko 950, a prihvatili oko 50 preparata. No posve je jasno da niti jedan od tih pedesetak prihvaćenih preparata nije zaista efikasniji, jer je svaki od 1000 testiranih preparata pripadao populaciji koja je identična populaciji dosadašnjih lijekova. U ovom smo slučaju, dakle, »lažno otkriće« učinili kod 100% prihvaćenih (usvojenih) lijekova (a nikako ne kod 5% od prihvaćenih lijekova).

Uzmimo sada drugi primjer. Zamislimo da također imamo na testiranju 1000 novih preparata, ali da svi oni pripadaju populaciji koja se zaista razlikuje od populacije dosadašnjih lijekova. Testiranjem tih preparata na nivou signifikantnosti od 5% (na dovoljno velikim uzorcima) naći ćemo kod velike većine njih da je razlika prema dosadašnjem lijeku statistički značajna, a samo jedan vrlo mali broj ćemo (neopravdano) odbaciti, jer nam račun nije pokazao statističku značajnost. (O p a s k a: u ovom slučaju nije moguće čak ni na izmišljenom primjeru točnije odrediti postotak prihvaćenih odnosno odbačenih preparata, jer šansa, da se neka razlika među aritmetičkim sredinama odbaci kao nerealna, pa makar ona među populacijama postoji, ovisi o tome koliko je velika stvarna razlika među aritmetičkim sredinama obih populacija. Kod vrlo velikih razlika gotovo da uopće nema šanse da u pokusu, osim kod vrlo malenih uzoraka, ne uspijemo odbaciti nul-hipotezu. Naprotiv, ako je razlika među aritmetičkim sredinama populacija malena, šansa, da je našim uzorcima nećemo uspjeti potvrditi, mnogo je veća. No to ustvari i nije naročito važno. Pogreška koju smo učinili sastoji se — kako vidimo — u tome da smo određeni (vjerojatno mali) postotak uspješnih i efikasnih lijekova odbacili, ali još uvijek je ostalo mnogo lijekova koje smo prihvatili.)

U ovom primjeru svi prihvaćeni lijekovi se efikasni, dakle nismo učinili niti jedno »lažno otkriće«, jer postotak neefikasnih lijekova među prihvaćenim lijekovima iznosi 0 (nula).

I tako smo konačno došli do toga da možemo dati odgovor na pitanje, postavljeno u naslovu ovog članka, a koje glasi: Koliki postotak prihvaćenih lijekova nije efikasan?

Kao što vidimo, odgovor glasi: to ovisi o tome kakvo je stvarno stanje u populacijama iz kojih uzimamo uzorke. Ako među populacijama nema razlike, svi lijekovi koje smo prihvatili pogrešno su prihvaćeni, pa je prema tome 100% prihvaćenih lijekova neefikasno. Ako naprotiv uzorke uzimamo iz populacija koje se stvarno razlikuju, niti jedan prihvaćeni lijek nije pogrešno prihvaćen, pa prema tome postotak neefikasnih lijekova iznosi 0%. U realnosti se vjerojatno nikada ne nalazimo na ovim ekstremima, pa će u realnosti postotak pogrešno prihvaćenih lijekova odgovarati nekoj vrijednosti između 0 i 100%, a ta će vrijednost biti determinirana stvarnim stanjem u populacijama, tj. omjerom stvarno efikasnih i stvarno neefikasnih preparata. Nažalost, vrlo lako će postotak neefikasnih lijekova (»lažnih otkrića«) biti mnogo veći od postotka za koji velik broj istraživača vjeruje da predstavlja količinu njihovih »lažnih otkrića«, tj. 5%. Razlog tome je u već spomenutoj činjenici da je postotak djelotvornih tvari među svim ispitivanim tvarima vrlo malen (oko

0,1 promil!). Tek u slučaju kad bi postotak djelotvornih tvari među ispitivanim tvarima iznosio oko 50% (pri razini značajnosti od 5%), imali bismo u najboljem slučaju (tj. pod pretpostavkom da smo uspjeli identificirati sve djelotvorne tvari) oko 5% »lažnih otkrića«, tj. nedjelotvornih lijekova među svim »otkrivenim« lijekovima. Evo dokaza: ako pretpostavimo da se među 1000 ispitivanih tvari nalazi polovica stvarno djelotvornih, i ako k tome pretpostavimo da smo uspjeli otkriti svih tih 500 djelotvornih tvari, preostaje još 500 nedjelotvornih, od kojih ćemo — ako radimo na nivou značajnosti od 5% — pogrešno »otkriti« još oko 25 »djelotvornih« (5% od 500). Ukupno ćemo dakle na 525 »otkrića« imati 25 »lažnih otkrića«, što je nešto manje od 5%.

No, ako je postotak djelotvornih tvari mnogo manji od 50% (što je, izgleda, realna slika stvarnog stanja stvari), postotak »lažnih otkrića«, čak i pri »strožem« nivou značajnosti, može biti velik. Ako na primjer neka grupa istraživača ispituje 5000 preparata uz razinu značajnosti od 1%, pa ako među tim ispitivanim tvarima ima 30 koje su stvarno djelotvorne, onda — pod pretpostavkom da svih 30 djelotvornih tvari bude otkriveno — bit će još oko 50 »lažnih otkrića« (1% od 4970). Dakle, postotak »lažnih otkrića« među svim otkrićima bit će 62% (tj. 50/80) (4).

Nije — nadamo se — potrebno dokazivati da ista logika vrijedi za sva istraživanja, pa prema tome i za eksperimente koje vršimo na bilo kojem drugom području. U vezi sa gore navedenim primjerom (ispitivanje lijekova) može se, doduše, staviti primjedba da se u stvarnosti jedna te ista supstancija testira nekoliko puta u uzastopnim pokusima, tako da je, zbog toga, ukupni rizik da smo pogriješili mnogo manji od 5%; ali, on je danas ipak nepoznat, jer ga nitko ne izračunava odnosno ne objavljuje. (Može se pokazati, da bi, s obzirom na spomenutu proporciju od 0,1 promil novih lijekova među ispitivanim tvarima, »ukupna« razina značajnosti trebala iznositi oko 0,000 005, kako bismo mogli tvrditi da među novootkrivenim lijekovima nema više od 5% lažnih lijekova, premda bi ta potrebna razina mogla biti i niža u slučaju da se znatan dio lijekova ne odbacuje zbog nedjelotvornosti nego zbog štetnih nuspojava, tj. u tom bi slučaju rizik smio biti veći od 0,000 005. Međutim, danas se ne zna ni koja je »ukupna« razina značajnosti zapravo potrebna, ni koja se doista postiže!)

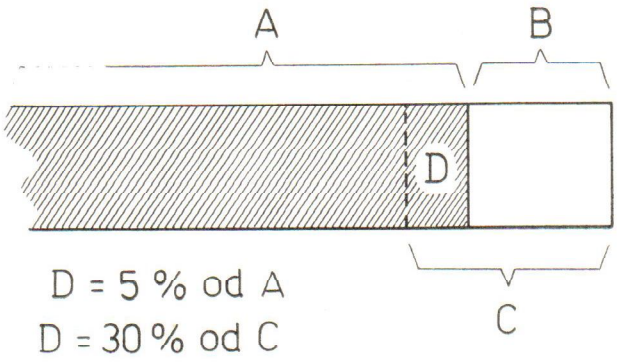
Sada se možemo zapitati zašto dolazi do te pogreške mišljenja i zaključivanja, tj. zašto većina istraživača, usprkos poznavanju statistike, izvodi o svojim istraživanjima pogrešan zaključak, tj. zaključak da »provjerenost« rezultata na razini od 5% znači da su u 5% značajnih rezultata pogrešno zaključili o postojanju neke razlike, koja u stvarnosti ne postoji, ali da su zato u preostalih 95% tih značajnih rezultata ispravno zaključili da stvarno postoji razlika i među populacijama.

Razlog toj pojavi sastoji se u tome da se pravila, koja se odnose samo na nul-hipotezu, primjenjuju i na situacije u kojima se ne radi o nul-hipotezi.

Kao što smo rekli, osnovna logika nul-hipoteze može se slobodnim riječima formulirati ovako: »Ako među populacijama iz kojih uzimamo uzorke nema razlike, onda — ako račune radimo na nivou značajnosti od 5% — naš rizik da smo pogriješili u zaključku (tj. da smo nepostojeću razliku proglasili postojećom) iznosi 5%. Iz toga slijedi da u oko 95% slučajeva nismo pogriješili (naime, u tih 95% slučajeva nismo proglasili da razlika postoji).« To se pra-

vilo potpuno neopravdano primjenjuje na skup svih naših znanstvenih otkrića, pa se zaključuje, da među tim našim otkrićima oko 5% nije ispravno, a 95% jest. No, kao što rekosmo, to se pravilo odnosi samo na situacije kada među populacijama nema razlike. Evo dokaza: ako među populacijama nema razlike, onda ćemo — radeći račune na nivou značajnosti od 5% — pogrešno u oko 5% slučajeva proglasiti da razlike među aritmetičkim sredinama postoje. To ujedno znači da smo u preostalih 95% slučajeva ispravno prihvatili nul-hipotezu i zaključili da ne možemo tvrditi da razlika postoji (a nismo učinili »ispravno otkriće«!), pa zato tih 95% »odbačenih« rezultata ne ulazi u sumu naših znanstvenih otkrića, te se, prema tome, logika o »5% neispravnih i 95% ispravnih zaključaka« ne može odnositi na inventar naših znanstvenih (značajnih) rezultata (npr. na inventar prihvaćenih lijekova).

Pogreška koju radimo dade se dosta dobro prikazati i grafički, što će možda za neke čitaoce biti i jasnije. Na slici 1. tamna površina A predstavlja pre-



Sl. 1. — *Skup eksperimenata (A+B), istinitih nul-hipoteza (A), istinitih alternativnih hipoteza (B) i skup statistički značajnih rezultata (»otkrića«) (C).*

parate koji nisu djelotvorni, a bijela površina B predstavlja djelotvorne preparate. Ispitujući sve te tvari (dakle A+B), pod pretpostavkom da smo uspjeli otkriti sve djelotvorne tvari (površina B), mi smo pogrešno proglasili uspješnima i oko 5% preparata iz grupe A, pa smo kao djelotvorne lijekove prihvatili površinu označenu slovom C. Budući da površina C predstavlja naša »znanstvena otkrića«, mi (pogrešno) tvrdimo da vjerojatno oko 5% tih otkrića nije ispravno. A kao što se iz slike vidi, tamna površina sektora C (označena slovom D) mnogo je veća od 5% i u ovom slučaju iznosi blizu 30%! (D je blizu 30% od C.) Površina D može prekrivati veći ili manji dio površine C, što ovisi o situaciji u populaciji (u skupu eksperimenata); ako u svakom eksperimentu postoji razlika između dvije populacije, tamne površine D uopće nema, tako da je cijela površina C svijetla, a ako ne postoje razlike među populacijama, u svim eksperimentima, onda je cijela površina C tamna, tj. svi naši eksperimenti su »lažna otkrića«. — Pravila u vezi s nul-hipotezom odnose se samo na površinu A. Mnogi, naprotiv, u realnom životu ta pravila pogrešno primjenjuju na površinu C!

Engleski autor *Oakes* (3) navodi da je skupini od 70 akademski obrazovanih psihologa postavio ovaj zadatak:

Pretpostavimo da provjeravate neki postupak, koji bi mogao imati efekt na neku pojavu. Usporedbom aritmetičkih sredina svoje eksperimentalne i kontrolne skupine (recimo da je po 20 ispitanika u svakoj od njih) našli ste, upotrebom t-testa, da je $t = 2{,}7$, $p < 0{,}01$. Označite svaki od donjih odgovora kao »točan« ili »netočan«.

1) Potpuno ste odbacili nul-hipotezu (po kojoj nema razlike među aritmetičkim sredinama obih populacija)
2) Našli ste vjerojatnost da je nul-hipoteza istinita
3) Potpuno ste dokazali svoju eksperimentalnu hipotezu (tj. da postoji razlika među aritmetičkim sredinama obih populacija)
4) Možete izvesti vjerojatnost da je eksperimentalna hipoteza istinita
5) Ako odlučite da odbacite nul-hipotezu, znate vjerojatnost da ste pogrešno odlučili
6) Posjedujete pouzdan eksperimentalni nalaz u smislu, da biste kod velikog broja ponavljanja ovog pokusa dobili u 99% slučajeva statistički značajan rezultat.

Iako su ispitanici bili postdiplomski studenti, nastavnici, istraživači sa najmanje dvije godine istraživačkog iskustva, ipak su dali odgovore koji su niže navedeni (prikazane su frekvencije odgovora »točno«, koje daju zbroj preko 70, jer su ispitanici više odgovora označili kao »točne«):

tvrdnja:	1)	2)	3)	4)	5)	6)
frekvencija:	1	25	4	46	60	42

Iako niti jedan od predloženih odgovora nije točan, kao što vidimo, preko 85% ispitanika (njih 60 od ukupno 70) smatralo je točnom tvrdnju 5). Ta je tvrdnja vrlo slična tvrdnji c) u primjeru našeg eksperimenta, navedenog na početku ovog članka!

Rekli smo da postotak neopravdano prihvaćenih eksperimentalnih rezultata (»otkrića«) ovisi o tome kakvo je stanje u populacijama. Budući da nama stanje u populacijama nikada nije poznato, izgledalo bi da smo posve nemoćni u utvrđivanju postotka »lažnih« odnosno »ispravnih« otkrića i u utjecanju na taj postotak, no to nije posve točno (uostalom, na to je već upozorio jedan od autora ovog napisa) (4, 5); ako razumno biramo probleme, tj. ako postavljamo znanstvene hipoteze kod kojih postoji opravdano teoretsko očekivanje da bi mogle biti istinite, imat ćemo posla s populacijama koje su vjerojatno u dosta velikom postotku stvarno različite. U tom slučaju i postotak »pravih otkrića« među prihvaćenima bit će također velik. Naprotiv, radimo li bez puno razmišljanja, »rutinski«, na način kakav se danas — zbog uspješnih tehničkih pronalazaka koji olakšavaju provođenje različitih analiza — može naći kod nekih istraživača, koji usvoje neku brzu tehniku, i onda je primjenjuju na sve moguće slučajeve što im stoje na raspolaganju, u tom slučaju radimo vjerojatno često s populacijama koje su samo u minimalnom postotku različite, te će prema tome vrlo velik postotak prihvaćenih »otkrića« biti lažna otkrića.

Upotreba različitih gotovih kompjuterskih programa omogućava danas veoma brzo računsko provjeravanje velikog broja eksperimentalnih rezultata, što zavodi pojedine istraživače na rutinsko iskušavanje velikog broja hipoteza u vezi s nekom usvojenom metodologijom, jer se sve one mogu relativno brzo »provjeriti«. Još prije petnaestak do dvadesetak godina, kada je takvih prilika za provjeravanje bilo relativno malo, istraživači su se odlučivali na eksperimente, za koje su imali ozbiljnih argumenata da smatraju da bi mogli dobiti očekivani rezultat, jer je računsko provjeravanje rezultatâ eksperimenata bio dugotrajan i mukotrpan posao. Danas, kada u većini slučajeva istraživač dobiva računski obrađene rezultate bez ikakvog svog udjela u tom poslu, i osim toga u vrlo kratkom vremenu, on je sklon da neku usvojenu tehniku primjenjuje na gotovo sve moguće situacije, pa prema tome i na one u kojima ne postoje gotovo nikakvi razumni izgledi da bi se zaista nešto »moglo naći«. Tim svojim postupkom (uz obrazloženje: »pa možda tu ipak nečega ima!«) on povećava »fond« populacija, koja se ne razlikuju, a time ujedno povećava — a da toga nije svjestan — i postotak »lažnih otkrića« među svojim otkrićima. — »... Istinitost rezultata naučnog rada ... ovisi o sposobnosti inteligentnog zaključivanja, shvaćanja, predviđanja i postavljanja dobrih hipoteza. Od statističkog provjeravanja loših, proizvoljnih hipoteza ne može se očekivati mnogo koristi, čak ni onda kada se dobivaju statistički visoko značajni rezultati« (5).

Iz svega iznesenoga naravno slijedi da to upozorenje vrijedi kako za jedan izolirani problem, koji istražujemo, tako i za skup različitih problema. Posve je, naime, svejedno da li ćemo (npr.) jednom kockom načiniti velik broj uzoraka od po 40 bacanja, pa tražiti u koliko se tih uzoraka pojavilo bar dvanaest »šesticâ« (to bi se trebalo dogoditi u nešto manje od 5% slučajeva), ili ćemo odjednom baciti 40 kocaka, i to ponoviti mnogo puta, pa tražiti u koliko smo takvih ponavljanja dobili bar dvanaest »šestica« istodobno na raznim kockama. Isto ćemo tako naći da je razlika »statistički značajna« u oko 5% eksperimenata (uz tu razinu značajnosti od 5%), premda u tim eksperimentima stvarno nema razlike među populacijama, bez obzira na to da li mnogo puta ispitujemo isti problem (npr. ljekovitost nekog preparata), ili mnoge, razne probleme (preparate) ispitujemo po jedanput. (U oba ta slučaja pretpostavili smo da su preparati u stvari neefikasni, tj. da nema stvarne razlike među populacijama, u svim tim eksperimentima.) Kao što je već objašnjeno, kad bi (npr.) od 1000 raznih, nezavisnih istraživanja gotovo sva (ili bar oko polovine njih) dala statistički značajne rezultate, mogli bismo zaista vjerovati da se velika većina tih značajnih rezultata nije pojavila slučajno, već da se radi o istinskim otkrićima; ali ako bi se u tih 1000 istraživanja (eksperimenata) dobilo samo oko 50 rezultata značajnih na razini 0,05, tada bi bilo opravdano vjerovati da su gotovo sva ta »otkrića« lažna — jer se u 1000 eksperimenata (uz tu razinu od 5%) upravo i očekuje oko 50 slučajnih značajnih rezultata (tj. 5% od 1000). Budući da je danas nepoznat taj omjer dobivenih otkrića i izvršenih eksperimenata, nemamo pravo dosadašnje objavljene statističke testirane eksperimentalne rezultate (»otkrića«) smatrati doista dovoljno provjerenima!

Prema tome, možemo reći da postizanje pojedinačnog statistički značajnog rezultata (pogotovo na razini kao što je 0,05 ili sl.) zapravo nije samo po sebi

dovoljno za odbacivanje nul-hipoteze, jer rizik lažnog otkrića ne ovisi samo o postignutoj razini značajnosti, već ovisi i o omjeru istinitih i neistinitih nul-hipoteza u velikom skupu istraživanja, a taj se omjer odražava i u omjeru uspješnih i neuspješnih istraživanja, tj. značajnih i neznačajnih rezultata!

Literatura

1. *Dunnet, C. W.:* Drug screening: the never-ending search for new and better drugs. U: Statistics: a guide to the unknown, ur. Tanur J. M., Mosteller F., Kruskal W. H., Link R. F., Pieters R. S., Rising G. R., San Francisco: Holden-Day Inc., 1972, str. 23—33.
2. *Vrhovac B. i sur:* Kliničko ispitivanje lijekova. Zagreb, Školska knjiga, 1984, str. 3.
3. *Oakes M. W.:* Statistical inference: a commentary for the social and behavioural sciences. Chichester-New York-Brisbane-Toronto-Singapore: John Wiley & Sons, 1986, str. 79—82.
4. *Sorić B.:* Poboljšanje metode i kontrola ispravnosti statističkog odlučivanja. Zdravstvo *23* (1981) 154—170.
5. *Sorić B.:* Kritika statističkog odlučivanja. Zdravstvo *23* (1981) 143—153.

Summary

WHAT IS THE PERCENTAGE OF »FALSE DISCOVERIES«?

The risk of discarding a true null hypothesis is assessed by means of statistical verification of the results of scientific investigations. The way of reasoning about this risk is mistakenly applied to the risk (percentage) of obtaining »false discoveries«, e. g. to the percentage of ineffective drugs among the accepted medicaments. The percentage of mistakenly accepted results in a set of statistically significant results (»discoveries«) depends not only on the level of statistical significance but also upon the unknown situation in the set of investigated populations, or, likewise, upon the ratio of significant and insignificant results, which could be estimated, but which is neglected today.

Health Centre »Pešćenica«, Zagreb
and Department for Psychology,
Faculty of Arts, University of Zagreb,
Zagreb

Received for publication
December 26, 1986