Analiza statističke snage testa u znanstvenom istraživanju

Power analysis in research

Mary L. McHugh

Fakultet sestrinstva, Sveučilište Indianapolisa, Indianapolis, Indiana, SAD
School of Nursing, University of Indianapolis, Indianapolis, Indiana, USA

Sažetak

Statistička snaga testa je mjera vjerojatnosti da će istraživač u uzorku naći statističku značajnost, ako učinak postoji u cijelokupnoj populaciji. Snaga testa je funkcija ovisna o tri primarna i jednom sekundarnom čimbeniku: veličini uzorka, veličini učinka, razini značajnosti i snazi korištenog statističkog testa. Najčešći razlog provođenja analize snage testa jest određivanje veličine uzorka potrebnog za određeno istraživanje. No, analiza snage testa može se izračunati i nakon završetka istraživanja kako bi se odredilo je li nedovoljna snaga bila razlog statistički neznačajnog učinka. Općenito, ne preporuča se *post hoc* analiza snage testa; ona bi se trebala provesti prije početka istraživanja. Njom se istražuje utjecaj veličine učinka, značajnosti, veličine uzorka i snage statističkih testova.

Ključne riječi: snaga testa; značajnost; veličina učinka; veličina uzorka; statistička snaga testa

Abstract

Statistical power is a measure of the likelihood that a researcher will find statistical significance in a sample if the effect exists in the full population. Power is a function of three primary factors and one secondary factor: sample size, effect size, significance level, and the power of the statistic used. The most common reason to conduct a power analysis is to determine the sample size needed for a particular study. However, power analysis may also be used after a study has been completed to determine if the reason an effect was not significant was insufficient power. Generally, however, *post hoc* power analysis is not suggested; that work should be done prior to beginning a study. The influence of effect size, significance, sample size, and the power of the statistic are explored.

Key words: power; significance; effect size; sample size; statistical power

Pristiglo: 1. srpnja 2008. *Received: July 1, 2008*

Prihvaćeno: 6. kolovoza 2008. *Accepted: August 6, 2008*

Uvod

Snaga testa je vrlo važan koncept za istraživače, budući da je ona stup na koji se naslanjaju postignuća statističke značajnosti. Statistička značajnost je čimbenik istraživanja koji istraživači rabe kako bi odredili je li intervencija promijenila rezultat. To se ne može postići testom nedovoljne snage. S druge strane, iznimno jaka snaga testa može utjecati na istraživača da pridoda mnogo veće značenje statističkom rezultatu nego što to opravdava klinička situacija. Svrha ovoga članka jest dati pregled temelja statističke snage testa te informacije o tome kako se ona rabi u svrhu povećanja vjerojatnosti dobivanja pouzdanih informacija iz istraživanja.

Introduction

Power is a critically important concept for researchers because it is the hub around which the achievement of statistical significance revolves. Statistical significance is the research factor that researchers use to determine if an intervention changes an outcome. That determination cannot be achieved with insufficient power. On the other hand, extremely high power might influence a researcher to give more weight to a statistical result than the clinical situation warrants. The purpose of this paper is to review the foundations of statistical power, and to provide information on how it is used to increase the probability of obtaining reliable information from research studies.

Značenje snage testa

U kontekstu istraživanja, snaga se odnosi na vjerojatnost da će istraživač naći značajan rezultat (učinak) u uzorku ako takav učinak postoji u populaciji koju ispituje (1). Uporabom nul-hipoteze istraživač postavlja pitanje o značajnom rezultatu. Nul-hipoteza uvijek iznosi hipotezu da ne postoji razlika između eksperimentalne i kontrolne skupine za varijable koje se ispituju. Nul-hipoteza je ono što sve inferencijske statistike testiraju.

Vrijednosti koje snaga može podnijeti sežu od 0,0 do 1,0. Te se vrijednosti ne mogu tumačiti izravno. Međutim, vjerojatnost pogreške tipa II. računa se kao 1 – snaga. Stoga vrijedi da, što je veća snaga, to je vjerojatnije da će se otkriti značajan učinak. Kada je snaga manja, nije vjerojatno da će istraživač naći učinak i time odbaciti nul-hipotezu, čak i kada postoji stvarna razlika između eksperimentalne i kontrolne skupine. Učinak koji istraživač želi naći jest alternativna hipoteza – što zapravo predstavlja hipotezu istraživanja. To se obično izražava ovim riječima: „Postoji razlika između eksperimentalne i kontrolne skupine". Opisana na drugi način, snaga je vjerojatnost da će lažna nul-hipoteza (odnosno, postoji učinak u cjelokupnoj populaciji) biti odbačena (Tablica 1.). Kada se odbaci nul-hipoteza, prihvaća se ona alternativna. Neizravno to znači da je snaga ključni čimbenik sposobnosti istraživača da izvuče točan zaključak iz podataka koje mu daje uzorak.

Problemi sa snagom mogu dovesti do različitih pogrešaka u tumačenju statističkih rezultata. Mogu navesti istraživača na zaključak da eksperimentalno liječenje nije polučilo učinak, a taj učinak zapravo postoji u populaciji. Mogu ga navesti na netočan zaključak kako postoji značajan učinak, koji doista i postoji, no toliko je malen da je beznačajan. Stoga je važno da svaki istraživač shvati značenje snage testa i čimbenika koji utječu na statističku snagu testa, kako bi statistički zaključci bili točniji i pouzdaniji.

Temelji statističke snage testa

Statistička snaga testa je funkcija ovisna o tri čimbenika (Slika 1.) i jednom dodatnom čimbeniku. Primarni su čim-

Meaning of power

In the context of research, power refers to the likelihood that a researcher will find a significant result (an effect) in a sample if such an effect exists in the population being studied (1). The way a researcher poses the question about a significant result is through use of the null hypothesis. The null hypothesis always proposes the hypothesis that there is no difference between the experimental and control groups for the variable being tested. The null hypothesis is what all inferential statistics test.

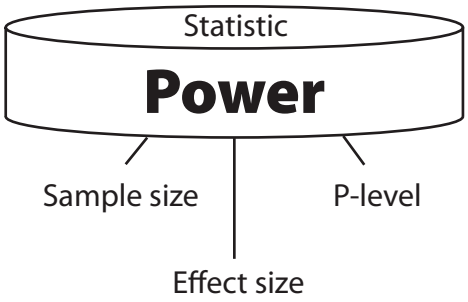
The values that Power can take range from 0.0 to 1.0. These values cannot be interpreted directly. However, the probability of a Type II error is calculated as 1-Power. Therefore, the higher the power, the more likely one is to detect a significant effect. When power is low, it is unlikely that the researcher will find an effect, and thus reject the null hypothesis, even when there is a real difference between the experimental and control groups. The effect the researcher is trying to find is the alternate hypothesis – which is, of course, the study hypothesis. That is typically worded in a fashion similar to this statement: "There is a difference between the experimental and control groups". Described in a different way, power is the likelihood that a false null hypothesis (that is, there is an effect in the full population), will be rejected (Table 1). When a null hypothesis is rejected, the alternate hypothesis is accepted. Indirectly, this means that power is a key factor in the researcher being able to draw correct conclusions from sample data.

Problems with power can lead to a variety of errors in interpretation of statistical results. They might lead the researcher to conclude there is no effect from an experimental treatment when in fact an effect does exist in the population. They might lead the researcher to incorrectly conclude that there is an important effect when the fact is that there is an effect, but it is so small as to be inconsequential. Therefore, it is important for every researcher to understand the meaning of power and the factors that affect statistical power so that statistical conclusions are more accurate and reliable.

Tablica 1. Pogreške tipa I. i II.

Table 1. Type I and type II errors

	Null hypothesis is true	Null hypothesis is false
Researcher accepts the null hypothesis	Correct decision	Type II error 1 – power (1 – β)
Researcher rejects the null hypothesis	Type I error alpha (α) or p-level (usually P < 0.05)	Correct decision



Slika 1. Tri komponente snage testa.

Figure 1. The three components of power.

benici veličina učinka i uzorka te razina značajnosti koja se rabi u istraživanju. Sekundarni čimbenik je snaga primijenjenih statističkih testova. Kada su poznata bilo koja dva primarna čimbenika iz njih se može izračunati treći, a kada su poznata sva tri primarna čimbenika može se izračunati snaga statističkog rezultata. Jednako važno je da se, kad su poznati snaga i samo jedan primarni čimbenik – veličina učinka, može izračunati veličina uzorka potrebna za postizanje statističke značajnosti.

Veličina uzorka

Prvi čimbenik – i čimbenik koji je pod najizravnijom kontrolom istraživača – jest veličina uzorka. Veličina uzorka je zapravo jedini čimbenik koji istraživač zaista može kontrolirati. Ona vrlo izravno i snažno utječe na statističku snagu testa u svakom istraživanju. Jednostavno rečeno, što je veći uzorak, to je veća statistička snaga. Suprotno tome, kada je uzorak malen, statistička snaga je slaba. To je logički istinito jer znamo da bi istraživač, kad bi mogao ispitati čitavu, cjelokupnu populaciju, imao potpunu moć naći bilo koji učinak koji postoji u populaciji za mjerene varijable. Zapravo bi tada inferencijska statistika bila nepotrebna. Inferencijska statistika dozvoljava istraživaču da iz uzorka izvede zaključak (procijeni) o veličini učinka u populaciji. Kad bi se ispitivala cijela populacija, ne bi bilo potrebno procjenjivati učinak, budući da bi njegova veličina bila odmah poznata. Drugim riječima, ako istraživač ispituje cjelokupnu populaciju, snaga statističkog testa je 100%, jer je tada svaki učinak otkriven.

Nadalje, ako istraživač ispituje cjelokupnu populaciju nema opasnosti da uzorak bude slaba procjena populacije. Iako uzorkovanje nije tema ovoga članka, važno je naglasiti da je inferencijska statistika točna u onolikoj mjeri u kojoj uzorak predstavlja populaciju. Stoga, niti jedna teorija koja podupire istraživanje uzorka ne vrijedi ako istraživač sakupi pristran uzorak (odnosno uzorak koji ne predstavlja populaciju). Kod istraživanja koje rabi cjelokupnu populaciju ne postoji opasnost od nereprezentativnog rezultata.

Foundations of statistical power

Statistical power is primarily a function of three factors (Figure 1), and secondarily of one additional factor. The primary factors are sample size, effect size and level of significance used in the study.

The secondary factor is the power of the statistic used. When any two of the primary factors are known, the third can be calculated from the other two. And when all three factors are known, the power of a statistical result can be calculated. Equally important, when power and just one of the primary factors – effect size – are known, the sample size needed to achieve statistical significance can be calculated.

Sample size

The first factor – and the factor most directly under the control of the researcher – is sample size. In fact, sample size is often the only factor that the researcher can realistically control. Sample size has a very direct and very strong effect on statistical power in any study. Simply put, the larger the sample, the greater the statistical power. Conversely, when sample size is small, power is weak. This is logically true because we know that if the researcher could measure an entire, large population, then the researcher would have complete power to find any effects that might exist in the population for the variables measured. In fact, inferential statistics would be unnecessary. Inferential statistics allow the researcher to infer (estimate) the effect size in the population from a sample. If the entire population were measured, there would be no need to estimate the effect because the effect size would be directly known. In other words, if a researcher measures the entire population, the power is 100% because any effect will be detected. Furthermore, if the researcher measures the entire population, there is no danger of the sample being a poor estimate of the population. Although sampling is not the topic of this paper, it is necessary to note that inferential statistics are only as accurate as

Obrnuto, dobro je poznato da su vrlo mali uzorci nepouzdani procjenjivači nekog populacijskog parametra. Niti jedan savjestan istraživač neće niti pokušati predvidjeti djelovanje novog lijeka na milijunsku populaciju uzimajući samo jednu osobu kao uzorak. Visoka vjerojatnost pogrešnog zaključka ako je „N = 1" je toliko dobro poznata da je postala kliše. Veličina uzorka od 5 ispitanika bi isto tako bila loša za testiranje djelovanja novog lijeka. Taj je uzorak premalen da bi mogao predstavljati široku populaciju. Zapravo, često rabljeno nepisano pravilo u istraživanjima kaže kako se uzorci manji od 30 ispitanika smatraju malima i da bi se trebali rabiti samo u probnim istraživanjima.

Tada se nameće pitanje: „Koja je veličina uzorka potrebna istraživaču kako bi otkrio učinak ako dotični postoji u populaciji?" Tipični način pronalaska odgovora na to pitanje zove se analiza statističke snage testa (engl. *power analysis*) i uključuje izvođenje matematičkih izračuna kako bi se odredilo koja je veličina uzorka potrebna za otkrivanje učinka određene veličine. Kako bi izračunao potrebnu veličinu uzorka istraživač mora znati veličinu učinka. Treba također naglasiti da ponekad istraživač otkrije kako umjerena veličina učinka nije statistički važna. U tom se slučaju može provesti analiza statističke snage, ako je problem statističke značajnosti bio nedovoljno snažan zbog nedostatne veličine uzorka.

Postoje razni programi dostupni na internetu koji pomažu istraživaču brzo odrediti veličinu uzorka. Jedan od najkorisnijih može se naći na mrežnoj stranici Sveučilišta u Iowi: http://www.stat.uiowa.edu/~rlenth/Power/index.html (2). Korisnik prepoznaje statistički test koji rabi, unosi podatke o veličini učinka i program će izračunati potrebnu veličinu uzorka te određenu razinu statističke snage. Na primjer, pretpostavimo da istraživač planira provesti istraživanje na dva slučajno odabrana uzorka od kojih je jedan bio na eksperimentalnom liječenju, a drugi nije. Tipični test za ispitivanje razlika između skupina je t-test. Početna stranica nudi izbornik s različitim statističkim testovima. Kad korisnik dvostrukim pritiskom na lijevu tipku miša odabere jedan od tih testova, na zaslonu se pojavi grafičko korisničko sučelje (engl. *graphical user interface,* GUI) s kalkulatorom (Slika 2.). Treba obratiti pozornost na to da je na slici 2. veličina učinka 0,50, ali je snaga tek 0,41. Te razine rezultiraju potrebnom veličinom uzorka od samo 25 ispitanika u svakoj skupini (ukupno N = 50). Međutim, ta je snaga preslaba za istraživanje, pa je na slici 3. snaga postavljena na 0,80, tako da se mišem klikne i povuče alatna traka u odjeljku *Power*. Zanimljivo je da je veličina uzorka po skupini potrebna za pronalaženje učinka veličine 0,50 i snage 0,80 narasla na N = 63. Pretpostavimo kako istraživač želi snagu od 0,80, ali sumnja da će veličina učinka biti samo 0,35. Slika 4. pokazuje da je veličina uzorka potrebna za pronalaženje tog učinka porasla na 129 ispitanika po skupini. Na taj način istraživač može rabiti mrežnu stranicu Sveučilišta u Iowi u određivanju veličine uzorka

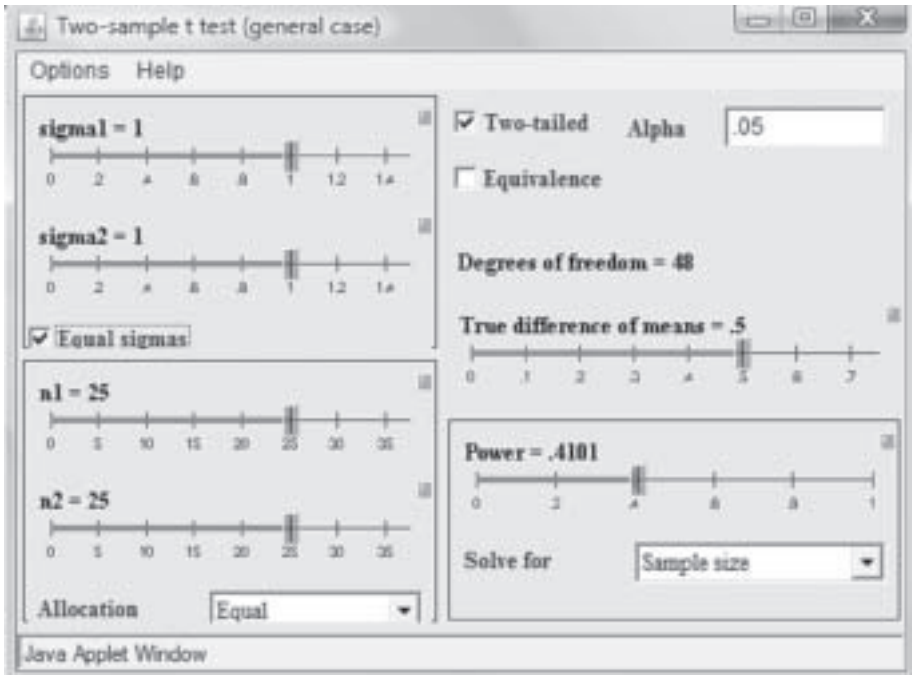
the sample is representative of the population. Therefore, none of the theories that support sample research apply if the researcher obtains a biased sample (that is, a sample that is not representative of the population). With a study that uses the entire population, there is no danger of an unrepresentative result.

Conversely, it is well known that very small sample sizes are unreliable estimators of a population parameter. No sensible researcher would try to predict the effect of a new drug on a population of millions by sampling one individual. The high likelihood of an erroneous conclusion with an "N of one" is so well known as to constitute a cliché. A sample size of 5 individuals would be almost as bad for testing the effects of a new drug. That sample size is too small to fully represent a large population. In fact, a heuristic often used in research is that samples of less than 30 are considered small sample sizes and should be used only for pilot studies.

The question then arises, "What sample size does a researcher need to detect an effect if it exists in the population?" The typical way to find the answer to that question is called "power analysis" and it involves performing mathematical calculations to determine what sample size is needed to detect an effect of a certain size. In order to calculate the sample size needed, the researcher needs to know the effect size. It must also be noted that sometimes a researcher discovers that a moderate effect size is not found to be statistically significant. A power analysis might be performed in this case to discover if the problem with statistical significance was insufficient power due to an inadequate sample size.

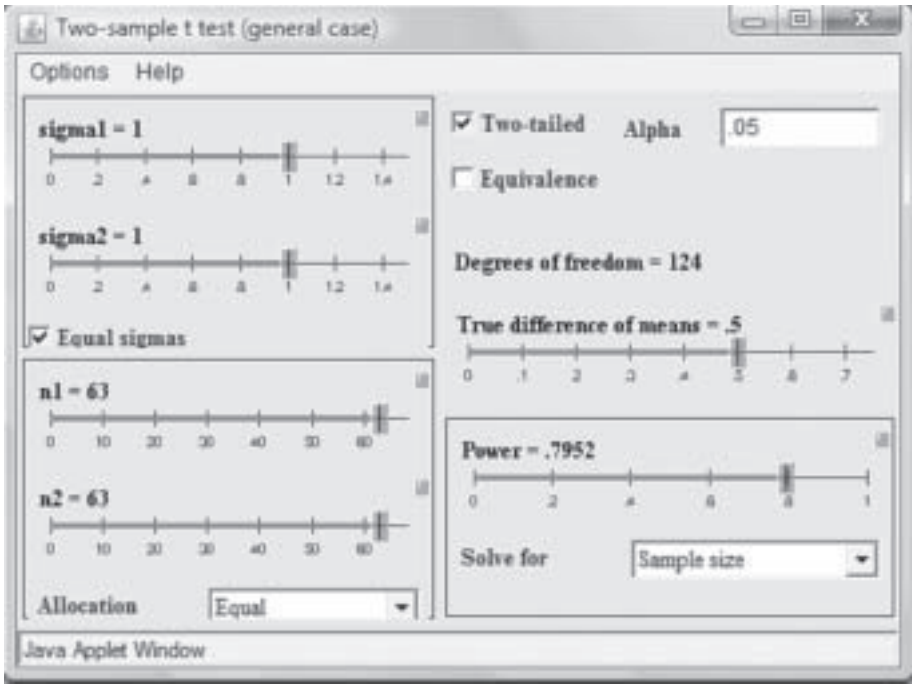
There are a variety of programs available *via* the Internet to assist the researcher to quickly determine sample size. One of the most useful can be found on the University of Iowa web site (2): http://www.stat.uiowa.edu/~rlenth/Power/index.html. The user identifies the statistic to be used, and inputs information about effect size and the program will calculate the sample size required for a particular power level. For example, suppose the researcher plans to run a study on two randomly assigned samples, one of which has received an experimental treatment and the other has not. The typical test used to test group differences is the t-test. The home screen offers a screen menu on the site with a variety of statistical tests. When the user double clicks on one of the statistics in that menu, the graphical user interface (GUI) calculator comes up on the screen (Figure 2).

Note on Figure 2 that effect size is 0.50 but power is only 0.41. Those levels result in a needed sample size of only 25 in each study group (total N = 50). However, that power is too weak to use in a research study, so in Figure 3, the power has been reset to 0.80 by simply clicking and dragging on the bar in the Power box. Notice that the *per*-group sample size required to find an effect size of



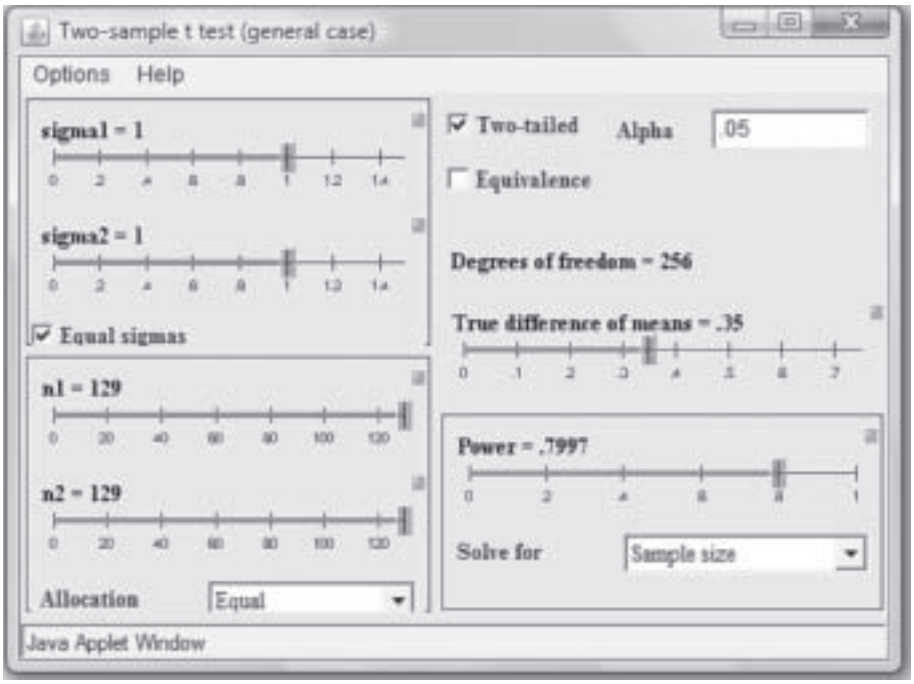
Slika 2. Kalkulator snage testa s mrežne stranice Sveučilišta u Iowi – t-test kalkulator.

Figure 2. University of Iowa online power calculator – test calculator.



Slika 3. Potrebna veličina uzorka kada je snaga povećana na 0,80.

Figure 3. Sample size needed with power changed to 0.80.



Slika 4. Promjena veličine uzorka zbog veličine učinka.

Figure 4. Sample size change due to change in effect size.

koji će biti potreban za postizanje značajnosti za određenu veličinu učinka i razinu snage. Upute za uporabu kalkulatora snage mogu se naći na mrežnoj stranici: http://hschealth.uchsc.edu/son/pdf3/PowerCalculatorsHowTo.pdf. Druga stranica s dodatnim izračunima snage i veličine uzorka mogu se naći na mrežnim stranicama Sveučilišta Harvard: http://hedwig.mgh.harvard.edu/sample_size/size.html.

Veličina učinka

Veličina učinka predstavlja veličinu razlike između liječene i neliječene skupine u istraživanju, odnosno predstavlja magnitudu učinka liječenja (3). Istraživači provode eksperimentalna istraživanja kako bi testirali veličinu učinka. Odnsno, tipično je da istraživači teže otkriti ima li liječenje učinak kod eksperimentalnih ispitanika, i ako je tako, koju veličinu učinka je ono proizvelo? Ljudi često misle na korelaciju kada govore o veličini učinka. To je normalna pojava budući da su korelacije mjere veličine učinka. Međutim, statistički testovi za ispitivanje razlika kao što su t-test i ANOVA također imaju svoju veličinu učinka. Zapravo, mjera veličine učinka za t-test je *point-biserijski* koeficijent korelacije, a za ANOVU *Eta-square* statistika.

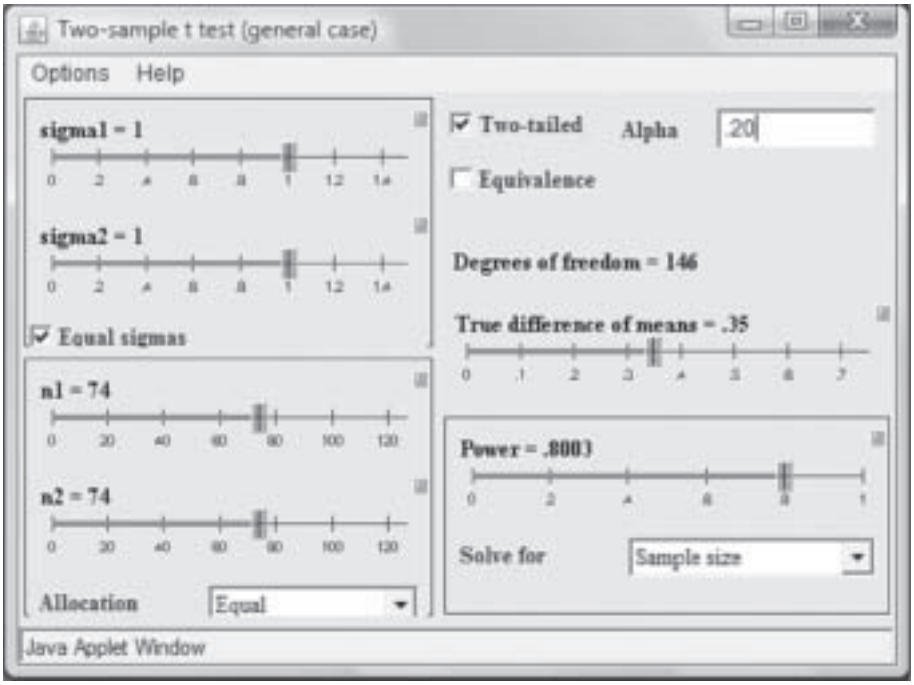
Svi statistički testovi koji se upotrebljavaju za mjerenje učinka liječenja – dakle, čitava inferencijska statistika – imaju odgovarajuću mjeru za veličinu učinka. Niti jedan istraživač ne bi trebao objavljivati značajnost, a da pritom

0.50 at 0.80 power has increased to N = 63. Now suppose the researcher wants a power of 0.80 but suspects the effect size will be only 0.35. Figure 4 shows the sample size required to find that effect has raised to 129 *per* group. In this way, the researcher can use the

University of Iowa site to determine the sample size needed to achieve significance for a particular effect size and power level. The reader should note that there is a set of directions for using the University of Iowa power calculator at the following web site: http://hschealth.uchsc.edu/son/pdf3/PowerCalculatorsHowTo.pdf. Another site with additional power and sample size calculations can be found at Harvard University's site: http://hedwig.mgh.harvard.edu/sample_size/size.html

Effect size

Effect size represents the size of the difference between the treated and untreated groups in a research study, that is, it represents the magnitude of the treatment effect (3). It is to test for effect size that researchers perform experimental studies. That is, researchers typically seek to discover if a treatment produces an effect in the experimental subjects, and if so, what size of an effect did the treatment produce? People often think of correlation when they think of effect size. This is natural because correlations are measures of effect size. However, difference statistics such as the t-test and ANOVA also have an effect



Slika 5. Promjena veličine uzorka zbog promijene u značajnosti

Figure 5. Sample size change due to change in alpha level.

ne objavi i veličinu učinka. Iako sofisticiran i statističkim znanjima i vještinama potkovan čitatelj može iz veličine uzorka i razine značajnosti sam procijeniti veličinu učinka, nema potrebe da on sam izvodi izračune. Odgovornost je istraživača pružiti čitatelju podatke potrebne za ispravno vrednovanje istraživanja.

Kao što je prethodno spomenuto, podatak o razini značajnosti i veličini uzorka može navesti čitatelja na krivi zaključak. To je zbog toga što će vrlo velik uzorak, dakle 1000 ili više ispitanika, dati značajan rezultat čak i za vrlo mali učinak. Pogledajmo što to znači na primjeru izvještavanja o značajnoj korelaciji između primjene nekog biljnog pripravka i kraćeg tijeka neke poznate bolesti, recimo obične prehlade. Čitatelji bi iz značajnog rezultata mogli pretpostaviti da će im se stanje znatno brže poboljšati ako samo taj biljni pripravak uzmu u trenutku kada im započne prehlada. Međutim, da je uzorak bio 2500 ispitanika i da je trajanje prehlade u skupini koja je dobila tu biljku kao lijek bilo samo 5 minuta dulje, taj bi rezultat bio statistički značajan dok klinički ne bi bio značajan. Bitno je da istraživač razumije kako će ekstremno visoka razina snage dati statistički značajne rezultate, čak i za krajnje male uzorke.

Postoji bitna razlika između statističke i kliničke značajnosti. Za prosječnu čitalačku publiku ova vrst razlike može izgledati vrlo složena. Međutim, istraživači bi trebali biti svjesni činjenice da veliki uzorci vrlo dobro polučuju

size. In fact, the effect size measure for the t-test is the point biserial correlation coefficient, and the eta-squared statistic is the effect size measure for ANOVA.

All statistics used to measure treatment effects – that is, all inferential statistics – have an associated effect size measure. No researcher should ever report significance without also reporting the effect size. While the statistically sophisticated reader can estimate effect size from the sample size and significance level, there should never be the need for a reader to perform that calculation. It is the responsibility of the researcher to provide the reader with the information the reader needs to properly evaluate the study.

As mentioned earlier, a significance level and sample size report can result in a misled reader. This is because a very large sample size, that is, 1,000 or more subjects, will produce significant results even for very small effect sizes. Suppose, for example, the researcher reports a significant correlation between the use of some herb and a shorter course of a common illness, such as common cold. Readers might assume from the significant result that if they only take the herb when they come down with a cold, they will get well much faster. However, if the sample size was 2,500 and the duration of the cold in the herb group only 5 minutes shorter, that result would be statistically significant. It would not be clinically significant. It is im-

pouzdane rezultate, no oni također daju značajne rezultate za gotovo sve veličine učinka. Kada su lijekovi, biljni pripravci i ostala kemijski aktivna sredstva predmet ispitivanja, bitno je ne uzimati u obzir samo statističku značajnost. Veličina učinka se također mora uzeti u obzir. Jako mali učinci (oni od 0,30 ili manje) trebali bi se uzeti u obzir s oprezom. Oni bi se prije mogli smatrati slučajnim nego pouzdanim učincima u širokoj populaciji. Oni kazuju kako je liječenje izazvalo mali učinak na zavisnoj varijabli. Ta bi se veličina učinka trebala kvadrirati kako bi se ocijenio postotak varijance u zavisnoj varijabli koji je proizvela nezavisna varijabla. Stoga, učinak veličine 0,30 znači da je liječenje objasnilo samo 9% razlike u zavisnoj varijabli; 91% učinka na zavisnoj varijabli nije objašnjeno nezavisnom varijablom. Stoga se može zaključiti kako je učinak liječenja bio premalen da bi se ljudima preporučilo da odvoje novac za liječenje – naročito stoga što će ta terapija (lijek ili biljni pripravak) gotovo sigurno imati i neželjene nuspojave kod nekih ljudi. Rizik od nuspojava nije vrijedan male moguće koristi.

U slučaju da istraživač možda ne zna koliki učinak treba očekivati od liječenja, kako onda može rabiti izračune za određivanje potrebne veličine uzorka? Postoje dva načina na koja istraživač može odrediti veličinu učinka: prijašnja istraživanja i minimalna veličina učinka od interesa.

Ako su provedena prethodna istraživanja, njihova veličina učinka može poslužiti istraživaču kao najbolja procjena veličine učinka kojim bi liječenje moglo rezultirati. Kad su takva istraživanja dostupna, njihova bi se veličina učinka trebala uzeti u obzir. Međutim, uobičajenija je situacija da ne postoje prethodna istraživanja za izvorno istraživanje ili se ta prethodna istraživanja previše razlikuju od onog izvornog. Moguće je da su prethodna istraživanja bila provedena na životinjskim vrstama različitima od onih koje predloženo istraživanje planira rabiti.

Za minimalnu veličinu učinka nema prihvaćenog standarda. Ona može varirati ovisno o situacijama. Na primjer, ako postoji ozbiljna bolest bez učinkovite terapije, minimalna veličina učinka može biti relativno mala. Ako novi lijek uzrokuje samo 10% poboljšanja u rezultatima, to bi moglo biti korisno za bolesnike. Kako bi se postiglo tih 10% veličina učinka mora biti 0,32, a to znači da se 32% promjene u zavisnoj varijabli može pripisati liječenju. Međutim, ako postoji prihvaćeno liječenje s poznatim učinkom, minimalna veličina učinka bi trebala, u većini slučajeva, biti veća za jednu cijelu veličinu učinka od učinka poznatog liječenja. Na primjer, ako poznato liječenje pokaže učinak veličine 0,45, novi lijek bi morao imati učinak od barem 0,55. Ili možda veličina učinka novog lijeka iznosi samo 0,45, ali on ima znatno manje nuspojava (ili su manje ozbiljne). Kao što se može vidjeti, odabir minimalne veličine učinka je rezultat istraživačeva poznavanja srodnih istraživanja i njegove dobre procjene.

portant for the researcher to understand that extremely high power levels will produce statistically significant results, even for minuscule effect sizes.

There is an important difference between statistical significance and clinical significance. It can be difficult for the average member of the public to understand that kind of a difference. However, researchers should be cognizant of the fact that while large sample sizes are very good for producing reliable results, they also produce significant results for almost every effect size. When drugs, herbal remedies, and other chemically active agents are the subject of research studies, it is important to consider not only statistical significance. Effect size must be considered as well. Very small effect sizes (effect sizes of 0.30 or less) should be viewed with skepticism. They may be random rather than reliable effects in a large population. And they mean that the treatment produced a small effect on the dependent variable. The effect size should be squared to evaluate the percentage of variance in the dependent variable produced by the independent variable. Thus, an effect size of 0.30 means that the treatment accounted for only 9% of the difference in the dependent variable. Ninety-one percent of the effect on the dependent variable was not accounted for by the independent variable. Therefore, the treatment effect was too small to recommend that people spend money on the treatment – especially since the treatment (drug or herb remedy) will almost certainly have deleterious side effects in some people. The risk of side effects is not worth the small potential benefit.

Given that the researcher may not know what effect size to expect from a treatment, how then shall the calculators be used to determine sample size needed? There are two ways that the researcher may select an effect size: prior studies and minimal effect size of interest.

If prior studies have been performed, the effect size reported may be the researcher's best estimate of the effect size likely to be caused by the treatment. When such studies are available, prior reports of the effect size should be considered. However, the more common situation for original research is that either there are no prior studies of the treatment effect, or the prior studies were too dissimilar to the proposed study. Or perhaps prior studies were performed in an animal species different from that the proposed study intends to use.

The minimal effect size has no accepted standard. It may differ among situations. For example, if there is a serious disease with no effective treatment, the minimal effect size may be relatively small. If the new drug accounts for only 10% of the improvement in outcomes, that may be worthwhile to patients. To achieve that 10%, the effect size must be 0.32 and this means that 32% of the change in the dependent variable can be attributed to the treatment. However, if there is an accepted treatment with a

U kliničkim istraživanjima istraživač određuje najmanju veličinu učinka koja bi bila klinički važna. Ako su rezultati zbog velikog uzorka statistički značajni, a klinički učinak u ljudskoj populaciji beznačajan, tada rezultati nisu klinički značajni. To je drugačiji standard od standarda za statističku značajnost.

Kao primjer uzmimo da medicinski istraživač proučava sepsu koja nije uzrokovana na meticilin rezistentnim zlatnim stafilokokom (*engl*. non-MRSA). Standardni lijek koji je u upotrebi ima stopu preživljenja od 60%. Novi lijek nudi stopu preživljenja od 62% u uzorku od 2204 ispitanika. Veličine učinka su 0,77 za standardni i 0,79 za novi lijek (iznosi su zaokruženi). Općenito, novi će lijek biti puno skuplji. Je li 2% poboljšanja rezultata vrijedno milijuna dolara godišnje više za troškove liječenja? Ako troškovi liječenja ostanu isti, jesu li nuspojave drugačije? Koju bi veličinu učinka istraživač zahtijevao u ovom tipu istraživanja lijeka, ako se dogodi da ili troškovi novog lijeka budu puno viši ili ako novi lijek ima neugodne ili opasne nuspojave? Ovo su pitanja koja istraživač mora uzeti u obzir kada odabire minimalnu veličinu učinka.

Kada jednom odredi potrebnu veličinu učinka, istraživač ju jednostavno unese u kalkulator s minimalnom veličinom učinka i kalkulator odredi veličinu uzorka potrebnu za određivanje veličine učinka. Ako učinak postoji, ali je manji od minimalne veličine učinka za istraživanje, neće postići značajnost. Ako postoji učinak veličinom isti ili veći od minimalnog učinka od interesa, rezultat će biti značajan. Na taj način istraživač može planirati probno istraživanje koje neće samo pomagati preliminarnim testiranjem instrumenata i postupaka prikupljanja podataka, nego će isto tako poboljšati vjerojatnost isplativosti provođenja glavnog istraživanja.

Razina značajnosti

Razina značajnosti istraživanja, također zvana i *P*-razinom, obično se postavlja znanstvenom konvencijom. Na primjer, u većini društveno znanstvenih istraživanja razina značajnosti trebala bi biti 0,50 ili manja. U nekim istraživanjima lijekova *P*-razina mora biti niža od 0,50 zbog nadzornih državnih propisa za djelotvornost i sigurnost. Značajnost predstavlja vjerojatnost pojave pogreške tipa I. To znači kako postoji vjerojatnost da će istraživač krivo tvrditi kako je postigao značajan učinak kad u populaciji nije bilo učinka (Tablica 1.).

S vrlo malim uzorkom, ili uzorkom koji je slabo zastupljen u populaciji, uvijek je visoka vjerojatnost da neće biti učinka ili suprotno, da bilo koji učinak pronađen u uzorku neće biti prisutan u cjelokupnoj populaciji. Stoga, kada se provode probna istraživanja s malim uzorkom, uobičajeno je da istraživač postavi razinu značajnosti više nego obično kako bi nadoknadio mali uzorak. Zato kada je konvencionalna razina značajnosti $P < 0,50$, u probnom bi se istraživanju mogla rabiti *P*-razina od 0,10 ili čak 0,20.

known effect, the minimum effect size should, in most cases, be an effect greater than the effect of the known treatment. For example, if the known treatment exhibits an effect size of 0.45, the new drug should have an effect of at least 0.55. Or perhaps its effect size is only 0.45 but the new drug produces substantially fewer (or less severe) side effects. As can be seen, the selection of a minimum effect size is a product of the researcher's knowledge of related research and good judgment.

In human clinical research, the researcher determines the smallest effect size that would be clinically important. If the results are statistically significant due to a large sample size, but the clinical effect in the human population is negligible, then the results are not clinically significant. This is a different standard than for statistical significance.

As an example, consider that a medical researcher is studying sepsis caused by non-MRSA *Staphylococcus aureus*. The standard drug used produces a survival rate of 60%. A new drug produces a survival rate of 62% and in a sample of 2,204 subjects the effect sizes are 0.77 and 0.79 respectively (rounded). Generally, the new drug will be much more expensive. Is a 2% change in the outcome worth millions of dollars a year more in treatment costs? If the treatment costs are the same, are the side effects different? What effect size would the researcher demand in this type of drug study if either the cost of the new drug were much higher or if it produced unpleasant or dangerous side effects? These are the kinds of questions that must be considered when the researcher selects a minimum effect size.

Once the requisite effect size has been determined, the researcher simply sets the effect size in the calculator to that minimal effect size and the calculator determines the sample size needed to detect that effect size. If an effect exists but the effect is less than the minimal effect size of interest, it will not achieve significance. If there is an effect at or larger than the minimal effect size of interest, the result will be significant. In this way, the researcher is able to plan a pilot study that will not only assist with pretesting instruments and data collection procedures, but will also improve the likelihood that the full study will be worth performing.

Significance level

The significance level, also called the *P*-level, of a study is typically set by scientific convention. For example, in most social science studies the significance level should be 0.05 or less. In some drug studies, the *P*-level must be much lower than 0.05 because of governmental review requirements for effectiveness and safety. Significance represents the likelihood of a Type I error. That is, it is the likelihood that the researcher will falsely claim a signifi-

Svrha više razine značajnosti u probnom istraživanju jest izbjegavanje odbacivanja onoga što bi inače moglo biti obećavajući ishod istraživanja temeljem probnog istraživanja u kojem nije nađen učinak liječenja. Zbog trenutne sklonosti urednika da objavljuju članke o probnim istraživanjima, čitatelji bi uvijek trebali imati na umu činjenicu da se istraživanja koja o učinaku izvještavaju na razini P < 0,10 ili višoj ne bi trebala primjenjivati na bolesničkoj populaciji, ili bi se na ljudskoj populaciji trebala primjenjivati uz najviši nadzor i oprez. Kod takvih je istraživanja vjerojatno da se učinci nađeni u uzorku jako razlikuju od onih u populaciji.

Također treba naglasiti, da kada istraživač objavi članak o probnom istraživanju uz primjenu povećane α-razine, veličina uzorka može biti dosta manja kako bi se sačuvala značajnost na istoj razini statističke snage i veličine učinka. Na slici 5. se vidi da je kod snage 0,80 i veličine učinka od 0,35, potreban uzorak od samo 74 ispitanika u svakoj skupini kako bi se sačuvala „značajnost" kad je *P*-razina postavljena na 0,20. S *P*-razinom od 0,05 u istom je istraživanju potrebno 129 ispitanika u svakoj skupini kako bi se postigla značajnost (vidi sliku 4.).

Manje je vjerojatno da će se otkriti pogreška tipa II. nego ona tipa I. Razlog tome je to što kada se dogodi pogreška tipa II., izvodi se zaključak da učinak ne postoji. Stoga se smjer istraživanja može napustiti. S pogreškom tipa I. vrlo je vjerojatno da će drugi istraživači testirati učinak o kojem se izvještavalo. Kada velik broj njih ne uspije dobiti učinak, priznat će se orginalna pogreška tipa I. Dakle, vjerojatnost pogreške tipa I. jednaka je razini značajnosti istraživanja. Nakon toga dolazi vjerojatnost pogreške tipa II. Ta se vjerojatnost računa kao 1-β. Budući da se statistička snaga vrlo često postavlja na 0,80, uobičajena vjerojatnost pogreške tipa II. je 1–0,80 ili 0,20. Dakle, dok je uobičajena vjerojatnost pogreške tipa I. 5%, tipična vjerojatnost pogreške tipa II. je 20%.

Statistički testovi

Sekundarni čimbenik koji utječe na statističku snagu je statistički test koji se rabi. Svaki test ima svoju razinu snage. Parametrijski testovi su sami po sebi jači od neparametrijskih, no to vrijedi samo ako se ispravno provode. Neparametrijski testovi su sami po sebi manje snažni od parametrijskih, no to vrijedi samo ako podaci i metoda istraživanja primijenjeni za dobivanje podataka podupiru upotrebu parametrijskih testova.

Parametrijski testovi povezani su s nizom pretpostavka o podacima. Kada se te pretpostavke prekrše, parametrijski testovi postaju nestabilni te mogu dati rezultate koji navode na krivi trag. Pretpostavke parametrijskih testova sadržavaju najčešće slijedeće: intervalna ili odnosna (omjerna) razina mjerenja barem zavisne varijable, slučajno raspoređivanje ispitanika u skupine, slučajno uzorkovanje ciljne populacije, jednake varijance među skupinama za

cant effect has been found when there is no effect in the population (Table 1).

With a very small sample size or a sample that poorly represents the population, there is always a high probability that no effect will be found, or conversely, that any effect found in the sample will not exist in the full population. Therefore, when performing pilot studies with small sample sizes, it is common for a researcher to set the significance level higher that usual in order to compensate for the small sample size. Thus, when the conventional significance level is P < 0.05, a pilot study might use a *P*-level of 0.10 or even 0.20. The purpose of the higher significance level in a pilot study is to avoid abandoning what might otherwise be a promising line of research on the basis of a pilot study that finds no effect for the treatment. Given the current tendency of editors to publish reports of pilot studies, readers should always keep in mind that studies reporting an effect at the P < 0.10 or higher levels should not be applied to patient populations, or should be applied to human populations only with the utmost oversight and care. Such studies are likely to result in population effects very different from the effects seen in the study sample.

It should also be noted that when the researcher publishes a report of a pilot study using an inflated alpha level, the sample size may be quite a bit smaller to obtain significance at the same power level and effect size. Note in Figure 5 that at a power of 0.80 and an effect size of 0.35, a sample of only 74 in each group was needed to obtain "significance" when the *P*-level was set to 0.20. With a p-level of 0.05, the same study requires a sample size of 129 in each group to achieve significance (see Figure 4).

A Type II error is less likely to be discovered than a Type I error. This is because when a Type II error is made, the conclusion is that there is no effect. Therefore, the line of research may be abandoned. With a Type I error, it is quite likely that other researchers will test the effect reported. When a number of them fail to find an effect, the original Type I error will be recognized. As noted, the probability of a Type I error is equal to the significance level of the study. What then, is the probability of a Type II error? That probability is calculated as 1-β. Since power is most often set at 0.80, the usual probability of a Type II error is 1–0.80 or 0.20. Thus, while there is usually only a 5% chance of a Type I error, there is typically a 20% probability of a Type II error.

The statistic

The secondary factor that affects power is the statistic used. Each statistic has an associated power level. Parametric statistics are inherently more powerful than non-parametric statistics, but this is true only when they are used correctly. Non-parametric statistics are inherently less powerful than parametric statistics, but that is true

zavisnu varijablu i ostale slične pretpostavke. Te se pretpostavke temelje na činjenici da se parametrijski testovi obično zasnivaju na metodi najmanjih kvadrata (linearna regresija), koja rabi srednju vrijednost kao temelj za računanje. U slučaju kada srednja vrijednost nije prikladna mjera centralne tendencije podataka, treba posegnuti za neparametrijskim testovima (ili testovima neovisnim o distribuciji podataka) za testiranje hipoteza. Neparametrijski testovi obično rabe medijan ili raspon podataka kao temelj svojih kalkulacija. Oni stoga imaju puno manje pretpostavaka nego parametrijski testovi.

Kada istraživač neprikladno upotrebljava parametrijske testove kako bi ispitao podatke koji nisu prilagođeni parametrijskoj statistici, tada se snaga rezultata dovodi u pitanje. Autorica je osobno vidjela niz slučajeva u kojima parametrijski testovi primijenjeni na podacima koji se odnose na skupu s ordinalnim varijablama (engl. *ordinal data*) nisu uspjeli naći značajan učinak, za razliku od neparametrijskih koji su ga pronašli. Moguće je također i suprotno. Ispravno primijenjen parametrijski test, budući da je snažniji, pronašao je značajan učinak liječenja koje analogni neparametrijski test nije pronašao. Kako bi snaga u istraživanju bila primjerena, bitno je da istraživač rabi statističke testove sukladno podacima za testiranje hipoteze.

Zaključak

Snaga je primarno funkcija veličine uzorka, veličine učinka i razine značajnosti (engl. *alpha-level, p-level*) i sekundarno primijenjenog statističkog testa za ispitivanje razlika između uzoraka. Čimbenik kojim istraživači najspremnije i najlakše barataju jest veličina uzorka. Primarna funkcija analize snage testa jest određivanje veličine uzorka potrebne za postizanje statističke značajnosti u istraživanju. Međutim, snaga se rabi i u probnim istraživanjima, kako bi se utvrdili učinci liječenja koji su preslabi da bi bili vrijedni istraživanja te da se odrede idealne razine značajnosti koje će se primijeniti u glavnom istraživanju. Postoji niz kalkulacija analize snage koje su dostupne na internetu i uporaba istih može biti korisna pomoć istraživačima u planiranju istraživanja. Trebalo bi im ući u naviku utvrditi potrebnu veličinu uzorka prije početka istraživanja. Može se naići na niz problema kod tumačenja rezultata ispitivanja ako istraživač ne shvati statističku snagu te način na koji se ona postiže. To uključuje i krivo tumačenje razultata zbog ili vrlo niske ili vrlo visoke snage te zbog neprikladnog odabira statističkog testa za testiranje hipoteze. Međutim, kada je snaga odgovarajuća i kada se ispravno rabe statistički testovi, tada se vjerojatnost ispravnog zaključka znatno povećava.

only if the data and research methods used to acquire the data support the use of parametric statistics.

Parametric statistics are associated with a number of assumptions about the data. When those assumptions are violated, the parametric statistics become unstable and may provide misleading results. Assumptions of parametric statistics most commonly include the following: interval or ratio level of measurement of at least the dependent variable, random assignment of subjects to study group, random sampling from the population of interest, equal variances among the study groups for the dependent variable, and other related assumptions. These assumptions are based on the fact that parametric statistics are usually founded in the least squares formula, which uses the mean as the basis for calculation. When the mean is not an appropriate measure of central tendency for the data, non-parametric (or distribution-free) statistics should be used to test the hypotheses. Non-parametric statistics usually use the median or rank order of the data as the basis of their calculation. They therefore have far fewer assumptions than parametric statistics.

When the researcher inappropriately uses parametric statistics to test data which are not appropriate for parametric statistics, the power of the results is called into question. The author has personally seen a number of cases in which parametric statistics used on ordinal data failed to find a significant effect but the non-parametric statistic did find a significant effect. The converse is also true. An appropriately applied parametric statistic, being more powerful, found a significant treatment effect that the analogous non-parametric statistic did not find. For power to be adequate in a study, it is essential that the researchers use statistics appropriate to the data for hypothesis testing.

Conclusions

Power is primarily a function of sample size, effect size and alpha-level, and secondarily of the statistic used to test sample differences. The factor most readily manipulated by the researcher is the sample size. Power analysis has as its primary function the determination of the sample size necessary to achieve statistical significance in a study. However, power can also be used in pilot tests to identify treatment effects too weak to be worth further pursuit, and to identify the ideal significance level to be used in the main study. There are a number of power analysis calculators available on the Internet and the use of these calculators can provide a useful tool to researchers planning studies. They should always be used to identify the necessary sample size prior to beginning a study. A number of problems with interpretation of research results can be encountered if the researcher does not understand statistical power and how it is achieved. These

include wrong interpretation of results due to either very low or very high power, and to inappropriate selection of a statistic to test the hypotheses. However, when power is adequate and the statistics are appropriately applied in hypothesis testing, the likelihood of correct conclusions is greatly improved.

Adresa za dopisivanje:

Mary L. McHugh, PhD, RN, BC
Dean and Professor
University of Indianapolis School of Nursing
1400 East Hanna Avenue
Indianapolis, Indiana 46227
U.S.A.
e-pošta: mchughm@uindy.edu
tel: +1 317 788-3206

Corresponding author:

Mary L. McHugh, PhD, RN, BC
Dean and Professor
University of Indianapolis School of Nursing
1400 East Hanna Avenue
Indianapolis, Indiana 46227
U.S.A.
e-mail: mchughm@uindy.edu
phone: +317 788-3206

Literatura / References

1. *Cohen J, Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, New Jersey: Lawrence Erlebaum Associates; 1998.*
2. *Lenth RV. Java Applets for Power and Sample Size [Computer software]; 2006 Downloaded on June 10, 2008, from http://www.stat.uiowa.edu/~rlenth/Power/index.html*
3. *Becker L, Effect size. Downloaded on June 3 from http://web.uccs.edu/lbecker/Psy590/es.htm.*