Strokovni članek ■

Nekaj malega o računanju velikosti vzorca

Janez Stare

Izvleček. V članku predstavim računanje velikosti vzorcev za nekatere preproste primere, ki pa jih v raziskovalni praksi pogosto srečamo.

Basics of Sample Size Calculations

Abstract. Calculation of the sample size for some most commonly used statistics is presented.

■ **Infor Med Slov**: 2007; 12(2): 29-33

Institucija avtorja: Inštitut za biomedicinsko informatiko, Medicinska fakulteta, Univerza v Ljubljani.

Kontaktna oseba: Janez Stare, Inštitut za biomedicinsko informatiko, Medicinska fakulteta, Univerza v Ljubljani, Vrazov trg 2, 1000 Ljubljana. email: janez.stare@mf.uni-lj.si.

Uvod

Da ima velikost vzorca nekaj besede pri tem, kako značilni bodo rezultati statistične analize podatkov, je skoraj splošno znano dejstvo. Za kakšen vpliv gre pa ve že malokdo. Vendar vse več strokovnih in znanstvenih medicinskih revij zahteva izračun potrebne velikosti vzorca oz., obrnjeno, izračun moči testa. Tudi na našem inštitutu čutimo porast takšnih zahtev, zaenkrat žal prepogosto šele po tem, ko je zbiranje podatkov že zaključeno.

Ko govorimo o potrebni velikosti vzorca, mislimo na določen namen. Najpreprostejši primer je ocenjevanje povprečja. Takrat želimo, da je vzorec dovolj velik, da bomo povprečje ocenili z določeno natančnostjo. Bolj pogosto pa nas zanima statistična značilnost. Recimo, želimo dovolj velika vzorca, da bo razlika med skupinama statistično značilna. Torej, če je med populacijama določena razlika, želimo, da jo naš test zazna. Ali jo bo vedno zaznal? Ne, lahko imamo smolo in izberemo vzorca, ki sta si zelo podobna, čeprav si populaciji nista. Na primer, vemo, da imajo starejši višji krvni pritisk kot mlajši, a če izberemo slučajna vzorca starejših in mlajših, se lahko zgodi, da sta si vzorčni povprečji blizu ali celo v obrnjenem razmerju kot v populaciji. Torej lahko govorimo le o *verjetnosti*, da bo naš test zaznal razliko, če ta obstaja. Tej verjetnosti pravimo *moč testa* in zanjo želimo, da je čim večja. To pa je odvisno od velikosti vzorca. Pogosto izbrana, in še sprejemljiva, moč testa je 0,8 oz. 80%. To pomeni, da bomo obstoječo razliko statistično zaznali v 80% primerov. In NE zaznali v 20% primerov!

Preden začnemo s konkretnimi primeri, naj spomnim še na tole: tudi če med populacijama ni razlik, je test lahko značilen. Kolikokrat? No, to res sodi v osnove statistike pa vseeno povem - v 5% primerov. Če smo mejo statistične značilnosti postavili pri 5% seveda, sicer pa pač ustrezno drugače.

V naslednjih treh razdelkih si bomo pobliže ogledali nekaj najpogostejših primerov izračunavanja velikosti vzorca. Za to potrebujemo nekaj statističnega znanja, ki ga v članku sicer ponovim, a ne razlagam. Ukvarjali se bomo samo z numeričnimi spremenljivkami, o atributivnih pa morda kdaj drugič.

Ocenjevanje povprečja

Kadar ocenjujemo povprečje neke numerične spremenljivke želimo predvsem čim večjo natančnost. To pomeni, da želimo, da je naša ocena z določeno verjetnostjo največ za d oddaljena od pravega povprečja. Da ne bi kompliciral, se dogovorimo, da bomo v vsakem razdelku izbrali neko konkretno verjetnost, v tem vzemimo 95%. To je sicer veliko, v praksi se ponavadi zadovoljimo z manjšo verjetnostjo. Zavedati se moramo, da bodo naše ocene v 5% primerov vendarle za več kot d oddaljene od pravega povprečja.

Najprej se spomnimo, da se povprečja porazdeljujejo normalno. To je popolnoma res, če je spremenljivka v populaciji porazdeljena normalno, a dovolj dobro res tudi, če ni. Naj bo pravo povprečje μ, prava standardna deviacija pa σ. Standardna deviacija porazdelitve povprečij, ki ji ponavadi rečemo standardna napaka, je $\sigma/\sqrt{n}$. Ker za normalno porazdeljeno spremenljivko velja, da je 95% njenih vrednosti znotraj intervala, ki seže 1,96 standardne deviacije levo in desno od povprečja in ker želimo, da je v 95% primerov vzorčno povprečje za manj kot d oddaljeno od povprečja populacije, mora torej biti $d=1{,}96\sigma/\sqrt{n}$ in odtod

$$n = \frac{1{,}96^2 \sigma^2}{d^2}.$$

Pozoren bralec seveda ne bo spregledal, da v gornji formuli nastopa populacijska σ, ki je v praksi praviloma ne bomo poznali. V formulo torej postavimo neko oceno in se pri tem zavedamo, da bo naš izračun pravilen le, če se pri oceni nismo zmotili. Če smo varianco precenili, ne bo hudega, saj bo zahtevani vzorec pač prevelik (toliko bolje!),

s podcenjevanjem variance pa seveda ne kaže računati velikosti vzorcev.

Primer: oceniti želimo povprečno vrednost sistoličnega krvnega pritiska v populaciji Slovencev. Pri tem hočemo, da se naša vzorčna ocena s 95% verjetnostjo ne bo razlikovala od prave vrednosti za več kot 2 mmHg. Potemtakem je $d=2$. Če privzamemo, da je $\sigma=15$, dobimo kot potrebno velikost vzorca $n=216$. Če bi bili zadovoljni z natančnostjo na 5 mmHg, pa bi potrebovali vsega 35 ljudi. Seveda izračun lahko tudi obrnemo in vprašamo, kako natančno bi ocenili povprečje na primer pri $n=50$. Dobili bi $d=4{,}2$. Naj še enkrat poudarim, da so ti rezultati zelo odvisni od tega, kaj smo privzeli za σ.

Testiranje hipoteze o povprečju populacije

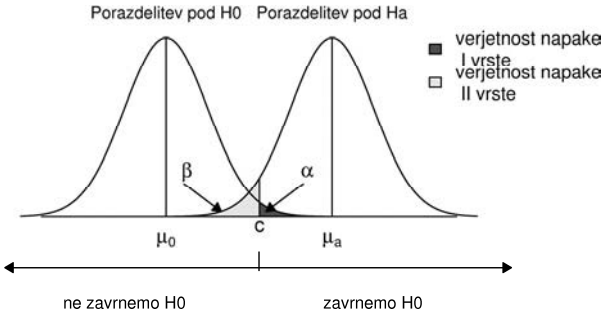
V prvem razdelku smo se ukvarjali z ocenjevanjem populacijskega povprečja. Šlo nam je za natančnost, ničesar nismo testirali. V nadaljevanju si bomo podrobneje pogledali, kako izračunamo potrebno velikost vzorca pri dveh najpogosteje uporabljanih testih: testiranju povprečja in testiranju razlike med dvema vzorcema. Recimo, da želimo preveriti hipotezo

$$H_0{:}\mu=\mu_0,$$

nasprotna hipoteza pa je

$$H_a{:}\mu>\mu_0.$$

Označimo dejansko povprečje v populaciji z μ_a. Napaka prve vrste (verjetnost, da zavrnemo pravilno ničelno hipotezo) naj bo α, napaka druge vrste (verjetnost, da sprejmemo napačno ničelno hipotezo) pa β, moč torej $1-\beta$. Privzemimo, da je standardna deviacija enaka σ, tako pod ničelno kot alternativno hipotezo. Izberimo točko c takole (glej Sliko 1):



Slika 1: Testiranje hipoteze o povprečju populacije z enostranskim testom.

Če velja ničelna hipoteza, naj bo desno od nje α (npr. 5%) vseh vrednosti pod vzorčno porazdelitvijo povprečij, če pa je pravilna alternativna hipoteza, naj β (npr. 10%) vseh vrednosti leži levo od c .

Potem je

$$c=\mu_0+z_\alpha\frac{\sigma}{\sqrt{n}}$$

in tudi

$$c=\mu_a-z_\beta\frac{\sigma}{\sqrt{n}}.$$

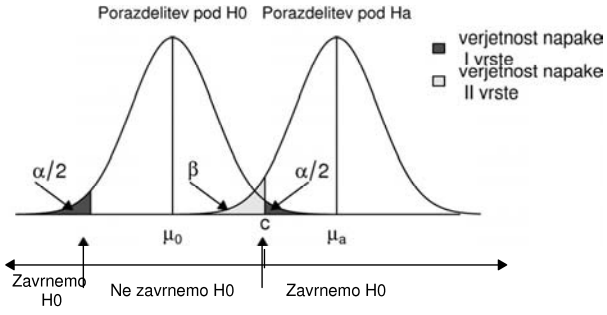
Izraza izenačimo in razrešimo na n:

$$n=\frac{\sigma^2(z_\alpha+z_\beta)}{(\mu_a-\mu_0)^2}.$$

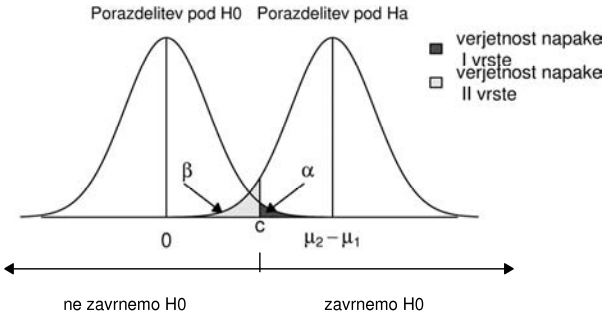
Če je alternativna hipoteza

$$H_a{:}\mu_a\neq\mu_0,$$

moramo z_α v zgornjih formulah nadomestiti z $z_{\alpha/2}$. Slika 2 ilustrira takšno situacijo.



Slika 2: Testiranje hipoteze o povprečju populacije z dvostranskim testom.



Slika 3: Dva vzorca, enostranski test za H_0:$\mu_1=\mu_2$ proti alternativni hipotezi, da je H_0:$\mu_2>\mu_1$.

Ocenjevanje razlike dveh povprečij

Problem je enak kot pri ocenjevanju povprečja populacije, le da je standardna deviacija porazdelitve razlik povprečij enaka

$$\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\mu_1^2}{n_1}+\frac{\mu_2^2}{n_2}}.$$

Če ponovno d predstavlja natančnost, je

$$d=z_{\alpha/2}\sqrt{\frac{\mu_1^2}{n_1}+\frac{\mu_2^2}{n_2}},$$

kar je izraz, iz katerega lahko izračunamo npr. n_2, če določimo n_1. Ponavadi se odločimo za razmerje med n_1 in n_2, torej $n_2=kn_1$, od koder potem sledi

$$n_1=\frac{z_{\alpha/2}^2(k\sigma_1^2+\sigma_2^2)}{kd^2}. \qquad (1)$$

Izraz se še nekoliko poenostavi, če lahko privzamemo enakost varianc in enako velikost obeh vzorcev.

Testiranje razlike povprečij dveh neodvisnih vzorcev

Recimo, da želimo preveriti hipotezo

$$H_0:\mu_1=\mu_2,$$

nasprotna hipoteza pa je

$$H_a:\mu_2>\mu_1.$$

Označimo razliko med populacijskima povprečjema z $\delta=\mu_1-\mu_2$. Potem lahko hipotezi zapišemo takole:

$$H_0:\delta=0,\ H_a:\delta>0.$$

Situacija je enaka kot pri testiranju enega vzorca, le da gre tukaj za razlike povprečij. Spet naj bo c točka, za katero velja (glej Sliko 3):

Če velja ničelna hipoteza, naj bo desno od nje α (npr. 5%) vseh vrednosti pod vzorčno porazdelitvijo razlik, če pa je pravilna alternativna hipoteza, naj β (npr. 10%) vseh vrednosti leži levo od c. Zdaj je

$$c=0+z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

in

$$c = \delta + z_\beta \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Od tu lahko izrazimo n_2 kot funkcijo n_1 (pa še σ_1, σ_2, z_α in z_β). Računanje si nekoliko olajšamo, če zopet postavimo $n_2=kn_1$, kar da

$$n_1 = \frac{(z_\alpha + z_\beta)^2 (k\sigma_1^2 + \sigma_2^2)}{k\delta^2}.$$

Če je alternativna hipoteza

$$H_a: \mu_1 \neq \mu_2,$$

moramo, tako kot prej, z_α v zgornjih formulah nadomestiti z $z_{\alpha/2}$.