Študijsko gradivo ■

A/B testiranje – najpreprostejša statistika za učinkovitejše spletne strani in druge poskuse

Gaj Vidmar

Izvleček. Gradivo razlaga eksperimentalni načrt in statistični test, ki zahtevata zgolj osnovnošolsko znanje matematike, a sta kljub temu uporabna pri strokovnem odločanju in raziskovalnem delu. Gre za primerjavo števila izidov ene od dveh možnih vrst med dvema enako velikima skupinama, v kateri so opazovane enote razvrščene naključno. Predstavljeni postopek je primeren za primerjavo dveh različic spletne strani oziroma uporabniškega vmesnika, s čimer se pogosto soočajo razvijalci spletnih strani oziroma aplikacij (tudi v medicinski informatiki). Na spletnem področju se je postopka prijelo ime A/B testiranje, že dolgo pa je znan v statistični literaturi in uporaben za najrazličnejše raziskave – od preučevanja vedenja živali do biomedicinskih kliničnih študij.

A/B Testing – the Simplest Statistics for More Effective Websites and Other Experiments

Abstract. The tutorial explains an experimental design and a statistical test requiring only elementary-school mathematical knowledge that are nevertheless useful in practical decision-making and research work. The task is to compare the number of outcomes of one of the possible kinds between to equally-sized groups to which the units under observation had been randomly assigned. The presented procedure is useful for comparing two alternative website or user interface designs, which is often encountered by web/application developers (including those in medical informatics). In the internet community the procedure has become known as A/B testing, but it has long been known in statistical literature and used in various research fields ranging from animal behaviour studies to biomedical clinical trials.

■ **Infor Med Slov**: 2013; 18(1-2): 37-42

Instituciji avtorja / Author's institutions: Univerzitetni rehabilitacijski inštitut Republike Slovenije – Soča; Univerza v Ljubljani, Medicinska fakulteta, Inštitut za biostatistiko in medicinsko informatiko.

Kontaktna oseba / Contact person: Gaj Vidmar, URI – Soča, Linhartova 51, SI-1000 Ljubljana. e-pošta / e-mail: gaj.vidmar@ir-rs.si.

Prejeto / Received: 19.12.2013. Sprejeto / Accepted: 23.12.2012. Recenzenta / Reviewers: doc. dr. Rok Blagus in doc. dr. Lara Lusa.

Uvod

Kdor načrtuje oziroma razvija spletne strani oziroma aplikacije in si želi pritegniti čim več obiskovalcev oziroma uporabnikov, se pogosto odloča med različnimi oblikovalskimi možnostmi. Če se le da, je k tej nalogi smiselno pristopiti empirično – s poskusom na reprezentativnem vzorcu. Če sta možnosti dve in če lahko vključene enote (v tem primeru obstoječe ali potencialne obiskovalce spletne strani oziroma uporabnike aplikacije) po naključju razdelimo med eksperimentalna pogoja, je priporočljiv postopek, ki se ga je v zadnjih letih prijelo ime A/B testiranje.[1-3] V nadaljevanju sta opisana njegov tehnični in računski del, sledi razlaga oziroma pojasnilo statističnega dela, zaključek pa obravnava uporabnost tovrstnega testiranja v biomedicinskih raziskavah.

Postopek

Če gre za spletno stran, pripravimo dve različici in nato programsko (z generatorjem psevdonaključnih števil) za vsakega obiskovalca izberemo, katero mu bomo pokazali. Različici lahko določa

- oblikovanje (npr. velikosti pisave ali oblika gumba) ali
- vsebina oziroma besedilo (npr. "Hočem poskusiti!" namesto "Kliknite tu za prenos brezplačne poskusne verzije").

Kombinacijam, tj. razlikam v več razsežnostih hkrati oziroma večjemu številu razlik, se raje izognemo, sicer bi morali izvesti večfaktorski poskus in podatke analizirati na mnogo bolj zapleten način (z eno od oblik analize variance).

Seveda moramo na ustrezen način meriti učinkovitost – npr. beležiti,

- ali je obiskovalec ostal na strani (tj. kliknil na katero od predvidenih povezav) ali jo zapustil;
- ali je opravil naslednji korak nakupnega postopka ali ne;
- ali je odgovoril na zastavljeno vprašanje ali ne ipd.

Če poimenujemo različici A in B, izid pa pozitiven ali negativen, zbrane podatke (v primeru, da sta bili skupini enako veliki) povzamemo v razpredelnici z dvema vrsticama in dvema stolpcema (tabela 1).

Tabela 1 Splošna oblika podatkov, zbranih z A/B testiranjem.

	izid		število
možnost	pozitiven	negativen	udeležencev
A	a	$x-a$	x
B	b	$x-b$	x

Oznake v tabeli 1 so izbrane tako, da poudarjajo, da za računski del potrebujemo le obe števili pozitivnih izidov (pod pogojem, da je število udeležencev v obeh skupinah enako). Statistični test opravimo v treh korakih, za katere zadošča računalo, "vgrajeno" v vsak sodoben telefon (in v vsak računalnik in v Google):

1. Seštejemo število pozitivnih izidov in vsoto označimo z N.
   $N = a + b$
2. Izračunamo razliko med številoma pozitivnih izidov in jo delimo z 2; rezultat označimo z D.
   $D = \frac{a-b}{2}$
3. Različici se statistično značilno razlikujeta glede izida, če je $D \times D$ več kot N.
   $D^2 > N \Rightarrow p < 0{,}05$

Primer

Denimo, da smo preizkus dveh različic spletne strani izvedli s štiridesetimi obiskovalci – dvajsetim smo prikazali statično obliko oglasa in dvajsetim animirano. Na statični oglas jih je kliknila četrtina

($a = 5$), na dinamičnega pa polovica ($b = 10$). Je razlika večja, kot je še smiselno pričakovati po naključju? (Tole sicer ni čisto korektna interpretacija statistične značilnosti, ampak približno v redu je – v skladu s samim statističnim testom, ki je, kot bomo kmalu videli, tudi približen.) Izračunajmo!

1. $N = 15$

2. $D = -2{,}5$

3. $D^2 = 6{,}25 < N \Rightarrow p > 5\%$

Razlika torej ni statistično značilna (na ravni tveganja 5%) – poskus nas ni prepričal, da bi bila ena različica oglasa učinkovitejša od druge. Kaj pa, če bi ga bili izvedli z večjim vzorcem, pri čemer bi bila deleža obiskovalcev, ki kliknejo na oglas, enaka? Poglejmo dva primera:

- če bi bilo obiskovalcev dvakrat toliko kot v začetnem primeru, torej po 40 v vsaki skupini, bi dobili $a = 10$, $b = 20$, $N = 30$, $D = -5$ in $D^2 = 25$, kar je še vedno manj kot N, torej razlika (pri dopustnem petodstotnem tveganju) še vedno ne bi bila statistično značilna;

- če bi bilo obiskovalcev trikrat toliko kot v začetnem primeru, torej po 60 v vsaki skupini, pa bi dobili $a = 15$, $b = 30$, $N = 45$, $D = -7{,}5$ in $D^2 = 56{,}25$, kar že več kot N, torej bi lahko sklenili, da je animirani oglas statistično značilno učinkovitejši od statičnega (s potrebnimi statističnimi pristavki: za vzorčeno populacijo na dani spletni strani pri dopustnem petodstotnem tveganju).

Hm, porečete, v vseh treh primerih je bila "stopnja preklikanja" (angl. *click-through rate*) enaka, a sklepi si (na videz) nasprotujejo! – Kaj je potemtakem res: je med oglasoma razlika ali je ni? – Statistika (oziroma znanost na sploh) seveda na tako vprašanje oziroma na tak način ne odgovarja in nasprotja seveda ni. K osnovam statističnega preizkušanja domnev (tj. testiranja hipotez) namreč sodi, da z večanjem vzorca proti neskončnosti vsaka (tudi še tako majhna) razlika "postane" statistično značilna (pri še tako nizki stopnji tveganja). Poleg samega statističnega sklepa je torej pomembno tudi vsebinsko vprašanje, tj. ali polovičen uspeh v primerjavi s četrtinskim predstavlja razliko, ki je pomembna v praksi? V danem primeru najbrž lahko brez oklevanja odgovorimo pritrdilno, saj animirana verzija oglasa najbrž ni povezana z znatnimi dodatnimi stroški (ali jih sploh ne prinaša), drugih "neželenih učinkov" (že vlečemo vzporednice s kliničnimi poskusi) pa prav tako nima.

S tem razmišljanjem smo se dotaknili pomembnega in obsežnega področja statistike, ki se mu reče moč testa in izračunavanje velikosti vzorca. Zainteresirani bralec si lahko več o njem prebere v kakšnem obsežnejšem sodobnem učbeniku (bio)statistike[4-5] ali preuči katero od knjig, ki so deloma[6] oziroma v celoti posvečene temu področju,[7-9] za sodobne študente, ki so jim od knjig ljubše dinamične in interaktivne spletne stvari, pa je priporočljiv vodnik iz projekta WISE.[10] (Slednji ima – kar je pri spletnih učnih pripomočkih prej izjema kot pravilo – tudi statistično korektno preverjeno učinkovitost[11]). V nadaljevanju si bomo za pokušino ogledali le zelo preprost poseben primer v zvezi z našim poskusom.

Ker smo videli, da je 80 udeležencev poskusa premalo, 120 pa že dovolj za dokazati statistično značilno razliko med stopnjama preklikanja 1/4 in 1/2, se torej vprašajmo, kolikšen bi bil najmanjši vzorec, ki bi že zadoščal?

- Veljati mora $D^2 = N$.

- Ker je $b = 2a$, je $N = 3a$ in $D^2 = \frac{a^2}{4}$.

- Ko obe strani enačbe množimo s 4 in delimo z a, ugotovimo, da je $a = 12$.

- Želena skupna velikost vzorca je torej $8a = 96$.

Ker smo se uspešno "matematično ogreli", nadaljujmo z izpeljavo testnega postopka –

pokažimo, kako deluje oziroma kakšna statistika se skriva za njim. Pred tem le še omenimo, da je A/B testiranje (spletna izvedba in analiza podatkov) podprto v specializiranih programskih orodjih. Prvi ga je nudil Google Website Optimizer, ki je kot samostojno orodje obstajal od leta 2008 do junija 2012,[12] odtlej pa je del njegovih zmožnosti vključen v orodje Google Analytics,[13] natančneje v storitev Google Analytics Content Experiments.[14] Vodilno komercialno orodje za A/B (in zapletenejše) testiranje različic spletnih strani je Optimizely.[15]

Izpeljava

Primerjamo dva deleža; ničelna domneva je, da sta enaka, torej da so pozitivni izidi enako porazdeljeni med skupini. Zato je primerno uporabiti χ^2 test prileganja:

$$\mathrm{X}^2 = \sum_{i=1}^{m} \frac{\left(f_{o_i} - f_{p_i}\right)^2}{f_{p_i}},$$

pri čemer je $m = 2$. Za opaženi frekvenci pozitivnih izidov že imamo oznaki: $f_{o_1} = a$ in $f_{o_2} = b$. Pričakovana frekvenca je za obe skupini enaka polovici vsote opaženih frekvenc, za to vsoto pa tudi že imamo oznako: $f_{p_1} = f_{p_2} = N/2$. Testna statistika je torej

$$\mathrm{X}^2 = \frac{\left(a - N/2\right)^2}{N/2} + \frac{\left(b - N/2\right)^2}{N/2}.$$

Števca obeh ulomkov sta enaka, in sicer sta oba enaka D^2. Velja namreč $\left(a - N/2\right)^2 = \left(2a/2 - (a+b)/2\right)^2 = \left((a-b)/2\right)^2$ in na enak način za drugi števec dobimo $\left((b-a)/2\right)^2$, kar pa je enako, saj je $(b-a)^2 = (a-b)^2$. Obrazec za testno statistiko se tako poenostavi v

$$\mathrm{X}^2 = \frac{4D^2}{N}.$$

Prišli smo do preprostega obrazca za izračun testne statistike, iz katerega bi (iz tabel ali s spletnim statističnim računalom ali s funkcijo v elektronski preglednici) dobili vrednost p. Toda obrazec lahko še poenostavimo, če vemo, da je kritična vrednost porazdelitve χ^2 za 1 prostostno stopnjo (za χ^2 test prileganja je namreč $df = m - 1$) pri stopnji tveganja 5% enaka 3,84. Če to zaokrožimo na 4 (s čimer bo test nekoliko "strožji" v smislu, da bo stopnja tveganja v resnici nekoliko nižja od 5%), dobimo poenostavljeno pravilo, da je razlika med skupinama statistično značilna, če je

$$\frac{4D^2}{N} > 4 \Leftrightarrow D^2 > N.$$

Dodajmo še, da je (ker je kritična vrednost χ^2 za $df = 1$ pri $\alpha = 0{,}01$ enaka 6,63) v primeru, da je D^2 vsaj dvakrat toliko kot N, razlika med skupinama statistično značilna na ravni tveganja, manjši kot 1%.

Širša uporabnost

Zadnjič smo spoznali "najpreprostejši statistični test"[16] za primerjavo dveh Poissonovo porazdeljenih števil dogodkov v randomiziranem poskusu z dvema enako velikima skupinama. Pot do testne statistike je bila nekoliko drugačna (trije ključni koraki so bili, da je razlika vzorčnih povprečij cenilka razlike med populacijskima povprečjema, da je varianca razlike dveh neodvisnih slučajnih spremenljivk vsota njunih varianc in da je varianca Poissonove slučajne spremenljivke enaka njenemu povprečju) in testna statistika je bila videti drugače:

$$z = \frac{a-b}{\sqrt{a+b}}.$$

Toda če izraz kvadriramo, dobimo obrazec, ki je identičen kot pri A/B testiranju:

$$\frac{(a-b)^2}{4(a+b)} \sim z^2 \Leftrightarrow \frac{D^2}{N} \sim \chi^2(df=1).$$

Definicija porazdelitve χ^2 z eno prostostno stopnjo je namreč, da je to porazdelitev kvadrirane vrednosti standardno normalno porazdeljene slučajne spremenljivke.

Skupaj s testom smo zadnjič navedli tudi primere kliničnih študij, v katerih so ga – oziroma bi ga bili lahko – uporabili. Omenili smo primerjavo učinkovitosti zdravila za srčno popuščanje s placebom,[17] metaanalizo študij revaskularizacije[17] in slovensko študijo s področja bolnišnične rehabilitacije.[18] Za razvedrilen konec jim tokrat dodajmo amatersko študijo vedenja oziroma prehranskih preferenc hrčka, iz katere si lahko na spletu polega številskih rezultatov ogledamo tudi videoposnetek.[19] Poskus je zaradi različnih vrst živil v resnici večfaktorski, predvsem pa bi bil moral avtor raziskave običajna in "organska" živila, med katerimi je hrček izbiral, po naključju postavljati na hrčkovo levo oziroma desno stran (ne pa, da so bila običajna vedno na hrčkovi desni in "organska" na levi); a ne glede na metodološke pomanjkljivosti si je udeleženec poskusa s hlačanjem, vohljanjem in grizljanjem prislužil že skoraj milijon spletnih ogledov …

Namesto zaključka

V prejšnji številki je avtor študijskega gradiva razkril svoj načrt za Nematematikove sprehode po matematiki in statistiki. Za zdaj načrtu sledi: "A/B sprehod" se vsebinsko navezuje na "Poissonovega" (preko "najpreprostejšega testa"), primerno pa je tudi, da dolgotrajnejšemu in zahtevnejšemu sprehodu sledi krajši in enostavnejši. Tokrat sicer gradiva ne spremlja Excelova preglednica, a je ena od prednosti predstavljenega postopka ravno to, da nam za izračun zadošča kakršnokoli računalo oziroma niti tega ne potrebujemo. Seveda bo avtor zelo vesel, če bo kakšen bralec sestavil in mu posredoval elektronski delovni zvezek, ki na prvem delovnem listu na podlagi opaženih frekvenc a in b izračuna vrednost p, na drugem delovnem listu pa na podlagi predvidenih deležev in želene stopnje tveganja izračuna potrebno velikost vzorca. Bo kdo poskusil? Bo morda kdo poročal o uspešnem primeru A/B testiranja v zdravstveni praksi (nemara v okviru razvoja storitev zdravja na daljavo)? V naprej hvala in prijeten sprehod!

Literatura

1. Cohen J: *Easy statistics for AdWords A/B testing, and hamsters*. http://blog.asmartbear.com/easy-statistics-for-adwords-ab-testing-and-hamsters.html, 2009.
2. Rocheleau J: *Guide to A/B testing with Google Website Optimizer*. http://www.hongkiat.com/blog/google-website-optimizer-ab-testing-guide/, 2011.
3. Mango A: *A simple approach to relevant A/B testing*. http://blog.lewispr.com/2013/09/a-simple-approach-to-relevant-ab-testing.html, 2013.
4. Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research* (4th ed.). Oxford 2002: Blackwell; sect. 4.6, p. 137-146. https://archive.org/details/StatisticalMethodsInMedicalResearch
5. Riffenburg, RH: *Statistics in medicine* (3rd ed.). London 2012: Academic Press; sec. 18.1-18.16, p. 365-388.
6. Ellis PD: The essential guide to effect sizes – statistical power, meta-analysis, and the interpretation of research results. Cambridge 2010: Cambridge University Press; partII, p. 45-85.
7. Murphy KR, Myors B: *Statistical Power Analysis – a simple and general model for traditional and modern hypothesis tests* (2nd ed.). London 2004: Lawrence Erlbaum.
8. Dattalo P: *Determining sample size – balancing power, precision, and practicality*. Oxford 2008: Oxford University Press.
9. Ryan TP: *Sample size determination and power*. New York 2013: John Wiley.
10. Claremont Graduate University, Web Interface for Statistics Education (WISE): *WISE power tutorial*. http://wise.cgu.edu/power/index.asp, 2012.
11. Aberson CL, Berger DE, Healy MR, and Romero VL. An interactive tutorial for teaching statistical power. *J Stat Educ* 2002; 10(3). http://www.amstat.org/publications/jse/v10n3/aberson.html

12. Google Website Optimizer. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2013: Wikimedia Foundation. http://en.wikipedia.org/wiki/Google_Website_Optimizer
13. Tzemah N: Helping to *create better websites: introducing content experiments*. http://analytics.blogspot.com/2012/06/helping-to-create-better-websites.html, 2012.
14. Google: *Overview of content experiments. Benefits of experiments*. https://support.google.com/analytics/answer/1745147, 2014.
15. Optimizely: *Optimizely – A-B testing software you'll actually use*. https://www.optimizely.com/, 2013.
16. Vidmar G: Poissonova porazdelitev – osnove, uporaba, nadgradnja. *Inf Med Slov* 2012; 17(2): 29-55.
17. Pocock SJ. The simplest statistical test: how to check for a difference between treatments. *BMJ* 2006; 332: 1256-1258.
18. Vidmar G: Primer uporabe najpreprostejšega statističnega testa: ali se zahtevnost rehabilitacije bolnišničnih pacientov povečuje? *Rehabilitacija* 2008; 7(2): 8-11.
19. Ken: *Hammy the hamster goes organic*. http://www.thecooksden.com/hamster, http://www.youtube.com/watch?v=8z8CWdRaQpw, 2009.