

A ali B: to je tu vprašanje

Jasna Urbančič

MaRS 2026

Uvod: kaj je A/B testiranje?

Pogosto imamo na voljo dve možnosti, **A** in **B**, in želimo razumeti, katera je boljša. *A/B testiranje* je okrajšava za preprost randomiziran kontroliran eksperiment, v katerem primerjamo več vzorcev (npr. A in B) ene same (vektorske) spremenljivke.

Ključni sestavni deli so: (i) vzorce naključno razdelimo v skupine, s čimer zagotovimo primerljivost; (ii) med skupinama spremenimo eno samo stvar; in (iii) opazujemo eno merljivo spremenljivko — na primer oceno na lestvici od 1 do 10. Odgovoriti želimo na dve vprašanji:

1. Ali je povprečna ocena višja za A ali za B?
2. Ali je razlika med njima **pomembna** (ali pa je le posledica naključja)?

Prvo vprašanje je preprosto: primerjamo povprečji. Pravo delo se skriva v drugem vprašanju, saj vsaka meritev nosi negotovost.

Vzorci in porazdelitve

Vsakemu vzorcu pripišemo skupino (A ali B) in izmerjeno oceno. Ocene izhajajo iz neke *verjetnostne porazdelitve*. Če narišemo histogram ocen za obe skupini, običajno dobimo dva prekrivajoča se zvonasta kupa. Kadar se kupa močno prekrivata, je razlika med A in B negotova; kadar sta jasno ločena, je razlika verjetno resnična. Prav to negotovost formalno opišemo z intervalom zaupanja.

Intervali zaupanja

Povprečje iz vzorca je le *ocena* oziroma približek pravega povprečja: če bi poskus ponovili, bi dobili nekoliko drugačno vrednost. Kako zelo povprečje niha, meri standardna napaka

$$SE = \frac{s}{\sqrt{n}}, \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2},$$

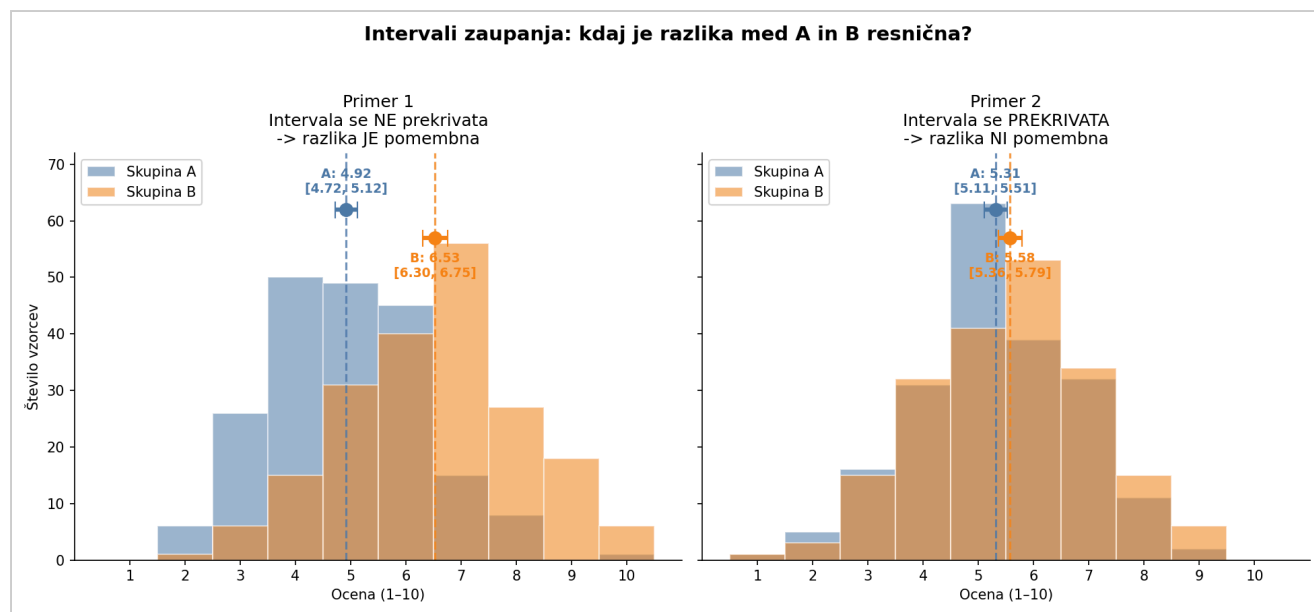
kjer je **s** standardni odklon ocen in **n** velikost vzorca. Večji vzorec pomeni manjšo standardno napako in s tem natančnejšo oceno.

Za velik vzorec iz približno normalne porazdelitve 95 % interval zaupanja za povprečje izračunamo kot

$$IZ_{95\%} = \bar{x} \pm 1,96 \cdot \frac{s}{\sqrt{n}}.$$

Faktor **1,96** izhaja iz normalne porazdelitve — 95 % vrednosti leži znotraj $\pm 1,96$ standardnega odklona od sredine. Interval razumemo takole: če bi poskus velikokrat ponovili, bi pravo povprečje padlo v ta razpon v približno 95 % primerov.

Uporaba pri A/B testu. Narišemo interval zaupanja za povprečje A in za povprečje B. Če se intervala *ne prekrivata*, je razlika pomembna; če se *prekrivata*, razlike ne moremo zanesljivo potrditi. Primer je prikazan na spodnji sliki. (Opomba: strožja različica gleda interval zaupanja *razlike* verjetnostnih porazdelitev $B - A$ in preveri, ali vsebuje ničlo.)

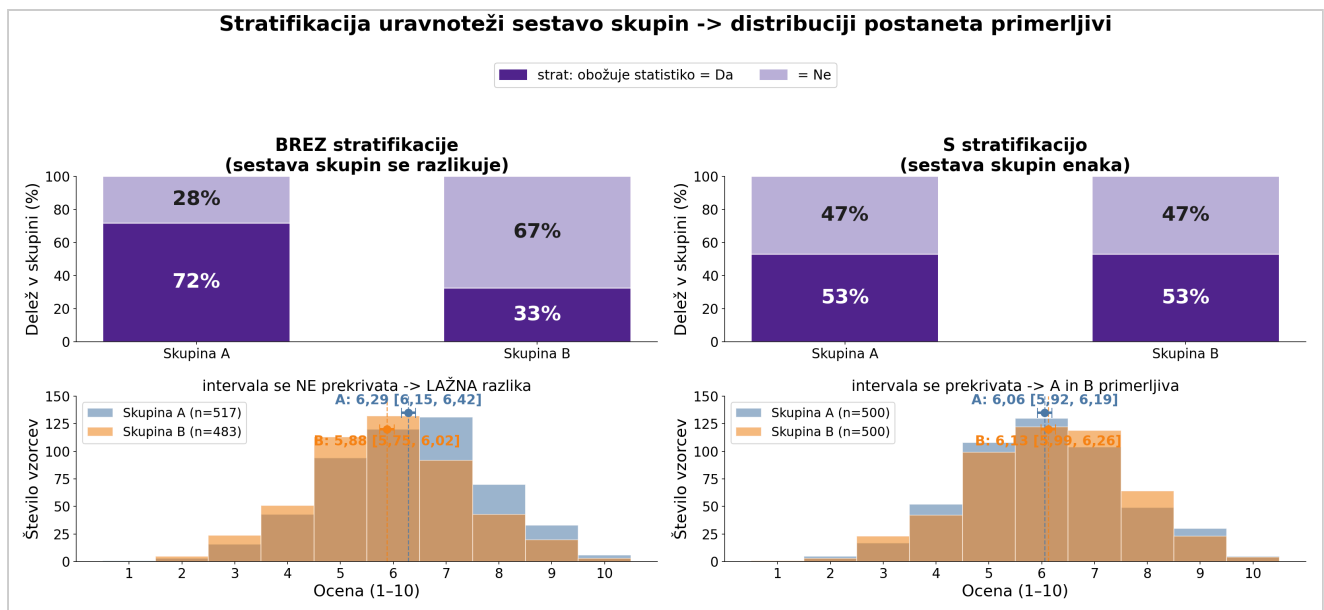


Dva primera A/B testa. *Levo:* povprečji sta jasno ločeni in intervala zaupanja se ne prekrivata — razlika je pomembna. *Desno:* povprečji sta blizu, intervala se prekrivata — čeprav ima B višje povprečje, razlike ne moremo potrditi.

Stratificirano vzorčenje

Naključna razdelitev v skupini je poštena le v povprečju. Pri manjših vzorcih (ali ob nesrečnem naključju) lahko ena skupina po sestavi močno odstopa od druge, na primer, če v skupino A pade veliko več ljudi z določeno lastnostjo, ki vpliva na oceno. Tedaj razlika v oceni ni posledica A proti B, temveč *različne sestave skupin*.

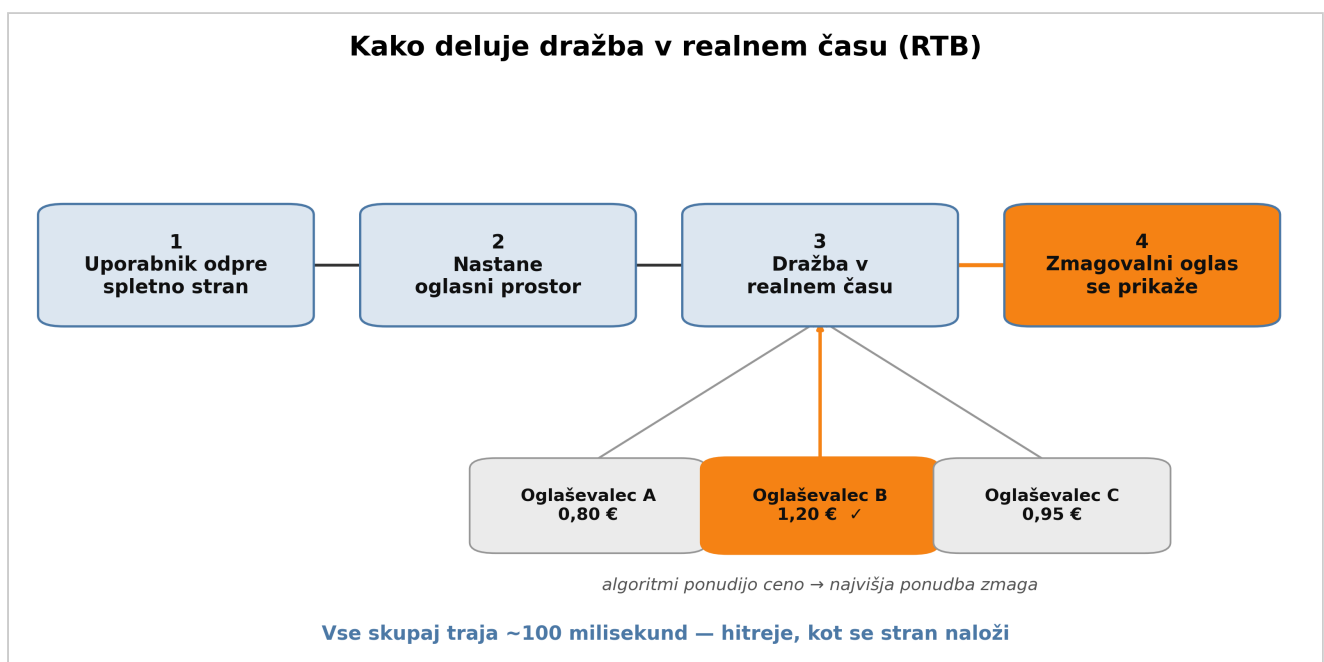
Rešitev je **stratificirano vzorčenje**: vzorce najprej razdelimo v *strate* — podskupine s podobnimi lastnostmi — in šele znotraj vsakega strata naključno razporedimo polovico v A in polovico v B. Tako sta skupini po sestavi skoraj enaki in primerjava postane poštena. Primer je prikazan na spodnji sliki.



Brez stratifikacije (levo) se sestavi skupin razlikujeta, kar ustvari lažno razliko v ocenah in ločena intervala zaupanja. S stratifikacijo (desno) sta sestavi enaki, distribuciji sovpadeta, intervala zaupanja pa se prekrivata in razlike med A in B ni.

A/B testiranje v spletnem oglaševanju

Na spletnih straneh del prostora ni namenjen vsebini, temveč **oglasom**; tako so strani za uporabnika brezplačne. Kateri oglas se prikaže, skoraj vedno določi **dražba v realnem času** (angl. *real-time bidding*, RTB): ko uporabnik odpre stran, nastane oglasni prostor, za katerega v delčku sekunde tekmuje več oglaševalcev. Najvišja ponudba zmaga, celoten postopek pa traja okrog 100 milisekund, saj morajo biti oglasi naloženi skupaj s spletno stranjo. Cene ne ponujajo ljudje, temveč **algoritmi**, ki ocenijo vrednost vsakega posameznega prikaza.



Shema dražbe v realnem času (RTB).

Ravno zato je A/B testiranje v tem okolju osrednje orodje: algoritme lahko nenehno izboljšujemo in med seboj primerjamo. A/B testiramo lahko:

- **oglase** — primerjamo razlike v sliki, barvah, naslovu ali pisavi; glavna opazovana metrika je delež klikov;
- **algoritme** — primerjamo razlike v značilkah, arhitekturi ali parametrih modela; opazovana metrika je odvisna od cilja (delež klikov, konverzij, zmag).

Izzivi so tu večji kot v učbeniškem primeru: kako sploh definirati vzorec (uporabnik, seja, prikaz, piškotek?), kako obvladati mešanje med skupinami, kako se spopasti z velikim oziroma prevelikim številom značilk za stratificirano vzorčenje in z dejstvom, da enako število zahtev še ne pomeni enakega proračuna. Poleg tega si pogosto želimo, da je testna skupina boljša iz *vseh* perspektiv hkrati, zato ene same metrike ni vedno mogoče izbrati.

Zaključek

A/B testiranje je preprosto v ideji, a zahtevno v podrobnostih. Intervali zaupanja nam povedo, kdaj je opažena razlika resnična in ne le naključje; stratificirano vzorčenje pa poskrbi, da primerjamo poštene, po sestavi primerljive skupine. Skupaj tvorita zanesljiv temelj za odločanje tako v učilnici kot v resničnem svetu spletnega oglaševanja.