Uvod

Literatura

1. Z. Lozanov Crvenković, Statistika PMF, 2012.
2. Julian J. Faraway, *Practical Regression and Anova using R*, 2002, http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf
3. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning, with Applications in R*, Springer, 2013, http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
4. Douglas C. Montgomery, *Design and analysis of experiments*, John Wiley & Sons Inc. 2001.
5. Andy Field, *Discovering Statistics using IBM SPSS Statistics*, SAGE, 2012, http://www.uk.sagepub.com/field4e/main.htm
6. N. R. Draper, H. Smith,1998: *Applied regression analysis.* Wiley-Interscience, New-York, 736 pp.
7. J. Kmenta,1997: *Počela ekonometrije.* Mate, Zagreb, 787 str.
8. S. Hadživuković, 1979*: Statistika.* Rad, Beograd, 244 str.

Statistika počinje problemom, nastavlja prikupljenjem podataka, obradom podataka, i završava se zaključkom. Česta je greška započeti složenu analizu, a ne obratiti pažnju na to šta su ciljevi, pa čak ni na to da li su podaci pogodni za predloženu analizu. Prvo treba podatke treba pažljivo pogledati.

Formulacija problema.

Da bi se problem dobro formulisao, potrebno je:

1. Razumeti fizičku pozadinu. Statističar sarađuje sa drugim ljudima, pa mora razumeti ponešto u predmetu istraživanja. Ovo treba posmatrati kao priliku da se nauči nešto novo, a ne kao obavezu.
2. Razumeti cilj. Saradnik sa kojim radite možda nema jasnu predastavu šta su ciljevi istraživanja. Čuvajte se "fishing expeditions" – ako dovoljno dugo i uporno tražite, skoro uvek ćete naći nešto što je statistički značajno, a možda je samo slučajnost.
3. Budite sigurni da znate šta vaš saradnik želi. Nekad statističar sprovede analizu koja je mnogo komplikovanija nego što je saradniku potrebno. Možda su mu potrebne samo deskriptivne statistike.
4. Formulisati problem u statističkim terminima. Ovo je ozbiljan i težak korak, kada se mogu napraviti nepopravljive greške. Kada se ovo dobro uradi, rešenje je rutinska stvar.

Prikupljanje podataka.

Važno je znati kako su podaci prikupljeni.

1. Da li su podaci iz studije preseka (cross sectional study, observational) ili iz dizajnirane studije (experimental, designed sample survey). Zaključci zavise od načina na koji su podaci prikupljani.
2. Da nedostaju podaci (missing value). Čest problem koji zahteva dosta vremena da se reši.
3. Koje su jedinice mere? Da li podaci imaju više decimala nego što je potrebno.
4. Proverite da li postoje greške u unosu podataka.

Vrste promenljivih.

Kvantitativne, numeričke: diskretne, neprekidne
Kvalitativne, atributivne, kategorijalne: ordinalne i nominalne.

Greška pri merenju.

Prikupljanje podataka se vrši merenjem promenljive koja nas intersuje. Vrlo često postoji razlika između stvarne vrednosti promenljive i vrednosti registrovane merenjem. Ovo se naziva greška pri merenju. Neke promenljive se mogu meriti direktno, kao što su težina, visina, profit, a neke moramo meriti indirektno, pomoću upitnika, izveštaja. U prvom slučaju merenje se vrše instrumentima koji se mogu kalibrirati, pa su greške pri merenju male, dok u drugom slučaju se mogu dešavati veće greške pri merenju. Da bi se greška primerenju svela na minimum, merni uređaj treba da poseduje dve osobine: validnost i pouzdanost.

Validnost znači da merni uređaj stvarno meri ono za sta je namenje. Metrom merimo visinu, međutim, nivo anksioznosti se u prihologiji meri, na primer, brzinom reakcije i provodljivošću kože. U ovom slučaju mogu postojati i drugi faktori koji utuču na anksioznost, što dovodi do toga da taj instrument nije validan merni instrument.

Pouzdanost znači da merni instrument daje isti rezultat pod istim okolnostima. Ako su okolnosti iste, treba da dobijemo isti rezultai primerenju.

Metode istraživanja.
Istraživanja se mogu podeliti na studije preseka (cross sectional research, correlational research) i eksperimentalna istraživanja (experimental research).

Studija preseka znači da posmatramo stvari koje se dešavanju prirodno, bez našeg uticaja na njih, ometanja. Ova studija može da se radi u jednom vremenskom momentu, ili tokom vremena – longitudinalni podaci. Na primer, merimo nivo zagađenje u nekoj reci i broj neke vrste riba koj živi u reci. Životne navike (pušenje, broj sati vežbanja u toku nedelje, način ishrane) i neko obolenje (rak, dijabetes). Studija preseka obezbeđuje prirodan uvid u problem koje istražujemo, jer prilikom istraživanja ne utičemo na ono što se dešava, pa istraživač ne utiče na pristrastnost (bias) merenja. Kako sve promenljive posmatramo u istom vremenskom momentu, simultano, ne može se odrediti šta je uzrok, a šta posledica. – Da li pušenje izaziva rak? Longitudinalna istraživanja mogu do nekle da reše ovaj problem. (Pitanje je kako se definišu uzrok i poseldica – David Hume)

U *eksperimentalnom istraživanju* (eksperiment) se odgovor o uzroku i posledici pokušava dobiti poređenjem dve (ili više) kontrolisane situacije (tretmani, uslovi): u kojima je pretpostavljeni uzrok prisutan i odsutan. U ovom istraživanju se manipuliše jednom, (nezavisnom) promenljivom, utiče se na nju, da bi se video njen uticaj na drugu, (zavisnu) promenljivu. Na primer, poredi se uspeh učenika koji uče primenom dve metode učenja, poredi se životni vek pacijenata primenom dve operativne metode.

Metode prikupljanja podataka. Kada se podaci prikupljaju u eksperimentalnom istraživanju, mogu se prikupljati na dva različita načina.

Jedan je korišćenjem dve različite grupe, od kojih je svaka podvrgnuta jednoj kontrolisanoj situaciji (tretmanu). Ovo se naziva nezavisni dizajn, dizajn između grupa, između subjekata (independent design, between groups, between subjects). Na prmer dve grupe učenika, od kojih svaka uči primenom jedne metode učenja, gde se metode razlikuju među sobom.

Drugi je korišćenjem jedne grupe subjekata, koja je izložena različitim kontrolisanim situacijama (tretmanima), u principu u dva različita vremenska momenta. Ovo se naziva dizajn unutar subjekata, dizajn sa ponovljenim merenjima (within-subjects, repeated measured design). Na primer jedna grupa učenika, koja uči prvo primenom jedne metode, a zatim primenom druge metode; ispituju se dve kreme za sunčanje: jedna se nanosi na levi obraz ispitanika, a druga na desni obraz ispitanika.

Vrste varijacije, varijabilnosti. Postoje dve osnovne vrste varijacije – sistematska i nesistematska varijacija.

Nesistematska varijacija nastaje kao posledica uticaja nepoznatih, slučajnih faktora koji dovode do malih razlika u tretmanima.

Sistematska varijacija nastaje kao posledica dejstva, uticaja, istraživača koji u jednom tretmanu nešto radi, a u drugom ne.

Randomizacija. U eksperimentalnom istraživanju (between subjects i u within subjects) važno je svesti nesistematsku varijaciju na minimum. Na taj način dobijamo osetljiviju meru uticaja kojim delujemo u eksperimentu. Ovo se postiže randomizacijom subjekata u istraživanju. Mnogi statistički testovi se zasnivaju na određivanju, a zatim na upoređivanju nesistematske i sistematske varijacije. Ovo poređenje omogućava da se vidi da li je eksperiment izazvao značajno više varijacije nego što bi se dobilo de nije bilo uticaja u eksperimentu. Randomizacija je važna jer eliminiše većinu drugih izvora sistematske varijacije i omogućuje da se utvrdi da je sistematska varijacija nastala zbog uticaja na nezavisnu promenljivu koji se vrši u eksperimentu.

Randomizacija se može vršiti na dva načina, u zavisnosti da li se radi o dizajnu između ili unutar grupa.

Početna analiza podataka.

Važan korak koji uvek teba sprovesti.

1. Numeričke promenljive- aritmetička sredina, medijana, moda, disperzija, standardna devijacija, asimetričnost, spljoštenost, korelacije.
2. Grafički prikaz: histogrami, dijagrami rasipanja.
3. Autlajeri, nedostajuće vrednosti, greške u unosu, asimetrične raspodele, neobične raspodele.

Primer.
https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
patients are females at least 21 years old of Pima Indian heritage.