SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

BIOLOŠKI ODSJEK

SEMINAR

Greške u prikupljanju, obradi i interpretaciji bioloških podataka

Ivna Ivanković

Preddiplomski studij molekularne biologije

(Undergraduate study of Molecular Biology)

Mentor: dr. sc. Goran Igaly

Zagreb, 2016.

Sadržaj

1. Uvod

Biologija je znanost koja proučava svojstva živih organizama stoga gotovo svaki biološki eksperiment zahtijeva prikupljanje, obradu i interpretaciju podataka. Statistika je neizostavan alat u svim znanostima, posebno u biološkim istraživanjima. Dobro istraživanje je prikladno dizajnirano, raspolaže s kvalitetnim mjerenjima, primjenjuje odgovarajuće statističke metode i korektno interpretira analitičke rezultate. Svaka od nabrojenih komponenata može biti nezadovoljavajuća što narušava i umanjuje vrijednost rezultata istraživanja. U biološkim i biomedicinskim istraživanjima koristi se biostatistika: *skup alata za statističku analizu koji omogućuje procjenu odnosa između rezultata* (Thiese i sur. 2015). Razvojem tehnologije raste i kompleksnost metoda obrade podataka čime raste i upotreba biostatistike. Iako one postaju točnije, brže i dostupnije, način na koji se koriste u mnogo slučajeva je pogrešan i dovodi do netočnih rezultata.

Većina slučajeva pogrešne upotrebe statistike rezultat je nedostatka poznavanja metoda, no istraživanje o falsifikaciji i zloupotrebi statistike u znanosti iz 2009. godine (Fanelli, 2009) utvrdilo je da 33.7% ispitanika priznaje kako su koristili upitne metode istraživanja i modificirali rezultate kako bi poboljšali ishod eksperimenta, interpretirali podatke izuzevši metodološke ili analitičke detalje te odbacivali neke uzorke zbog osjećaja da su neispravni. Neovisno radi li se o namjernoj ili nenamjernoj zloupotrebi statistike rezultati obaju slučajeva su jednaki: pogrešni rezultati vode do krivih saznanja na koje se oslanja ostatak znanstvene zajednice što u konačnici vodi do razvoja znanosti u neispravnom smjeru. Kako bi se smanjila pogrešna upotreba statističkih metoda, potrebno je ukazati na najčešće greške, upozoriti na moguće propuste te predložiti kako ih je moguće ispraviti u budućim istraživanjima.

2. Prikupljanje bioloških podataka

Dobri uzorci temelj su dobre znanosti (Whitlock i Schluter, 2015).

Uzorak je selektirani skup jedinki iz populacije, a proces prikupljanja uzoraka naziva se uzorkovanje. Prvi korak svakog prikupljanja bioloških podataka jest odabir ciljne populacije nakon čega slijedi uzorkovanje. U biološkim istraživanja uzorak vrlo često predstavlja jedinku ispitivane populacije. Osim toga, uzorak može biti nukleotidni ili proteinski slijed te grupa organizama poput kaveza miševa ili bakterijske kolonije (1 uzorak = 1 kolonija). Populacije čini vrlo velik broj jedinica koje je sve gotovo nemoguće prikupiti i izmjeriti zbog vremenskih i financijskih ograničenja istraživanja. Stoga prikupljamo uzorak koji predstavlja populaciju od interesa. Korištenjem statističkih metoda možemo procijeniti koliko naš uzorak odstupa od stvarnih vrijednosti za cijelu populaciju, ali da bi procjena bila pouzdana neophodno je prikupiti dobar uzorak. Sve procjene temeljene na uzorcima odstupaju od stvarnih vrijednosti karakteristika populacije zbog neizbježnog utjecaja slučajnosti. Ovakvo odstupanje naziva se greška uzorkovanja.

2.1 Veličina uzorka

Mnogo je slučajeva u kojima je skup uzoraka premalen za analizu. Što je skup manji, veća je sklonost pristranih (eng. *biased*) rezultata. Npr. pri ispitivanju učestalosti određene bolesti u seoskim četvrtima zabilježena je puno manja pojavnost (<10), dok je u gradskim četvrtima pojavnost veća (>100). Ovakav rezultat ne mora nužno značiti da je bolest zastupljenija kod stanovnika iz gradskih četvrti, već da je njih naprosto više. Kako bi se izbjeglo pogrešno zaključivanje, dobra je ideja izračunati stopu pojave bolesti na 1000 stanovnika svake četvrti.

2.2 Slučajno uzorkovanje

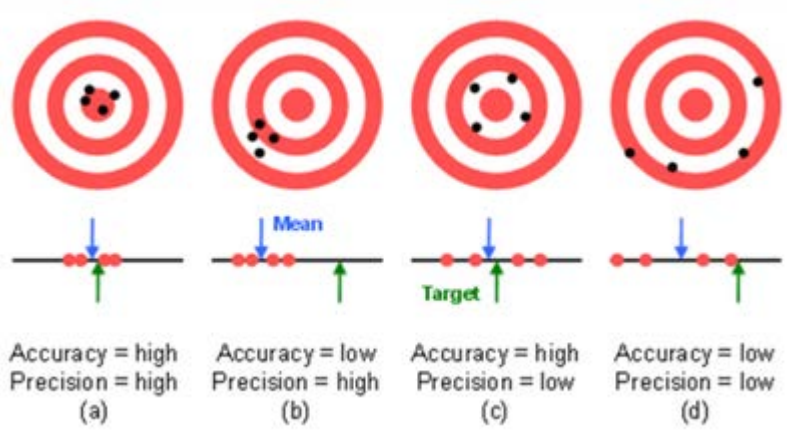
Za većinu statističkih metoda u biologiji od iznimne je važnosti da su podaci za analizu dobiveni slučajnim uzorkovanjem jer u suprotnom dolazi do pogrešnih rezultata. Uzorkovanje je slučajno ako zadovoljava dva kriterija:

1. svaki član populacije ima jednaku vjerojatnost biti uzorkovan
2. odabir članova mora biti nezavisan (Whitlock i Schluter, 2015).

Poštivanje prvog kriterija često nije moguće zbog teške dostupnosti jedinki iz populacije. Primjerice, kad se na izrazito nepristupačnom terenu uzorkuju samo biljke koje rastu uz cestu ili dostupnija područja unutar terena, ne dobivaju se informacije o ostatku populacije. Ako se karakteristike biljaka koje rastu dalje od ceste razlikuju od prikupljenih, istraživač raspolaže s nepouzdanim uzorkom koji će dovesti do pristranih (eng. *biased*) rezultata. Poštivanje drugog kriterija narušava se ako već uzorkovana jedinka na bilo koji način utječe na odabir druge jedinke. Kada se u istraživanju radi sa slučajnim uzorkom minimizirano je dobivanje pristranih rezultata i moguće je pouzdano izračunati grešku uzorkovanja, no takav uzorak u većini istraživanja nije moguće prikupiti te je stoga bitno da znanstvenici ukažu na moguće propuste u prikupljanju. Greške koje rezultiraju neslučajnim uzorkom mogu biti i rezultat namjernog propusta znanstvenika koji uzorkuju samo malenu skupinu jedinki iz populacije za koje znaju da pokazuju željena svojstva. Rezultati upitnika koje su ispunili dobrovoljci, a ne nasumično izabrani ispitanici, također nisu dobiveni slučajnim uzorkovanjem.

2.3 Točnost i preciznost

Točnost govori koliko su uzorkovane vrijednosti blizu stvarnih vrijednosti, dok preciznost daje informaciju o raspršenosti mjera. Drugim riječima, točnost opisuje razliku između stvarne vrijednosti i vrijednosti dobivenih uzorkovanjem i mjerenjem, dok preciznost opisuje varijaciju između ponavljajućih mjerenja iste varijable (Barry, 2012). Vizualizacija ovih razlika prikazuje Slika 1.



Slika 1. prikaz razlike između točnosti (engl. *accuracy*) i preciznosti (engl. *precision*). Preuzeto s: www.wfsscience.org/uploads/1/0/0/4/10044856/2680888.png?398

Preciznost rezultata mjerenja može zavarati o njihovoj točnosti. Tako kada u trima uzastopnim titracijama rezultat utrošene kiseline iznosi: 5.0 ml, 5.1 ml, 5.0 ml mogli bismo zaključiti da je taj volumen točan, ali ako je došlo do greške u pripravljanju koncentracije kemikalija s kojima radimo ili subjektivne pogreške kao očitanja volumena, ovaj rezultati neovisno o preciznosti mogu biti potpuno krivi jer bi očekivani volumen trebao iznositi 7 ml kiseline.

3. Obrada bioloških podataka

Postoji mnoštvo statističkih testova za različite tipove podataka i namjene. Kako bi se iz dobrih podataka došlo do dobrog zaključka potrebno je izabrati odgovarajući statistički test. Osim što može doći do pogreške zbog pogrešnog odabira testa, svaki test ima više verzija stoga i izbor verzije mora biti u skladu s traženim pretpostavkama istraživanja.

3.1 Prikaz osnovnih podataka

Još u svom istraživanju iz 1977. godine o krivoj upotrebi statističkih metoda, Gore i suradnici ukazuju na važnost prikaza podataka. Medijan, mod (engl. *mode*) i aritmetička sredina uzorka daju informacije o "središtu" vrijednosti uzoraka. Ali takav prikaz u znanstvenim radovima ne daje nikakve informacije o raspršenosti stoga je za ispravnu vizualizaciju podataka potrebno pružiti informacije o nekoj mjeri raspršenja, npr. standardnoj devijaciji (Gore, 1977).

3.2 T-test

Studentov t-test koristi se za usporedbu srednjih vrijednosti dvaju setova kontinuiranih uzoraka, a upareni t-test koristi se kod uparenih setova podataka (kada opažanje u jednom uzorku ima pridruženo opažanje u drugom) (Thiese 2015).

Ovaj test često se odabire bez prethodne potvrde da su zadovoljene pretpostavke koje će omogućiti ispravno korištenje metode. Te pretpostavke su:

1. distribucija uzorka je normalna
2. uzorak je statistički nezavisan
3. varijance su jednake (Gore 1977).

Postoje brojni oblici t-testova koji se temelje na varijancama uzoraka. Glantz je u svom radu proučavajući dva časopisa otkrio da je u polovici članaka koji su koristili statistiku došlo do primjene t-testa u slučajevima kada je ispravno bilo upotrijebiti test za višestruku usporedbu (Glantz 1980). Istraživanje koje su proveli Williams i suradnici otkrilo je da gotovo polovica članaka u časopisu *American Journal of Physiology* koristi upareni ili

neupareni t-test, a od tih članaka otprilike je u 17% njih pogrešno korišten t-test za višestruku usporedbu modifikacijom testa korekcijskim metodama (Williams, 1997).

3.3 ANOVA

Analiza varijance (ANOVA) testira hipotezu da su aritmetičke sredine dvije ili više populacije jednake. Ispravno provođenje ovog testa zahtijeva kontinuirane varijable i barem jedan kategorijski faktor s dva ili više stupnja. Kao i t-test, tako i statistički test ANOVA u većini slučajeva zahtijeva korekciju ovisno o vrsti uzoraka s kojima se izvodi istraživanje. Nekorištenje ili pogrešno korištenje korekcija i u ovom slučaju dovodi do pogrešnih zaključaka prema istraživanju Gorea i suradnika (Gore 1977).

4. Interpretacija bioloških podataka

Donošenje zaključaka iz rezultata uvijek je zahtjevan proces podložan raspravama. Kriva interpretacija ispravnih rezultata dobivenih ispravnom analizom rezultirat će krivim zaključkom eksperimenta stoga je ona jednako bitna za istinitost rezultata. Točnost zaključaka, osim kvalitetno prikupljenih i obrađenih podataka, zahtijeva ispravno rasuđivanje i izbjegavanje najčešćih zabluda koje su navedene dalje u tekstu.

4.1 Korelacija i kauzalnost

Identifikacija uzroka ispitivanih događaja je ključan proces u znanosti. Prvi korak je pronalaženje obrazaca koji mogu ukazati na povezanost (uzajamnost, korelaciju) između događaja. Već su naši preci primijetili da žvakanje kore vrbe umanjuje glavobolju i temperaturu, a danas znamo da je povezanost između žvakanja kore vrbe i smanjenja bolova objašnjena prisutnošću salicilne kiseline u kori jer blokira otpuštanje lipidnih spojeva koji su odgovorni za prijenos boli na upaljenim ili oštećenim mjestima (prostaglandini).

Korelacija je dobar indikator postojanja moguće uzročne veze između varijabli. Prvotnim opažanjem korelacije (penicilin smanjuje rast bakterija) može se otkriti kauzalnost (penicilin uzrokuje smrt bakterija), no ovdje je iznimno lako učiniti pogrešku rasuđivanja jer korelacija između dvije varijable može postojati bez nužne kauzalnosti. Da nije tako, moglo bi se pogrešno zaključiti da kriminalne radnje povećavaju želju za slatkišima ili da slatkiši uzrokuju kriminalnu radnju iz podataka koji pokazuju pozitivan trend konzumacije slatkiša i broja ubojstava u određenom gradu. Isto tako tvrdnja „Djevojke u ženskim školama imaju bolje ocjene od djevojaka u mješovitim školama, stoga su ženske škole bolje za djevojke.“ (Smith 2015) podložna je oštrim raspravama. Nije teško pronaći podatke koji bi poduprli prvi dio tvrdnje, a drugi dio tvrdnje je izveden zaključak koji implicira kauzalnost. No možemo li taj zaključak izveden samo na temelju ove korelacije smatrati točnim? Bez dodatnih informacija ne jer postoje i drugi faktori koje treba uzeti u obzir, a samo neka pitanja su:

1. Razlikuje li se nastavni sadržaj koji se obrađuje u ženskim školama od sadržaja koji se obrađuje u mješovitim?

2. Jesu li ženske škole selektivnije od mješovitih prilikom odabira učenica?

3. Plaća li se i ako da, koliko se plaća školarina u ženskim školama?

Uspostavljanje uzročnih veza prilično je zahtjevan proces, ali može se olakšati dobrim postavljanjem eksperimenta, ispravnom i detaljnom obradom podataka te ponovnim i ponovnim izvođenjem eksperimenta. Korelacije za koje se još ne zna pokazuju li kauzalnost, često su dobri putokazi u novim istraživanjima.

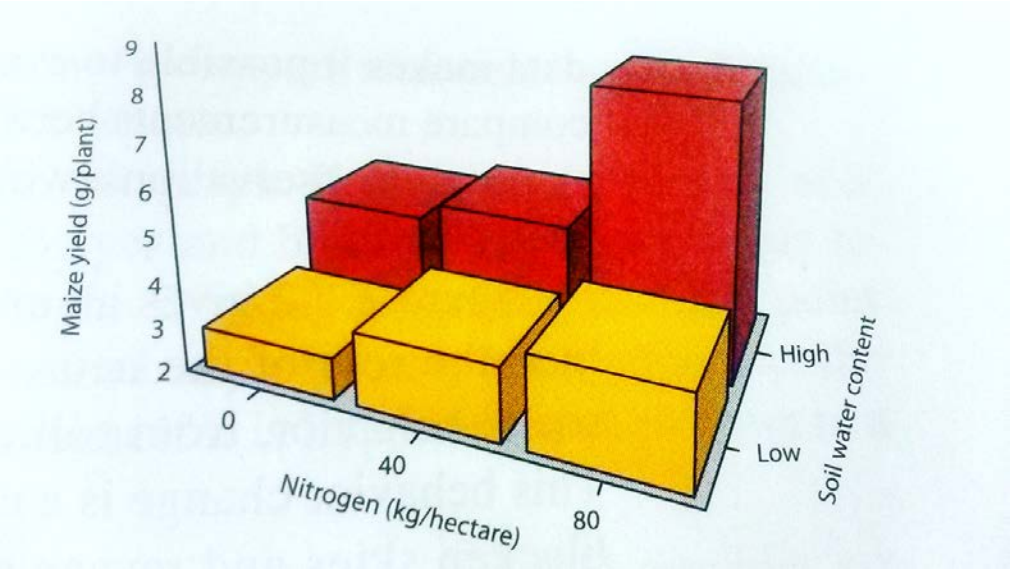
4.2 Nasumičnost i slučajnost (engl. *randomness and chance*)

Nasumičnost i slučajnost su također blisko povezani pojmovi stoga može doći do pogrešnog rasuđivanja da je svaki slučajni događaj nasumičan. Utjecaj veličine uzorka na rezultate često je zanemaren i krivo navodeći. To se može pokazati primjerom iz knjige *Statistical Analysis Handbook* (Smith 2015): pretpostavimo da se u velikoj bolnici rodi u prosjeku 40 djece u jednom danu od kojih je 50% njih muško. U manjoj bolnici u prosjeku se rodi 10 djece u danu i također je njih 50% muško. Neke dane taj postotak biti će manji, a neke veći. Pitanje koje se postavlja je: „U kojoj bolnici je za očekivati više dana u kojima je postotak rođenja muške djece najmanje 60%?“ Intuicija navodi na zaključak u većoj, ali u manjoj bolnici jer promjena s 5 na 6 rođenja muške djece već podiže postotak na 60% rođenja prema muškom spolu, a kod veće bolnice trebalo bi se roditi najmanje 4 više dječaka od djevojčica kako bi rezultat bio 60% rođenih dječaka što je manje vjerojatan događaj.

Sličan primjer je zabluda tužitelja (engl. *prosecutor's fallacy*): tužitelj pozove svjedoka koji tvrdi da određeni dokaz (npr. preklapanje krvi osumnjičenog s krvi koja je pronađena na mjestu zločina, a radi se o rijetkoj krvnoj grupi) pruža poveznicu s optuženim koja bi se mogla dogoditi samo jednom u milijun puta. Stoga tužitelj govori kako je u tom slučaju šansa da je optuženi nevin samo jedan naspram milijun. Budući da ne znamo je li optuženi stvarno kriv (nije ispravno polaziti od pretpostavke da je), potrebno je vidjeti koliko drugih ljudi u populaciji bi mogli pokazati ovakvu povezanost. Ako se radi o 10-20 ljudi, bez drugih dokaza možemo smatrati da je optuženih kriv, ali to možemo tvrditi sa sigurnošću od samo 5-10%.

4.3 Pogrešna interpretacija vizualizacija

Vizualizacija podataka bitna je u prezentaciji, ali i analizi podataka jer omogućuje identifikaciju uzoraka koji se pojavljuju u podacima. Nepotpuni ili loše napravljeni grafički prikazi vode do pogrešnih čitanja podataka što može uzrokovati pogrešnu interpretaciju podataka. Naglasak u vizualnim prikazima stavlja se na jednoznačnost i jednostavnosti u prikazivanju podataka, označavanje osi te pošteno korištenje dimenzija.



Slika 2. Primjer lošeg grafičkog prikaza srednjih vrijednosti visine kukuruza uzgojenih u posudama, rasli su u zemlji s različitim količinama vode i dušika. Izvor: *The Analysis of biological data* (Whitlock i Schluter, 2015), figure 2.1.-1, str. 27

Slika 1. prikazuje primjer lošeg grafičkog prikaza naknadno konstruiranog iz podataka istraživanja Quayea i suradnika (Quaye 2009). U ovakvom prikazu nedostaju informacije o podacima: svaki stupac prikazuje prosječan prinos kukuruza uzgojenih u posudama, ali nigdje nisu vidljive eksperimentalne jedinice, odnosno prinos svake pojedine posude čime nedostaju informacije o varijaciji prinosa po posudama. Ovakvim prikazom sakrivena su neobična opažanja u podacima koja su se mogla dogoditi. Također, vrlo je teško vidjeti uzorke (obrasce) koji se javljaju u podacima jer trodimenzionalna struktura i nagnutost grafa narušuju perspektivu promatrača i onemogućuju jednoznačnu usporedbu visina stupaca. Osim toga, veličina na vertikalnoj osi grafa ne počinje od nule što visinu stupaca stavlja izvan proporcija istinitih dimenzija i grafički elementi nisu dovoljno veliki (nazivi osi) da bi se jasno vidjeli.

5. Literatura

Fanelli D. 2009. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS ONE* 4(5): e5738.

Glantz SA. 1980. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1,1-7.

Gore SM, Jones IG, Rytter EC. 1977. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976 *British Medical Journal*;1,85-87.

Quaye AK, Laryea KB, Abeney-Mickson S. 2009. Soil Water and Nitrogen Effects on Maize (*Zea mays* L.) Grown on a Vertisol. *Journal of Forestry Horticulture and Soil Science:* 3,1.

Smith MJ. 2015. STATSREF: Statistical Analysis Handbook - a web-based statistics resource. The Winchelsea Press, Winchelsea, UK.

Thiese MS, Arnold ZC, Walker Skyler D. 2015. The misuse and abuse of statistics in biomedical research. *Biochemia Medica* 25(1):5-11.

Williams JL, Hathaway CA, Kloster KL, Layne BH. 1977. Low power, type II errors, and other statistical problems in recent cardiovascular research. *American Journal of Physiology* 273:H487–93

Whitlock MC, Schluter D. 2015. The analysis of biological data. Roberts and Company Publishers, Greenwood Village, Colorado.

blog.minitab.com/blog/real-world-quality-improvement/

plato.stanford.edu/entries/chance-randomness/

support.minitab.com/minitab/17/topic-library/modeling-statistics/anova/basics/what-is-anova/

wfsscience.org/uploads/1/0/0/4/10044856/2680888.png?398

yourhormones.info/hormones/prostaglandins.aspx

6. Sažetak

Biostatistika je neizostavan alat u biologiji koja ispravnim korištenjem omogućuje vidjeti veze između različitih fenomena te predvidjeti buduće trendove. Gotovo svako istraživanje zahtijeva prikupljanje, obradu te interpretaciju podataka kako bi se došlo do novih spoznaja. Kako bi spoznaje bile ispravne, potrebno je koristiti statistiku na ispravan način. Dostupnost velikih broja metoda znanstvenike navodi na pogreške, a u nekim slučajevima pogrešne se metode namjerno koriste kako bi došlo do dobivanja krivih, ali željenih rezultata. U ovom radu navedene su i opisane neke od najčešćih grešaka u prikupljanju, obradi te interpretaciji bioloških podataka kako bi se ukazalo na problem neispravne manipulacije uzorkom i podacima od samog početka do kraja analize podataka u biološkim istraživanjima. Također je sugerirano kako unaprijediti analizu te na koje je stvari potrebno obratiti pažnju kako bi došlo do minimiziranja grešaka u istraživanjima i kvalitetnog razvoja znanosti.

7. Summary

Biostatistics is a fundamental tool in biology that allows to visualize the data and observe connections between different phenomena. Also, in some cases it makes possible to predict future trends of our subject of interest. Every research requires collecting, processing and interpretation of data. Only proper use of statistics can guarantee accurate results and lead to reliable conclusions. This paper describes the most common mistakes and errors made during data collection, processing and interpretation. The paper also suggest how to improve data analysis in order to improve the development of biological sciences.