SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Doris Zelić

VIŠESTRUKO TESTIRANJE HIPOTEZA I SORIĆEVA METODA

Diplomski rad

Voditelj rada:
prof. dr. sc. Bojan Basrak

Zagreb, rujan 2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Velika hvala mom mentoru, prof. dr. sc. Bojanu Basraku na pruženoj pomoći i usmjeravanju pri izradi diplomskog rada. Zahvaljujem se i svojim roditeljima na velikoj podršci i hrabrenju u posrnulim danima te svekrvi i svekru koji su uvijek bili spremni pomoći. Velika hvala šogoru Domagoju na uloženom trudu i vremenu za pomoć pri nastajanju ovoga rada, a najveće hvala suprugu Sebastijanu i kćerkici Eni na velikom strpljenju i odricanju tijekom pisanja ovog diplomskog rada.

Sadržaj

Uvod

Tema ovog diplomskog rada posvećena je problematici višestrukog testiranja statističkih hipoteza. Problem višestrukog testiranja je ozbiljan problem u znanstvenom istraživanju koji može dovesti do pogrešnih zaključaka u znanosti ukoliko se prethodno ne izvrši odgovarajuća prilagodba takvom testiranju.

U radu govorimo o razlozima zbog kojih dolazi do problema prilikom višestrukog testiranja hipoteza te o načinima rješavanja istih koje su predložili poznati matematičari i statističari. Navodimo nekoliko primjera koji ukazuju na važne stvarne posljedice ukoliko problemi prilikom višestrukog testiranja ostanu neriješeni. U radu također govorimo o metodi dr. Branka Sorića i tzv. FDR teoriji (engl. false discovery rate) o kojima govori njegov članak *Statistical "discoveries" and effect size estimation* iz 1989. godine. FDR teorija se pokazala mnogo uspješnijom u rješavanju problema višestrukih uspoređivanja od dotadašnjih metoda. O tome govori i sama činjenica da je rad Benjaminija i Hochberga *Controlling the false discovery rate: a practical and powerful approach to multiple testing* iz 1995. godine u kojem su matematički precizno iznijeli Sorićevu ideju, postao jedan od najcitiranijih statističkih, i ne samo statističkih, radova u povijesti.

Ovim radom želimo ukazati na probleme do kojih dovodi neprilagođeno višestruko testiranje hipoteza te potaknuti čitatelja na prosuđivanje o istinitosti nekih rezultata u znanosti i dobivenih zaključaka, želimo dati pregled nekih starih i novih procedura za kontrolu pogrešaka do kojih dolazi pri višestrukom testiranja te njihovu usporedbu dajemo u simulacijskoj studiji.

U prvom poglavlju podsjećamo čitatelja na osnovne pojmove iz statistike kojima se služimo prilikom testiranja statističkih hipoteza, iznosimo i neke primjere testiranja hipoteza te dajemo uvod u problem višestrukog testiranja pokazujući na primjerima njegove važne posljedice. Nakon toga iznosimo pregled kontrolnih procedura za kontrolu pogrešaka pri višestrukom testiranju. Drugo poglavlje posvećeno je FDR metodi. Govorimo o njezinom razvitku te je formalno definiramo i opisujemo vrlo značajne procedure za kontrolu FDR-a. Posljednje, treće poglavlje čini simulacijska studija u kojoj uspoređujemo praktične posljedice triju procedura za kontrolu grešaka. Dajemo grafički prikaz dobivenih rezultata te iznosimo zaključke.

Poglavlje 1

Višestruko testiranje hipoteza

1.1 Testiranje hipoteza

Osnovni pojmovi

Kada govorimo o testiranju hipoteze podrazumijevamo donošenje odluke o istinitosti, odnosno, neistinitosti slutnje koju nazivamo hipoteza. *Statistička hipoteza* jest konkretna pretpostavka o (populacijskoj) razdiobi nekog statističkog obilježja X. Kažemo da je statistička hipoteza *jednostavna* ukoliko jednoznačno određuje razdiobu od X. U suprotnom kažemo da je *složena*.

Primjer 1.1.1. *Navedimo primjer jedne složene i jedne jednostavne hipoteze. Hipoteza*

$$H_1 : X \text{ ima normalnu razdiobu},$$

jest složena hipoteza budući da ne određuje jednoznačno razdiobu od X dok je hipoteza

$$H_2 : X \sim N(0, 2),$$

primjer jednostavne hipoteze prema kojoj je razdioba statističkog obilježja X normalna s očekivanjem 0 i varijancom 2 pa je time jednoznačno određena.

Postupak donošenja odluke o odbacivanju ili ne odbacivanju statističke hipoteze zove se *testiranje statističkih hipoteza*. Takav postupak možemo podijeliti u nekoliko koraka. Najprije identificiramo hipotezu koju želimo testirati. Tu hipotezu nazivamo *osnovnom* ili *nul* hipotezom koju označavamo sa H_0. U postupak testiranja uzimamo i njoj *alternativnu* hipotezu H_1 u kojoj navodimo tvrdnju za koju mislimo da je istinita ukoliko je nul hipoteza neistinita. Drugi korak jest odabir kriterija na temelju kojeg donosimo odluku u postupku testiranja hipoteza. Odabrati kriterij znači definirati pravilo za odbacivanje nul hipoteze.

U trećem koraku na osnovi realizacije slučajnog uzorka za X računamo testnu statistiku i p-vrijednost, ukoliko je to moguće. Pritom je, grubo govoreći, *p-vrijednost* vjerojatnost ne odbacivanja nul hipoteze uz uvjet da je ona istnita. U zadnjem koraku zaključujemo jesu li podaci iz uzorka u skladu s nul hipotezom, odnosno, donosimo odluku o (ne)odbacivanju nul hipoteze. Budući da sve odluke zasnovane na uzorcima iz populacije nisu 100% pouzdane, ni zaključak (odluka) statističkog testa nije sasvim pouzdan. Dakle, može se dogoditi da je zaključak testa pogrešan. Sljedeća tablica nam prikazuje koje su moguće pogreške prilikom donošenja odluke o nul hipotezi.

Hipoteza H_0	Istinita	Neistinita
Odbacujemo	Pogreška 1. vrste	Pravilan zaključak
Ne odbacujemo	Pravilan zaključak	Pogreška 2. vrste

Dakle, pogreška koja nastaje pri odbacivanju nul hipoteze H_0 u slučaju kada je ona istinita nazivamo *pogreška prve vrste*. Takva se pogreška događa u slučaju kada „vidimo" učinak kojeg zapravo nema. Najveću vjerojatnost počinjenja pogreške 1. vrste koju dopuštamo nazivamo *razina značajnosti* i označavamo ju sa α. Kada je p-vrijednost manja od vrijednosti α, tada se rezultat naziva *statistički značajnim na razini* α. *Pogreška druge vrste* nastaje kada ne odbacujemo nul hipotezu H_0 u slučaju kada ona nije istinita. Vjerojatnost pogreške 2. vrste označavamo β. Osim vjerojatnosti α i β mogućih pogrešaka pri donošenju odluke za nul hipotezu, uvodimo još i pojam *snage* testa. Pod snagom testa podrazumijevamo vjerojatnost odbacivanja hipoteze H_0 kada je ona zaista neistinita (vjerojatnost pravilnog odbacivanja) i ona je jednaka $1-\beta$ (uočimo da može ovisiti o nepoznatom parametru).

Test (hipoteze H_0 u odnosu na alternativu H_1) je preslikavanje $\tau\colon \mathbb{R}^n \to \{0,1\}$. Ako je za realizaciju uzorka x funkcija $\tau(x) = 1$, tada odbacujemo H_0 u korist H_1, a ako je $\tau(x) = 0$, tada ne odbacujemo H_0 u korist H_1. Tada je

$$C := \tau^{-1}(1) = \{x \in \mathbb{R}^n : \tau(x) = 1\}$$

područje realizacija uzoraka za koje se H_0 odbacuje u korist H_1. Skup C nazivamo *kritično područje* za test τ. Ovisno o tome koje su nam osnovne pretpostavke (iz koje razdiobe dolazi slučajni uzorak, koji su nam poznati, odnosno nepoznati parametri), koja nam je nulta, odnosno alternativna hipoteza, razlikujemo statističke testove. Važniji statistički testovi su: z-test, t-test, χ^2-test, F-test. Navedimo sada nekoliko primjera testiranja statističkih hipoteza.

Primjeri

Primjer 1.1.2. *Pretpostavimo da želimo testirati hipotezu da djeca do druge godine života prebole 5 upala uha sa pretpostavljenom standardnom devijacijom od 2 upale uha. Identi-*

ficirajmo nul hipotezu i njoj alternativnu hipotezu.

$$H_0 : \mu = 5$$

$$H_1 : \mu \neq 5.$$

U uzorak su uzeti podaci o broju preboljenih upala uha do druge godine života desetero djece. Prosječan broj preboljenih upala uha u uzorku je 3. Želimo na razini značajnosti od 5% donijeti odluku o nul hipotezi. Odnosno, postavljeno pitanje može glasiti možemo li na razini značajnosti od 5% prihvatiti pretpostavku da je uzorak izabran iz osnovnog skupa (gdje je $\mu = 5$). Drugi korak jest odabir kriterija prema kojemu donosimo odluku o nul hipotezi. U ovom slučaju, to je naša zadana razina značajnosti $\alpha = 0.05$. Treći korak, računamo testnu statistiku i p-vrijednost na osnovu opaženih podataka u uzorku. Na osnovu centralnog graničnog teorema testna statistika je

$$Z = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} = \frac{3 - 5}{2} \cdot \sqrt{10} = -\sqrt{10} \approx -3.16.$$

Budući da se ovdje radi o dvostranom Z-testu, iz tablica vrijednosti funkcije distribucije za jediničnu normalnu razdiobu čitamo

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96.$$

Kritično područje za ovaj dvostrani test može biti

$$C = \langle -\infty, -z_{0.025}] \cup [z_{0.025}, +\infty\rangle = \langle -\infty, -1.96] \cup [1.96, +\infty\rangle.$$

Računamo i p-vrijednost koristeći tablice te dobivamo

$$p = 2 \cdot \mathbb{P}(Z > |z|) = 2 \cdot \mathbb{P}(Z > \sqrt{10}) = 2 \cdot (0.0008) = 0.0016.$$

Zaključak sada možemo donijeti na dva načina. Koristeći vrijednost testne statistike uspoređujemo

$$|z| = \sqrt{10} \approx 3.16 > 1.96 = z_{0.025}.$$

Budući da nam vrijednost $|z| \approx 3.16$ upada u kritično područje, odbacujemo hipotezu H_0. Drugi način jest da usporedimo dobivenu p-vrijednost s α, a kako je

$$p = 0.0016 < 0.05 = \alpha,$$

dolazimo do istog zaključka, odnosno odbacujemo hipotezu H_0, tj. odbacujemo hipotezu da je prosječan broj preboljenih upala uha u djece mlađe od 2 godine jednak 5.

U prethodnom primjeru imali smo zadanu teorijsku distribuciju (N(5,4)) te smo na osnovu uzorka testirali podudara li se dobiveno očekivanje uzorka zaista sa teorijskom distribucijom uz razinu značajnosti od 5%. Tu smo imali dvostranu hipotezu. U idućem primjeru imamo zadanu teorijsku distribuciju no očekivanje nam je nepoznato. Na osnovu uzorka testiramo vrijednost očekivanja i tu nam je alternativna hipoteza jednostrana.

Primjer 1.1.3. *Neka je X normalna varijabla s varijancom* $\sigma^2 = 4$. *Na osnovi uzorka duljine* 64 *želimo testirati nul hipotezu*

$$H_0 : \mu = 20,$$

u odnosu na alternativnu hipotezu

$$H_1 : \mu = 19$$

uz razinu značajnosti $\alpha = 0.05$. *Testna statistika za ovaj primjer jest*

$$Z = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} = \frac{\bar{X} - 20}{2} \cdot \sqrt{64} = \frac{\bar{X} - 20}{2} \cdot 8.$$

Ovdje se radi o jednostranoj hipotezi pa iz tablica čitamo

$$z_\alpha = z_{0.05} = 1.64,$$

iz čega slijedi da je kritično područje

$$C = \langle -\infty, -z_{0.05}] = \langle -\infty, -1.64].$$

Dakle, H_0 *odbacujemo ukoliko je* $z \leqslant -1.64$, *a ako je* $z > -1.64$, *tada nul hipotezu ne odbacujemo. U tom slučaju je vjerojatnost pogreške 2. vrste jednaka*

$$\begin{aligned} \beta &= \mathbb{P}(Z > -1.64|H_1) = \\ &= \mathbb{P}(\frac{\bar{X} - 19}{2}\sqrt{64} > \frac{20 - 19}{2}\sqrt{64} - 1.64|H_1) = \\ &= \mathbb{P}(\bar{X}^* > 2.36) = \\ &= 1 - \Phi(2.36) = \\ &= 0.0091. \end{aligned} \tag{1.1}$$

Snaga testa je

$$1 - \beta = 1 - 0.0091 = 0.9909.$$

Pogledajmo sada jedan primjer u kojem testiramo jednakost očekivanja dviju grupa koje su normalno distribuirane s jednakim varijancama.

Primjer 1.1.4. *Na jednom fakultetu u grupi od 20 studenata uzeti su podaci o tjednom broju sati bavljenja nekom fizičkom aktivnosti za pojedinog studenta. U tablici se nalaze podaci:*

spol	*broj sati*											
ženski	*3*	*10*	*5*	*1*	*4*	*8*	*4*	*2*				
muški	*4*	*8*	*9*	*10*	*10*	*3*	*8*	*10*	*2*	*4*	*6*	*3*

Uz pretpostavku da je broj sati normalno distribuiran s jednakom varijancom, zanima nas možemo li na razini značajnosti od 5% tvrditi da su muškarci u prosjeku fizički aktivniji od žena. Prvo identificirajmo hipoteze:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 < \mu_2,$$

pri čemu je μ_1 prosječan broj sati bavljenja nekom fizičkom aktivnošću tjedno za žene, a μ_2 prosječan broj sati fizičke aktivnosti za muškarce. U ovom primjeru koristimo jednostrani t-test za nezavisne uzorke budući da uspoređujemo prosjek određene veličine između dviju nezavisnih grupa (u ovom slučaju je to broj sati fizičke aktivnosti tjedno koji uspoređujemo između muškaraca i žena). Pretpostavka o normalnosti i jednakosti varijanci mora biti zadovoljena kako bi mogli provesti t-test. Testna statistika je

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_d} \cdot \frac{1}{\sqrt{1/n_1 + 1/n_2}},$$

gdje je n_1 broj podataka za žene, n_2 broj podataka za muškarce, a S_d je standardna devijacija. Izračunajmo sada veličine koje nam trebaju za određivanje testne statistike.

$$n_1 = 8,\ \bar{x_1} = 4.625,\ s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x_1})^2 = 9.125$$

$$n_2 = 12,\ \bar{x_2} = 6.417,\ s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_i - \bar{x_2})^2 = 9.538$$

$$S_d = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{7 \cdot 9.125 + 11 \cdot 9.538}{18}} = 3.062$$

Iz tablica iščitavamo

$$t_\alpha(n_1 + n_2 - 2) = t_{0.05}(18) = 1.734.$$

Dakle, kritično područje nam je

$$C = \langle -\infty, -1.734],$$

a realizacija statistike T

$$t = \frac{4.625 - 6.417}{3.062} \cdot \frac{1}{\sqrt{1/8 + 1/12}} = -1.282.$$

Budući da t ne upada u kritično područje, ne odbacujemo hipotezu H_0 u korist alternative H_1. Dakle, na razini značajnosti od 5% ne možemo tvrditi da su muškarci u prosjeku fizički aktivniji od žena.

1.2 Uvod u višestruko testiranje hipoteza

Kako bismo dobili uvid u problem višestrukog testiranja, pogledajmo jedan jednostavan primjer.

Primjer 1.2.1. *(Bacanje novčića) Pretpostavimo da želimo testirati hipotezu da je novčić simetričan u odnosu na alternativu da nije simetričan, odnosno, želimo testirati hipoteze*

$$H_0 : p = 1/2$$

$$H_1 : p > 1/2,$$

na razini značajnosti 5%, gdje p označava vjerojatnost da na novčiću padne glava. Neka u 10 bacanja novčića glava padne barem 9 puta. Računamo testnu statistiku

$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1 - p_0)}} \cdot \sqrt{n} = \frac{0.95 - 0.5}{\sqrt{1/4}} \cdot \sqrt{10} = 2.846.$$

Kritično područje opet izvodimo iz centralnog graničnog teorema

$$C = [z_{0.05}, +\infty\rangle,$$

a budući da je $z_{0.05} = 1.65$, dakle vrijednost naše testne statistike upada u kritično područje pa odbacujemo hipotezu da je novčić simetričan. Problem višestrukog testiranja nastaje ako se ovaj test, koji je prikladan za testiranje simetričnosti jednog novčića, želi koristiti u testiranju simetričnosti više novčića. Zamislimo da želimo testirati simetričnost 100 novčića ovom metodom. Tada bi vjerojatnost da pogrešno odbacimo hipotezu da je barem jedan novčić simetričan bila

$$\alpha_{100} = 1 - (1 - 0.05)^{100} \approx 0.994,$$

a imali bi

$$100 \cdot 0.05 = 5$$

statistički značajnih rezultata.

U statistici dolazi do problema višestrukih usporedbi ili višestrukog testiranja kada se u obzir uzima skup statističkih zaključaka istovremeno. U tom slučaju su greške u zaključivanju puno češće. Problem se javlja i u konstrukciji pouzdanih intervala koji ne sadrže svoje odgovarajuće populacijske parametre ili kod testiranja hipoteza koja pogrešno odbacuju nultu hipotezu. Kako bi se takve greške spriječile, razvijeno je nekoliko statističkih tehnika dopuštajući da se izravno uspoređuju razine značajnosti za pojedinačna i višestruka testiranja. Takve tehnike općenito zahtijevaju manju razinu značajnosti za individualna uspoređivanja kako bi se uzeo u obzir broj zaključaka dobivenih izravnom usporedbom. Izraz "uspoređivanja" u višestrukim uspoređivanjima se odnosi na usporedbu dviju grupa, grupe pod tretmanom i kontrolne grupe. "Višestruka uspoređivanja" nastaju kada statistička analiza obuhvaća niz formalnih usporedbi s pretpostavkom da se obrati pozornost na najveće razlike među svim napravljenim usporedbama. Neuspjeh u korekciji višestrukih uspoređivanja može imati važne stvarne posljedice, kao što je prikazano u sljedećim primjerima:

Primjer 1.2.2. *Pretpostavimo da uspoređujemo novi način poučavanja pisanja za učenike sa standardnim načinom poučavanja pisanja. Učenike, koji su ovisno o načinu poučavanja pisanja podijeljeni u dvije grupe, možemo uspoređivati u vidu gramatike, izgovora, organizacije, sadržaja itd. Što se više svojstava uspoređuje, to je veća vjerojatnost da će tretirana i kontrolna grupa pokazati razlike u barem jednom odabranom svojstvu.*

Primjer 1.2.3. *Pretpostavimo da razmatramo učinkovitost lijeka u smislu smanjenja bilo kojeg od simptoma bolesti. Što više simptoma uzimamo u obzir, veća je vjerojatnost da će lijek pokazati napredak u odnosu na postojeće lijekove, odnosno, veća je vjerojatnost da će primjena novog lijeka rezultirati manjom uzoračkom frekvencijom barem za jedan od simptoma bolesti.*

Primjer 1.2.4. *Pretpostavimo sada da ispitujemo sigurnost lijeka u pogledu pojave različitih vrsta nuspojava. Što više tipova nuspojava uzimamo u obzir, veća je vjerojatnost da će novi lijek pokazati manju sigurnost u odnosu na postojeće lijekove, odnosno, veća je vjerojatnost da će primjena novog lijeka rezultirati većom uzoračkom frekvencijom barem za jednu od nuspojava.*

U sva tri primjera kako broj usporedbi raste, raste i vjerojatnost da će se uspoređivane grupe razlikovati u barem jednom svojstvu. Na primjer, ako se izvodi jedan test na razini značajnosti od 5% , tada je samo 5% vjerojatnosti za pogrešno odbacivanje nul hipoteze ukoliko je ona istinita. Za 100 testova gdje su sve nul hipoteze istinite očekivani broj pogrešnih odbacivanja je 5. Ako su testovi nezavisni, vjerojatnost najmanje jednog pogrešnog odbacivanja hipoteze jest 99.4% (vidi *primjer 1.2.1*). Ovo su greške prve vrste koje smo definirali na početku. Problem se također javlja kod pouzdanih intervala. Primijetimo da će jedan interval pouzdanosti s 95%−tnom vjerojatnosti pokrivanja vjerojatno

sadržavati populacijski parametar koji bi trebao sadržavati, odnosno, 95% pouzdanih intervala izgrađenih na taj način će sadržavati pravi parametar populacije. Ako se u obzir uzme 100 pouzdanih intervala istovremeno, s vjerojatnošću pokrivanja 0.95 svaki, velika je vjerojatnost da barem jedan interval neće sadržavati svoj populacijski parametar. Očekivani broj takvih nepokrivenih intervala jest 5 i ako su oni nezavisni, vjerojatnost da barem jedan interval ne sadrži populacijski parametar jest 99.4%. Razvijene su tehnike za kontrolu omjera broja hipoteza koje su pogrešno odbačene i ukupnog broja hipoteza. Slično, razvijene su tehnike za namještanje intervala pouzdanosti tako da je vjerojatnost da barem jedan od intervala ne pokriva svoju određenu vrijednost manja od određene granice.

Klasifikacija testova s m hipoteza

Sljedeća tablica nam daje broj počinjenih grešaka prilikom testiranja m nul hipoteza:

	Istinite nul hipoteze	Istinite alternativne hipoteze	Ukupno
Odbačene nul hipoteze	V	S	R
Ne odbačene nul hipoteze	U	T	$m - R$
Ukupno	m_0	$m - m_0$	m

· m - ukupan broj testiranih hipoteza
· m_0 - broj istinitih nul−hipoteza
· $m - m_0$ - broj istinitih alternativnih hipoteza
· V - broj lažno pozitivnih/grešaka tipa I
· S - broj istinito pozitivnih
· T - broj lažno negativnih/grešaka tipa II
· U - broj istinito negativnih
· R - broj odbačenih nul−hipoteza

Uočimo, slučajna varijabla R je opažena, a S, T, U i V nisu.

Definicija 1.2.5. ***Familywise error rate (FWER)****, u oznaci $\bar{\alpha}$ je vjerojatnost pojave jednog ili više lažnih otkrića, odnosno, pogrešaka 1. vrste među svim hipotezama pri testiranju više hipoteza,*

$$\bar{\alpha} = \mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V = 0).$$

Dakle, osiguravanjem $\bar{\alpha} \leq \alpha$, vjerojatnost i jedne pogreške 1. vrste kontroliramo na razini α. Ako izvodimo n nezavisnih usporedbi, $\bar{\alpha}$ je dana s

$$\bar{\alpha} = 1 - (1 - \alpha_{po\_usporedbi})^n.$$

Stoga, osim ako su testovi potpuno zavisni, $\bar{\alpha}$ raste s rastom broja usporedbi. Ako ne pretpostavljamo da su usporedbe nezavisne, tada imamo gornju ogradu za $\bar{\alpha}$, $\bar{\alpha} \leq n \cdot \alpha_{po\_usporedbi}$ što slijedi iz Booleove nejednakosti jer

$$\mathbb{P}(V \geq 1) = \mathbb{P}(\bigcup_{i=1}^{n}(V = i)) \leq \sum_{i=1}^{n} \mathbb{P}(V = i) = n \cdot \alpha_{po\_usporedbi}.$$

Primjer 1.2.6. *Neka je broj usporedbi* $n = 6$ *i neka je* $\alpha_{po\_usporedbi} = 0.05$*, tada imamo*

$$\bar{\alpha} = 1 - (1 - 0.05)^6 = 0.2649 \leq 6 \cdot 0.05 = 0.3$$

Kako bi zadržali propisanu stopu pogrešaka FWER, u analizi koja uključuje više od jedne usporedbe, stopa pogreške za svaku pojedinu usporedbu mora biti manja od α. Booleova nejednakost implicira da ako se svaki test izvodi tako da mu je pogreška 1. vrste α/n, ukupna stopa pogrešaka neće prelaziti α. Ova metoda se naziva *Bonferroni korekcija* i jedna je od najčešće korištenih pristupa u višestrukim uspoređivanjima. U nekim situacijama, Bonferroni korekcija je znatno konzervativnija, tj. stvarni FWER je mnogo manji od zadane razine α. To se događa kada su testne statistike jako zavisne, a u ekstremnim slučajevima, kada su testovi potpuno zavisni, FWER bez prilagodbe višestrukom uspoređivanju, odnosno ne održavajući $\bar{\alpha}$ ispod odgovarajuće granice, jednak je stopi pogreške za pojedinačni test. Zbog konzervativnosti Bonferroni korekcije i sličnih jednostavnih tehnika, velika se pozornost usmjerila na razvoj boljih tehnika tako da se može održati ukupna stopa lažno pozitivnih, odnosno, grešaka prve vrste bez povećavanja lažnonegativnih, grešaka druge vrste. Takve metode se mogu podijeliti u više kategorija. Jednu kategoriju čine metode u kojima se može dokazati da ukupna stopa pogreške nikada ne prelazi određenu vrijednost α, odnosno za bilo koju konfiguraciju istinitih i neistinitih nul-hipoteza. Za takve metode kažemo da kontroliraju FWER *u jakom smislu*. Sljedeću kategoriju čine metode koje kontroliraju FWER samo u slučaju kada su sve nul-hipoteze istinite, tj. kada je $m_0 = m$. Tada govorimo o metodama koje kontroliraju FWER *u slabom smislu*. Empirijske metode koje kontroliraju omjer grešaka 1. vrste koristeći korelacije i karakteristike distribucija promatranih podataka čine treću kategoriju. Pojavom računalnih metoda uzorkovanja kao što su bootstrap i Monte Carlo simulacije, poboljšale su se metode u zadnjoj kategoriji.

1.3 Kontrola grešaka

Slijedi kratak pregled nekih starih rješenja koja osiguravaju snažnu razinu α za kontrolnu veličinu FWER-a te neke novije metode.

Stare metode

Bonferronijeva procedura

Neka za $i = 1, ..., m$ vrijede hipoteze H_i i neka su p-vrijednosti $p_i \sim U(0, 1)$ za $i = 1, ..., m$.

Teorem 1.3.1. *(Bonferronijeva procedura) Ako hipotezu H_i za $i = 1, ..., m$ odbacujemo kada je $p_i \leq \alpha/m$, gdje je p_i p-vrijednosti za testiranje H_i, tada FWER za simultana testiranja $H_1, ..., H_m$ zadovoljava nejednakost*

$$FWER \leq \alpha.$$

Dokaz. Neka je $I \subseteq \{1, ..., m\}$. Tada je

$$\begin{aligned} FWER &= P\{\text{odbacimo } H_i, \text{ za } i \in I\} \leq \sum_{i \in I} P\{\text{odbacimo } H_i\} = \\ &= \sum_{i \in I} P\{p_i \leq \alpha/m\} \leq \sum_{i \in I} \alpha/m \leq |I|\alpha/m \leq \alpha. \end{aligned} \tag{1.2}$$

□

Šidákova procedura

Šidákova je procedura primjenjiva kada su testne statistike nezavisne. Tada testiramo svaku hipotezu na nivou

$$\alpha_{SID} = 1 - (1 - \alpha)^{1/n}.$$

Ovaj test je snažniji od Bonferronijevog, ali dobitak je mali, a postupak je manje općenit od Bonferronijevog jer zahtijeva nezavisnost.

Tukeyeva procedura

Tukeyeva se procedura može primjenjivati samo za sparene usporedbe. Ona podrazumijeva nezavisnost opažanja koja se testiraju te jednako odstupanje u svim opažanjima (homoskedastičnost). Ovaj postupak za svaki par opažanja izračunava raspon studentizirane statistike: $\frac{y_A - y_B}{SE}$ gdje je $y_B \leq y_A$, a SE je standardna pogreška u ispitivanim podacima.

Nove metode

Holmova procedura

Holmova se procedura uobičajeno navodi u terminima p-vrijednosti $p_{(1)}, ..., p_{(m)}$ m individualnih testova. Neka su $p_{(1)} \leq ... \leq p_{(m)}$ uređene p-vrijednosti i neka su $H_{(1)}, ..., H_{(m)}$ odgovarajuće nul hipoteze. Holmova procedura je definirana korak po korak, *engl.stepwise* na sljedeći način:

1. korak Ako je $p_{(1)} \geq \alpha/m$, prihvati $H_{(1)}, ..., H_{(m)}$ i stani. Ako je $p_{(1)} < \alpha/m$, odbaci $H_{(1)}$ i testiraj ostalih $m-1$ hipoteza na razini $\alpha/(m-1)$.
2. korak Ako je $p_{(1)} < \alpha/m$, a $p_{(2)} \geq \alpha/(m-1)$, prihvati $H_{(2)}, ..., H_{(m)}$ i stani. Ako je $p_{(1)} < \alpha/m$ i $p_{(2)} < \alpha/(m-1)$, odbaci $H_{(2)}$ uz $H_{(1)}$ i testiraj ostalih $m-2$ hipoteza na razini $\alpha/(m-2)$.
...

Teorem 1.3.2. *Holmova procedura zadovoljava nejednakost*

$$FWER \leq \alpha.$$

Dokaz. Pretpostavimo da je H_i za $i \in I$ skup istinitih hipoteza. Neka je j najmanji, slučajno odabrani indeks tako da je zadovoljeno

$$p_{(j)} = \min_{i \in I} p_i.$$

Primijetimo da je $j \leq m - |I| + 1$. Holmova procedura pogrešno odbacuje hipotezu ako je

$$p_{(1)} \leq \alpha/m, p_{(2)} \leq \alpha/(m-1), ..., p_{(j)} \leq \alpha/(m-j+1),$$

što povlači da je

$$\min_{i \in I} p_i = p_{(j)} \leq \alpha/(m-j+1) \leq \alpha/|I|.$$

Stoga, prema Bonferronijevoj nejednakosti, vjerojatnost lažnog odbacivanja je ograničena odozgo s

$$\mathbb{P}\{\min_{i \in I} p_i \leq \alpha/|I|\} \leq \sum_{i \in I} \mathbb{P}\{p_i \leq \alpha/|I|\} \leq \alpha.$$

□

Osim stepwise procedure, koriste se još i tzv. Holmove *stepdown* i *stepup* procedure. Opišimo stepdown proceduru. Neka su

$$\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_m$$

konstante. Ako je $p_{(1)} \geq \alpha_1$, prihvatimo sve hipoteze. Inače, za $r = 1, ..., m$, odbacimo hipoteze $H_1, ..., H_m$ ako je

$$p_1 < \alpha_1, ..., p_r < \alpha_r.$$

Dakle, stepdown procedura počinje s najznačajnijom p-vrijednošću i nastavlja s odbacivanjem hipoteza sve dok su odgovarajuće p-vrijednosti male. Holmova procedura koristi $\alpha_i = \alpha/(m-i+1)$. Ova procedura je bolja od Bonferronijeve iz razloga što kontrolira FWER za svih m hipoteza na nivou α u jakom smislu. U ovoj se proceduri testira svaki presjek hipoteza pomoću jednostavnog Bonferronijevog testa. Takav postupak testiranja

naziva se *postupak zatvorenog testiranja* ili *engl. closed testing procedure*. S druge strane, stepup procedura započinje s najmanje značajnom p-vrijednošću. Neka su

$$\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_m$$

konstante. Odbacimo sve hipoteze ako je $p_{(m)} < \alpha_m$. Inače, za $r = m, m-1, ..., 1$, odbacimo hipoteze $H_1, ..., H_m$ ako je

$$p_{(m)} \geq \alpha_m, ..., p_{(r+1)} \geq \alpha_{r+1}, \text{ a} p_{(r)} < \alpha_r.$$

Hochbergova step-up procedura

Opišimo sada Hochbergovu proceduru koja je vrlo slična Holmovoj. Neka su $p_{(1)}, ..., p_{(m)}$ uređene p-vrijednosti i neka su $H_{(1)}, ..., H_{(m)}$ odgovarajuće hipoteze. Za zadani α, neka je r najveći k takav da vrijedi

$$p_k \leq \frac{\alpha}{m+1-k}.$$

Odbacimo nul-hipoteze $H_{(1)}, ..., H_{(r)}$. Hochbergova procedura je snažnija od Holmove. Ipak, dok se Holmova temelji na Bonferronijevoj korekciji bez ograničenja na zajedničke distribucije testnih statistika, Hochbergova procedura se koristi samo u slučaju nezavisnosti (i za neke oblike pozitivne zavisnosti).

Dunnettova korekcija

Dunnettova korekcija je još jedna od metoda za kontrolu FWER-a, poznata je i kao *Dunnettov test*. Ova je procedura manje konzervativna od Bonferronijeve korekcije. U njoj uspoređujemo k grupa s istom kontrolnom grupom.

Ostali postupci

Ostale naprednije procedure koje osiguravaju snažnu razinu α kontrole FWER-a, uključuju *Maximum modulum test*. Također, treba napomenuti da postoji mnogo alternativa u pokušaju kontroliranja FWER-a. Najpoznatija među njima jest *false discovery rate* (FDR) koja rješava mnoge probleme zaključivanja velikog razmjera na praktičniji način. O toj metodi ćemo govoriti u idućem poglavlju.

Primjer 1.3.3. *Promotrimo randomizirano kliničko ispitivanje novog lijeka protiv depresije. Osobe koje su podvrgnute ispitivanju podijeljene su u tri grupe:*
· grupa koja je dobila postojeći lijek
· grupa koja je dobila novi lijek
· grupa koja je dobila placebo

U takvom bi dizajnu mogli testirati više stvari. Primjerice, smanjuju li se simptomi depresije u većoj mjeri za one osobe koje koriste novi lijek u odnosu na stari lijek. Nadalje, mogli bi ispitivati jesu li primjećene neke nuspojave pri uzimanju novog lijeka (npr. pospanost, smanjen seksualni nagon, suha usta). U tom slučaju identificiramo dvije familije:
1. učinak lijeka na simptome depresije
2. pojava nuspojava
Odredimo nivo značajnosti α (uglavnom je to 0.05) za svaku familiju i kontroliramo FWER korištenjem odgovarajuće procedure za višestruko uspoređivanje. Za prvu familiju, učinak antidepresiva na simptome depresije, mogli bi zajednički kontrolirati uspoređivanje po parovima između skupina korištenjem tehnika kao što je Tukeyev range test. Također, ovdje bi mogli primijeniti i Bonferronijevu korekciju jer imamo samo tri testa (tri usporedbe za simptome depresije). Kod druge familije imamo tri usporedbe za svaku nuspojavu i to, omogućujući svakoj nuspojavi vlastitu razinu α, zbog

$$\bar{\alpha} = 1 - (1 - 0.05)^9 = 1 - 0.63 = 0.37,$$

rezultira s vjerojatnošću od 37% *za počinjenje najmanje jedne greške tipa I. Imajući ukupno* 9 *hipoteza, Bonferronijeva korekcija bi u ovom slučaju bila previše konzervativna; moćniji bi alati bili Tukeyev range test ili Holmova metoda. Tada bi α podijelili na* 3, $0.05/3 = 0.0167$ *i dodijelili bi* 0.0167 *svakoj proceduri za višestruko uspoređivanje nuspojava. U slučaju Tukeyeva range testa, kritična vrijednost q, studentizirana rang statistika, bi na taj način bila temeljena na vrijednosti α od* 0.0167.

Poglavlje 2

FDR metoda

2.1 Statističke zablude i razvitak FDR-a

U svom članku *Statistical "discoveries" and effect-size estimation* objavljenom 1989. godine Branko Sorić skreće pažnju na neke zablude u statistici koje dovode do velikog broja netočnih rezultata istraživanja. U svom radu on koristi termin statističko "otkriće" koji odgovara odbacivanju nul hipoteze, odnosno dobivanju pouzdanih intervala koji ne sadrže nulu. Pretpostavimo da istraživači naprave veliki broj n nezavisnih eksperimenata uz odabranu razinu statističke značajnosti α i da je dobiveno r otkrića, tj. statistički značajnih rezultata s p-vrijednosti manjom ili jednakom α. Nadalje, pretpostavimo da je u n eksperimenata broj istinitih nul hipoteza a koji je nepoznat. Sorić nas upozorava na to da će dobivenih r otkrića istraživači objaviti, no tu postoji opasnost da bi veliki dio znanosti zbog toga mogli smatrati neistinitim osim ako bi se moglo dokazati da je omjer lažnih otkrića i svih proglašenih otkrića zanemarivo mali. Taj omjer označava sa Q i uz navedene oznake, jednak je

$$Q = \frac{\alpha \cdot a}{r}.$$

Omjer Q je nepoznat budući da je broj istinitih nul hipoteza a nepoznat. U slučaju pouzdanih intervala, Sorić napominje da ako imamo velik broj n 95%-tnih pouzdanih intervala, tada 95% njih sadrži populacijski parametar (npr. razliku između aritmetičkih sredina populacija), ali isto ne vrijedi i za podskup od r 95%-tnih pouzdanih intervala. Odnosno, udio pogrešne procjene parametara u skupu od n c%-tnih pouzdanih intervala jest α, ali nije nužno α u skupu od r c%-tnih pouzdanih intervala koji ne sadrže nulu. Udio pogrešne intervalne procjene veličine učinka u skupu c%-tnih pouzdanih intervala koji ne sadrže nulu Sorić označava sa E i označava

$$E = \frac{\alpha \cdot n}{r}.$$

Omjer Q kojeg je uveo Branko Sorić zapravo je ekvivalentan omjeru čiju su kontrolu predložili Benjamini i Hochberg u svom radu *Controlling the false discovery rate: a practical and powerful approach to multiple testing* iz 1995 godine. Taj omjer nazvan je stopom lažnih otkrića, *engl. false discovery rate*, kraće FDR i njegova je kontrola zamijenila dotadašnju kontrolu FWER-a. FDR procedure su dizajnirane kako bi kontrolirale očekivani udio pogrešno odbačenih nul hipoteza, odnosno lažnih otkrića. Dakle, za razliku od procedura koje kontroliraju FWER i nastoje smanjiti vjerojatnost pojave ijednog lažnog otkrića, procedure koje kontroliraju FDR nastoje smanjiti očekivani udio lažnih otkrića. Na taj način FDR procedure imaju veću snagu na uštrb povećane stope pogrešaka 1. vrste. Vjeruje se da moderna raširena upotreba FDR-a potječe i da je motivirana razvojem tehnologija koje dopuštaju prikupljanje i analizu velikog broja različitih varijabli u pojedinim slučajevima (jedinkama). Primjerice, stupanj ekspresije svakog od 10000 različitih gena u 100 osoba. Do kasnih osamdesetih i devedesetih godina prošlog stoljeća razvojem znanosti došlo je do brzog prikupljanja podataka te s rastom računalne snage omogućeno je jednostavno izvođenje stotine i tisuće statističkih testova na danom skupu podataka. Kako su tehnologije visoke propusnosti postajale sve uobičajenije, tehnološka i/ili financijska ograničenja su dovela istraživače do prikupljanja skupova podataka s relativno malom veličinom uzorka, točnije, samo se par jedinki testiralo, a mjerio se velik broj varijabli po uzorku. U tim skupovima podataka premalo je mjerenih varijabli pokazalo statističku značajnost nakon klasične prilagodbe višestrukom testiranju sa standardnim procedurama za višestruke usporedbe. To je u znanstvenoj zajednici stvorilo potrebu za napuštanjem FWER-a neprilagođenog višestrukim testiranjem hipoteza kako bi na druge načine istaknuli i svrstali u publikacije one varijable koje pokazuju statističku značajnost u pojedinačnom slučaju ili tretmane koji bi inače bili odbačeni kao neznačajni. Kao odgovor na taj problem, predložene su razne stope pogrešaka, manje konzervativne od FWER-a u označavanju eventualno bitnih zapažanja i koje su postale često korištene. Kao nuspojava, standardna prilagodba višestrukom testiranju je doslovno nestala iz svih publikacija osim onih koje iznose rezultate s velikim veličinama uzorka. Koncept stope pogrešnih otkrića formalno su opisali Yoav Benjamini i Yosef Hochberg kao manje konzervativan i vjerojatno više odgovarajući pristup za identificiranje važnijih među mnogim trivijalnim testiranim učincima. Kako je FDR bio prva alternativa FWER-u koja je stekla široko prihvaćanje u mnogim znanstvenim područjima, osobito u prirodnim znanostima, bio je veoma utjecajan. Benjaminijev i Hochbergov rad iz 1995. godine jedan je od najcitiranijih statističkih radova u povijesti, ali i u znanosti općenito.

2.2 Definicija

Promotrimo tablicu iz poglavlja 1 koja nam daje broj počinjenih grešaka pri testiranju m nul hipoteza. Omjer grešaka počinjenih pogrešnim odbacivanjem nul hipoteza vidi se iz

slučajne varijable

$$Q = \frac{V}{V+S} = \frac{V}{R}.$$

Prirodno definiramo $Q = 0$ za $V + S = 0$ pa ponekad pišemo

$$Q = \frac{V}{R}\mathbb{1}_{\{R>0\}}.$$

Definiramo FDR kao

$$FDR := E(Q) = E(\frac{V}{V+S}) = E(\frac{V}{R}).$$

Iz ove definicije možemo primijetiti da je FDR jednak FWER-u ako su sve nul hipoteze istinite. Budući da su nam tada realizirane vrijednosti $s = 0$ i $v = r$, u slučaju kada je $v = 0$ tada je i $Q = 0$, a ako je $v > 0$ onda je $Q = 1$. Dakle, u tom slučaju vrijedi

$$FWER = P(V \geq 1) = E(Q) = FDR.$$

Prema tome, kontrola FDR-a implicira kontrolu FWER-a u slabom smislu. U slučaju kada nisu sve nul hipoteze istinite, odnosno, kada je $m_0 < m$ slijedi da je FDR $\leq$ FWER. Zaista, u tom slučaju, za $v > 0$ vrijedi

$$\frac{v}{r} \leq 1 \text{ povlači da je } \mathbb{1}_{\{V\geq 1\}} \geq Q.$$

Računanjem očekivanja obiju strana dobivamo

$$P\{V \geq 1\} \geq Q_e.$$

Kao rezultat dobivamo da svaka procedura koja kontrolira FWER također kontrolira i FDR. Procedura koja kontrolira samo FDR može biti manje konzervativna i može se očekivati dobitak na snazi procedure. Posebice, što je veći broj neistinitih nul hipoteza, to je veći S pa je veća i razlika između omjera grešaka, FDR-a i FWER-a. Dakle, potencijal za rast u snazi procedure je veći što je veći broj neistinitih nul hipoteza. Vratimo se sada na problem koji je uočio Sorić. Iako se Q ne može egzaktno izračunati, Sorić nam daje gornju granicu za Q koju označava s Q_{max} i vrijedi

$$Q_{max} = Q^{\alpha}_{max} = \frac{\alpha}{1-\alpha}(\frac{m}{R}\mathbb{1}_{\{R>0\}} - 1).$$

Primijetimo da za sve α, uz $R > 0$ vrijedi

$$Q_{max} = Q^{\alpha}_{max} = \frac{\frac{\alpha}{1-\alpha}}{\frac{\frac{R}{m}}{1-\frac{R}{m}}}.$$

Sorić je pokazao da za veliki broj m vrijedi

$$Q \leq Q_{max}.$$

Pokažimo da ista nejednakost vrijedi i uzimanjem očekivanja.

Propozicija 2.2.1. *Za svaki* $\alpha \in (0, 1)$ *vrijedi*

$$\mathbb{E}(Q) \leq \mathbb{E}(Q_{max}).$$

Dokaz. Očito vrijedi nejednakost

$$m_0 - V \leq min\{m - 1, m - R\}.$$

Uzimanjem očekivanja dobivamo

$$m_0(1 - \alpha) \leq m - \mathbb{E}(R \vee 1).$$

Primjenom Jensenove nejednakosti vrijedi

$$\mathbb{E}(\frac{\mathbb{E}(R \vee 1)}{R \vee 1}) \geq 1$$

pa stoga imamo

$$\mathbb{E}Q = \mathbb{E}(\frac{m_0\alpha}{R \vee 1}) \leq \frac{\alpha}{1-\alpha}\mathbb{E}(\frac{m}{R \vee 1} - \frac{\mathbb{E}(R \vee 1)}{R \vee 1}) \leq \frac{\alpha}{1-\alpha}\mathbb{E}(\frac{m}{R \vee 1} - 1) = \mathbb{E}Q_{max}.$$

□

Pogledajmo jedan konkretan primjer za izračunavanje veličine Q_{max}.

Primjer 2.2.2. *Neka je zadana razina značajnosti* $\alpha = 0.05$*, neka je ukupan broj testiranja* $m = 10000$ *i neka je nađeno* $r = 7000$ *statističkih otkrića. Tada je*

$$Q_{max} = \frac{\frac{0.05}{1-0.05}}{\frac{\frac{7000}{10000}}{1-\frac{7000}{10000}}} \approx 0.023.$$

Dakle, možemo reći da svih r nul hipoteza možemo odbaciti poznavajući maksimalan rizik, odnosno Q_{max}*.*

2.3 Važnost kontrole FDR-a

Sljedeći primjeri pokazuju važnost kontrole FDR-a u čestim situacijama. Zajedničko im je da indiciraju poželjnost velikog broja statističkih otkrića. Jedan tip problema kod višestrukog uspoređivanja uključuje cjelokupnu odluku koja je temeljena na višestrukom zaključivanju. Primjer problema ovog tipa jest "problem višestrukih krajnjih točaka". Radi se o uspoređivanju između kontrolne i tretirane grupe koje rezultira različitim pogledima na neki efekt. Primjerice, novi lijek za poticanje cirkulacije pokazuje veću djelotvornost no uzrokuje veće

opterećenje srca. U ovakvim primjerima problem je u donošenju konačne odluke, je li bolje preporučiti novi tretman ili ostati pri standardnom. Ovdje su otkrića odbacivanja nul hipoteza kojima se tvrdi da tretman nije bolji od standardnog na određenim krajnjim točkama. Zaključci o različitim pogledima na dobrobit novog tretmana su od interesa sami po sebi, no do konačnog se zaključka glede novog tretmana dolazi uzimanjem u obzir svih otkrića. Prema tome, želimo napraviti što je više moguće otkrića zadržavajući kontrolu FDR-a koji će dovesti do zaključka u korist novog tretmana. Kontrola vjerojatnosti grešaka nije potrebna budući da mali omjer grešaka neće promijeniti ispravnost konačnog zaključka. Drugi tip problema uključuje donošenje višestrukih pojedinačnih zaključaka gdje nije potrebno donošenje konačne odluke. Primjer ovog tipa jest problem podgrupa gdje se uspoređuju 2 tretmana u različitim podgrupama i zasebne preporuke na određenim tretmanima se moraju izvršiti u svim podgrupama. I u ovom slučaju želimo otkriti što više značajnih razlika koji će nas dovesti do odgovarajućih odluka, ali uz dopuštanje predodređenih omjera grešaka, odnosno korištenjem FDR kontrolne procedure. Treći tip uključuje probleme gdje su višestruki potencijalni efekti uklonjeni kako bi uništili nul efekte. Jedan primjer je uklanjanje raznih kemikalija radi potencijalnog razvitka lijekova. Drugi primjer je testiranje više faktora u eksperimentalnom dizajnu (2^k). U takvim primjerima želimo dobiti što je moguće više otkrića (kandidate za razvitak lijekova, faktore koji utječu na kvalitetu proizvoda), ali ponovo želimo kontrolirati FDR jer bi preveliki udio lažnih zaključaka otežao iduću fazu analize utvrđivanja.

2.4 Procedura za kontrolu FDR-a

Neka su $H_1, ..., H_m$ nul hipoteze koje testiramo s odgovarajućim p-vrijednostima $p_1, ..., p_m$. Neka su $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ uređene p-vrijednosti i označimo sa $H_{(i)}$ nul hipotezu koja odgovara $p_{(i)}$. Definirajmo sljedeću proceduru za višestruko testiranje: *neka je k najveći* $i = 1, 2, ..., m$ *za koji je* $p_{(i)} \leq \frac{i}{m} \cdot q^*$. *Odbacujemo sve* $H_{(i)}$ *za* $i = 1, 2, ..., k$.

Teorem 2.4.1. *Za nezavisne statističke testove i za bilo koju strukturu lažnih nul hipoteza definirana procedura kontrolira FDR na nivou* q^*.

Napomena 2.4.2. *Pod pojmom "procedura kontrolira FDR na nivou* q^**" podrazumijevamo da je FDR*$\leq q^*$.

Iskažimo i dokažimo sada lemu iz koje slijedi ovaj teorem.

Lema 2.4.3. *Za bilo kojih* $0 \leq m_0 \leq m$ *nezavisnih p-vrijednosti koje odgovaraju istinitim nul hipotezama i za bilo kojih* $m_1 = m - m_0$ *p-vrijednosti koje odgovaraju lažnim nul hipotezama, procedura za višestruka uspoređivanja definirana iznad zadovoljava nejednakost*

$$\mathbb{E}(Q|P_{m_0+1} = p_1, ..., P_m = p_{m1}) \leq \frac{m_0}{m} \cdot q^*.$$

Sada pretpostavimo da je $m_1 = m - m_0$ hipoteza neistinito. Koju god da zajedničku distribuciju $P_1^n, ..., P_{m1}^n$ koja odgovara ovoj lažnoj hipotezi uzmemo, integriranjem zadnje nejednakosti, dobivamo

$$\mathbb{E}(Q) \leq \frac{m_0}{m} \cdot q^*$$

i FDR na taj način kontroliramo.

Dokaz. Dokazujemo matematičkom indukcijom po m. Kako je baza indukcije za $m = 1$ zadovoljena, pretpostavimo da tvrdnja vrijedi za $m' \leq m$. Pokazat ćemo da vrijedi i za $m + 1$. Ako je $m_0 = 0$, sve nul hipoteze su neistinite pa je $Q = 0$ i

$$\mathbb{E}(Q|P_1 = p_1, ..., P_m = p_m) = 0 \leq \frac{m_0}{m+1} \cdot q^*.$$

Ako je $m_0 > 0$, označimo s P_i' za $i = 1, 2, ..., m_0$ p-vrijednosti koje odgovaraju istinitim nul hipotezama i neka je $P'_{(m_0)}$ najveća među tim p-vrijednostima. P_i' su nezavisne slučajne varijable i $P_i' \sim U(0, 1)$. Radi jednostavnije notacije pretpostavimo da je $p_1 \leq p_2 \leq ... \leq p_{m_1}$, gdje su $p_1, ..., p_{m_1}$ p-vrijednosti koje odgovaraju neistinitim nul hipotezama. Označimo s j_0 najveći j, $0 \leq j \leq m_1$ koji zadovoljava nejednakost

$$p_j \leq \frac{m_0 + j}{m+1} \cdot q^*. \tag{2.1}$$

Označimo desnu stranu ove nejednakosti s p'' za $j = j_0$. Sada, uz $P'_{(m_0)} = p$ imamo

$$\begin{aligned}\mathbb{E}(Q|P_{m_0+1} = p_1, ..., P_m = p_{m_1}) = &\int_0^{p''} \mathbb{E}(Q|P'_{(m_0)} = p, P_{m_0+1} = p_1, ..., P_m = p_{m_1}) \cdot f'_{P_{(m_0)}}(p)\mathrm{d}p \\ &+ \int_{p'''}^{1} \mathbb{E}(Q|P'_{(m_0)} = p, P_{m_0+1} = p_1, ..., P_m = p_{m_1}) \cdot f'_{P_{(m_0)}}(p)\mathrm{d}p,\end{aligned} \tag{2.2}$$

gdje je $f'_{P_{(m_0)}}(p) = m_0 \cdot p^{m_0-1}$. U prvom članu desne strane relacije (2.2) imamo da je $p \leq p''$. Stoga su sve $m_0 + j_0$ nul hipoteze odbačene i $Q = m_0/(m_0 + j_0)$. Računanjem integrala pa primjenom nejednakosti (2.1) dobivamo

$$\frac{m_0}{m_0 + j_0}(p'')^{m_0} \leq \frac{m_0}{m_0 + j_0} \frac{m_0 + j_0}{m+1} q^*(p'')^{m_0-1} = \frac{m_0}{m+1} q^*(p'')^{m_0-1}. \tag{2.3}$$

U drugom članu desne strane relacije (2.2) razmotrimo posebno svaki $p_{j_0} < p_j \leq P'_{(m_0)} = p < p_{j+1}$, zajedno s $p_{j_0} \leq p'' < P'_{(m_0)} = p < p_{j_0+1}$. Primijetimo da se zbog definicije j_0 i p''

nijedna hipoteza čije su odgovarajuće p-vrijednosti $p, p_{j+1}, p_{j+2}, ..., p_{m_1}$ ne može odbaciti. Stoga, kada uzmemo sve nul hipoteze zajedno, i istinite i neistinite i poredamo njihove p-vrijednosti po veličini, hipoteza $H_{(i)}$ se može odbaciti samo ako postoji k, $i \leq k \leq m_0 + j - 1$ za koji je $p_{(k)} \leq [k/(m+1)]q^*$ ili, ekvivalentno,

$$\frac{p_{(k)}}{p} \leq \frac{k}{m_0 + j - 1} \frac{m_0 + j - 1}{(m+1)p} q^*. \tag{2.4}$$

Uz $P'_{(m_0)} = p$, P'_i/p za $i = 1, 2, ..., m_0 - 1$ su nezavisne slučajne varijable iz $U(0,1)$ i p_i/p za $i = 1, 2, ..., j$ su brojevi između 0 i 1 koji odgovaraju neistinitim nul hipotezama. Koristeći nejednakost (2.4) za testiranje $m_0 + j - 1 = m' \leq m$ hipoteza ekvivalentno je korištenjem procedure za kontrolu FDR-a, s time da konstanta $[(m_0 + j - 1)/(m+1)p]q^*$ preuzima ulogu q^*. Korištenjem hipoteze indukcije, imamo

$$\mathbb{E}(Q|P'_{m_0} = p, P_{m_0+1} = p_1, ..., P_m = p_{m_1}) \leq \frac{m_0 - 1}{m_0 + j - 1} \frac{m_0 + j - 1}{(m+1)p} q^* = \frac{m_0 - 1}{(m+1)p} q^*. \tag{2.5}$$

Granica u nejednakosti (2.5) ovisi o p, ali ne ovisi o intervalu $p_j < p < p_{j+1}$ za koji je izračunata, dakle

$$\begin{aligned} \int_{p''}^{1} \mathbb{E}(Q|P'_{(m_0)} = p, P_{m_0+1} = p_1, ..., P_m = p_{m_1}) f'_{P_{(m_0)}}(p)\mathrm{d}p &\leq \int_{p''}^{1} \frac{m_0 - 1}{(m+1)p} q^* m_0 p^{(m_0-1)} \mathrm{d}p \\ &= \frac{m_0}{m+1} q^* \int_{p''}^{1} (m_0 - 1) p^{(m_0-2)} \mathrm{d}p = \frac{m_0}{m+1} q^* (1 - p''^{(m_0-1)}). \end{aligned} \tag{2.6}$$

Sada zbrajanjem nejednakosti (2.3) i (2.6), lema je dokazana. □

Napomena 2.4.4. *Primijetimo da za dokaz teorema 2.4.1 nije potrebna nezavisnost testnih statistika koje odgovaraju neistinitim nul hipotezama.*

Poglavlje 3

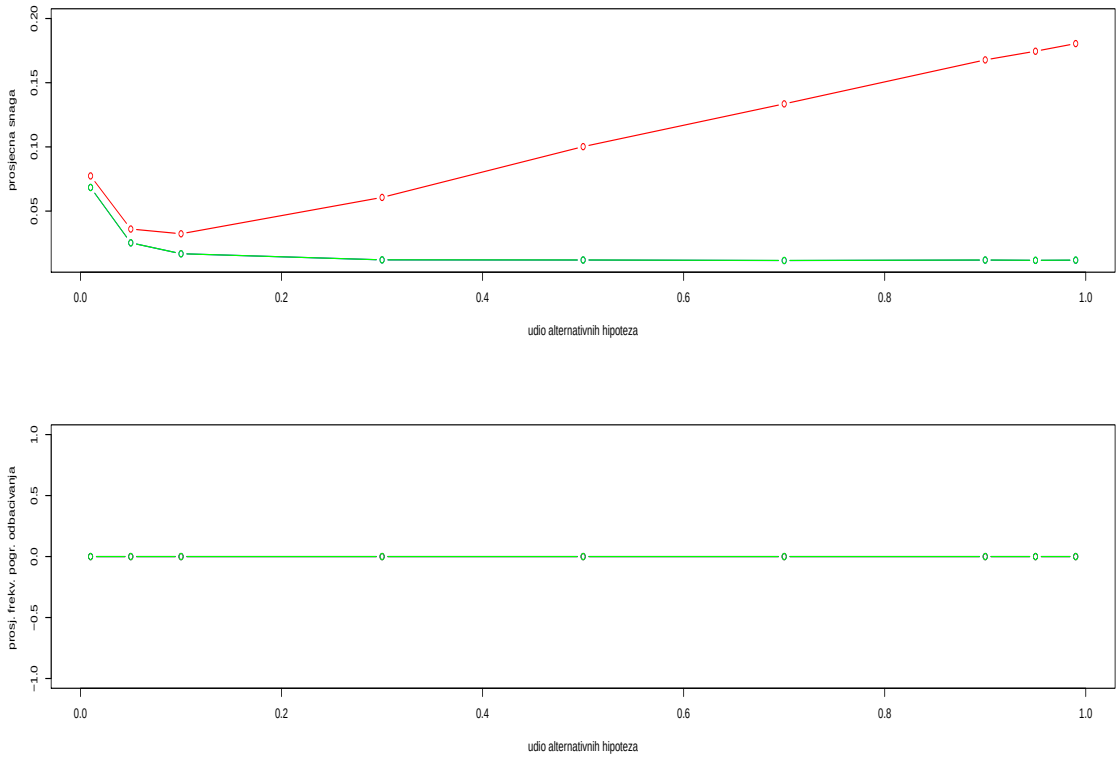
Simulacijska studija

U ovoj simulacijskoj studiji uspoređujemo snagu procedure za kontrolu FDR-a, snagu Bonferronijeve procedure za kontrolu FWER-a te snagu Holmove procedure za kontrolu FWER-a te uspoređujemo i prosječne frekvencije pogrešno odbačenih nul hipoteza za svaku proceduru. U pripremi ove simulacijske studije koristili smo programski jezik R. Uspoređivanje kontrole FDR-a i FWER-a vršimo na nivou q^*, odnosno α i to za $q^* = \alpha$ redom 0.01, 0.05, 0.1 i 0.2. Za svaki nivo variramo omjer $1 - m_0/m$ gdje m_0 označava broj istinitih nul hipoteza, a m ukupan broj hipoteza koji nam je jednak 1000. Za m_0 uzimamo redom 10, 50, 100, 300, 500, 700, 900, 950 i 990. Označimo p-vrijednosti sa p_i, $i = 1, ..., m$. Pretpostavimo da testne statistike za nul hipoteze i alternativne hipoteze imaju $N(\mu, 1)$ distribuciju, tj. imaju normalnu distribuciju s nepoznatim očekivanjem μ i poznatom varijancom $\sigma^2 = 1$. Pod nul hipotezom podrazumijevamo $\mu = 0$, a pod alternativnom $\mu_A = -2$. P-vrijednosti dobivamo testiranjem

$$H_0 : \mu = 0$$

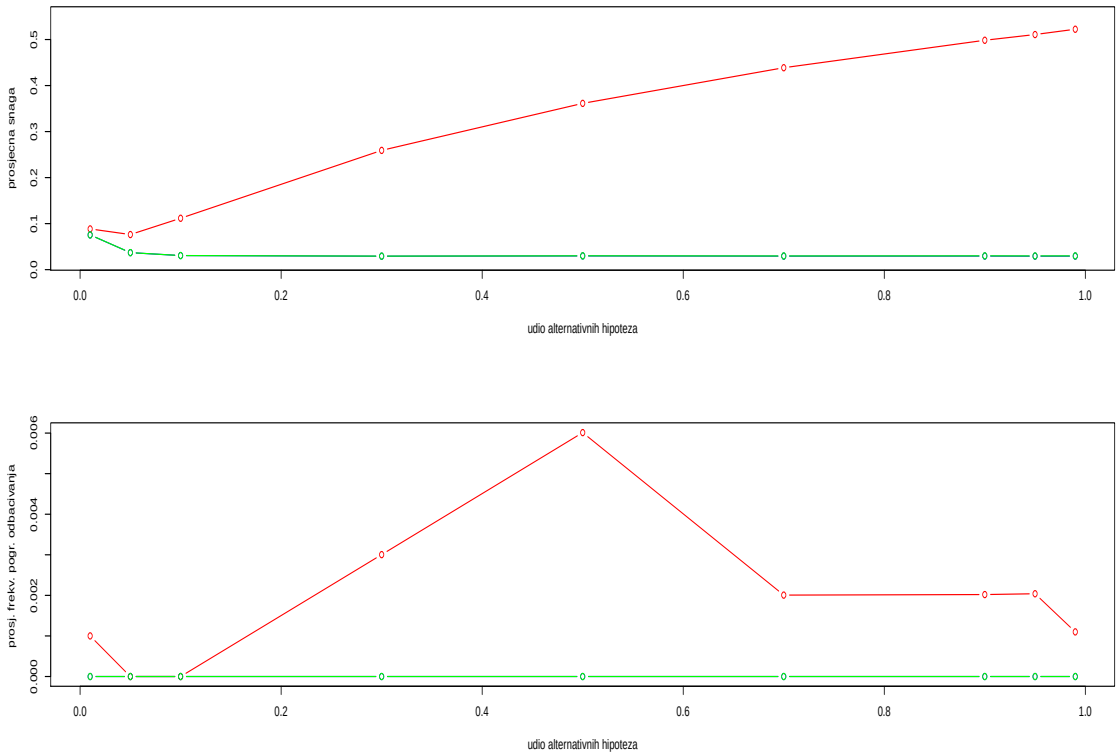
$$H_1 : \mu < 0$$

i p_i za $i \leq m_0$ odgovaraju istinitim nul hipotezama, a za $i > m_0$ odgovaraju istinitim alternativnim hipotezama. Za svaku grupu od $m = 1000$ testova radimo 1000 ponavljanja i na grafu prikazujemo prosječnu snagu svake procedure te prosječnu frekvenciju pogrešnih odbacivanja nul hipoteza koju napravi svaka procedura. Iz grafova možemo vidjeti da je za svaki omjer $1 - m_0/m$, koji neprecizno zovemo udio alternativnih hipoteza u nastavku, i za svaki α prosječna snaga procedure za kontrolu FDR-a veća od prosječne snage procedura za kontrolu FWER-a, uz navedene parametre. No, pri tom je prosječna frekvencija pogrešno odbačenih nul hipoteza veća kod procedure za kontrolu FDR-a nego kod ostalih dviju procedura.



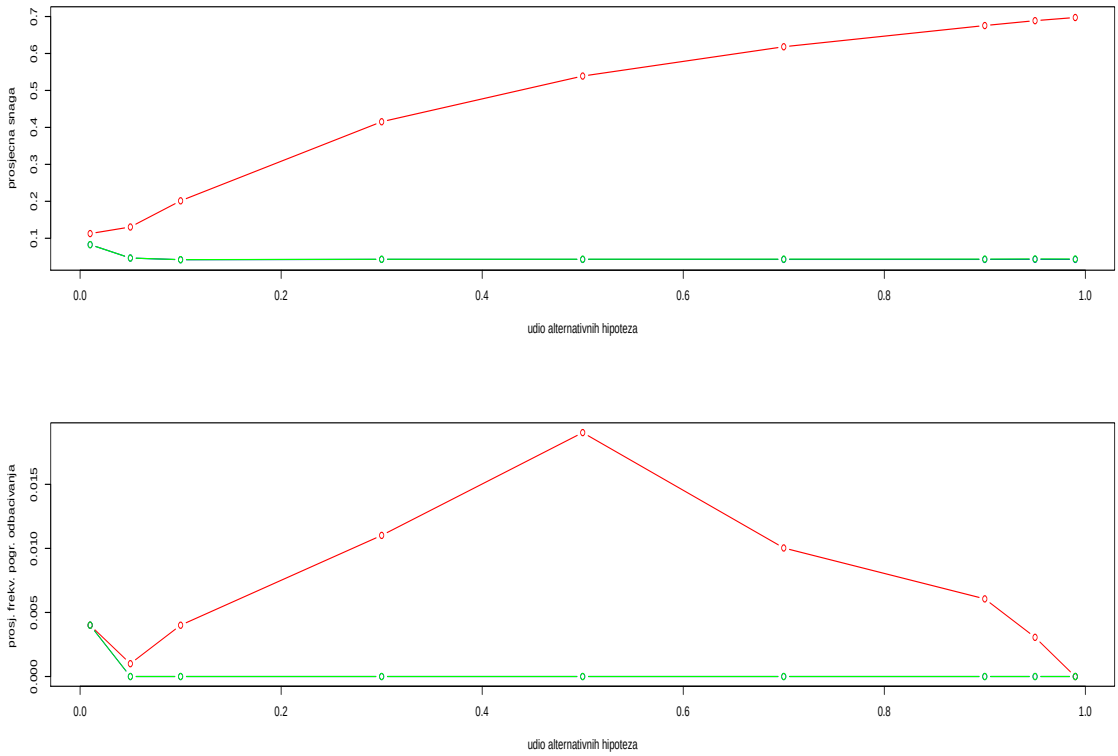
Slika 3.1: prosječna snaga i prosječna frekvencija pogrešno odbačenih nul hipoteza za $q^* = 0.01$, odnosno $\alpha = 0.01$ (crvena linija↔ FDR, zelena linija↔ Holm's i FWER)

Na slici 3.1 gornji graf prikazuje prosječne snage procedura za kontrolu FDR-a, odnosno FWER-a za $q^* = \alpha = 0.01$. Iako smo u simulacijskoj studiji implementirali i Bonferronijevu proceduru za kontrolu FWER-a, rezultati se u odnosu na FDR ne razlikuju značajno od Holmove procedure u navedenom primjeru stoga te rezultate ne možemo vidjeti na grafu. Vidimo da je prosječna snaga procedure za kontrolu FDR-a veća od prosječne snage Holmove (a ujedno i Bonferronijeve) procedure za kontrolu FWER-a. Najmanja razlika u snagama tih procedura jest za omjer $1 - m_0/m$ između 0.01 i 0.05 gdje snage padaju približno istom brzinom. Prosječna snaga procedure za FDR na tom mjestu pada od 0.08 do 0.03, a prosječna snaga Holmove procedure od 0.07 do 0.02. Za udio alternativnih hipoteza od 0.05 do 0.1 snage još uvijek padaju, ali sporije. Od 0.1 nadalje snaga procedure za FDR raste do 0.18 dok snaga Holmove procedure pada do približno 0.02. Na donjem su grafu prosječne frekvencije pogrešno odbačenih nul hipoteza jednake 0 za svaku proceduru te za svaku varijaciju omjera $1 - m_0/m$.



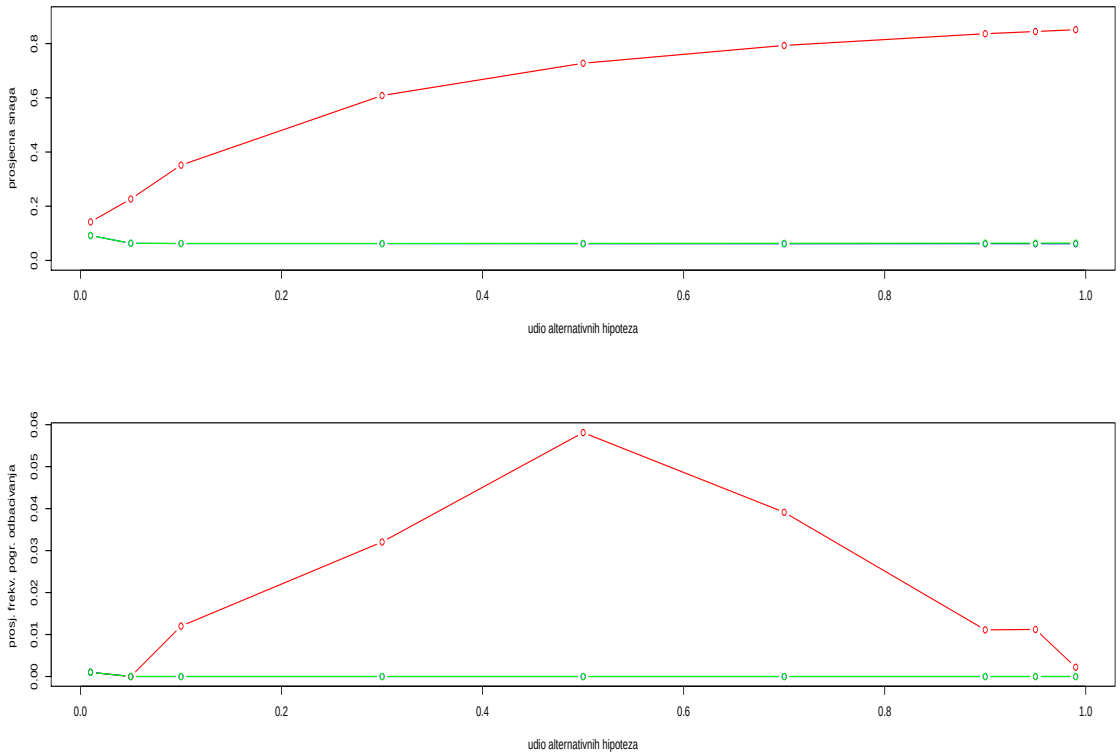
Slika 3.2: prosječna snaga i prosječna frekvencija pogrešno odbačenih nul hipoteza za $q^* = 0.05$, odnosno $\alpha = 0.05$ (crvena linija↔ FDR, zelena linija↔ Holm's i FWER)

Slika 3.2 također na gornjem grafu prikazuje prosječne snage procedura za kontrolu FDR-a, odnosno FWER-a, ali za $q^* = 0.05$, odnosno $\alpha = 0.05$. Prosječna snaga procedure za kontrolu FDR-a pada od 0.09 do 0.08 za udio alternativnih hipoteza od 0.01 do 0.05, a nakon toga nadalje raste sve do 0.51. Prosječna snaga Holmove procedure za kontrolu FWER-a pada s povećanjem udjela alternativnih hipoteza, i to od 0.08 do 0.04 za $1 - m_0/m$ između 0.01 i 0.05, a nakon toga polako pada sve do vrijednosti od 0.03. Donji graf prikazuje prosječnu frekvenciju pogrešno odbačenih nul hipoteza za svaku proceduru i to također za $q^* = 0.05$, odnosno $\alpha = 0.05$. Vidimo da frekvencija pogrešno odbačenih nul hipoteza za svaki udio alternativnih hipoteza za Holmovu (i Bonferronijevu) proceduru iznosi 0, a za proceduru koja kontrolira FDR frekvencija se kreće između 0 (za $1 - m_0/m = 0.05$ i $1 - m_0/m = 0.1$) i 0.006 (za $1 - m_0/m = 0.5$).



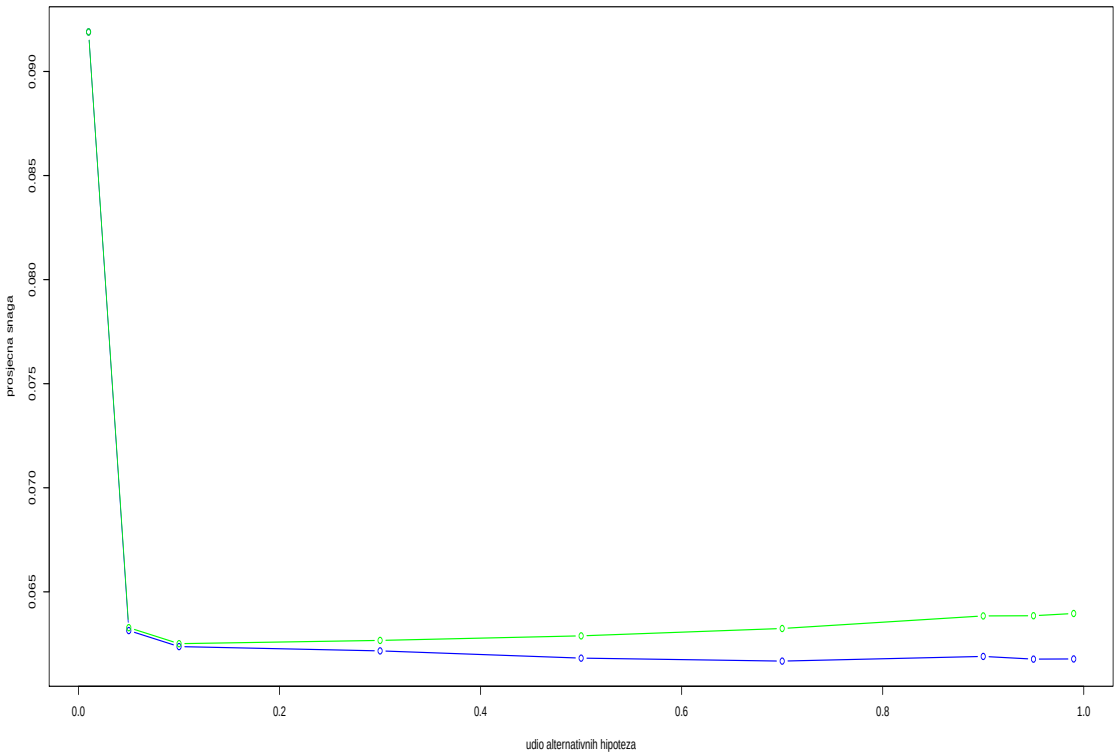
Slika 3.3: prosječna snaga i prosječna frekvencija pogrešno odbačenih nul hipoteza za $q^* = 0.1$, odnosno $\alpha = 0.1$ (crvena linija↔ FDR, zelena linija↔ Holm's i FWER)

Vidimo da je situacija na slici 3.3 vrlo slična onoj na slici 3.2. Ovdje nam je $q^* = \alpha = 0.1$. Prosječna snaga procedure za kontrolu FDR-a raste od 0.1 do 0.7 s porastom udjela alternativnih hipoteza dok prosječna snaga Holmove (i Bonferronijeve) procedure pada od 0.09 do 0.02 za udio alternativnih hipoteza od 0.01 do 0.05 te snaga ostaje na vrijednosti 0.02 s porastom omjera $1 - m_0/m$. Prosječna frekvencija pogrešno odbačenih nul hipoteza od vrijednosti omjera $1 - m_0/m = 0.05$ naviše iznosi 0 za Holmovu (i Bonferronijevu) proceduru,a za vrijednost omjera 0.01 prosječne snage svih procedura iznose 0.004. Procedura za kontrolu FDR-a u prosjeku sve više pogrešnih odbacivanja nul hipoteza radi od omjera $1 - m_0/m = 0.05$ do $1 - m_0/m = 0.5$ gdje postiže vrijednost 0.019, a nakon tog omjera smanjuje se frekvencija do 0 za $1 - m_0/m = 0.99$.



Slika 3.4: prosječna snaga i prosječna frekvencija pogrešno odbačenih nul hipoteza za $q^* = 0.2$, odnosno $\alpha = 0.2$ (crvena linija↔ FDR, zelena linija↔ Holm's i FWER)

Na slici 3.4 za $q^* = \alpha = 0.2$ vidimo da procedura koja kontrolira FDR ima najveću snagu do sada, raste s porastom udjela alternativnih hipoteza od 0.15 za $1 - m_0/m = 0.01$ do 0.85 za $1 - m_0/m = 0.99$. Prosječna snaga Holmove procedure iznosi 0.1 za $1 - m_0/m = 0.01$, a od vrijednosti omjera $1 - m_0/m = 0.05$ naviše snaga joj iznosi 0.06. Frekvencija pogrešno odbačenih nul hipoteza za proceduru koja kontrolira FDR-a raste od 0 (za $1 - m_0/m = 0.05$) do 0.059 (za $1 - m_0/m = 0.5$), a zatim počinje padati do 0.003 (za $1 - m_0/m = 0.99$). Frekvencija pogrešno odbačenih nul hipoteza za Holmovu (i Bonferronijevu) proceduru u našem primjeru ima konstantnu vrijednost 0 za sve varijacije omjera $1 - m_0/m$ osim za $1 - m_0/m = 0.01$ gdje sve procedure imaju jednaku frekvenciju 0.001.



Slika 3.5: prosječne snage Bonferronijeve i Holmove procedure za kontrolu FWER-a za $\alpha = 0.2$ (plava linija↔ Bonferronijeva procedura, zelena linija↔ Holmova procedura)

Slika 3.5 prikazuje usporedbu prosječnih snaga Bonferronijeve i Holmove procedure same za $\alpha = 0.2$. Vidimo da im se snage počinju razlikovati od omjera $1 - m_0/m = 0.05$. S porastom udjela alternativnih hipoteza raste i razlika u snagama tih dviju procedura. Primijetimo da Holmova procedura ima veću snagu od Bonferronijeve, a njihova najveća razlika, dakle za $1 - m_0/m = 0.99$, iznosi približno 0.005. Dakle, iako se snage Bonferronijeve i Holmove procedure razlikuju, njihova međusobno najveća razlika svejedno je vrlo mala u odnosu na razliku u snazi koju čine te dvije procedure sa procedurom za kontrolu FDR-a.

Na kraju ove simulacijske studije možemo zaključiti da je, za naše parametre koje smo naveli na početku ovog poglavlja, snaga procedure koja kontrolira FDR znatno veća od snaga Bonferronijeve i Holmove procedure koje su približno jednakog i to vrlo malog iznosa obzirom na snagu procedure za FDR. Isto tako, razlika u snagama procedure za kontrolu FDR-a i ostalih dviju procedura raste s povećanjem q^*, odnosno α u našim primje-

rima. Osim toga, snaga procedure za kontrolu FDR-a raste s porastom broja alternativnih hipoteza dok su snage ostalih dviju procedura približno konstantne i to vrijedi za svaki q^*, odnosno α. Primijetimo i da u ovoj simulacijskoj studiji snaga procedure za kontrolu FDR-a raste brže za manji udio alternativnih hipoteza, a sporije za veći udio što su q^*, odnosno α veći. Ali, prosječna frekvencija pogrešno odbačenih nul hipoteza kod FDR-a je zato veća nego kod ostalih dviju procedura kojima iznosi 0 u gotovo svim konfiguracijama. No, to nije ništa neočekivano jer za razliku od FWER-a koji predstavlja omjer pogrešno odbačenih nul hipoteza među svim nul hipotezama, FDR označava omjer pogrešno odbačenih nul hipoteza među svim odbačenim nul hipotezama. Uočimo da procedura za kontrolu FDR-a u našim primjerima najviše pogrešaka napravi za omjer alternativnih hipoteza jednak 0.5 dok najmanje napravi za omjere $1 - m_0/m = 0.05$ i $1 - m_0/m = 0.99$.

Bibliografija

[1] B. Basrak, *On the erroneous discovery rate*, neobjavljeni rad, 2013.

[2] Y. Benjamini, Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. R. Statist. Soc. B, 57, 289-300, 1995.

[3] E. L. Lehmann, *Testing statistical hypotheses*, Springer, 2005.

[4] B. Sorić, *Statistical "discoveries" and effect size estimation*, J. Am. Statist. Ass., 84, 608-610, 1989.

[5] http://en.wikipedia.org/wiki/False_discovery_rate (ožujak 2014.).

[6] http://en.wikipedia.org/wiki/Multiple_comparisons_problem (ožujak 2014.).

Sažetak

U ovom radu govorimo o problemu višestrukog testiranja hipoteza. Na početku rada dajemo uvod u problem testiranja statističkih hipoteza, navodimo osnovne pojmove te primjere testiranja. Zatim objašnjavamo kako, odnosno zašto dolazi do problema višestrukog testiranja hipoteza te neka predložena rješenja u statističkoj literaturi za taj problem. Upoznajemo se s osnovnim pojmovima koji se javljaju kada govorimo o višestrukom testiranju kao što su pogreške tipa I i II, snaga testa, FWER (engl. familywise error rate) te navodimo neke od procedura za kontrolu grešaka. Velik dio ovog rada posvećujemo FDR-u (engl. false discovery rate) čiju je ideju dao dr. Branko Sorić. Definiramo FDR metodu, uspoređujemo ju sa FWER-om te govorimo o njenoj važnosti. Opisujemo i FDR proceduru koja je dizajnirana kako bi kontrolirala očekivani udio pogrešno odbačenih nul hipoteza, a u simulacijskoj studiji ilustriramo prednosti FDR metode koje su razlog njene velike popularnosti.

Naglasimo da smo ovim radom samo zagrebali po površini ove vrlo važne tematike te da je literatura o višestrukom testiranju hipoteza vrlo bogata i da se u njoj konstantno javljaju novi doprinosi.

Summary

In this thesis we talk about the multiple testing problem. At the beginning of the thesis we introduce the problem of statistical hypotheses testing, its basic concepts and give some examples of testing. Then we explain how and why the problem occurs and present some solutions suggested in statistics literature. We introduce the basic concepts that arise in multiple testing, such as errors of type I and II, the power of the test, FWER (familywise error rate) and we list some of the procedures for error control. A big part of this thesis is dedicated to FDR (false discovery rate) whose idea goes back to dr. Branko Sorić. We define the FDR method, we compare it with FWER and we talk about its importance. We also describe the FDR procedure which is designed to control the expected proportion of incorrectly rejected null hypotheses. In a simulation study, we illustrate the benefits of the FDR method which are the reason for its great popularity.

In this thesis we have only scratched the surface of this very important topic. The literature on multiple testing problem is very rich. New contributions and developments appear continuously.

Životopis

Rođena sam 18. svibnja 1988. godine u Zagrebu. Prvih pet razreda osnovne škole pohađala sam u OŠ Eugena Kvaternika u Velikoj Gorici, a zadnja tri nastavila u OŠ Nikole Hribara također u Velikoj Gorici. U to sam vrijeme paralelno pohađala i OGŠ Franje Lučića u istom gradu gdje mi je glavni predmet bio flauta. Od 2003. do 2007. godine pohađala sam opću gimnaziju u Velikoj Gorici te srednju glazbenu školu Vatroslava Lisinskog u Zagrebu. 2007. godine upisala sam 1. godinu preddiplomskog sveučilišnog studija Matematika na PMF-MO, inženjerski smjer u Zagrebu te potom nastavila i diplomski sveučilišni studij Matematičke statistike u Zagrebu od 2011. godine. Iste godine sam se i udala, a 2012. godine moju je malu obitelj proširila jedna prekrasna djevojčica.