Sveučilište u Zagrebu

PMF - Matematički odjel

Ime i prezime

Naslov rada

Diplomski rad

mjesto, mjesec godina

Sveučilište u Zagrebu

PMF - Matematički odjel

Ime i prezime

Naslov rada

Diplomski rad

Voditelj rada:
Mentor

mjesto, mjesec godina

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Često nas zanima konvergira li neki niz slučajnih varijabli po distribuciji te, ukoliko konvergira, prema kojoj distribuciji. Osim teorijskih vrijednosti ovakvi su rezultati značajni i u praktičnoj primjeni. U ovom ćemo radu pokazati da su Wilcoxonove statistike sume rangova (eng. *Wilcoxon rank-sum statistic*) i rangova s predznacima (eng. *Wilcoxon signed-rank statistic*) asimptotski normalne (teže po distribuciji prema normalnoj razdiobi) uz određene uvjete.

1 Wilcoxonova statistika sume rangova

1.1 Rangovi u usporedbi dva tretmana

U različitim se područjima često pojavljuje problem je li predložena inovacija bolja od trenutnog rješenja. "Produžuje li novi lijek život pacijenata s malignim tumorima?", "Smanjuje li se štetan učinak cigareta uporabom određenog filtera?" i "Smanjuje li novi sustav prijevoza troškove?" samo su neki od primjera. Kroz sljedeći primjer pokušati ćemo ilustrirati način na koji, barem intuitivno, možemo testirati ovakva pitanja.

Primjer 1.1

Bolnica za ljude s mentalnim poteškoćama želi testirati novi lijek koji navodno ima pozitivan utjecaj na neki mentalni poremećaj. U bolnici se trenutno nalazi 5 pacijenata s tim poremećajem u, otprilike, istom stadiju. Od ovih pet pacijenata tri su odabrana na slučajan način da prime novi lijek dok će preostala dva služiti kao kontrolna grupa. Oni će primiti običnu tabletu bez ikakvih aktivnih sastojaka (takav "lijek" zovemo *placebo*). Pacijenti (po mogućnosti i osoblje) neće znati da li su dobili novi lijek. Ovo eliminira mogućnost psihološkog utjecaja.

Nakon nekog vremena nezavisni psihijatar pregledava pacijente te ih rangira po stupnju njihovog stanja. Pacijent za čije se stanje procijeni da je najgore dobije rang 1, sljedeći 2, ..., do ranga 5.

Pretpostavimo da lijek nema utjecaja, tj. da se pacijentovo stanje ne mijenja bez obzira je li primio novi lijek ili placebo. Ovu tvrdnju nazovimo *nultom hipotezom*. Kako pod nultom hipotezom rang koji pacijenti dobivaju ne ovisi o lijeku, već samo o stanju pacijenta, jasno je da izbor pacijenata koji će primiti novi lijek ne utječe na rangiranje. Dakle, možemo promatrati svaki od rangova kao određen i prije nego što smo izabrali testnu i kontrolnu grupu. Izbor pacijenata koji će biti u testnoj grupi tada u potpunosti određuje i rangove. Svaki izbor testne grupe dijeli rangove u dvije skupine, rangove koji pripadaju testnoj grupi i one koji pripadaju kontrolnoj grupi. Sve kombinacije prikazane su u sljedećoj tablici.

Testna grupa	(3,4,5)	(2,4,5)	(1,4,5)	(2,3,5)	(1,3,5)
Kontrolna grupa	(1,2)	(1,3)	(2,3)	(1,4)	(2,4)
Testna grupa	(2,3,4)	(1,3,4)	(1,2,4)	(1,2,3)	(1,2,5)
Kontrolna grupa	(1,5)	(2,5)	(3,5)	(4,5)	(3,4)

(1.1.1)

Kao što se vidi i iz (1.1.1) pacijenti, a time i njihovi rangovi, mogu se podijeliti na deset različitih načina. Pod pretpostavkom da smo 3 pacijenta za testnu grupu izabrali na slučajan način mislimo da će bilo koji od ovih 10 načina biti jednako vjerojatan, tj. imati će vjerojatnost $\frac{1}{10}$. Budući da je svaki od načina jednako vjerojatan veliki rangovi kod testne grupe ukazivati će na uspješnost novog lijeka (više o ovoj tvrdnji biti će priloženo u sljedećem odlomku).

Razmatranja iz Primjera 1.1 lako se generaliziraju. Pretpostavimo da nam je dano N pacijenata za testiranje i da je na slučajan način izabrano n od početnih N, koji će primiti tretman dok će ostalih $m = N - n$ služiti kao kontrolna grupa. Iz elementarne teorije prebrojavanja slijedi da je broj izbora jednak $\binom{N}{n}$. Po pretpostavci, n pacijenata koji će primiti novi tretman izabrano je na slučajan način (misleći ovdje, a i ubuduće, da su svi izbori jednako vjerojatni) pa svaki od izbora ima vjerojatnost $1/\binom{N}{n}$. Na kraju testiranja pacijenti su rangirani (po mogućnosti od nepristranog promatrača) s obzirom na ciljani rezultat. Kao i prije, uz ispravnost nulte hipoteze, na rangiranje ne utječe koji su pacijenti u testnoj grupi. Rang svakog od pacijenata može se promatrati kao određen (iako nama nepoznat) prije nego li je izvršeno particioniranje na testnu i kontrolnu grupu. Stoga, uz ispravnost nulte hipoteze, svako pridruživanje rangova jednako je vjerojatno sa vjerojatnošću $1/\binom{N}{n}$.

Zapišimo gornju tvrdnju malo formalnije. Neka su $S_1, \ldots, S_n$ rangovi testne grupe. Tada je

$$P_{H_0}(S_1 = s_1, \ldots, S_n = s_n) = \frac{1}{\binom{N}{n}} \tag{1.1.2}$$

gdje je $\{s_1, \ldots, s_n\}$ n-člani podskup od $\{1, 2, \ldots, N\}$, a P_{H_0} označava vjerojatnost uz istinitost nulte hipoteze.

Tijekom cijelog odlomka pretpostavljali smo da ne izabiremo pacijente već da su nam dani te da smo na slučajan način odabrali testnu grupu. Ovakav ćemo model, u kojem slučajnost ulazi samo kroz odabir testne grupe, zvati ***randomizacijski model*** (eng. *randomization model*). Model u kojem se N pacijenata izabire na određeni način iz neke populacije zovemo ***populacijski model***. U ovom modelu slučajnost ulazi i kroz izbor pacijenata.

Iako nas uglavnom zanima prijenos rezultata na populaciju postoje primjeri gdje je randomizacijski model dovoljan. Npr. farmer koji želi testirati donosi li novo gnojivo napredak na njegovih nekoliko polja i ne zanima ga donosi li napredak na ostalim poljima.

1.2 Randomizacijski model

U Primjeru 1.1 rečeno je da će dovoljno veliki rangovi ukazivati na uspješnost novog lijeka. Međutim, kada su rangovi $(S_1, \ldots, S_n)$ dovoljno veliki? Takvi se zaključci uglavnom donose pomoću neke testne statistike čije velike vrijednosti odgovaraju velikim rangovima. Jedna od takvih test statistika je i

$$W_S = S_1 + \cdots + S_n. \tag{1.2.1}$$

Statistiku W_S definiranu relacijom (1.2.1) zovemo ***Wilcoxonova statistika sume rangova*** (eng. *Wilcoxon rank-sum statistic*), a test definiran s W_S ***Wilcoxonov test sume rangova*** (eng. *Wilcoxon rank-sum test*). Ime su dobili po poznatom američkom statističaru Franku Wilcoxonu koji ih 1945. g. prvi puta upotrijebio u svom radu [9].

Time smo dobili način za provjeru jesu li nam rangovi dovoljno veliki. Da bismo izračunali kritično područje, ili p-vrijednost, potrebno je poznavati, uz H_0, distribuciju od W_S. Za Primjer 1.1 gdje je $N = 5,\ n = 3$ distribucija se može dobiti iz tablice (1.1.1). Svakom od izbora testne grupe odgovara jedna vrijednost w testne statistike W_S kao što je prikazano u tablici (1.2.2).

3,4,5	2,4,5	1,3,5	2,3,5	1,3,5	2,3,4	1,3,4	1,2,4	1,2,3	1,2,5
12	11	10	10	9	9	8	7	6	8

(1.2.2)

Budući da je vjerojatnost svakog izbora jednaka $\frac{1}{10}$ vrijedi

w	6	7	8	9	10	11	12
$P_{H_0}(W_S = w)$	0.1	0.1	0.2	0.2	0.2	0.1	0.1

(1.2.3)

Ove vjerojatnosti tvore distribuciju od W_S uz hipotezu H_0. Distribuciju uz uvjet istinitosti od H_0 nazivamo *nulta distribucija*. Iz (1.2.3) slijedi da za Primjer 1.1. vrijedi

$$P_{H_0}(W_S \geq 12) = P_{H_0}(W_S = 12) = 0.1.$$

Za razinu značajnosti $\alpha = 0.1$ nultu hipotezu odbacujemo samo kada je $W_S = 12$, tj. kada su rangovi testne grupe 3, 4 i 5.

2 Wilcoxonova statistika rangova s predznacima

2.1 Randomizacijski model za sparene uzorke

U prvom smo poglavlju uspoređivali dva tretmana kada je N subjekata dano za testiranje i podijeljeno na slučajan način na testnu i kontrolnu grupu (randomizacijski model) ili kada se N subjekata odabire metodom jednostavnog slučajnog uzorkovanja iz neke populacije (populacijski model). Kod ovakvih modela problem nastaje kada se subjekti značajno razlikuju (drastične razlike u stupnju bolesti npr.) jer tada razlika može umanjiti ili poništiti učinak tretmana. U takvim slučajevima učinkovitost usporedbe može se povećati razlaganjem subjekata u homogenije grupe.

Često će biti lakše pronaći male homogene grupe, nego velike i uglavnom se subjekti dijele na homogene grupe po dvoje (parove). Usporedbe s homogenim grupama od dvoje subjekata nazivamo ***sparene usporedbe*** (eng. *paired comparisons*). Primjer je proučavanje blizanaca gdje se svaka grupa sastoji od dva blizanca. Dijeljenje na parove nije ograničeno samo na situacije gdje imamo prirodno sparivanje, već se može postići i detaljnim proučavanjem subjekata. Primjer su pacijenti koji su u istom stadiju bolesti ili populacije koje imaju istu geografsku lokaciju.

3 Asimptotski rezultati

3.1 Očekivanje i varijanca

Da bismo mogli koristiti asimptotske rezultate za neke slučajne varijable potrebno je poznavati njihovo očekivanje i varijancu. Osnovna svojstva matematičkog očekivanja i varijance ovdje ćemo koristiti pretpostavljajući da je čitatelj već upoznat s njima. Formalni iskazi i dokazi mogu se naći u [6] i [7].

Primjer 3.1

Pretpostavimo da se populacija sastoji od N brojeva $v_1, \ldots, v_N$. Neka je na slučajan način izabran jedan od tih brojeva i označimo ga s V. Za v_i različite vrijedi $V \sim \begin{pmatrix} v_1 & \cdots & v_N \\ \frac{1}{N} & \cdots & \frac{1}{N} \end{pmatrix}$ i iz definicije matematičkog očekivanja slijedi

$$\mathrm{E}(V) = \frac{v_1 + \cdots + v_N}{N} = \overline{v}\ , \tag{3.1.1}$$

gdje s $\overline{v}$ označavamo aritmetičku sredinu. Ako vrijednosti v_i nisu različite, pretpostavimo da ih je točno n_1 jednako a_1,..., n_c jednako a_c. Tada je $V \sim \begin{pmatrix} a_1 & \cdots & a_c \\ \frac{n_1}{N} & \cdots & \frac{n_c}{N} \end{pmatrix}$ pa iz definicije matematičkog očekivanja slijedi

$$\mathrm{E}(V) = \frac{n_1 a_1 + \cdots + n_c a_c}{N} = \overline{v}\ . \tag{3.1.2}$$

Vidimo da je očekivanje u oba slučaja jednako aritmetičkoj sredini vrijednosti $v_1, \ldots, v_N$.

Izračunajmo sada varijancu. Neka su vrijednosti v_i različite. Tada iz (3.1.1) i definicije varijance slijedi

$$\mathrm{Var}(V) = \tau^2 \tag{3.1.3}$$

gdje je

$$\tau^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \overline{v})^2 \tag{3.1.4}$$

ili, koristeći $\mathrm{Var}(V) = \mathrm{E}(V^2) - \mathrm{E}(V)^2$,

$$\tau^2 = \frac{1}{N}\sum_{i=1}^{N} {v_i}^2 - \overline{v}^2. \tag{3.1.5}$$

Ako vrijednosti v_i nisu različite, pretpostavimo da ih je točno n_1 jednako a_1,..., n_c jednako a_c. Tada iz definicije varijance i (3.1.2) slijedi

$$\mathrm{Var}(V) = \frac{n_1}{N}(a_1 - \overline{v})^2 + \cdots + \frac{n_c}{N}(a_c - \overline{v})^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \overline{v})^2 = \tau^2. \tag{3.1.6}$$

Lagano se pokaže da vrijedi i formula (3.1.5).

A CGT za još neke rang statistike

Ovdje ćemo pokazati asimptotsku normalnost statistika

$$S_N = \sum_{j=1}^{N} z_{Nj} a_N(R_{Nj}) \tag{A.1}$$

gdje su $z_{N1},\ldots,z_{NN}$ i $a_N(1),\ldots,a_N(N)$ konstante a $R_{N1},\ldots,R_{NN}$ permutacije brojeva $1,\ldots,N$ sve jednako vjerojatne (sa vjerojatnosti $1/N!$). Zbog jednostavnosti često ćemo izostavljati indeks N kod z, a i R te pisati

$$S_N = \sum_{j=1}^{N} z_j a(R_j) \ . \tag{A.2}$$

Važan primjer dobijemo za $z_j = j$ i $a(j) = 1$ kada je $1 \leq j \leq n$ te $a(j) = 0$ inače. Tada statistika S_N jest Wilcoxonova statistika W_S uz nultu hipotezu.

Primijetimo da se distribucija od S_N neće promijeniti ako indekse zamijenimo. Stoga, bez smanjenja općenitosti, možemo pretpostaviti da su konstante $a(j)$ (ili z_j ili oboje) u rastućem poretku. Distribucija od S_N neće se promijeniti ni ako zamijenimo $a(j)$ i z_j jer možemo pisati $S_N = \sum_{j=1}^{N} a(j) z_{R'_j}$ gdje je R'_j inverzna permutacija od R_j.

Bibliografija

[1] T.S. Ferguson: *A Course in Large Sample Theory*, Chapman & Hall/CRC, London, 1996.

[2] J. Hajek: *Some Extensions of Wald-Wolfowitz-Noether Theorem*, Ann. Math. Statistics, Vol. 32, 505-523, 1961.

[3] E.F. Harding: *An Efficient, Minimal-storage Procedure for Calculating the Mann-Whitney U, Generalized U and Similar Distributions*, Applied Statistics, Vol. 33, No. 1, 1-6. 1984.

[4] E.L. Lehmann: *Nonparametrics: Statistical Methods Based on Ranks*, Revised First Edition, Springer, New York, 2006.

[5] H.B. Mann, D.R. Whitney: *On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other*, Ann. Math. Statistics, Vol. 18, 50-60, 1947.

[6] N. Sarapa: *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[7] A.N. Shiryaev, *Probability*, Second Edition, Springer, New York, 1996.

[8] P. Sprent, N.C. Smeeton, *Applied Nonparametric Statistical Methods*, Fourth Edition, Chapman & Hall/CRC, Boca Raton, 2007.

[9] F. Wilcoxon: *Individual Comparisons by Ranking Methods*, Biometrics, Vol. 1, 80-83, 1945.