Asimptotska razdioba statistika baziranih na rangovima

Jurilj, Ena

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ena Jurilj

ASIMPTOTSKA RAZDIOBA STATISTIKA BAZIRANIH NA RANGOVIMA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Huzak

Zagreb, studeni, 2015

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. _____

Diplomski rad posvećujem svojoj obitelji. Posebno se želim zahvaliti svome mentoru, prof. dr. sc. Miljenku Huzaku, na mnogim korisnim savjetima koji su mi jako pomogli u pisanju diplomskoga rada.

Sadržaj

Uvod

Kod neparametarskih metoda često susrećemo statistike koje se mogu zapisati kao funkcije rangova. Primjerice, znamo da se kod permutacijskog t-testa za dva uzorka, kojeg ćemo opisati u zadnjem poglavlju, koristi upravo takva statistika. Računajući egzaktnu distribuciju spomenute statistike nailazimo na problem već za jako male uzorke. Naime, za uzorak veličine n treba provesti 2^n računanja. Kako takvo računanje često nije vremenski prihvatljivo, koristi se asimptotski rezultat kao odlična alternativa. Ključ našega istraživanja je upravo dokazati asimptotsku normalnost statistika koje se mogu zapisati kao funkcije rangova, uz određene uvjete.

Na početku ćemo se prisjetiti nekih osnovnih pojmova i tvrdnji iz teorije vjerojatnosti. Zatim ćemo u drugom poglavlju opisati neke testove čije statistike zadovoljavaju uvjete Lindeberg-Fellerova teorema koji ćemo tu i iskazati.

U trećem i četvrtom poglavlju bavit ćemo se osnovnom temom rada, a to je dokaz asimptotske normalnosti statistika koje se baziraju na rangovima, i to preko Lindeberg-Fellerova teorema iz drugog poglavlja, te njegovom primjenom na neke poznatije statistike.

Poglavlje 1

Konvergencija po distribuciji

Na početku ćemo se prisjetiti nekih osnovnih vrsta konvergencija nizova slučajnih varijabli. To će biti uvod u Slutskyjev teorem. Teorem ćemo u ovom poglavlju iskazati, a njegove važnije posljedice i dokazati.

Neka je $\mathcal{X} = (X_n, n \in \mathbb{N})$ niz slučajnih varijabli. Znamo da niz $\mathcal{X}$ **konvergira gotovo sigurno (g.s.)** prema slučajnoj varijabli X ako je

$$P\left\{\omega \in \Omega \ : \ X(\omega) = \lim_{n\to\infty} X_n(\omega)\right\} = 1,$$

što označavamo sa $(g.s.)\lim_{n\to\infty} X_n = X$ ili $X_n \xrightarrow{g.s.} X$. Zatim, niz $\mathcal{X}$ slučajnih varijabli **konvergira po vjerojatnosti** prema slučajnoj varijabli X ako $\forall \epsilon > 0$ vrijedi

$$\lim_{n\to\infty} P\left\{|X_n - X| \geq \epsilon\right\} = 0,$$

uz oznake $(P)\lim_{n\to\infty} X_n = X$ ili $X_n \xrightarrow{P} X$.

U dokazima se često umjesto konvergencije po vjerojatnosti dokazuje konvergencija u srednjem reda p koja implicira onu po vjerojatnosti. Dakle, neka je $1 \leq p < \infty$ i neka su X_n, X slučajne varijable takve da je $EX_n < \infty$, $EX < \infty$, $\forall n \in \mathbb{N}$. Niz $(X_n, n \in \mathbb{N})$ **konvergira u srednjem reda *p*** prema X ako vrijedi

$$\lim_{n\to\infty} E\left[|X_n - X|^p\right] = 0.$$

U tom slučaju pišemo $(m^p)\lim_{n\to\infty} X_n = X$ ili $X_n \xrightarrow{m^p} X$. Kao najvažniju u ovom radu navodimo konvergenciju po distribuciji. Niz $\mathcal{X} = (X_n, n \in \mathbb{N})$ **konvergira po distribuciji** prema slučajnoj varijabli X ako je

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \forall x \in C(F_X),$$

gdje je $C(F_X)$ skup točaka na kojima je funkcija F_X neprekidna. Najčešće pišemo $X_n \xrightarrow{D} X, \quad n \to \infty$.

Sljedeći teorem je izuzetno bitan jer daje glavne veze među kovergencijama. Mi ćemo ga samo iskazati, a dokaz se može pronaći u [4].

Teorem 1.0.1. *Vrijede sljedeće implikacije:*

1. $X_n \xrightarrow{g.s.} X \Rightarrow X_n \xrightarrow{P} X$,
2. $X_n \xrightarrow{m^p} X \Rightarrow X_n \xrightarrow{P} X, \quad (1 \leq p < \infty)$,
3. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$.

1.1 Slutskyjev teorem

Zanima nas uz koje su uvjete dva niza slučajnih varijabli asimptotski ekvivalentna. Potrebne rezultate daje nam sljedeći teorem, i to u terminima slučajnih vektora.

Teorem 1.1.1. *(Slutsky) Neka je $(\boldsymbol{X}_n, n \in \mathbb{N})$ niz slučajnih vektora. Tada vrijede sljedeće tvrdnje:*

1. *Ako je $\boldsymbol{X}_n \in \mathbb{R}^d$, $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ i ako je $f : \mathbb{R}^d \to \mathbb{R}^k$ takva da je $\mathbb{P}(\boldsymbol{X} \in C(f)) = 1$, gdje je $C(f)$ skup na kojem je f neprekidna, tada*

$$f(\boldsymbol{X}_n) \xrightarrow{D} f(\boldsymbol{X}).$$

2. *Ako $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ i $(\boldsymbol{X}_n - \boldsymbol{Y}_n) \xrightarrow{P} 0$, tada*

$$\boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X}.$$

3. *Ako $\boldsymbol{X}_n \in \mathbb{R}^d$, $\boldsymbol{Y}_n \in \mathbb{R}^k$, $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ i $\boldsymbol{Y}_n \xrightarrow{D} c$, (c je konstanta), tada*

$$\begin{pmatrix} \boldsymbol{X}_n \\ \boldsymbol{Y}_n \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \boldsymbol{X} \\ c \end{pmatrix}.$$

Dokaz prethodnog teorema može se pogledati u [2]. U ovom ćemo se radu pozivati na posljedice teorema. U nastavku iskazujemo tri važnija korolara uz pripadajuće dokaze. Od sada pa nadalje promatramo nizove slučajnih varijabli.

Korolar 1.1.2. *Neka su* $(X_n, n \in \mathbb{N})$ *i* $(Y_n, n \in \mathbb{N})$ *nizovi slučajnih varijabli s konačnim pozitivnim varijancama. Ako je*

$$\frac{X_n - EX_n}{\sqrt{VarX_n}} \xrightarrow{D} X \quad i \quad Corr(X_n, Y_n) \to 1,$$

tada vrijedi

$$\frac{Y_n - EY_n}{\sqrt{VarY_n}} \xrightarrow{D} X.$$

Dokaz.

Neka su U_n i V_n nizovi slučajnih varijabli s očekivanjem 0 i varijancom 1, takvih da je

$$Corr(U_n, V_n) \to 1, \quad n \to \infty.$$

Budući da je

$$Corr(U_n, V_n) = \frac{Cov(U_n, V_n)}{\sqrt{VarU_n}\sqrt{VarV_n}} = \frac{E[(U_n - EU_n)(V_n - EV_n)]}{\sqrt{VarU_n}\sqrt{VarV_n}} = \frac{E[U_nV_n]}{1},$$

pa zbog

$$\begin{aligned} E[(U_n - V_n)^2] &= E[U_n^2] - 2E[U_nV_n] + E[V_n^2] \\ &= VarU_n - 2Corr(U_n, V_n) + VarV_n \\ &= 2(1 - Corr(U_n, V_n)) \end{aligned}$$

slijedi da je

$$E\left[(U_n - V_n)^2\right] = 2 \cdot (1 - Corr(U_n, V_n)) \to 0, \quad \text{kada } n \to \infty,$$

(jer tada $Corr(U_n, V_n) \to 1$), to jest $U_n - V_n$ konvergira u srednjem reda 2 prema 0. Iz toga slijedi da $U_n - V_n$ konvergira po vjerojatnosti, pa su U_n i V_n asimptotski ekvivalentne i iz teorema 1.1.1 (druga tvrdnja) zaključujemo da imaju asimptotski isti zakon razdiobe. Ovo opažanje se može direktno primijeniti na nizove

$$U_n = \frac{X_n - EX_n}{\sqrt{VarX_n}} \quad i \quad V_n = \frac{Y_n - EY_n}{\sqrt{VarY_n}}$$

uz $Corr(X_n, Y_n) \to 1$, budući da je $Corr(U_n, V_n) = Corr(X_n, Y_n)$. □

Pretpostavimo sada da su X_n i Y_n slučajne varijable s očekivanjem nula i jednakim varijancama, te neka je

$$X_n \xrightarrow{D} X \quad \text{i} \quad Corr(X_n, Y_n) \to 1, \quad n \to \infty.$$

Pitamo se vrijedi li tada nužno

$$Y_n \xrightarrow{D} X \ ?$$

Sljedećim ćemo primjerom pokazati da X_n i Y_n ne moraju nužno biti asimptotski ekvivalentne.

Neka su U i V različite slučajne varijable, obje s očekivanjem 0. Neka su dalje U i V definirane na intervalu $[-1, 1]$, $Var(U) = Var(V)$, te neka je W nezavisna o U i V, i

$$P(W = -1) = P(W = 1) = \frac{1}{2}.$$

Neka je vektor (X_n, Y_n) jednak

$$(X_n, Y_n) \sim \begin{pmatrix} (nW, nW) & (U, V) \\ 1/n & (n-1)/n \end{pmatrix}.$$

Stoga vrijedi

$$X_n \xrightarrow{D} U \quad i \quad Y_n \xrightarrow{D} V, \quad n \to \infty,$$

ali

$$Var(X_n) = n + \frac{n-1}{n} Var(U), \quad Var(Y_n) = n + \frac{n-1}{n} Var(V) \quad i \quad Cov(X_n, Y_n) = n,$$

iz čega slijedi

$$Corr(X_n, Y_n) \to 1.$$

Korolar 1.1.3. *Neka su* $(X_n, n \in \mathbb{N})$ *i* $(Y_n, n \in \mathbb{N})$ *nizovi slučajnih varijabli.*

1. *Ako vrijedi*

$$\frac{E\left[(X_n - Y_n)^2\right]}{VarX_n} \to 0,$$

tada

$$Corr(X_n, Y_n) \to 1, \quad n \to \infty.$$

2. *Ako vrijedi*

$$\frac{X_n - EX_n}{\sqrt{VarX_n}} \xrightarrow{D} X \quad i \quad \frac{E\left[(X_n - Y_n)^2\right]}{VarX_n} \to 0,$$

tada je

$$\frac{Y_n - EY_n}{\sqrt{VarY_n}} \xrightarrow{D} X, \quad n \to \infty.$$

Dokaz. Vrijedi:

$$0 \leq Var(X_n - Y_n) = Var(X_n) - 2Cov(X_n, Y_n) + Var(Y_n) =$$

$$= E\left[(X_n - Y_n)^2\right] - \left[E\left(X_n - Y_n\right)\right]^2 \leq E\left[(X_n - Y_n)^2\right].$$

Iz nejednakosti Cauchy-Schwarz-Bunjakovskog (CSB) imamo da je

$$Cov(X_n, Y_n) = E\left[(X_n - EX_n)(Y_n - EY_n)\right] \overset{(CSB)}{\leq} \sqrt{E\left[(X_n - EX_n)^2\right]}\sqrt{E\left[(Y_n - EY_n)^2\right]} =$$

$$= \sqrt{Var(X_n)}\sqrt{Var(Y_n)}.$$

Stoga je

$$Var(X_n) - 2\sqrt{Var(X_n)}\sqrt{Var(Y_n)} + Var(Y_n) \leq Var(X_n) - 2Cov(X_n, Y_n) + Var(Y_n),$$

odnosno

$$\left(\sqrt{Var(X_n)} - \sqrt{Var(Y_n)}\right)^2 \leq E\left[(X_n - Y_n)^2\right].$$

Ukoliko posljednju nejednadžbu podijelimo s $Var(X_n)$, dobijemo

$$\left(1 - \frac{\sqrt{Var(Y_n)}}{\sqrt{Var(X_n)}}\right)^2 \leq \frac{E\left[(X_n - Y_n)^2\right]}{Var(X_n)}.$$

Budući da po pretpostavci korolara desna strana teži u nulu, vrijedi

$$\frac{\sqrt{Var(Y_n)}}{\sqrt{Var(X_n)}} \to 1, \quad n \to \infty.$$

Nakon što nejednakost s početka podijelimo s $\sqrt{Var(X_n)}\sqrt{Var(Y_n)}$, iskoristimo prethodnu konvergenciju i pretpostavku korolara, dobijemo da desna strana nejednakosti

$$0 \leq \frac{\sqrt{Var(X_n)}}{\sqrt{Var(Y_n)}} - 2Corr(X_n, Y_n) + \frac{\sqrt{Var(Y_n)}}{\sqrt{Var(X_n)}} \leq \frac{E\left[(X_n - Y_n)^2\right]}{Var(X_n)} \cdot \frac{\sqrt{Var(X_n)}}{\sqrt{Var(Y_n)}}$$

teži u nulu. Iz toga nužno slijedi da je

$$Corr(X_n, Y_n) \to 1, \quad n \to \infty.$$

Time smo dokazali prvu tvrdnju korolara, dok druga tvrdnja slijedi lagano iz prve tvrdnje i korolara 1.1.2.

□

Korolar 1.1.4. *Ako*

$$X_n \xrightarrow{D} X > 0 \quad i \quad \frac{X_n}{Y_n} \xrightarrow{P} 1$$

tada

$$Y_n \xrightarrow{D} X.$$

Dokaz. Iz prve tvrdnje Slutskyjevog teorema dobijemo

$$log(X_n) \xrightarrow{D} log(X), \quad n \to \infty, \tag{1.1}$$

budući da je *log* neprekidna funkcija na $(0, \infty)$. Znamo iz teorije vjerojatnosti da je

$$\frac{X_n}{Y_n} \xrightarrow{P} 1 \quad \Leftrightarrow \quad \frac{X_n}{Y_n} \xrightarrow{D} 1$$

(jer je slučajna varijabla kojoj niz $\frac{X_n}{Y_n}$ konvergira konstanta), stoga iz (1.1) slijedi $log(X_n) - log(Y_n) \xrightarrow{D} 0, \quad n \to \infty$, odnosno ponovnom primjenom gornje ekvivalencije dobijemo

$$log(X_n) - log(Y_n) \xrightarrow{P} 0, \quad n \to \infty.$$

Primjenom druge tvrdnje Slutskyjevog teorema dobijemo

$$log(Y_n) \xrightarrow{D} log(X), \quad n \to \infty,$$

te ponovo iz prvog dijela tog istog teorema slijedi

$$Y_n \xrightarrow{D} X, \quad n \to \infty.$$

□

Poglavlje 2

Klasični centralni granični teorem. Primjene

U ovom poglavlju ćemo iskazati klasični teorem o konvergenciji slučajnih vektora po distribuciji (dokaz u [2]), potom ćemo se fokusirati na Lindeberg-Fellerov[1] teorem koji ćemo primijeniti na neke statistike.

Teorem 2.0.5. *Neka je* $(\boldsymbol{X}_n, n \in \mathbb{N})$ *niz nezavisnih jednako distribuiranih slučajnih vektora s očekivanjem* $\boldsymbol{\mu}$ *i konačnom kovarijacijskom matricom* $\boldsymbol{\Sigma}$*. Tada vrijedi*

$$\sqrt{n}\,(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad n \to \infty. \tag{2.1}$$

Za primjene u statisci važnija je proširena verzija klasičnog centralnog graničnog teorema na nezavisne, ali ne nužno jednako distribuirane, slučajne varijable. Stoga navodimo Lindeberg-Fellerov teorem u jednoj dimenziji bez dokaza. Dokaz se može pogledati u [4]. Korisno ga je iskazati u terminima slučajnih varijabli koje možemo poredati u obliku beskonačnog trokuta

$$\begin{array}{cccc} X_{11} & & & \\ X_{21} & X_{22} & & \\ X_{31} & X_{32} & X_{33} & \\ X_{41} & X_{42} & X_{43} & X_{44} \\ \ldots, & & & \end{array} \tag{2.2}$$

tako da su slučajne varijable u svakom redu nezavisne s očekivanjem nula i konačnom varijancom. Slijedi iskaz Lindeberg-Fellerovog teorema.

[1] Jarl Waldemar Lindeberg (1876.-1932.) finski matematičar; Vilim (William) Feller (1906.-1970.) američko-hrvatski matematičar.

Teorem 2.0.6. (*Lindeberg-Feller*) *Neka je* $\left\{X_{nj}, n \in \mathbb{N}, j = 1, 2, ..., n\right\}$ *niz slučajnih varijabli takvih da su za svaki* $n \in \mathbb{N}$, $X_{n1}, X_{n2}, ..., X_{nn}$ *nezavisne slučajne varijable s očekivanjem* $EX_{nj} = 0$ *i konačnim varijancama* $VarX_{nj} = \sigma_{nj}^2$, *te neka je* $S_n = X_{n1} + X_{n2} + ... + X_{nn}$ *i* $B_n^2 = VarS_n = \sum_{j=1}^n \sigma_{nj}^2 > 0$, $(n \in \mathbb{N})$. *Ako je zadovoljen Lindebergov uvjet*

$$\frac{1}{B_n^2} \sum_{j=1}^{n} E\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|>\epsilon B_n\}}\right] \to 0, \quad kada\ n \to \infty, \quad za\ sve\ \epsilon > 0, \tag{2.3}$$

tada

$$\frac{S_n}{B_n} \xrightarrow{D} \mathcal{N}(0,1), \quad kada\ n \to \infty.$$

Obratno, ako

$$\frac{\max\limits_{j\le n} \sigma_{nj}^2}{B_n^2} \to 0, \quad kada\ n \to \infty,$$

(tj. ako niti jedan član sume B_n^2 *ne igra značajnu ulogu u limesu) i*

$$\frac{S_n}{B_n} \xrightarrow{D} \mathcal{N}(0,1),$$

tada je zadovoljen Lindebergov uvjet.

Posebno je važan slučaj kada imamo jedan niz $X_1, X_2, ...$ nezavisnih jednako distribuiranih slučajnih varijabli s očekivanjem i konačnom varijancom jednakim

$$EX_j = \mu \quad i \quad VarX_j = \sigma^2 > 0, \quad \forall j \in \mathbb{N}.$$

Sljedeći korolar govori o asimptotskoj normalnosti takvog niza.

Korolar 2.0.7. *Neka je* $X_1, X_2, ...$ *niz nezavisnih jednako distribuiranih slučajnih varijabli s očekivanjem* μ *i konačnom varijancom* $\sigma^2 > 0$. *Neka je* $T_n = z_{n1}X_1 + z_{n2}X_2 + ... + z_{nn}X_n$, *gdje su* z_{nj} *zadani brojevi, j=1,2,...,n, koji nisu svi nula, za sve* $n \in \mathbb{N}$, *i takvi da je*

$$\frac{\max\limits_{j\le n} z_{nj}^2}{\sum_{j=1}^n z_{nj}^2} \to 0, \quad n \to \infty.$$

Neka je $\mu_n = ET_n$ *i* $\sigma_n^2 = VarT_n$. *Tada vrijedi*

$$\frac{T_n - \mu_n}{\sigma_n} \xrightarrow{D} \mathcal{N}(0,1), \quad kad\ n \to \infty. \tag{2.4}$$

Dokaz. Da bismo dokazali tvrdnju iz korolara, raspisat ćemo čekivanje i varijancu od T_n.

$$\mu_n = ET_n = \sum_{j=1}^{n} E\left(z_{nj}X_j\right) = \mu \sum_{j=1}^{n} z_{nj},$$

$$\sigma_n^2 = Var(T_n) = \sum_{j=1}^{n} Var(z_{nj}X_j) = \sigma^2 \sum_{j=1}^{n} z_{nj}^2 > 0.$$

Kako je $T_n - \mu_n = \sum_{j=1}^{n} z_{nj}(X_j - \mu)$, koristit ćemo Lindeberg-Fellerov teorem uz $X_{nj} = z_{nj}(X_j - \mu)$. Prema tome, slijedi da je $EX_{nj} = 0$, $VarX_{nj} = \sigma_{nj}^2 = \sigma^2 z_{nj}^2$, $S_n = T_n - \mu_n$ i $B_n^2 = \sigma_n^2 = \sigma^2 \sum_{j=1}^{n} z_{nj}^2$. Dakle, ukoliko je zadovoljen Lindebergov uvjet, dobit ćemo željeni rezultat. Neka je $\epsilon > 0$ proizvoljan. Tada imamo

$$\begin{aligned} \tfrac{1}{B_n^2}\textstyle\sum_{j=1}^{n} E\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|>\epsilon B_n\}}\right] &= \tfrac{1}{B_n^2}\textstyle\sum_{j=1}^{n} z_{nj}^2 E\left[(X_j-\mu)^2 \mathbb{1}_{\left\{|X_j-\mu|>\frac{\epsilon B_n}{|z_{nj}|}\right\}}\right] \\ &\leq \tfrac{1}{B_n^2}\textstyle\sum_{j=1}^{n} z_{nj}^2 E\left[(X_j-\mu)^2 \mathbb{1}_{\left\{|X_j-\mu|>\frac{\epsilon B_n}{\max\limits_{i\leq n}|z_{ni}|}\right\}}\right]. \end{aligned}$$

Kako su X_j, $j \in \mathbb{N}$, jednako distribuirane, očekivanje s desne strane nejednakosti ne ovisi o j. Uz to je $B_n^2 = \sigma^2 \sum_{j=1}^{n} z_{nj}^2$, pa slijedi

$$\tfrac{1}{B_n^2}\textstyle\sum_{j=1}^{n} E\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|>\epsilon B_n\}}\right] \leq \tfrac{1}{\sigma^2} E\left[(X_1-\mu)^2 \mathbb{1}_{\left\{|X_1-\mu|>\frac{\epsilon B_n}{\max\limits_{j\leq n}|z_{nj}|}\right\}}\right].$$

Ukoliko L_n definiramo sa

$$L_n := (X_1-\mu)^2 \mathbb{1}_{\left\{|X_1-\mu|>\frac{\epsilon B_n}{\max\limits_{j\leq n}|z_{nj}|}\right\}},$$

dobijemo da je

$$\lim_{n\to\infty} L_n := (X_1-\mu)^2 \lim_{n\to\infty} \mathbb{1}_{\left\{|X_1-\mu|>\frac{\epsilon B_n}{\max\limits_{j\leq n}|z_{nj}|}\right\}} = 0,$$

budući da je $\max\limits_{j\leq n} z_{nj}^2/B_n^2 \to 0$ po pretpostavci. Kako je

$$|L_n| \leq (X_1-\mu)^2,$$

a znamo da je varijanca od X_1 konačna, zaključujemo da je $|L_n|$ ograničena odozgo integrabilnom funkcijom za svaki prirodan broj n, pa stoga prema Lebesgueovom teoremu o dominiranoj konvergenciji slijedi da je

$$\lim_{n\to\infty} E(L_n) = 0.$$

Teorem o sendviču sada daje da je zadovoljen Lindebergov uvjet, pa je prema Lindeberg-Fellerovom teoremu

$$\frac{S_n}{B_n} = \frac{T_n - \mu_n}{\sigma_n} \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty.$$

□

Prelazimo na primjere pomoću kojih ćemo pokazati kako je asimptotski rezultat pripadnih statistika zapravo odlična alternativa standardnom načinu računanja. Sve to uz primjenu iskazanog Lindeberg-Fellerovog teorema i njegovog korolara.

Primjer 2.0.8. ***Permutacijski t-test za sparene uzorke.*** *Uzeti su uzorci vrhnja iz 10 mljekara (od A do J). Uzorak iz svake mljekare podijeljen je na dva dijela. Jedan je dio poslan na ispitivanje u Laboratorij I , a drugi u Laboratorij II. Pitanje je: postoji li značajna razlika u rezultatima mjerenja? Dobivene vrijednosti su zabilježene u tablici (tisuće bakterija u ml^{-1}) :*

Mljekara	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Laboratorij I	*11.7*	*12.1*	*13.3*	*15.1*	*15.9*	*15.3*	*11.9*	*16.2*	*15.1*	*13.6*
Laboratorij II	*11.1*	*11.9*	*13.4*	*15.4*	*14.8*	*14.8*	*12.3*	*15.0*	*14.2*	*13.1*

Primijetimo kako su ovdje prisutna dva zavisna uzorka, onaj koji se šalje u *Laboratorij I* i onaj koji se šalje u *Laboratorij II*; zavisna u smislu da se podrazumijeva da se svo vrhnje u pojedinoj mljekari proizvodi u jednakim uvjetima (zavisnost unutar para). Upravo zbog toga nije bitno koji dio iz mljekare će se slati u koji laboratorij, pa je taj odabir slučajan. Jedan od testova koji nam može dati odgovor na pitanje iz primjera je premutacijski t-test za sparene uzorke. Prvo ćemo se pozabaviti opisom njegove provedbe, pa ga potom primijeniti na naš problem.

U opisu testova, kadagod je to moguće, koristit ćemo se općenitim slikovitim primjerom pojave i testiranja novog lijeka kod liječenja određene bolesti. Jedan od načina kako provesti istraživanje o učinkovitosti novog lijeka je taj da promatramo dani uzorak od n parova pacijenata, takvih da su pacijenti unutar para u otprilike jednakom stadiju bolesti. Iz svakog para na slučajan način odaberemo pacijenta koji će uzimati novi lijek, dok će drugi primati standardnu terapiju. Drugim riječima, dijelimo uzorak na onaj dio njih koji će biti podvrgnuti tretmanu i onaj dio koji će služiti kao kontrola. Nakon određenog vremena provođenja terapije liječnik pregledava sve pacijente i bilježi rezultate.

Označimo sa (X_j, Y_j) rezultate mjerenja j-tog para, gdje je X_j rezultat od pacijenta podvrgnutog tretmanu, a Y_j rezultat kontrolnog pacijenta, $j = 1, 2, ..., n$. Hipoteze su sljedeće:

H_0 : nema razlike u mjerenju,
H_1 : postoji rezlika.

Ukoliko sa $Z_j = X_j - Y_j$ označimo razlike u mjerenjima, hipoteze prelaze u:

$$\begin{array}{ll} H_0: & Z_j \text{ su simetrične oko 0,} \\ H_1: & \text{ne } H_0. \end{array} \tag{2.5}$$

Problem dva uzorka sveli smo na problem jednog uzorka. Analiza permutacijskog t-testa za sparene uzorke provodi se uvjetno na opažene vrijednosti od Z_j. Prema tome, uz H_0 uvjetno na opažene vrijednosti, slučajne varijable $Z_1, Z_2, ..., Z_n$ su nezavisne, jednako distribuirane s razdiobom

$$Z_j \sim \begin{pmatrix} -|z_j| & |z_j| \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad j = 1, 2, ..., n,$$

gdje su $|z_j| = |x_j - y_j|$, $j = 1, 2, ..., n$, apsolutne vrijednosti razlika pripadnih realizacija slučajnog uzorka $(X_1, Y_1), ..., (X_n, Y_n)$. Permutacijski t-test koristi T statistiku jednog uzorka,

$$T = \sqrt{n-1}\frac{\overline{Z}_n}{s_z}, \tag{2.6}$$

gdje je

$$s_z^2 = \frac{1}{n}\sum_{j=1}^{n}\left(Z_j - \overline{Z}_n\right)^2$$

uzoračka varijanca, odnosno s_z standardna devijacije uzorka, te

$$\overline{Z}_n = \frac{1}{n}\sum_{j=1}^{n} Z_j$$

uzoračka aritmetička sredina.

Promatrajmo slučajni vektor $(Z_1, Z_2, ..., Z_n)$. Iz njegove definicije je jasno da je broj mogućih vrijednosti koje može poprimiti jednak 2^n, s tim da je svaka od realizacija oblika $(\pm|z_1|, ..., \pm|z_n|)$ jednako vjerojatna (s vjerojatnošću $\frac{1}{2^n}$). Test se provodi na sljedeći način:

- Za svaku od 2^n različitih vrijednosti vektora $(Z_1, Z_2, ..., Z_n)$ izračunamo vrijednost T statistike po formuli (2.6);
- Dobivene vrijednosti poredamo uzlazno te ih označimo redom sa $T_1, T_2, ..., T_{2^n}$;
- Izračunamo realizaciju T statistike na temelju opaženih vrijednosti $z_1, z_2, ..., z_n$ i rezultat označimo sa $T_{realizacija}$;
- Uz odabranu (ili zadanu) razinu značajnosti odredimo kritično područje ili izračunamo p-vrijednost testa, te iznesemo zaključke.

Vratimo se sada na primjer 2.0.8. Neka je X_j rezultat mjerenja *Laboratorija I*, a Y_j rezultat mjerenja *Laboratorija II*, ($j = 1, 2, ..., 10$). Tada $Z_j = X_j - Y_j$ označava razliku u mjerenju između dva laboratorija. Testiramo hipoteze (2.5). Dakle, promatramo sljedeću tablicu:

Mljekara	A	B	C	D	E	F	G	H	I	J
z_j	0.8	0.2	-0.1	-0.3	1.1	0.5	-0.4	1.2	0.9	0.5

Slijedeći opis provedbe testa, u programskom jeziku R-u izračunali smo $2^{10} = 1024$ vrijednosti T statistika. Poredali smo uzlazno dobivene vrijednosti,

$$T_1 \leq T_2 \leq T_3 \leq T_4 \leq \cdots \leq T_{1021} \leq T_{1022} \leq T_{1023} \leq T_{1024}, \tag{2.7}$$

te izračunali realizaciju testne statistike opaženog uzorka:

$$T_{realizacija} = T_{1000} = T_{1001} = T_{1002} = 2.4246.$$

Budući da je riječ o dvostranom testu (zanima nas samo postoji li razlika u mjerenju, a ne i koji laboratorij daje manje/veće vrijednosti), te je realizacija testne statistike opaženog uzorka pozitivna, p-vrijednost testa dobijemo iz jednakosti

$$p_1 = 2 \cdot P(T \geq T_{realizacija}|H_0) = 2 \cdot \frac{25}{1024} \approx 0.0488.$$

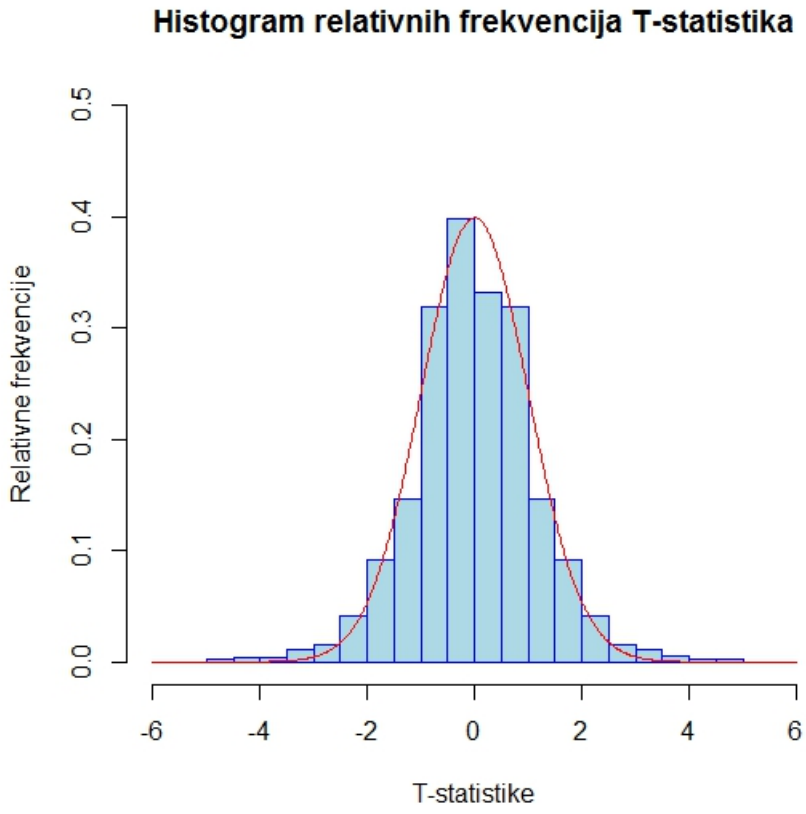
Dakle, na temelju izračunate p-vrijednosti $p_1 = 0.0488$ zaključujemo da hipotezu H_0 o nepostojanju razlika u mjerenju između dvaju laboratorija ne odbacujemo na razini značajnosti 0.01, dok ju odbacujemo u korist alternative na razini značajnosti 0.05.

Na drugi način zaključujemo da, ukoliko promatramo kritično područje uz razinu značajnosti od 5%,

$$\{T_1, T_2, \ldots, T_{26}\} \cup \{T_{999}, T_{1000}, \ldots, T_{1024}\},$$

gdje su $T_{26} = -2.25$ i $T_{999} = 2.25$, primijećujemo da realizacija testne statistike upada u kritično područje, pa odbacujemo nultu hipotezu u korist alternative.

Očito je da ovaj postupak iziskuje mnogo vremena. Već za uzorak duljine 10 trebali smo izračunati $2^{10} = 1024$ vrijednosti T statistike. Kada bismo raspolagali uzorkom duljine 15, imali bismo $2^{15} = 32768$ potrebnih računanja (premda se još uvijek radi o relativno "malenom" uzorku). Problem je u eksponencijalnom rastu funkcije $n \longmapsto 2^n$. Stoga se radije koriste asimptotski rezultati. Obratimo pažnju na sliku 2.1 na kojoj vidimo kako funkcija gustoće standardne normalne razdiobe lijepo opisuje ponašanje T statistike iz primjera 2.0.8. Sada je već intuitivno jasno čemu ćemo težiti u daljnjim razmatranjima.



Slika 2.1: Histogram relativnih frekvencija T statistike i krivulja standardne normalne razdiobe.

Vratimo se na definiciju slučajne varijable Z_j, te promotrimo cijeli niz takvih slučajnih varijabli $Z_1, Z_2, \ldots$ koje su nezavisne i jednako distribuirane. Tada $sign(Z_j)$, uz H_0, uvjetno na opažene vrijednosti, ima sljedeću distribuciju:

$$sign(Z_j) \sim \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{2.8}$$

Mi ćemo promatrati niz $(sign(Z_j), j \in \mathbb{N})$ nezavisnih jednako distribuiranih slučajnih varijabli definiranih sa (2.8), s očekivanjem i varijancom jednakim

$$E\left[sign(Z_j)\right] = 0, \quad Var\left[sign(Z_j)\right] = 1.$$

Ako stavimo da je $z_{nj} := |z_j|$, $j = 1, 2..., n$, za svaki $n \in \mathbb{N}$, tada je

$$Z_j = z_{nj} \cdot sign(Z_j) = |z_j| \cdot sign(Z_j).$$

Definirajmo

$$S_n = Z_1 + Z_2 + ... + Z_n = \sum_{j=1}^{n} z_{nj} \cdot sign(Z_j),$$

pa je varijanca od S_n jednaka

$$B_n^2 = VarS_n = \sum_{j=1}^{n} Var\left[z_{nj} \cdot sign(Z_j)\right] = \sum_{j=1}^{n} z_j^2,$$

dok je očekivanje jednako $\mu_n = ES_n = 0$. Budući da je razumno pretpostaviti da je

$$\frac{\underset{j\leq n}{max}|z_j|^2}{\sum_{j=1}^{n} z_j^2} \to 0, \quad n \to \infty, \tag{2.9}$$

prema korolaru 2.0.7 slijedi

$$\frac{S_n}{B_n} \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty,$$

odnosno

$$\frac{Z_1 + ... + Z_n}{\sqrt{z_1^2 + ... + z_n^2}} = n\frac{\overline{Z}_n}{B_n} = n\frac{\overline{Z}_n}{\sqrt{n}\alpha_n} = \sqrt{n}\frac{\overline{Z}_n}{\alpha_n} \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty,$$

gdje je $\alpha_n^2 = \frac{1}{n}B_n^2$. U praksi se u permutacijskom t-testu češće upotrebljava statistika $T = \sqrt{n-1} \cdot \overline{Z}_n / s_z$ od statistike $\sqrt{n}\overline{Z}_n/\alpha_n$. Kako smo već pokazali asimptotsku normalnost od $\overline{Z}_n$, preostaje nam objasniti vezu između $\sqrt{n}\frac{\overline{Z}_n}{\alpha_n}$ i $\sqrt{n-1}\frac{\overline{Z}_n}{s_z}$. Može se pokazati da su dva permutacijska testa koja koriste ove dvije statistike ekvivalentna. A to je zbog toga što je T rastuća funkcija od $v = \sqrt{n}\overline{Z}_n/\alpha_n$. Uočimo da su T i v uvijek istog predznaka, te da je

$$v^2 = n\frac{\overline{Z}_n^2}{\alpha_n^2} = n\frac{\overline{Z}_n^2}{s_z^2 + \overline{Z}_n^2} = n\left(\frac{n-1}{T^2} + 1\right)^{-1}.$$

Opet se vraćamo na primjer 2.0.8. Već smo izračunali da je $T_{realizacija} = 2.4246$, a p-vrijednost je iznosila $p_1 = 0.0488$. Sada možemo koristiti asimptotski rezultat T statistike,

$$\sqrt{n-1}\frac{\overline{Z}_n}{s_z} \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty,$$

iz kojeg dobijemo novu p-vrijednost testa:

$$p_2 = 2 \cdot (1 - \phi(T_{realizacija})) \approx 0.0153.$$

To što je vrijednost p_2 znatno manja od p_1 možemo pripisati tome što je uzorak relativno mali. Dakle, prema p_2 zaključujemo da nultu hipotezu odbacujemo na razinama značajnosti većima od 0.0153. Međutim, na razini značajnosti od 1% ne odbacujemo H_0.

■

Primjer 2.0.9. ***Wilcoxonov test predznaka rangova za sparene uzorke***. *Sastavljeno je 9 parova studenata pri čemu su studenti unutar svakog para podjednako uspješni. Jedan iz svakog para slučajnim je odabirom pohađao uobičajena predavanja, dok je drugi pohađao novi tečaj CAL. Na kraju semestra svim studentima su podijeljeni jednaki testovi. Zabilježeni su sljedeći rezultati (od 100 maksimalnih bodova):*

Par	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
CAL	*50*	*56*	*51*	*46*	*88*	*79*	*81*	*95*	*73*
Lekcije	*25*	*58*	*65*	*38*	*91*	*32*	*31*	*13*	*49*

Jesu li studenti koji su pohađali novi tečaj uspješniji?

Kod ovog primjera koristit ćemo test predznaka rangova za sparene uzorke. Sličan je permutacijskom t-testu. Glavna razlika je u tome što se ovdje ne promatraju opažene vrijednosti, već pripadni rangovi. Krenimo s opisom testa.

Dakle, ponovno raspolažemo s $2n$ opservacija sparenih kao kod permutacijskog t-testa. Neka je (X_j, Y_j) rezultat vezan uz j-ti par, $j = 1, 2, ..., n$, te neka je $Z_j = X_j - Y_j$ razlika u mjerenju. Neka su $z_j = x_j - y_j$, $j = 1, 2, ..., n$, opažene vrijednosti (u tablici).

Hipoteze koje testiramo

$$\begin{aligned} H_0 &: \text{ nema razlike između } X_j \text{ i } Y_j, \\ H_1 &: \text{ vrijednosti } X_j \text{ su veće od vrijednosti } Y_j, \end{aligned}$$

prelaze u

$$\begin{aligned} H_0 &: Z_j \text{ simetrične oko 0}, \\ H_1 &: \text{Pozitivne vrijednosti od } Z_j \text{ su vjerojatnije}. \end{aligned}$$

Za razliku od premutacijskog t-testa sparenih uzoraka, gdje smo promatrali stvarne vrijednosti $z_1, ..., z_n$, kod primjene ovog testa koristimo njihove rangove. Bez smanjenja općenitosti možemo pretpostaviti da su vrijednosti $|z_1|, |z_2|, ..., |z_n|$ uzlazno sortirane. Pretpostavljamo i to da je svaki $|z_j|$ različit od nule i da su međusobno različiti, odnosno

$$0 < |z_1| < |z_2| < ... < |z_n|.$$

Dodijelimo rangove sortiranim apsolutnim vrijednostima od 1 do n, potom dodamo rangovima odgovarajuće predznake (one koje su imali z_j prije apsoultne vrijednosti). Statistika koja se koristi kod ovog testa je sljedeća:

$$W_+ = \sum_{j=1}^{n} j \mathbb{1}_{\{Z_j>0\}}, \tag{2.10}$$

tj. zbrojimo pozitivne rangove. Test provodimo na sljedeći način:

- Za svaku od 2^n različitih vrijednosti vektora $(Z_1, Z_2, ..., Z_n)$ izračunamo vrijednost statistike W_+ po formuli (2.10);
- Dobivene vrijednosti poredamo uzlazno te ih označimo redom sa $W_1, W_2, ..., W_{2^n}$;
- Izračunamo realizaciju W_+ statistike na temelju opaženih vrijednosti $z_1, z_2, ..., z_n$ (odnosno vrijednosti pripadnih rangova) i rezultat označimo sa $W_{realizacija}$;
- Uz odabranu (ili zadanu) razinu značajnosti izračunamo kritično područje ili izračunamo p-vrijednost testa, te iznesemo zaključke.

Riješimo sada primjer 2.0.9. U daljnjem ćemo tekstu, radi jednostavnosti, rezultate studenata koji su pohađali tečaj CAL jednostavno zvati CAL, a rezultate studenata koji su pohađali standardna predavanja zvati Lekcije. Ukoliko sa X_j označimo rezultate CAL-a, a sa Y_j Lekcija, $Z_j = X_j - Y_j$ predstavlja razliku u rezultatima jednog para, $j = 1, 2, ..., 9$. Testiraju se sljedeće hipoteze:

H_0 : Oba programa su podjednako uspješna,
H_1 : Uspješniji su studenti koji su pohađali novi tečaj.

U programskom jeziku R-u izračunali smo da je

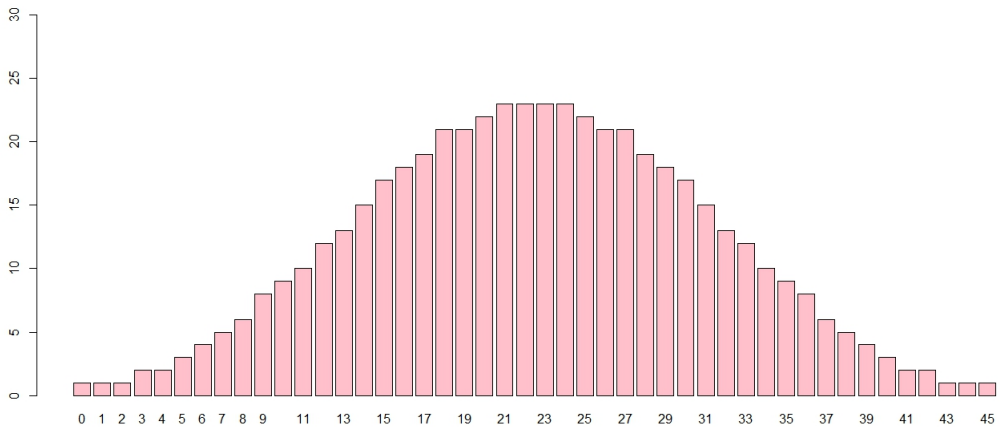
$$W_{realizacija} = 38,$$

dok je p-vrijednost testa

$$p_1 = 0.0742.$$

Prema tome, zaključujemo da ne možemo odbaciti pretpostavku H_0 na razinama značajnosti manjima od vrijednosti p_1.

Budući da je postupak podjednako kompleksan kao u primjeru 2.0.8, objasnit ćemo kako jednostavnije doći do rješenja. Može nam kao motivacija poslužiti stupičasti dijagram iz primjera 2.0.9 (na slici 2.2), koji nas navodi na misao da bi statistika W_+ mogla imati asimptotski normalnu razdiobu.



Slika 2.2: Stupičasti dijagram frekvencija W_+ statistike.

Kako su uz H_0, $\mathbb{1}_{\{Z_j>0\}}$ nezavisne jednako distribuirane Bernoullijeve slučajne varijable, $j = 1, 2, ..., n$, imamo da je

$$EW_+ = \frac{1}{2}\sum_{j=1}^{n} j = \frac{n(n+1)}{4}, \tag{2.11}$$

$$VarW_+ = \frac{1}{4}\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{24}. \tag{2.12}$$

U duhu teorema 2.0.6, odnosno korolara iza, pitamo se vrijedi li

$$\frac{W_+ - EW_+}{\sqrt{VarW_+}} \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty \ ? \tag{2.13}$$

Bitno je uočiti da se (2.10) može zapisati u formi permutacijskog t-testa. To ćemo objasniti uz pomoć statistike

$$W_n = \sum_{j=1}^{n} j\mathbb{1}_{\{Z_j>0\}} - \sum_{j=1}^{n} j\mathbb{1}_{\{Z_j<0\}} = \sum_{j=1}^{n} j \cdot sign(Z_j) = 2W_+ - \sum_{j=1}^{n} j$$

koja predstavlja razliku sume svih pozitivnih rangova i sume svih negativnih rangova. Definicija od W_n pokazuje kako su statistike W_+ i W_n kolinearne. No, $(1/n)W_n$ je upravo u formi od $\overline{Z}_n$ iz permutacijskog t-testa, uz $|z_j| = j$. Preostaje nam još pokazati da niz $|z_j| = j$

zadovoljava uvjet (2.9). Budući da je

$$\max_{j\le n} j^2 = n^2 \quad \text{i} \quad \sum_{j=1}^{n} j^2 = \frac{n(n-1)(2n+1)}{6}$$

(nazivnik puno brže raste), vrijedi

$$\frac{\max_{j\le n} j^2}{\sum_{j=1}^{n} j^2} \to 0, \quad n \to \infty,$$

pa iz korolara 2.0.7 slijedi da su W_n, pa onda i W_+, asimptotski normalne, te vrijedi (2.13).

Koristeći asimptotski rezultat, dobijemo novu p-vrijednost

$$p_2 = 0.0663,$$

za čiju bismo vrijednost mogli reći da je približno jednaka vrijednosti p_1. Dakle, asimptotski rezultat se pokazuje kao iznimno koristan kod opisanog problema.

■

Primjer 2.0.10. ***Rekordi.***

Neka je $(Z_n, n \in \mathbb{N})$ niz nezavisnih jednako distribuiranih neprekidnih slučajnih varijabli. Kažemo da se **rekord postiže** u k ako je $Z_k > \max_{j<k} Z_j$. Neka je dalje $(R_n, n \in \mathbb{N})$ niz slučajnih varijabli takvih da je $R_k = 1$ ukoliko se rekord postigao u k, te $R_k = 0$ ukoliko nije. Može se pokazati[2] da je $(R_n, n \in \mathbb{N})$ niz nezavisnih Bernoullijevih slučajnih varijabli s $P(R_k = 1) = 1 - P(R_k = 0) = \frac{1}{k}$, $k \in \mathbb{N}$. Neka

$$S_n = \sum_{k=1}^{n} R_k \tag{2.14}$$

broji koliko je puta postignut rekord u prvih n opažanja. Želimo odrediti ES_n, $VarS_n$ i provjeriti vrijedi li za statistiku S_n sljedeće:

$$\frac{S_n - ES_n}{\sqrt{VarS_n}} \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty \ ?$$

Krenimo od ključne statistike

$$R_k \sim \begin{pmatrix} 0 & 1 \\ 1-\frac{1}{k} & \frac{1}{k} \end{pmatrix},$$

[2] Vidjeti Example 2.3.2. u Durrett (u [1]), str. 60

s očekivanjem i varijancom redom

$$ER_k = \frac{1}{k} \quad \text{i} \quad VarR_k = E\left[(R_k - ER_k)^2\right] = \left(\frac{1}{k}\right)\left(1 - \frac{1}{k}\right).$$

Slijedi da su očekivanje i varijanca od S_n jednaki

$$ES_n = \sum_{k=1}^{n} \frac{1}{k} \quad \text{i} \quad B_n^2 = VarS_n = \sum_{k=1}^{n} \left(\frac{1}{k}\right)\left(1 - \frac{1}{k}\right).$$

Stavimo da je X_{nj} iz Lindeberg-Fellerova teorema jednak $R_j - (1/j)$. Tada je i $(R_k - (1/k), k \in \mathbb{N})$ niz nezavisnih slučajnih varijabli, te $EX_{nj} = 0$ i $B_n^2 = VarS_n$. U terminima trokutova X_{nj} zapisujemo kao

$$\begin{array}{llll} R_1 - 1 & & & \\ R_1 - 1 & R_2 - (1/2) & & \\ R_1 - 1 & R_2 - (1/2) & R_3 - (1/3) & \\ R_1 - 1 & R_2 - (1/2) & R_3 - (1/3) & R_4 - (1/4) \\ \ldots & & & \end{array}$$

Ispitajmo Lindebergov uvijet:

$$\begin{aligned} \tfrac{1}{B_n^2} \textstyle\sum_{j=1}^{n} E\left[\left(R_j - \tfrac{1}{j}\right)^2 \mathbb{1}_{\left\{|R_j - \frac{1}{j}| \geq \epsilon B_n\right\}}\right] &\leq \tfrac{1}{B_n^2} \textstyle\sum_{j=1}^{n} E\left[\left(R_j - \tfrac{1}{j}\right)^2 \mathbb{1}_{\{1 \geq \epsilon B_n\}}\right] \\ &= \mathbb{1}_{\{1 \geq \epsilon B_n\}} \\ &= \mathbb{1}_{\left\{\frac{1}{B_n} \geq \epsilon\right\}}. \end{aligned}$$

Za dovoljno veliki n, $\mathbb{1}_{\left\{\frac{1}{B_n} \geq \epsilon\right\}}$ je jednako 0, jer za fiksni $\epsilon > 0$ niz B_n divergira. Teorem o sendviču sada daje da je zadovoljen Lindebergov uvjet, pa prema teoremu 2.0.6 slijedi asimptotska normalnost statistike (2.14).

■

Primjer 2.0.11. ***Kendallov*** τ

Neka je $(Z_n, n \in \mathbb{N})$ niz nezavisnih jednako distribuiranih neprekidnih slučajnih varijabli. Označimo sa X_k broj vrijednosti Z_i, $i < k$, takvih da je $Z_i > Z_k$. Formalno možemo pisati

$$X_k = \sum_{i=1}^{k-1} \mathbb{1}_{\{Z_i > Z_k\}}, \quad k \in \mathbb{N}.$$

Može se pokazati[3] da su $(X_n, n \in \mathbb{N})$ nezavisne slučajne varijable i da X_k ima diskretnu uniformnu razdiobu na skupu $\{0, 1, ..., k-1\}$, $k \in \mathbb{N}$. Statistika

$$T_n = \sum_{k=1}^{n} X_k \tag{2.15}$$

predstavlja ukupan broj razlika u poretku. Ukoliko su opservacije poredane uzlazno, statistika T_n postiže minimalnu vrijednost 0, dok joj je maksimalna vrijednost $\sum_{k=1}^{n}(k-1) = \frac{n(n-1)}{2}$ kad su opservacije u silaznom poretku. Može se koristiti kao statistika neparametarskog permutacijskog testa uz hipoteze o postojanju rastućeg ili padajućeg trenda kod opservacija.

Statistika

$$\tau_n = 1 - \frac{4T_n}{n(n-1)} \in [-1, 1]$$

zove se **Kendallov koeficijent korelacije rangova**. Vidimo da su T_n i τ_n kolinearne. Promatrat ćemo statistiku T_n i preko nje dokazati asimptotsku normalnost obiju statistika. Dakle, izračunat ćemo ET_n i $VarT_n$, te pokazati da je

$$\frac{T_n - ET_n}{\sqrt{VarT_n}} \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty.$$

Razdioba od X_k je

$$X_k \sim \begin{pmatrix} 0 & 1 & \dots & k-1 \\ \frac{1}{k} & \frac{1}{k} & \dots & \frac{1}{k} \end{pmatrix},$$

iz čega izračunamo očekivanje i varijancu od X_k,

$$EX_k = \frac{1}{k}\sum_{i=1}^{k-1} i = \frac{1}{k}\frac{k(k-1)}{2} = \frac{k-1}{2} \quad \text{i} \quad VarX_k = \frac{1}{k}\sum_{i=1}^{k-1} i^2 - \left(\frac{k-1}{2}\right)^2 = \frac{k^2-1}{12}.$$

Kako promatramo statistiku $T_n = X_1 + X_2 + ... + X_n$, računamo i njezino očekivanje i varijancu:

$$ET_n = \sum_{k=1}^{n} \frac{k-1}{2} = \frac{n(n-1)}{4},$$

$$VarT_n = \sum_{k=1}^{n} \frac{k^2-1}{12} = \frac{1}{12}\left[\frac{n(n+1)(2n+1)}{6} - n\right] = \frac{n(n-1)(2n+5)}{72}.$$

[3] Vidjeti Remark 2.3.4. u Durrett (u [1]), str. 61

U terminima Lindeberg-Fellerova teorema stavimo da je $X_{nj} = X_j - ((j-1)/2)$. To nam daje niz nezavisnih slučajnih varijabli $(X_j - ((j-1)/2), j \in \mathbb{N})$, $EX_{nj} = 0$ i $B_n^2 = VarT_n$. Budući da je X_j ograničena s 0 i $j-1$, imamo sljedeće

$$|X_{nj}| \leq |X_j - \frac{j-1}{2}| \leq |j - 1 - \frac{j-1}{2}| = \frac{|2j-2-j+1|}{2} \leq \frac{j-1}{2} \leq \frac{n-1}{2}, \quad \text{za } j \leq n.$$

Prema tome je

$$\begin{aligned} \tfrac{1}{B_n^2} \textstyle\sum_{j=1}^n E\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}| \geq \epsilon B_n\}}\right] &\leq \tfrac{1}{B_n^2} \textstyle\sum_{j=1}^n E\left[X_{nj}^2 \mathbb{1}_{\{(n-1)/2 \geq \epsilon B_n\}}\right] \\ &= \mathbb{1}_{\{(n-1)/2 \geq \epsilon B_n\}} \\ &= \mathbb{1}_{\{(n-1)/(2B_n) \geq \epsilon\}}. \end{aligned}$$

Za fiksni $\epsilon > 0$ desna strana gornje nejednakosti teži u nulu za $n \to \infty$ (jer je B_n reda $n^{(3/2)}$). Dakle, T_n, a time i njemu kolinearan τ_n, imaju asimptotski normalnu razdiobu prema Lindeberg-Fellerovom teoremu.

■

Za kraj ćemo riješiti primjer koji će nam približiti već opisane rekorde i Kendallov τ.

Primjer 2.0.12. *Jedna od ponajboljih hrvatskih sportašica, atletičarka Blanka Vlašić, može se pohvaliti višestrukim zlatnim, srebrnim i brončanim medaljama osvojenima na raznim svjetskim natjecanjima u disciplini skoka u vis. Povrh svega toga, Blanka ističe da je za nju obaranje vlastitog rekorda važnije od same pobjede. Sa službene stranice Blanke Vlašić preuzeli smo podatke o najboljim godišnjim rezultatima koje je postizala u razdoblju od 1999. do 2013.godine. Rezultati su prikazani u tablici:*

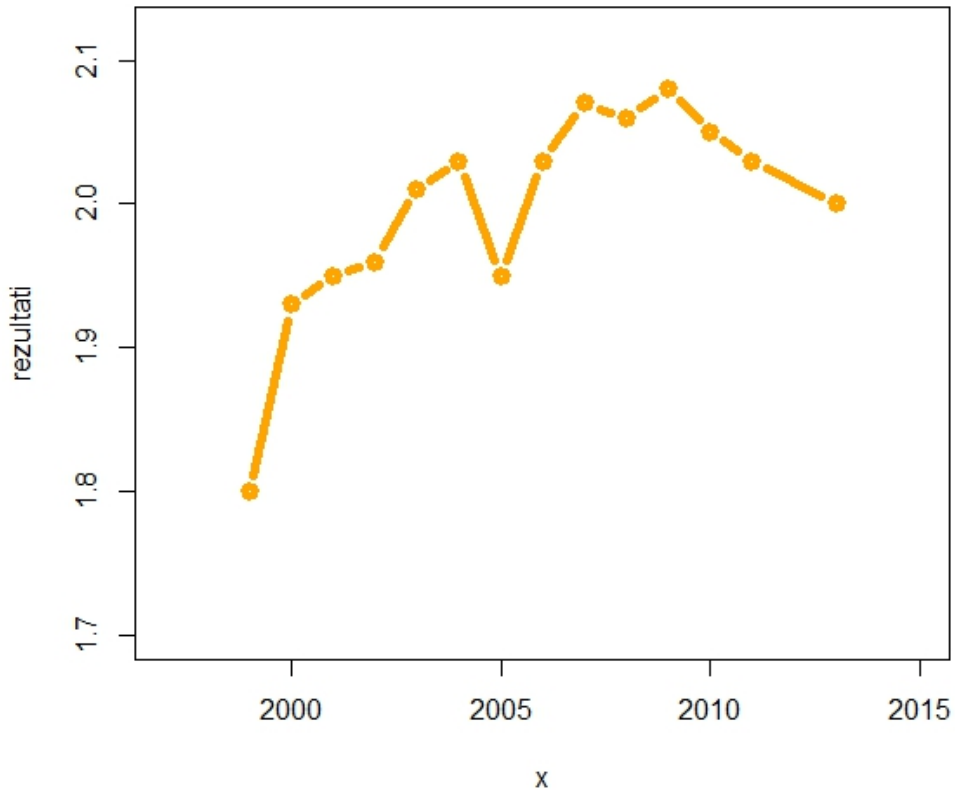
Godina	*1999.*	*2000.*	*2001.*	*2002.*	*2003.*	*2004.*	*2005.*
Rezultat (m)	*1.80*	*1.93*	*1.95*	*1.96*	*2.01*	*2.03*	*1.95*

Godina	*2006.*	*2007.*	*2008.*	*2009.*	*2010.*	*2011.*	*2013.*
Rezultat (m)	*2.03*	*2.07*	*2.06*	*2.08*	*2.05*	*2.03*	*2.00*

Pitamo se je li prisutan pozitivan trend u obaranju vlastitog rekorda tokom godina. Testiramo na razini značajnosti od 5%.

Ispitujemo sljedeće hipoteze

H_0 : Rezultati su simetrični oko medijana,
H_1 : Postoji rastući trend.



Slika 2.3: Napredovanje rezultata.

Podaci su prikazani na grafu 2.3. Na njemu se vidi kako je Blanka do 2004. godine konstantno napredovala, 2005. za nju je bila nešto lošija godina, zatim se ponovno vraća u formu i postiže svoj konačni osobni rekord 2009. godine preskočivši 208 centimetara. Nakon 2009. lagano opadaju visine, no i dalje skače iznad 2 metra.

Za testiranje koristimo statistiku (2.15). Uz $n = 14$, $T_{14} = 19$, $ET_{14} = (14 \cdot 13)/4$ i $Var(T_{14}) = (14 \cdot 13 \cdot 33)/72$, realizacija normalizirane testne statistike je -2.9015. P-vrijednost testa je

$$p = \phi(-2.9014) \approx 0.0018.$$

Jasno je da na razini značajnosti od 0.05 odbacujemo nultu hipotezu u korist alternative

koja kaže da postoji rastući trend. Kendallov τ iznosi

$$\tau_{14} = 1 - \frac{4 \cdot 19}{13 \cdot 14} \approx 0.582 = 58.2\%.$$

Iz ovoga također vidimo da se tokom vremena visina povećava.

■

Poglavlje 3

Centralni granični teorem za statistike bazirane na rangovima

U ovom ćemo poglavlju iskazati i dokazati naš glavni rezultat, a to je poseban oblik centralnog graničnog teorema. Taj teorem primjenjujemo na određene statistike koje se baziraju na rangovima. Dokaz je vrlo jednostavan, međutim, to ne umanjuje njegovu vrijednost u neparametarskoj statistici. Naprotiv, teorem je izuzetno bitan jer se može primijeniti na veliki broj statistika, a mi ćemo navesti samo njih nekoliko. Bitno je spomenuti i to da je njegov dokaz zasnovan na ispunjavanju Lindebergovog uvjeta i primjeni Lindeberg-Fellerova teorema. Ovdje će biti riječ o posebnoj vrsti statistika koje se mogu zapisati kao suma S_N, malo kasnije definirana kao funkcija rangova. Novim teoremom izbjeći ćemo direktnu primjenu Lindeberg-Fellerova teorema, za razliku od prethodnog poglavlja.

Na razvoju rezultata iz ovog poglavlja radili su Wald i Wolfowitz (1944), Noether (1949) i Hoeffding (1952). Mi ćemo slijediti postupak koji je zaokružio Hájek (1961), a on se upravo poziva na teorem 2.0.6.

3.1 Teorem

Neka je $R_{N1}, R_{N2}, ..., R_{NN}$ slučajna permutacija skupa $\{1, 2, ..., N\}$. Dakle, sve su permutacije jednako vjerojatne. Znamo da takvih permutacija ima ukupno $N!$, a vjerojatnost svake je $\frac{1}{N!}$. U ovom poglavlju ispitujemo asimptotsku razdiobu zbroja funkcija u formi

$$S_N = \sum_{j=1}^{N} z_{Nj} a_N\left(R_{Nj}\right), \tag{3.1}$$

gdje su $z_{N1}, ..., z_{NN}$ i $a_N(1), ..., a_N(N)$ zadani skupovi brojeva. Radi jednostavnijeg zapisa, na desnoj strani izostavljamo indeks N, pa (3.1) prelazi u

$$S_N = \sum_{j=1}^{N} z_j a\left(R_j\right). \tag{3.2}$$

U daljnjem tekstu N će uglavnom biti fiksan, stoga ćemo se oslanjati na zapis (3.2). Kada pustimo da N teži u beskonačnost, naglasit ćemo da distribucija od R ovisi o N, te da možda i z i a ovise o N.

Primijetimo da se distribucija od S_N iz (3.1) ne mijenja ukoliko promijenimo poredak sumacije. Prema tome, možemo bez smanjenja općenitosti pretpostaviti da je $a(j)$ (ili z_j ili oba) u rastućem poretku. Također, distribucija od S_N ostaje nepromijenjena ukoliko ju zapišemo kao

$$S_N = \sum_{j=1}^{N} a(j) z_{R_j'},$$

gdje je R_j' inverzna permutacija od R_j, tj. $R_j' = i$ ako i samo ako $R_i = j$.

Vratimo se na definiciju od S_N i uočimo da su očekivanje i varijanca od $a(R_j)$ jednaki

$$Ea(R_j) = \frac{1}{N} \sum_{i=1}^{N} a(i) = \overline{a}_N,$$

$$Var\left(a(R_j)\right) = \frac{1}{N} \sum_{i=1}^{N} \left(a(i) - \overline{a}_N\right)^2 = \sigma_a^2,$$

te da ne ovise o j. Očekivanje i varijancu od z_j redom obilježavamo sa

$$\overline{z}_N = \frac{1}{N} \sum_{j=1}^{N} z_j,$$

$$\sigma_z^2 = \frac{1}{N} \sum_{j=1}^{N} \left(z_j - \overline{z}_N\right)^2.$$

Kao prvu statistiku koju možemo zapisati u formi (3.1) navest ćemo statistiku za uzorkovanje. Nama je bitna zato što je jednostavna i zato što su mnoge druge statistike njezin malo složeniji oblik. Važno svojstvo statistike uzorkovanja i preostalih koje ćemo navesti je to da je a monotona funkcija po j.

Primjer 3.1.1. ***Uzorkovanje.***

Pretpostavimo da uzimamo slučajan uzorak bez ponavljanja, fiksne duljine $n \geq 1$, iz populacije s vrijednostima $\{z_1, ..., z_N\}$. Ukoliko S_N označava zbroj vrijednosti iz uzorka, tada taj zbroj možemo pisati u formi (3.1) uz

$$a(j) = \begin{cases} 1, & \text{ako je } 1 \leq j \leq n \\ 0, & \text{ako je } (n+1) \leq j \leq N \end{cases} \tag{3.3}$$

Podrazumijeva se da se permutiranje odnosi na elemente skupa $\{z_1, ..., z_N\}$, te da njih prvih n predstavlja elemente uzorka. Dakle, $R_j \leq n$, ukoliko je $R_1, R_2, ..., R_N$ slučajna premutacija od $1, 2, ..., N$.

■

Za početak ćemo odrediti formule za očekivanje i varijancu za općeniti zapis od S_N.

Lema 3.1.2. *Vrijedi sljedeće:*

$$ES_N = N\overline{z}_N\overline{a}_N$$

i

$$VarS_N = \frac{N^2}{N-1}\sigma_z^2\sigma_a^2. \tag{3.4}$$

Dokaz. Raspišimo ES_N:

$$ES_N = \sum_{j=1}^{N} z_j Ea(R_j) = \sum_{j=1}^{N} z_j\overline{a}_N = N\overline{z}_N\overline{a}_N.$$

Uočimo da $Cov(a(R_j), a(R_k))$ ne ovisi o j i k za $j \neq k$. Kako je

$$P(R_1 = i, R_2 = j) = \frac{1}{N(N-1)}, \quad \text{za sve } i \neq j,$$

imamo sljedeće

$$\begin{aligned} Cov(a(R_1), a(R_2)) &= \tfrac{1}{N(N-1)} \sum \sum_{i \neq j} (a(i) - \overline{a}_N)(a(j) - \overline{a}_N) \\ &= -\tfrac{1}{N(N-1)} \sum_{i=1}^{N} (a(i) - \overline{a}_N)^2 \\ &= -\tfrac{1}{N-1}\sigma_a^2. \end{aligned} \tag{3.5}$$

Iz ovoga slijedi da za varijancu od S_N vrijedi

$$\begin{aligned} VarS_N &= \sum_{j=1}^{N} z_j^2 Var\, a(R_j) + \sum\sum_{j\neq k} z_j z_k Cov(a(R_j), a(R_k)) \\ &= \sigma_a^2 \left[\sum_{j=1}^{N} z_j^2 - \frac{1}{N-1}\sum\sum_{j\neq k} z_j z_k\right] \\ &= \sigma_a^2 \left[\sum z_j^2 - \frac{1}{N-1}\left(\sum z_j\right)^2 + \frac{1}{N-1}\sum z_j^2\right] \\ &= \sigma_a^2 \left[\frac{N}{N-1}\sum z_j^2 - \frac{1}{N-1}\left(\sum z_j\right)^2\right] \\ &= \sigma_a^2 \left[\frac{N}{N-1}\sum z_j^2 - \frac{N^2}{N-1}\bar{z}_N^2\right] \;=\; \frac{N}{N-1}\sigma_a^2 \left[\sum z_j^2 - N\bar{z}_N^2\right] \\ &= \frac{N^2}{N-1}\sigma_z^2\sigma_a^2. \end{aligned}$$

□

S obzirom na to da ćemo u dokazu teorema primijeniti verziju Lindeberg-Fellerova teorema iskazanu u prethodnom poglavlju, želimo raditi sa statistikom očekivanja nula. Radi toga ćemo prijeći na novu statistiku $S_N^{'}$, koja predstavlja sumu nezavisnih slučajnih varijabli, te zadovoljava Lindebergov uvjet. Sjetimo se korolara 1.1.2 koji kaže da su dvije normalizirane statistike asimptotski ekvivalentne ukoliko im korelacija teži u 1. Dakle, cilj će nam biti dokazati da vrijedi $\frac{S_N^{'}}{\sqrt{VarS_N^{'}}} \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty$.

U tu svrhu , neka su $U_1, U_2, \ldots, U_N$ nezavisne jednako distribuirane slučajne varijable takve da je $U_j \sim \mathcal{U}(0,1)$, za $j = 1, 2, \ldots, N$. $\mathcal{U}(0,1)$ označava neprekidnu uniformnu razdiobu na intervalu $[0,1]$. Neka R_j označava rang od U_j u poretku $U_1, U_2, \ldots, U_N$ od najmanjeg do najvećeg. Tada je $(R_1, \ldots, R_N)$ slučajna permutacija skupa $\{1, 2, \ldots, N\}$ i može se koristiti u formuli (3.1). Povrh toga, može se dokazati da vrijedi

$$Corr(U_j, \frac{R_j}{N}) \to 1\, quad\text{za}\, N \to \infty,$$

što znači da je R_j/N prilično blizu U_j. Prema tome, težit ćemo tome da u sumi (3.1) zamijenimo R_j sa $\lceil NU_j \rceil$ kako bismo dobili sumu nezavisnih jednako distribuiranih slučajnih varijabli bez značajne promjene u vrijednosti sume. Pišemo

$$S_N - ES_N = \sum_{j=1}^{N} (z_j - \bar{z}_N)(a(R_j) - \bar{a}_N),$$

i definiramo

$$S_N^{'} = \sum_{j=1}^{N} (z_j - \bar{z}_N)(a(\lceil NU_j \rceil) - \bar{a}_N).$$

Tada je $ES_N^{'} = 0$, i

$$Var(S_N^{'}) = \sum_{j=1}^{N}(z_j - \bar{z}_N)^2 Var(a(R_1)), \tag{3.6}$$

jer su $\lceil NU_j \rceil$ nezavisne jednako distribuirane slučajne varijable, s jednakom distribucijom kao i R_1, dakle uniformno su distribuirane na skupu $\{1, 2, ..., N\}$.

Prije nego dokažemo da je $Corr(S_N, S_N^{'}) \to 1, \quad N \to \infty$, pojednostavit ćemo zapis korelacije u sljedećoj lemi.

Lema 3.1.3. *Vrijedi:*

$$Corr(S_N, S_N^{'}) = \sqrt{\frac{N}{N-1}} Corr\left(a\left(R_1\right), a\left(\lceil NU_1 \rceil\right)\right). \tag{3.7}$$

Dokaz. Znamo da je

$$Corr(S_N, S_N^{'}) = \frac{Cov(S_N, S_N^{'})}{\sqrt{VarS_N \cdot VarS_N^{'}}},$$

stoga ćemo prvo raspisati $Cov(S_N, S_N^{'})$:

$$Cov(S_N, S_N^{'}) = \sum_{j=1}^{N}\sum_{k=1}^{N}(z_j - \bar{z}_N)(z_k - \bar{z}_N) Cov(a(R_j), a(\lceil NU_k \rceil)).$$

Vrijednost $c_1 = Cov(a(R_j), a(\lceil NU_j \rceil))$ ne ovisi o j, te vrijednost $c_2 = Cov(a(R_j), a(\lceil NU_k \rceil))$ ne ovisi o j i k za $j \neq k$. Imamo da je

$$\begin{aligned} Cov(S_N, S_N^{'}) &= c_1 \textstyle\sum_{j=1}^{N}(z_j - \bar{z}_N)^2 + c_2 \sum_{j=1}^{N}\sum_{k=1}^{N}(z_j - \bar{z}_N)(z_k - \bar{z}_N) \\ &= (c_1 - c_2) \textstyle\sum_{j=1}^{N}(z_j - \bar{z}_N)^2. \end{aligned} \tag{3.8}$$

Kako je $\sum_{j=1}^{N} a(R_j)$ konstanta, slijedi da je

$$0 = Cov\left(\textstyle\sum_{j=1}^{N} a(R_j), a(\lceil NU_k \rceil)\right) = \sum_{j=1}^{N} Cov\left(a(R_j), a(\lceil NU_k \rceil)\right)$$

$$= c_1 + (N-1)c_2.$$

Ovo pokazuje kako je

$$c_2 = -\frac{c_1}{N-1}.$$

Nakon što c_2 uvrstimo u (3.8), dobijemo jednakost

$$Cov(S_N, S_N^{'}) = \frac{Nc_1}{N-1}\sum_{j=1}^{N}(z_j - \bar{z}_N)^2.$$

Uvrstimo u definiciju korelacije s početka dokaza zadnju dobivenu jednakost za Cov, te jednakosti varijance otprije (formule (3.4) i (3.6)). Nakon što malo sredimo izraz, dobijemo

$$Corr(S_N, S_N^{'}) = \sqrt{\frac{N}{N-1}} Corr\left(a\left(R_1\right), a\left(\lceil NU_1 \rceil\right)\right).$$

□

Obratimo pažnju na jednakost dokazanu u prehodnoj lemi. Ako pustimo da N teži u beskonačnost, jasno je da $\sqrt{\frac{N}{N-1}} \to 1$, pa će biti dovoljno pokazati da $Corr\left(a\left(R_1\right), a\left(\lceil NU_1 \rceil\right)\right)$ konvergira prema 1. Iz jednakosti

$$\frac{E\left[(a(R_1) - a(\lceil NU_1 \rceil))^2\right]}{Var\left(a(R_1)\right)} = 2\left(1 - Corr\left(a\left(R_1\right), a\left(\lceil NU_1 \rceil\right)\right)\right),$$

koju dobijemo iz

$$E\left[(a(R_1) - a(\lceil NU_1 \rceil))^2\right] = E\left[(a(R_1) - \overline{a}_N + \overline{a}_N - a(\lceil NU_1 \rceil))^2\right] =$$

$$= E\left[(a(R_1) - \overline{a}_N)^2\right] - 2E\left[(a(R_1) - \overline{a}_N)\left(a(\lceil NU_1 \rceil) - \overline{a}_N\right)\right] + E\left[(a(\lceil NU_1 \rceil) - \overline{a}_N)^2\right] =$$

$$= 2Var(a(R_1)) - 2Cov(a(R_1), a(\lceil NU_1 \rceil)),$$

slijedi da će korelacija $Corr\left(a\left(R_1\right), a\left(\lceil NU_1 \rceil\right)\right)$ konvergirati prema 1 ukoliko dokažemo da $E\left[(a(R_1) - a(\lceil NU_1 \rceil))^2\right] \to 0$,

Iako nejednakost koju u sljedećoj lemi dokazujemo vrijedi generalno, mi ćemo iznijeti dokaz samo za slučaj kada je $a(j)$ zadan kao u (3.3). Za takav $a(j)$ iz statistike za uzorkovanje vrijedi da je $\overline{a}_N = n/N$. Time smo pokrili i statistike koje se koriste u permutacijskom t-testu za dva uzorka, te Wilcoxonovu testu sume rangova, čiji se zapisi mogu svesti na oblik iz uzorkovanja. O tim će testovima biti riječi nešto kasnije. U zadnjem ćemo poglavlju pokazati da nejednakost iz sljedeće leme vrijedi i kada je $a(j) = j$. Ovako definiran $a(j)$ bit će spomenut kod zadnja dva testa koja ćemo obraditi. Sada ćemo iznijeti dokaz leme samo za prvi slučaj.

Lema 3.1.4. *(Hájek) Pretpostavimo da je $a(j)$ monotona funkcija po j. Tada*

$$E\left[(a\left(R_1\right) - a\left(\lceil NU_1 \rceil\right))^2\right] \leq \frac{2\sqrt{2}}{N} \max_j |a(j) - \overline{a}_N| \sqrt{\sum_{i=1}^{N} (a\left(i\right) - \overline{a}_N)^2}. \tag{3.9}$$

Dokaz. Za $a(j)$ kao u (3.3), imamo da je

$$\max_j |a(j) - \overline{a}_N| = max\left\{\frac{n}{N}, \frac{N-n}{N}\right\},$$

$$\sum_{j=1}^{N}(a(i)-\overline{a}_N)^2 = n\left(1-\frac{n}{N}\right)^2 + (N-n)\left(\frac{n}{N}\right)^2 =$$

$$= \frac{n(N-n)(N-n+n)}{N^2} = n\frac{N-n}{N},$$

stoga nam preostaje dokazati da je

$$E\left[(a(R_1)-a(\lceil NU_1\rceil))^2\right] \leq \frac{2\sqrt{2}}{N} max\left\{\frac{n}{N}, 1-\frac{n}{N}\right\}\left[n\left(1-\frac{n}{N}\right)\right]^{\frac{1}{2}}.$$

Pokazat ćemo da je zadovoljena nešto snažnija nejednakost

$$E\left[(a(R_1)-a(\lceil NU_1\rceil))^2\right] \leq \frac{1}{N}\left[n\left(1-\frac{n}{N}\right)\right]^{\frac{1}{2}}. \tag{3.10}$$

Ovo očekivanje ćemo izračunati uvjetno na dani poredak statistika $U_{(1)} < U_{(2)} < \cdots < U_{(N)}$. Ključno je svojstvo to da rangovi $(R_1, ..., R_N)$ ne ovise o poretku statistika $U_{(.)} = (U_{(1)}, \ldots, U_{(N)})$. Naime, za proizvoljan j, svaki od $N!$ različitih rangova jednako je vjerojatan da bude stvarni rang od U_j. Ukoliko je R_1 rang od U_1, tada ćemo pisati $U_1 = U_{(R_1)}$. Prema tome,

$$\begin{aligned} E\left[(a(R_1)-a(\lceil NU_1\rceil))^2\right] &= E\left[E\left\{(a(R_1)-a(\lceil NU_{(R_1)})\rceil)\right)^2 |U_{()}\right\}\right] \\ &= E\left[(1/N)\textstyle\sum_{j=1}^{N}(a(j)-a(\lceil NU_{(j)}\rceil))^2\right]. \end{aligned}$$

Svaki od članova sume $S = \sum_{j=1}^{N}(a(j)-a(\lceil NU_{(j)}\rceil))^2$ je ili 1 ili 0, a S predstavlja broj odstupanja. Ukoliko je točno n onih U_j koji su manji od (n/N), tada je S jednak nuli. No, ukoliko se taj broj povećava ili raste za jedan, tada S raste za 1. Prema tome, $S = |K-n|$, gdje je K broj onih U_j koji su manji od (n/N). K ima binomnu distribuciju s parametrima N i n/N ($K \sim B(N, n/N)$). Stoga imamo sljedeće:

$$\begin{aligned} E\left[(a(R_1)-a(\lceil NU_1\rceil))^2\right] &= (1/N)E\left[|K-n|\right] \\ &\leq (1/N)\left(E(K-n)^2\right)^{(1/2)} \\ &= (1/N)\left(n(N-n)/N\right)^{(1/2)}. \end{aligned}$$

Time smo dokazali da vrijedi (3.10), a budući da je

$$\frac{1}{N}\left[n\left(1-\frac{n}{N}\right)\right]^{\frac{1}{2}} \leq \frac{2\sqrt{2}}{N}\max_j |a(j)-\overline{a}_N| \sqrt{\sum_{i=1}^{N}(a(i)-\overline{a}_N)^2},$$

dokazana je i glavna nejednakost (3.9). □

Došli smo do ključnog teorema. Kako sada limes ovisi o N, vraćamo ga u indeks na mjestu gdje smo ga u prethodnim razmatranjima bili izostavili.

Teorem 3.1.5. *Ako je*

$$\delta_N = N\frac{\max\limits_j (z_{N_j} - \overline{z}_N)^2}{\sum_{j=1}^N (z_{N_j} - \overline{z}_N)^2}\frac{\max\limits_j (a_N(j) - \overline{a}_N)^2}{\sum_{j=1}^N (a_N(j) - \overline{a}_N)^2} \to 0, \quad N \to \infty, \tag{3.11}$$

tada vrijedi

$$\frac{S_N - ES_N}{\sqrt{VarS_N}} \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty.$$

Dokaz. Budući da je

$$1 \leq N\frac{\max\limits_j (z_{N_j} - \overline{z}_N)^2}{\sum_{j=1}^N (z_{N_j} - \overline{z}_N)^2},$$

iz prepostavke teorema (3.11) slijedi da je tada nužno

$$\frac{\max\limits_j (a_N(j) - \overline{a}_N)^2}{\sum_{j=1}^N (a_N(j) - \overline{a}_N)^2} \to 0, \quad N \to \infty.$$

Ukoliko bez smanjenja općenitosti pretpostavimo da je $a(j)$ neopadajući niz, prema lemi 3.1.4 imamo nejednakost

$$\frac{E\left[(a(R_1) - a(\lceil NU_1 \rceil))^2\right]}{\frac{1}{N}\sum_{j=1}^N (a(i) - \overline{a}_N)^2} \leq \frac{2\sqrt{2}}{N}\frac{\max\limits_j |a_N(j) - \overline{a}_N|}{\frac{1}{N}\sum_{j=1}^N (a(i) - \overline{a}_N)^2}\sqrt{\sum_{i=1}^N (a(i) - \overline{a}_N)^2} =$$

$$= 2\sqrt{2}\frac{\max\limits_j |a_N(j) - \overline{a}_N|}{\sqrt{\sum_{j=1}^N (a(i) - \overline{a}_N)^2}}.$$

Dakle, pretpostavka teorema daje da desna strana teži u 0 kad $N \to \infty$, pa imamo sljedeće:

$$\frac{E(a(R_1) - a(\lceil NU_1 \rceil))^2}{Var(a(R_1))} \to 0, \quad N \to \infty. \tag{3.12}$$

Iz jednakosti prije iskaza leme 3.1.4 slijedi da je tada

$$Corr(a(R_1), a_N(\lceil NU_1 \rceil)) \to 1, \quad N \to \infty,$$

pa je prema lemi 3.1.3

$$Corr(S_N, S_N^{'}) \to 1, \quad N \to \infty.$$

Sada nam korolar 1.1.2 daje da statistike

$$\frac{S_N^{'}}{\sqrt{VarS_N^{'}}} \quad i \quad \frac{S_N - ES_N}{\sqrt{VarS_N}}$$

imaju jednaku asimptotsku razdiobu. Dokaz teorema završavamo primjernom Lindeberg-Fellerova teorema kako bismo pokazali da uvjet (3.11) implicira

$$\frac{S_N^{'}}{\sqrt{VarS_N^{'}}} \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty.$$

Neka je $B_N^2 = Var(S_N^{'})$. Uočimo da je očekivanje od $X_{Nj} = (z_{Nj} - \overline{z}_N)(a_N(\lceil NU_j \rceil) - \overline{a}_N)$ jednako 0. Ispitajmo Lindebergov uvjet. Neka je $\epsilon > 0$.

$$\frac{1}{B_N^2} \sum_{j=1}^{N} E\left[X_{Nj}^2 \mathbb{1}_{\{|X_{Nj}| \geq \epsilon B_N\}}\right]$$

$$= \frac{1}{B_N^2} \sum_{j=1}^{N} E\left[(z_{Nj} - \overline{z}_N)^2 (a_N(\lceil NU_j \rceil) - \overline{a}_N)^2 \mathbb{1}_{\{(z_{Nj} - \overline{z}_N)^2 (a_N(\lceil NU_j \rceil) - \overline{a}_N)^2 \geq \epsilon^2 B_N^2\}}\right]$$

$$\leq \frac{1}{B_N^2} (z_{Nj} - \overline{z}_N)^2 E\left[(a_N(\lceil NU_j \rceil) - \overline{a}_N)^2 \mathbb{1}_{\{\delta_N \geq \epsilon^2\}}\right] = \mathbb{1}_{\{\delta_N \geq \epsilon^2\}}$$

Zbog pretpostavke (3.11) desna strana teži u nulu kad $N \to \infty$, i time smo dokazali teorem. □

U nastavku ćemo opisati četiri poznata neparametarska testa. Primjenom prethodnog teorema pokazat ćemo asimptotsku normalnost pripadajućih statistika.

Poglavlje 4

Primjene

Opisat ćemo nekoliko neparametarskih testova i pomoću teorema 3.1.5 pokazati asimptotsku normalnost njihovih pripadajućih statistika.

Objasnimo primjenu teorema prvo na statistici za uzorkovanje iz primjera 3.1.1. $a_N(j)$ definiran je s (3.3), gdje n ovisi o N. Vrijedi da je $\overline{a}_N = n/N$ i

$$\begin{aligned}(1/N)\sum_{j=1}^{N}(a_N(j)-\overline{a}_N)^2 &= Var(a_N(R_1)) \\ &= Var(a_N(\lceil NU_1\rceil))\rceil \\ &= (n/N)(1-(n/N)).\end{aligned}$$

Kako je $\frac{1}{4} \leq \underset{j}{max}\,(a_N(j)-\overline{a}_N)^2 \leq 1$, uvjet (3.11) je ekvivalentan s

$$N\frac{\underset{j}{max}(z_{N_j}-\overline{z}_N)^2}{\sum_{j=1}^{N}(z_{N_j}-\overline{z}_N)^2}\frac{N}{n(N-n)} \to 0, \quad N\to\infty. \tag{4.1}$$

Posebno, uvjet (4.1) će biti zadovoljen ukoliko vrijedi da je ili $min\,(n, N-n) \to \infty$ i

$$N\frac{\underset{j}{max}(z_{N_j}-\overline{z}_N)^2}{\sum_{j=1}^{N}(z_{N_j}-\overline{z}_N)^2} \quad \text{je omeđeno,}$$

ili ako je $min\,(n, N-n)/N$ omeđen odozdo pozitivnom konstantom i

$$\frac{\underset{j}{max}(z_{N_j}-\overline{z}_N)^2}{\sum_{j=1}^{N}(z_{N_j}-\overline{z}_N)^2} \to 0, \quad N\to\infty.$$

Ono što možemo iz ovoga zaključiti je

$$\frac{S_N - ES_N}{\sqrt{Var(S_N)}} = \frac{\sqrt{n}\frac{S_N - \bar{z}_N}{n}}{\sqrt{\sigma_z^2\left(1 - \frac{n}{N}\right)}} \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty.$$

Možemo procijeniti σ_z^2 sa varijancom opaženih vrijednosti s_z^2. Ovo nadalje vodi do standardne procedure koja se primjenjuje u teoriji uzorkovanja za dobivanje pouzdanog intervala za populacijsko očekivanje. Međutim, pri tome moramo koristiti procjenu populacijske varijance pomoću varijance uzorka. Tada je uvjet za asimptotsku normalnost taj da N i $N - n$ trebaju biti veliki, a $(N/(n(N-n)))max_j(z_j - \bar{z}_N)^2/s_z^2$ malen. Kako potonji uvjet uključuje neopažene z-ove, ovo se mora pretpostaviti.

■

Primjer 4.0.6. ***Permutacijski t-test dva uzorka***.

Dajemo primjer testiranja na dva nezavisna uzorka. Ovaj test je najjednostavnije objasniti na primjeru pojave novog lijeka i njegova testiranja. Recimo da u bolnici imamo N pacijenata koji boluju od određene bolesti. Pojavio se novi obećavajući lijek i želimo testirati njegovu učinkovitost. Na slučajan način odaberemo $m < N$ pacijenata koji će uzimati novi lijek, dok će ostalih $n = N - m$ pacijenata i dalje primati standardnu terapiju.

Neka slučajne varijable $X_1, X_2, ..., X_m$ opisuju rezultate grupe koja je primala novi lijek, a $Y_1, Y_2, ..., Y_n$ predstavljaju rezultate kontrolne skupine (standardna terapija).Testiraju se hipoteze

H_0 : Nema razlike u mjerenju,
H_1 : Razlika postoji.

Uobičajena statistika koja se koristi kod permutacijskog t-testa dva uzorka je sljedeća:

$$T = \overline{X}_m - \overline{Y}_n, \tag{4.2}$$

gdje je

$$\overline{X}_m = \frac{1}{m}\sum_{j=1}^{m} X_j$$

uzoračka sredina eksperimentalne skupine, dok je

$$\overline{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

uzoračka sredina kontrolne skupine. Permutacijski t-test provodi se uvjetno na opažene vrijednosti. Naime, ako opažene vrijednosti od $X_1, ..., X_m$ i $Y_1, ..., Y_n$ označimo sa $z_1, ..., z_N$,

tada prema svojstvu permutacije, svaki podskup veličine m ima jednaku vjerojnatnost da bude rezultat od $X_1, ..., X_m$. Takvih kombinacija ima $\binom{N}{m}$. Test dalje provodimo na sljedeći način:

- Za svaku od $\binom{N}{m}$ različitih vrijednosti koje može poprimiti $X_1, ..., X_m$, izračunamo vrijednost testne statistike (4.2);
- Dobivene vrijednosti poredamo uzlazno te ih označimo redom s $T_1, T_2, ..., T_{\binom{N}{m}}$;
- Izračunamo realizaciju T statistike na temelju opaženih vrijednosti $z_1, z_2, ..., z_N$ i rezultat označimo sa $T_{realizacija}$;
- Izračunamo p-vrijednost testa ili odredimo kritično područje na temelju zadane razine značajnosti, te iznesemo zaključke.

Statistiku (4.2) možemo pisati u formi (3.1) uz

$$a(j) = \begin{cases} 1/m, & \text{ako je } 1 \leq j \leq m, \\ -1/n, & \text{ako je } (m+1) \leq j \leq N. \end{cases} \tag{4.3}$$

Za razliku od uzorkovanja, za statistiku permutacijskog t-testa za dva uzorka poznate su sve opservacije, pa je primjena teorema 3.1.5 egzaktna. Asimptotski rezultat za permutacijski t-test dva uzorka direktno slijedi iz opisanog postupka kod uzorkovanja. Naime,

$$\frac{S_N - ES_N}{\sqrt{VarS_N}} \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty,$$

ukoliko je zadovoljen uvjet (4.1). U ovom se slučaju σ_z^2 može izračunati, pa nema potrebe za procjenom.

■

Primjer 4.0.7. ***Wilcoxonov test sume rangova***.

Ovaj test se provodi slično kao i permutacijski t-test (uz iste hipoteze), s tim da ovdje raspolažemo rangovima umjesto stvarnim vrijednostima. Dakle, u bolnici se liječi N pacijenata od određene bolesti. Nakon što na slučajan način odaberemo $m < N$ pacijenata koji će uzimati novi lijek (dok će ostalih $n = N - m$ i dalje primati standardnu terapiju), provedemo eksperiment kao kod permutacijskog t-testa, te rezultate rangiramo od najboljeg do najgoreg (na skali od 1 do N). Neka su $R_1, R_2, ..., R_m$ rangovi skupine pacijenata koji su uzimali novi lijek. Tada je pripadna testna statistika

$$W = \sum_{j=1}^{m} R_j. \tag{4.4}$$

Test se provodi uvjetno na opažene vrijednosti rangova. Naime, uz H_0, svaki od $\binom{N}{m}$ različitih m-teročlanih podskupova od $\{1, 2, ..., N\}$ ima jednaku vjerojatnost da bude skup rangova ekperimentalne skupine, odnosno $R_1, ..., R_m$. Dalje je postupak sličan kao kod permutacijskog t-testa.

- Za svaku od $\binom{N}{m}$ različitih vrijednosti rangova $R_1, ..., R_m$ izračunamo vrijednost testne statistike (4.4);
- Dobivene vrijednosti poredamo uzlazno te ih označimo redom sa $W_1, W_2, ..., W_{\binom{N}{m}}$;
- Izračunamo realizaciju W statistike na temelju opaženih vrijednosti $R_1, R_2, ..., R_N$ i rezultat označimo sa $W_{realizacija}$;
- Izračunamo p-vrijednost testa ili odredimo kritično područje na temelju zadane razine značajnosti, te iznesemo zaključke.

Statistiku (4.4) možemo pisati u formi (3.1) uz

$$z_j = j \quad \text{i} \quad a(j) = \begin{cases} 1, & \text{ako je } 1 \leq j \leq m, \\ 0 & \text{ako je } (m+1) \leq j \leq N. \end{cases} \tag{4.5}$$

Test sume rangova poseban je slučaj uzorkovanja, uz $z_j = j$ i n zamijenjen s m. Kako je $\bar{z}_N = (N+1)/2$,

$$\sum_{j=1}^{N} z_j^2 = \frac{N(N+1)(2N+1)}{6},$$

$$\sum_{j=1}^{N} (z_j - \bar{z}_N)^2 = \frac{N(N-1)(N+1)}{12},$$

i $max_j (z_j - \bar{z}_N)^2 = (N-1)^2/4$, uvjet (3.11) će biti ispunjen ukoliko $min\,(n, N-m) \to \infty$, jer je

$$N \frac{\max\limits_j (z_j - \bar{z}_N)^2}{\sum_{j=1}^{N} (z_j - \bar{z}_N)} \quad \text{omeđeno.}$$

Iz leme 3.1.2 slijedi

$$ES_N = N \frac{N+1}{2} \frac{m}{N} = \frac{m(N+1)}{2}$$

i

$$Var(S_N) = \frac{N}{N-1} \frac{N(N-1)(N+1)}{12} \frac{m(N-m)}{N^2} = \frac{m(N-m)(N+1)}{12}.$$

Prema tome je

$$\frac{S_N - ES_N}{\sqrt{Var(S_N)}} \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty.$$

■

Primjer 4.0.8. ***Permutacijski test o trendu.***

Kada uzorak uzimamo sekvencijalno, često nas zanima radi li se o slučajnom poretku opaženih vrijednosti ili postoji određeni trend (rastući ili padajući) kroz vrijeme. Jednostavna testna statistika koju koristimo u ovom testu bazira se na umnošku trenutka opažanja s vrijednošću opservacije,

$$S_N = \sum_{j=1}^{N} jX_j. \tag{4.6}$$

Prema nultoj hipotezi, kako je to i inače u permutacijskim testovima, pretpostavlja se da je svaki od $N!$ različitih poredaka jednako vjerojatan. Jasno je da i statistiku (4.6) možemo zapisati u formi (3.1) uz $a(j) = j$, dok su z_j opažene vrijednosti u nekom poretku.

Za statistiku koja se koristi u permutacijskom testu o trendu pokazat ćemo da vrijedi asimptotska normalnost ukoliko je zadovoljeno sljedeće:

$$\frac{\max_j (z_{Nj} - \overline{z}_N)^2}{\sum_{j=1}^{N}(z_{Nj} - \overline{z}_N)^2} \to 0.$$

Naime, za $a(j) = j$ imamo

$$\overline{a}_N = \frac{N+1}{2} \quad \text{i} \quad max_j(a(j) - \overline{a}_N)^2 = \frac{(N-1)^2}{4},$$

te

$$\sum_{j=1}^{N}(a(j) - \overline{a}_N)^2 = \frac{N(N+1)(N-1)}{12},$$

pa je

$$N\frac{\max_j (a(j) - \overline{a}_N)^2}{\sum_{j=1}^{N}(a(j) - \overline{a}_N)^2} = \frac{N-1}{12(N+1)}$$

omeđeno. Ovime smo pokazali svoju tvdrnju.

■

Primjer 4.0.9. ***Spermanov*** ρ**.**

Imamo još jedan neparametarski model za testiranje trenda. Sličan je permutacijskom testu o trendu, s tim da se ovdje umjesto opaženih vrijednosti promatraju njihovi rangovi. Statistika koju koristimo kod ovog testa je

$$S_N = \sum_{j=1}^{N} jR_j, \tag{4.7}$$

gdje je R_j rang j-te opservacije. Ovo je povezano sa Spermanovim koeficijentom korelacije za rangove ρ_N, koji se definira kao koeficijent korelacije između trenutka opažanja i ranga opažene vrijednosti. Koeficijent korelacije se može zapisati kao

$$\rho_N = \frac{12\frac{1}{N}\sum_{j=1}^{N} jR_j - \left(\frac{N+1}{2}\right)^2}{N^2-1}.$$

Ova statistika i statistika τ_n (Kendallov τ) su dosta slične i koriste se kod jednakih problema.

Ukoliko u prehodno razmatranje o testu o trendu uvedemo $z_j = j$, lagano se pokaže asimptotska normalnost za Spermanov ρ, budući da je

$$\frac{\max\limits_j (z_j - \bar{z}_N)^2}{\sum_{j=1}^{N}(z_j - \bar{z}_N)^2} = \frac{N-1}{12N(N+1)} \to 0, \quad N \to \infty.$$

Očekivanje i varijanca od S_N su

$$ES_N = N\bar{z}_N\bar{a}_N = \frac{N(N+1)^2}{4} \quad \text{i} \quad VarS_N = \frac{N^2(N-1)^2(N+1)^2}{12^2(N-1)} \approx \frac{N^5}{12^2}.$$

Stoga možemo zaključiti da je

$$12\sqrt{N}\left(\frac{1}{N}\sum_{j=1}^{N}\frac{j}{N}\frac{R_j}{N} - \frac{1}{4}\right) \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty.$$

Spermanov koeficijent korelacije rangova ρ_N je koeficijent korelacije između stvarnog ranga j i opaženog ranga R_j, a formula je

$$\rho_N = \frac{12}{N^2-1}\left[\frac{1}{N}\sum_{j=1}^{N} jR_j - \frac{(N+1)^2}{4}\right].$$

Iz ovoga slijedi da uz hipotezu H_0 o slučajnom rangiranju vrijedi

$$\sqrt{N}\rho_N \xrightarrow{D} \mathcal{N}(0,1), \quad N \to \infty.$$

■

Posljednjim primjerom ćemo pokazati rezultate iz primjera 4.0.6 i 4.0.7 na konkretnim podacima.

Primjer 4.0.10. *Zaposlenici tekstilne tvrtke, odjel za sportsku obuću, traže da se promijeni trenutni raspred dozvoljenih pauza. Naime, oni radije žele odlaziti na pauze svaki put kada osjete umor nego pratiti fiksan raspored. Radnici smatraju da će to povećanje slobode povećati njihovu produktivnost, pa su nagovorili vlasnika tvrtke da obavi istraživanje na tu temu. Grupa od 11 zaposlenika toga odjela na slučajan je način razdijeljena na dva dijela. Prvi dio grupe odlazio je na fiksne pauze, dok je drugoj grupi dozvoljeno da odlaze na pauze kad god požele. Tokom vremena ispitivanja bilježila se produktivnost svakog zaposlenika (količina proizvoda). Od njih 11, 6 ih je bilo na fiksnim pauzama, a 5 na proizvoljnim. Rezultati su prikazani u tablici:*

Slobodne paze	*351*	*316*	*480*	*446*	*470*	
Fiksan raspored	*357*	*347*	*380*	*259*	*342*	*282*

Test ćemo provesti na razini značajnosti od 5%. Prvo ćemo promatrati opažene vrijednosti $z_1, z_2, ..., z_{11}$. Očito je u ovom slučaju najbolji odabir permutacijski t-test za dva uzorka. Testiramo hipoteze:

H_0: Način na koji se prakticiraju pauze ne utječe na efikasnost zaposlenika,
H_1: Zaposlenici koji proizvoljno biraju pauze su efikasniji.

Kao što je to uobičajeno u praksi, odabiremo $m = 5$ jer je manji od ponuđenih brojeva 5 i 6. U programskom jeziku R-u osmislili smo kod koji računa svih $\binom{11}{5}$ vrijednosti statistike $T = \overline{X}_5 - \overline{Y}_6$. Budući da se radi o jednostranom testu, nakon što uzlazno sortiramo izračunate vrijednosti statistike,

$$T_1 \leq T_2 \leq T_3 \leq T_4 \leq \cdots \leq T_{459} \leq T_{460} \leq T_{461} \leq T_{462},$$

odaberemo gornjih 5% njih koji će predstavljati kritično područje testa. Prema tome, kritično područje je

$$\{T_{439}, T_{440}, \ldots, T_{462}\},$$

s tim da je $T_{439} = 70.83$. Izračunata realizacija testne statistike T na temelju opaženog uzorka je

$$T_{realizacija} = 84.76667$$

i upada u kritično područje, pa na razini značajnosti od 5% odbacujemo nultu hipotezu u korist alternative koja kaže da su zaposlenici koji odlaze na slobodne pauze efikasniji u obavljanju posla od ostalih.

P-vrijednost testa iznosi

$$p_1 = P(T \geq T_{realizacija}|H_0) = \frac{12}{462} \approx 0.0260,$$

iz čega ponovno zaključujemo da na razini značajnosti od 0.05 odbacujemo hipotezu H_0 u korist alternative. Međutim, na razini značajnosti od 1% ne odbacujemo hipotezu H_0, stoga poslodavac treba biti oprezan kod odabira razine značajnosti na kojoj će obavljati testiranje.

Znamo dalje da se testna statistika $T = \overline{X}_m - \overline{Y}_n$ može zapisati u obliku statistike S_N iz prethodnog poglavlja, pa ćemo iskoristiti formule za očekivanje i varijancu testne statistike S_N, odnosno T. Dobijemo

$$ET = 0 \quad \text{i} \quad VarT = 1930.667,$$

te nakon što normaliziramo podatke $T_1, ..., T_{462}$ pomoću formule

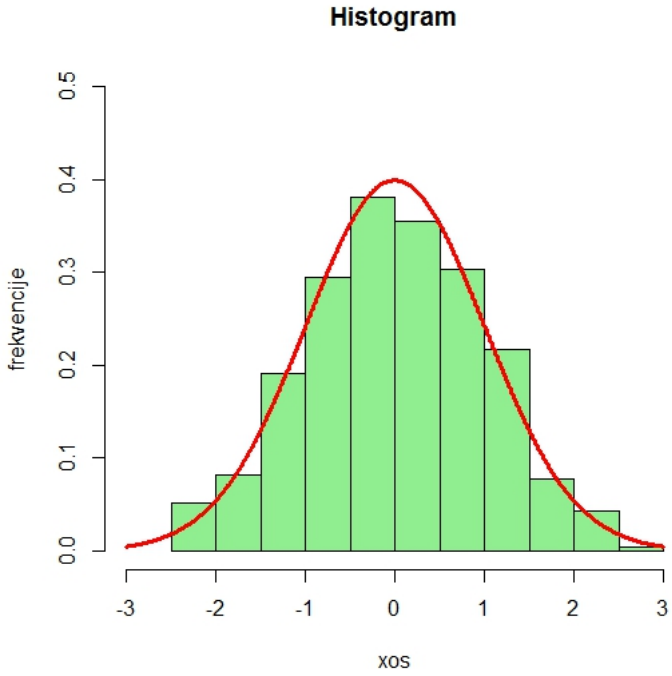
$$\frac{T - ET}{\sqrt{VarT}},$$

dobijemo podatke koje smo prikazali na slici 4.1 u obliku histograma. Crvenom smo bojom obilježili graf funkcije gustoće jedinične normalne razdiobe. Ova nam slika zapravo intuitivno značajno približava razmatranja iz prethodnog poglavlja jer je očigledno da se normalizirani podaci ponašaju kao podaci iz normalne distribucije. Vidjet ćemo sada koliko nam je olakšan posao ukoliko iskoristimo asimptotsku normalnost statistike T.

Znamo da je na razini značajnosti od 5% kritično područje kod jedinične normalne razdiobe interval $[1.65, +\infty\rangle$, dok je vrijednost normalizirane T statistike na temelju opaženih vrijednosti jednaka 1.92. Budući da ta vrijednost upada u kritično područje, odbacujemo hipotezu H_0 u korist alternative. Na ovaj način izračunata p-vrijednost iznosi

$$p_2 = P\left(\frac{T - ET}{\sqrt{VarT}} \geq 1.92|H_0\right) \approx 0.0269.$$

Zaključci su jednaki kao i za p_1.



Slika 4.1: Histogram relativnih frekvencija T statistike i krivulja standardne normalne razdiobe.

Za ovaj problem upotrijebit ćemo i test sume rangova. Opaženim vrijednostima iz početne tablice dodat ćemo odgovarajuće rangove i formulirati novu tablicu.

Slobodne pauze	351	316	480	446	470	
Rangovi	6	3	11	9	10	
Fiksan raspored	357	347	380	259	342	282
Rangovi	7	5	8	1	4	2

Testiraju se jednake hipoteze, pa uz H_0 svaki peteročlani podskup skupa $\{1, 2, ..., 11\}$ ima jednaku vjerojatnost da bude skup rangova grupe ispitanika koji su odlazili na slobodne pauze. Ponovno imamo 462 vrijednosti statistike

$$W = \sum_{j=1}^{5} R_j,$$

gdje su $R_1, ..., R_5$ rangovi testne skupine.

U R-u smo izračunali svih 462 vrijednosti statistike W, te rezultate sortirali uzlazno kako bismo jednostavnije odredili kritično područje,

$$W_1 \leq W_2 \leq W_3 \leq W_4 \leq \cdots \leq W_{459} \leq W_{460} \leq W_{461} \leq W_{462}.$$

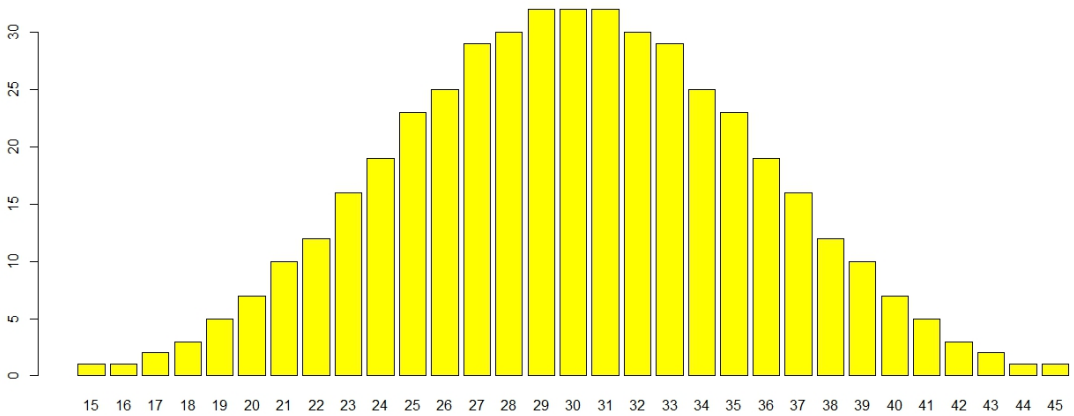
Dakle, kritično područje ovog jednostranog testa za razinu značajnosti 1%, uz H_0, je

$$\{W_{458}, W_{459}, \ldots, W_{462}\}.$$

Kako je $W_{456} = W_{457} = W_{458} = 42$, obuhvatit ćemo i W_{456} i W_{457} u kritično područje. Uz neznatno veću razinu značajnosti i izračunatu realizaciju opažene testne statistike

$$W_{realizacija} = 6 + 3 + 11 + 9 + 10 = 39$$

koja ne upada u kritično područje zaključujemo da ne odbacujemo početnu pretpostavku u korist alternative.



Slika 4.2: Stupičasti dijagram statistike W .

P-vrijednost testa je

$$p_a = \frac{29}{462} \approx 0.0628,$$

stoga na bitnim razinama značajnosti ne možemo odbaciti hipotezu H_0 (osim uz recimo 10%). Očigledno se uvođenjem rangova gube mnoge informacije, stoga ćemo prednost dati zaključcima koje smo iznijeli primjenom permutacijskog t-test.

Pogledajmo što se događa kada iskoristimo asimptotsku normalnost statistike W. U prilog nam ide i stupičasti dijagram 4.2 koji pokazuje kako se vrijednosti W_i ponašaju kao normalno distribuirani podaci. Izračunamo očekivanje, varijancu, pa i realizaciju normalizirane testne statistike:

$$EW = 30, \quad VarW \approx 27.73 \quad \text{i} \quad \frac{W - EW}{\sqrt{VarW}} \approx 1.7092.$$

Na razini značajnosti od 5% odbacujemo nultu hipotezu jer realizacija upada u kritično područje $[1.65, +\infty\rangle$. P-vrijednost testa izračunata koristeći asimptotsku normalnost je

$$p_b = 0.0437,$$

iz čega zaključujemo da nultu hipotezu odbacujemo na razini značajnosti od 5%, dok ju ne odbacujemo na razini značajnosti od 1%.

■

Bibliografija

[1] R. Durrett, *Probability: Theory and Examples*, https://www.math.duke.edu/~rtd/PTE/PTE4-1.pdf, (Stranici pristupljeno 8.11.2015.g.)

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

[3] M. K. Pelosi i T. M. Sandifer, *Elementary Statistics : From Discovery to Decision*, John Wiley and Sons, Hoboken, 2003.

[4] N. Sarapa, *Teorija vjerojatnosti*, Školska Knjiga, Zagreb, 2002.

[5] P. Sprent i N. C. Smeeton, *Applied Nonparametric Statistical Methods*, Chapman and Hall/CRC, Boca Raton, 2009.

[6] Blanka Vlašić, *Official page: Nothing but vision*, http://www.blanka-vlasic.hr/rezultati/, (Stranici pristupljeno 1.8.2015.g.)

Sažetak

U prvom se poglavlju prisjećamo osnovnih tipova konvergencija slučajnih varijabli. Navodimo nekoliko novih tvrdnji od kojih neke i dokazujemo.

U drugom poglavlju iskazujemo Lindeberg-Fellerov centralni granični teorem i primjenjujemo ga direktno na neke statistike. Ovdje također iznosimo neke primjere koji koriste statistike koje zadovoljavaju Lindebergov uvjet.

Glavni rezultat obrađujemo u trećem poglavlju. Iskazujemo i dokazujemo posebnu verziju Lindeberg-Fellerova teorema za statistike bazirane na rangovima.

U posljednjem poglavlju primjenjujemo prethodni teorem na statistike koje se pojavljuju u nekim neparametarskim metodama, kao npr. statistika kod permutacijskog t-testa dva uzorka, statistika kod testa sume rangova, itd.

Summary

In the first chapter we recall to basic types of convergences of random variables. Here we state a few claims about convergences, some of them with their proofs.

In the second chapter we state Lindeberg-Feller Central Limit Teorem without proof. Then we apply it directly on some statistics. We also state some interesting examples that use statistics satisfying Lindeberg Condition.

The main result is in the third chapter. There we state a special version of Lindeberg-Feller Theorem for the statistics based on ranks, with its proof.

In the last chapter we apply this theorem to statistics that we use in some nonparametric methods, such as statistic for the Two-Sample Randomisation t-Test, or statistic for the Rank-Sum Test, etc.

Životopis

Rođena sam 8. lipnja 1990. godine u Mostaru. Osnovnu školu završila sam u Širokom Brijegu 2005. godine. U trećem razredu osnovne škole upisujem osnovnu glazbenu školu koju završavam 2005.godine na odjelu za glasovir Osnovne glazbene škole Široki Brijeg. Nakon završene osnovne škole upisujem opću gimnaziju fra. Dominika Mandića u Širokom Brijegu. Maturirala sam 2009. godine, nakon čega upisujem preddiplomski studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Godine 2013. stekla sam titulu prvostupnice matematike, nakon čega u rujnu iste godine upisujem diplomski studij Matematička statistika na istom fakultetu.