U-statistike i primjene

Brdarić, Nikolina

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike: Financijska matematika i statistika

Nikolina Brdarić

U-statistike i primjena

Diplomski rad

Osijek, *2016.*

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike: Financijska matematika i statistika

Nikolina Brdarić

U-statistike i primjena

Diplomski rad

Mentor: prof.dr.sc. Mirta Benšić
Lektor: Gordana Lušić, prof.

Osijek, *2016.*

Sadržaj

1 Uvod

U ovome radu bavit ćemo se pojmom U-statistike, statistike koja je razvijena kao nepristran procjenitelj statističkog funkcionala.

Definiciju i osnove teorije o U-statistikama dao je Wassily Hoeffding 1948. godine. U-statistike posebno su važne u statističkoj teoriji procjene, grani statistike koja se bavi procjenom vrijednosti parametara na temelju izmjerenih podataka. Koristimo ih pri određivanju asimptotske normalnosti i računanju varijance onih statistika koje se mogu prikazati pomoću U-statistika.

Kako bismo u potpunosti mogli razumjeti teoriju iza U-statistika, u drugom poglavlju ćemo dati pregled osnovnih pojmova i teorema u teoriji vjerojatnosti i statistici. Posebno važni bit će pojmovi nepristranog procjenitelja i nepristranog procjenitelja minimalne varijance kao i pojam asimptotski normalnog procjenitelja. U zasebnom potpoglavlju definirat ćemo statistički test, statističke hipoteze, kritično područje i p-vrijednost.

U trećem poglavlju prvo ćemo se upoznati s konceptom U-statistike te ćemo ga ilustrirati na nekoliko konkretnih primjera. Nakon što uvedemo pojam U-statistike, u sljedećim potpoglavljima bavit ćemo se njenom varijancom i asimptotskom distribucijom.

Četvrto poglavlje ćemo posvetiti Wilcoxonovom testu ranga i predznaka te ćemo na njemu demonstrirati primjenu U-statistika.

2 Osnovni rezultati iz vjerojatnosti i statistike

U statistici zaključujemo o populacijama. Međutim, populacija je često jako velik skup pa podatke mjerimo na jednom dijelu populacije koji nazivamo *uzorak*. Ukoliko je svaka jedinka populacije imala jednaku vjerojatnost da bude izabrana u uzorak, govorimo o *slučajnom uzorku*. Analiziranjem uzorka donosimo zaključke o populaciji.

Pretpostavimo da se realizacija nekog slučajnog uzorka sastoji od $n \in \mathbb{N}$ mjerenja. Dakle, raspolažemo sa $\mathbf{x} = (x_1, \ldots, x_n)$ podataka koje možemo shvatiti kao jednu realizaciju slučajnog vektora $\mathbf{X} = (X_1, \ldots, X_n)$ kojim modeliramo slučajni uzorak.

Često svojstva populacije modeliramo slučajnom varijablom X koja ima distribuciju F. Tada podatke $\mathbf{x} = (x_1, \ldots, x_n)$ shvaćamo kao n nezavisnih realizacija te slučajne varijable. Dakle, podaci $\mathbf{x} = (x_1, \ldots, x_n)$ predstavljaju realizaciju slučajnog vektora $\mathbf{X} = (X_1, \ldots, X_n)$ čije su komponente $X_i, i = 1, \ldots, n$ nezavisne i jednako distribuirane slučajne varijable s distribucijom F. Definirajmo sada *statistički model*.

Definicija 2.1. *Statistički model $\mathcal{F}$ je familija dopuštenih funkcija distribucije F slučajnog vektora $\mathbf{X} = (X_1, \ldots, X_n)$ za koji podaci $\mathbf{x} = (x_1, \ldots, x_n)$ čine jednu realizaciju. Ako je familija $\mathcal{F}$ poznata do na neki k-dimenzionalni parametar distribucije $\theta \in \Theta \subseteq \mathbb{R}^k$, onda kažemo da je model parametarski. U suprotnom kažemo da je model neparametarski.*

2.1 Procjena parametra

U ovom potpoglavlju dat ćemo osnovne pojmove i teoreme vezane uz procjenu nepoznatog parametra θ.

Definicija 2.2. *Neka je $t : \mathbb{R}^n \to \mathbb{R}^k$ izmjeriva funkcija i $\mathbf{X} : \Omega \to \mathbb{R}^n$, $\mathbf{X} = (X_1, \ldots, X_n)$, slučajni vektor. Kompoziciju funkcija $T = t \circ \mathbf{X} : \Omega \to \mathbb{R}^k$ nazivamo statistika.*

Definicija 2.3. *Ako za izmjerivu funkciju t iz Definicije 2.2. vrijedi da $t : \mathbb{R}^n \to \Theta \subseteq \mathbb{R}^k$, gdje je $\Theta \subseteq \mathbb{R}^k$ skup svih dopuštenih vrijednosti nepoznatog parametra θ, statistiku $T = t(X_1, \ldots, X_n)$ nazivamo procjenitelj parametra θ i označavamo s*

$$\hat{\theta} = t(X_1, \ldots, X_n).$$

Kao primjer statistike navodimo *uređajnu statistiku*, jednu od najkorištenijih statistika u neparametarskim statističim modelima.

Primjer 2.1. *Sortiranjem slučajnog uzorka* $(X_1, \ldots, X_n)$ *tako da vrijedi*

$$X_{(1)} \leq \ldots \leq X_{(n)},$$

dobivamo statistiku $t(X_1, \ldots, X_n) = (X_{(1)}, \ldots, X_{(n)})$*, koju nazivamo uređajna statistika.*

Svaka statistika predstavlja način redukcije podataka. Umjesto korištenja cijelog uzorka, zaključujemo na temelju statistike. To je ono što želimo od procjenitelja, a potragu za dobrim procjeniteljima počinjemo s *dovoljnim statistikama.*

Definicija 2.4. *Neka je* $\mathbf{X} = (X_1, \ldots, X_n)$ *slučajni uzorak iz statističkog modela* $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ *i* $T = t(X_1, \ldots, X_n), t : \mathbb{R}^n \to \mathbb{R}^k$ *statistika. Kažemo da je* T *dovoljna statistika za parametar* θ*, ako za svaki* $\mathbf{t} = t(x_1, \ldots, x_n) \in \mathbb{R}^k$ *uvjetna distribucija od* $\mathbf{X}$*, uz dani* $T = \mathbf{t}$*, ne ovisi o* θ*.*

Teorem 2.1 (Neymanov teorem o faktorizaciji)**.** *Neka je* $\mathbf{X} = (X_1, \ldots, X_n)$ *slučajni uzorak iz statističkog modela* $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$*. Statistika* $T = t(X_1, \ldots, X_n)$ *je dovoljna za* θ *ako i samo ako se funkcija gustoće* $f(\mathbf{X}, \theta)$ *može faktorizirati kao* $f(\mathbf{X}, \theta) = g_\theta(t(\mathbf{X}))h(\mathbf{X})$*, gdje* $h : \mathbb{R}^n \to [0, \infty\rangle$ *i* $g_\theta : \mathbb{R}^k \to [0, \infty\rangle, \forall \theta \in \Theta$*.*

Za dokaz pogledati ([6] Poglavlje 2., str. 45., Korolar 2.6.1.).

Uočimo, ako su $X_1, \ldots, X_n$ nezavisne i jednako distribuirane slučajne varijable s funkcijom distribucije F_θ i funkcijom gustoće f_θ, tada je

$$f_\theta(X_1, \ldots, X_n) = f_\theta(X_1) \cdot \ldots \cdot f_\theta(X_n) = f_\theta(X_{(1)}, \ldots, X_{(n)}) = g_\theta(t(X_1, \ldots, X_n)) \cdot 1$$

te je prema Teoremu 2.1 uređajna statistika dovoljna statistika za θ.

Dovoljnih statistika ima mnogo te bismo htjeli odabrati onu koja najbolje reducira podatke. Stoga, uvodimo pojam *potpune statistike.*

Definicija 2.5. *Neka je* $\mathbf{X} = (X_1, \ldots, X_n)$ *slučajni uzorak iz statističkog modela* $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ *i* $T = t(X_1, \ldots, X_n)$*,* $t : \mathbb{R}^n \to \mathbb{R}^k$*, statistika.*
Kažemo da je T *potpuna statistika za nepoznati parametar* θ*, ako za svaku izmjerivu funkciju* g *za koju je* $E_\theta[g(T)] = 0$*,* $\forall \theta \in \Theta$*, vrijedi da je* $P_\theta(g(T) = 0) = 1$*.*

Može se pokazati kako je uređajna statistika nezavisnih i jednako distribuiranih slučajnih varijabli potpuna za familiju svih neprekidnih funkcija distribucije. Za dokaz pogledati ([6] Poglavlje 4., str. 118., Primjer 4.3.4.).
Radi jednostavnosti ćemo u daljnjem tijeku rada nepoznati parametar θ koji procjenjujemo smatrati jednodimenzionalnim, tj. pretpostavit ćemo da je $\theta \in \Theta \subseteq \mathbb{R}$. Često ćemo koristiti pojam *nepristranog procjenitelja.*

Definicija 2.6. *Neka je $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ statistički model. Statistika $T : \Omega \to \Theta$ je nepristran procjenitelj za θ, ako vrijedi da je $E_\theta[T] = \theta, \forall\theta \in \Theta$.*
Procjenitelj koji nije nepristran naziva se pristrani procjenitelj te je njegova pristranost dana s $b_\theta(T) = E_\theta[T - \theta]$.

Osim što je poželjno da je očekivana vrijednost procjenitelja jednaka pravoj vrijednosti parametra, želimo i da procjenitelj ima što je moguće manju varijancu. Takvog procjenitelja nazivamo *nepristrani procjenitelj minimalne varijance*. Sljedeći teorem govori nam u kojem smjeru krenuti kada želimo smanjiti varijancu nepristranog procjenitelja.

Teorem 2.2 (Rao-Blackwell)**.** *Neka je $\mathbf{X} = (X_1, \ldots, X_n)$ slučajni uzorak iz statističkog modela $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ i neka je $T = t(X_1, \ldots, X_n)$ dovoljna statistika za θ. Neka je $S = s(X_1, \ldots, X_n)$ nepristran procjenitelj za $g(\theta), g : \Theta \to \mathbb{R}$ konačne varijance za sve $\theta \in \Theta$.*
Tada za procjenitelja $S^* = E_\theta[S|T]$ vrijedi:

(i) S^* je nepristrani procjenitelj za $g(\theta)$

(ii) $Var_\theta(S^*) < Var_\theta(S)$, osim ako je $P_\theta(S^* = S) = 1$.

Za dokaz pogledati ([5] Poglavlje 1., str. 47., Teorem 7.8.).

Teorem 2.2 sugerira da nepristrani procjenitelj treba biti funkcija dovoljne statistike. Ako nije, uvjetovanjem nepristranog procjenitelja na dovoljnu statistiku možemo konstruirati procjenitelja manje varijance. Sljedeći teorem govori nam kakva mora biti dovoljna statistika kojom uvjetujemo procjenitelja da bismo dobili najmanju moguću varijancu.

Teorem 2.3 (Lehmann-Scheffe)**.** *Neka je $T = t(X_1, \ldots, X_n)$ potpuna dovoljna statistika za θ i neka je $S = s(X_1, \ldots, X_n)$ nepristran procjenitelj za $g(\theta), g : \Theta \to \mathbb{R}$ konačne varijance za sve $\theta \in \Theta$.*
Tada procjenitelj $S^* = E_\theta[S|T]$ ima najmanju varijancu među svim nepristranim procjeniteljima konačne varijance za $g(\theta)$ i jedinstven je P_θ gotovo sigurno, za sve $\theta \in \Theta$.

Za dokaz pogledati ([8] Poglavlje 1., str. 1., Teorem 2.).

Dakle, ako je nepristran procjenitelj konačne varijance funkcija potpune dovoljne statistike, npr. uređajne statistike dane Primjerom 2.1., onda je on i nepristrani procjenitelj minimalne varijance te je gotovo sigurno jedinstven za taj parametar.

2.2 Asimptotska teorija

Slučajni uzorak koji smo do sada promatrali bio je veličine n. Asimptotska teorija pretpostavlja kako veličina uzorka može rasti u nedogled, tj, da $n \to \infty$. U primjeni smatramo kako su asimptotski rezultati približno točni i za konačne slučajne uzorke čija je veličina n dovoljno velika.
U teoriji vjerojatnosti razlikujemo više vrsta konvergencije nizova slučajnih varijabli.

Definicija 2.7. *Neka je* $1 \leq p < \infty$ *i neka je* $(X_n, n \in \mathbb{N})$ *niz slučajnih varijabli na vjerojatnosnom prostoru* $(\Omega, \mathcal{P}, P)$ *takav da* $E[|X_n|^p] < \infty, \forall n \in \mathbb{N}$. *Ovaj niz konvergira u srednjem reda p prema slučajnoj varijabli* X *na* $(\Omega, \mathcal{P}, P)$ *ako vrijedi*

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

Pišemo: $X_n \xrightarrow{m^p} X, n \to \infty$.

Definicija 2.8. *Niz slučajnih varijabli* $(X_n, n \in \mathbb{N})$ *na vjerojatnosnom prostoru* $(\Omega, \mathcal{P}, P)$ *konvergira po vjerojatnosti prema slučajnoj varijabli* X *na* $(\Omega, \mathcal{P}, P)$ *ako vrijedi*

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0, \quad \forall \varepsilon > 0.$$

Pišemo: $X_n \xrightarrow{P} X, n \to \infty$.

Definicija 2.9. *Niz slučajnih varijabli* $(X_n, n \in \mathbb{N})$ *na vjerojatnosnom prostoru* $(\Omega, \mathcal{P}, P)$ *konvergira po distribuciji prema slučajnoj varijabli* X *ako vrijedi*

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

$\forall x \in \mathbb{R}$ *u kojem je* F_X *neprekidna funkcija.*

Pišemo: $X_n \xrightarrow{D} X, n \to \infty$.

Sljedeći teorem daje nam poveznicu među raznim tipovima konvergencija slučajnih varijabli. Iskazat ćemo i druge potrebne teoreme iz teorije vjerojatnosti.

Teorem 2.4. *Neka* $n \to \infty$. *Vrijede sljedeće implikacije*

(*i*) $X_n \xrightarrow{m^p} X \Rightarrow X_n \xrightarrow{P} X \quad (1 \leq p < \infty)$;

(*ii*) $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$.

Za dokaz pogledati ([7] Poglavlje 10., str. 322., Teorem 10.12.).

Teorem 2.5 (Slutsky)**.** *Neka $n \to \infty$ te neka $X_n \xrightarrow{D} X$ i $Y_n \xrightarrow{D} c, c \in \mathbb{R}$. Tada vrijedi*

(i) $X_n + Y_n \xrightarrow{D} X + c;$

(ii) $X_n Y_n \xrightarrow{D} Xc;$

(iii) $\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}, c \neq 0.$

Za dokaz pogledati ([3] Poglavlje 2., str. 45., Teorem 2.34.).

Teorem 2.6 (Centralni granični teorem - Lindberg-Levy)**.** *Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih slučajnih varijabli s očekivanjem μ i varijancom $\sigma^2 < \infty$.*
Tada niz slučajnih varijabli $(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}, n \in \mathbb{N})$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, konvergira po distribuciji prema standardnoj normalnoj slučajnoj varijabli, tj.

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{D} Z \sim \mathcal{N}(0,1), \quad n \to \infty.$$

Za dokaz pogledati ([7] Poglavlje 14., str. 507., Teorem 14.1.).

Pretpostavimo sada da proučavamo niz statistika $(T_n, n \in \mathbb{N})$, gdje je $T_n = t(X_1, \ldots, X_n), t : \mathbb{R}^n \to \mathbb{R}$. Dakle, niz statistika $(T_n, n \in \mathbb{N})$ je niz slučajnih varijabli te je vidljivo kako one ovise o veličini uzorka. Definirajmo *asimptotsku normalnost* niza procjenitelja.

Definicija 2.10. *Neka je $(T_n, n \in \mathbb{N})$ niz procjenitelja za parametar $\theta \in \Theta \subseteq \mathbb{R}$. Kažemo da je T_n asimptotski normalan procjenitelj ako postoji niz nenegativnih brojeva $(a_n, n \in \mathbb{N})$ takav da*

$$\frac{T_n - \theta}{a_n} \xrightarrow{D} Z \sim \mathcal{N}(0,1), \quad n \to \infty.$$

Teorem 2.7 (Slabi zakon velikih brojeva)**.** *Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih slučajnih varijabli s očekivanjem μ i varijancom $\sigma^2 < \infty$.*
Tada niz slučajnih varijabli $(\bar{X}_n, n \in \mathbb{N})$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, konvergira po vjerojatnosti prema svome očekivanju, tj.

$$\bar{X}_n \xrightarrow{P} \mu, \quad n \to \infty.$$

Za dokaz pogledati ([4] Poglavlje 2., str. 49., Teorem 2.1.2.).

2.3 Statistički test

Statistička hipoteza je pretpostavka o populacijskoj distribuciji promatrane slučajne varijable ili vektora na temelju kojih smo definirali statistički model $\mathcal{F}$, označavamo ju s $\mathcal{H}$. U statističkom testu razlikujemo dvije hipoteze, *nul-hipotezu* $\mathcal{H}_0$ i *alternativnu hipotezu* $\mathcal{H}_1$. Svaka statistička hipoteza $\mathcal{H}$ je podskup statističkog modela $\mathcal{F}$, a za $\mathcal{H}_0$ i $\mathcal{H}_1$ vrijedi da je

$$\mathcal{H}_0 \cup \mathcal{H}_1 = \mathcal{F}, \quad \mathcal{H}_0 \cap \mathcal{H}_1 = \emptyset.$$

Kažemo da je statistička hipoteza jednostavna ako je njome jednoznačno određena populacijska distribucija, npr. $\mathcal{H} : \theta = 0$. U suprotnom, hipoteza je složena, npr. $\mathcal{H} : \theta > 0$.

Statističkim testom zapravo testiramo istinitost nul-hipoteze, odnosno donosimo odluku o neodbacivanju ili odbacivanju $\mathcal{H}_0$. Ako odbacimo $\mathcal{H}_0$, prihvaćamo alternativnu hipotezu $\mathcal{H}_1$. Dakle, *statistički test* je pravilo temeljeno na realizaciji slučajnog uzorka iz populacije na temelju kojeg donosimo odluku o odbacivanju ili neodbacivanju nul-hipoteze $\mathcal{H}_0$.

2.3.1 Kritično područje

Označimo sa $\mathcal{X}$ skup svih mogućih realizacija $\mathbf{x}$ slučajnog uzorka. Statistički test dijeli $\mathcal{X}$ na dva disjunktna skupa $\mathcal{C}_r$ i $\mathcal{C}_r^c$. Skup $\mathcal{C}_r \subseteq \mathcal{X}$ nazivamo *kritično područje* te ako realizacija uzorka pripada kritičnom području, hipoteza $\mathcal{H}_0$ se odbacuje. Skup $\mathcal{C}_r$ odabire se tako da sadrži one realizacije uzorka u kojima dolazi do značajnog odstupanja od hipoteze $\mathcal{H}_0$. Ako se $\mathcal{C}_r$ može izraziti u terminima neke statistike T, onda tu statistiku nazivamo *test-statistika*.

2.3.2 Funkcija jakosti i pogreške statističkog testa

Statistički test provodi se uz toleranciju malih vjerojatnosti pogrešne odluke. Uvodimo *funkciju jakosti testa* $\pi : \mathcal{F} \to [0, 1]$ definiranu s

$$\pi(F) := P_F(\mathbf{X} \in \mathcal{C}_r).$$

Jakost testa predstavlja vjerojatnost odbacivanja $\mathcal{H}_0$ kada je F prava distribucija populacije.

Razlikujemo dva tipa pogreške:

- pogreška I. tipa
- pogreška II. tipa

Počiniti pogrešku I. tipa znači odbaciti $\mathcal{H}_0$ ako je ona istinita. Ako ne odbacimo $\mathcal{H}_0$ u uvjetima istinitosti hipoteze $\mathcal{H}_1$, činimo pogrešku II. tipa. U statističkom testu želimo postići što manju vjerojatnost pogrešaka I. i II. tipa. Vjerojatnost pogreške I. tipa možemo iskazati kao preslikavanje $\alpha : \mathcal{H}_0 \to [0, 1]$ definirano s

$$\alpha(F_0) := \pi(F_0) = P_{F_0}(\mathbf{X} \in \mathcal{C}_r).$$

Preslikavanje $\beta : \mathcal{H}_1 \to [0, 1]$ definirano s

$$\beta(F_1) := \pi(F_1) = P_{F_1}(\mathbf{X} \in \mathcal{C}_r)$$

naziva se *snaga statističkog testa*. Vjerojatnost pogreške II. tipa iznosi $P_{F_1}(\mathbf{X} \in \mathcal{C}_r^c) = 1 - \beta(F_1)$.
Budući da je

$$\pi(F) = P_F(\mathbf{X} \in \mathcal{C}_r) = 1 - P_F(\mathbf{X} \in \mathcal{C}_r^c),\ F \in \mathcal{F},$$

smanjivanjem vjerojatnosti pogreške I. tipa povećava se vjerojatnost pogreške II. tipa i obrnuto. Statistički test kreiran je tako da dopušta odabir maksimalne vjerojatnosti pogreške I. tipa. Tu vjerojatnost nazivamo *razina značajnosti* statističkog testa i označavamo s α. Najčešća razina značajnosti je $\alpha = 0.05$.

Dodatno, ako za $F_0 \in \mathcal{H}_0$ vrijedi da $P_{F_0}(\mathbf{X} \in \mathcal{C}_r) \to \alpha$ tada kažemo da statistički test ima *asimptotsku razinu značajnosti* α.

2.3.3 p-vrijednost

Pretpostavimo da je nul-hipoteza $\mathcal{H}_0$ statističkog testa jednostavna te da kritično područje toga testa možemo izraziti u terminima test-statistike na sljedeći način:

$$\mathcal{C}_r = \{\mathbf{x} : t(\mathbf{x}) > c_\alpha\},$$

pri čemu je $T = t(\mathbf{X}) : \mathbb{R}^n \to \mathbb{R}$ test-statistika i $\alpha = P_{F_0}(T > c_\alpha)$ razina značajnosti, dok $c_\alpha \in \mathbb{R}$ nazivamo kritična vrijednost testa.
Vidimo da, ako je realizacija $t(\mathbf{x})$ statistike T veća od c_α, onda ta realizacija pripada kritičnom području $\mathcal{C}_r$ te tada odbacujemo $\mathcal{H}_0$. Međutim, c_α je često teško izračunati. Alternativno, možemo izračunati vjerojatnost $p := P_{F_0}(T > t(\mathbf{x}))$ koju nazivamo *p-vrijednost* i odbaciti $\mathcal{H}_0$ onda kada je $p < \alpha$.

3 U-statistika

3.1 Parametarski i neparametarski statistički modeli

Iz Definicije 2.1. jasno je da u parametarskim statističkim modelima pretpostavljamo kako slučajan vektor $\mathbf{X}$ iz kojeg dolaze podaci ima distribuciju F_θ koja je poznata do na točno određeni parametar θ fiksne dimenzije k. Dakle, u parametarskim statističkim modelima, na osnovi slučajnog uzorka kojim raspolažemo, procjenjujemo parametar θ.

S druge strane, u neparametarskim statističkim modelima oblik distribucije F slučajnog vektora $\mathbf{X}$ je u potpunosti nepoznat. Pretpostavke o distribuciji u ovakvim modelima vrlo su slabe poput pretpostavke o neprekidnosti distribucije ili postojanju momenata distribucije. U tom slučaju parametre koje procjenjujemo opisujemo pomoću realnih funkcija definiranih na nekoj familiji distribucija $\mathcal{F}$.

U nastavku rada bavit ćemo se neparametarskim statističkim modelima.

3.2 Statistički funkcional

Neka su $X_1, \ldots, X_n$ nezavisne i jednako distribuirane slučajne varijable s funkcijom distribucije $F \in \mathcal{F}$. Pretpostavimo kako je $\mathcal{F}$ familija svih neprekidnih funkcija distribucije F za koje postoje momenti.

Definicija 3.1. *Funkciju $H : \mathcal{F} \to \mathbb{R}$ nazivamo statistički funkcional.*

Pomoću statističkog funkcionala možemo opisati parametar koji želimo procijeniti te za nepoznatu funkciju distribucije $F \in \mathcal{F}$ želimo saznati vrijednost poznatog statističkog funkcionala $H(F)$.

Primjer 3.1. *Promotrimo neke primjere statističkih funkcionala:*

a. *Statistički funkcional*
$$H(F) = F(c) = P_F(X \leq c),$$
gdje je $c \in \mathbb{R}$ konstanta i X slučajna varijabla s distribucijom F, svakoj funkciji distribucije F pridružuje $P_F(X \leq c)$.

b. *Statistički funkcional*
$$H(F) = F^{-1}(p), \quad p \in [0,1] \quad \textit{konstanta},$$
svakoj funkciji distribucije F pridružuje njen p-ti kvantil.

c. *Statistički funkcional*

$$H(F) = E_F[X],$$

gdje je X slučajna varijabla s distribucijom F, svakoj funkciji distribucije F pridružuje njeno očekivanje, ako ono postoji.

Prirodno se nameću sljedeća pitanja:

- Postoji li procjenitelj za $H(F)$ koji je nepristran neovisno o funkciji distribucije F? Možemo li karakterizirati statističke funkcionale $H(F)$ za koje je odgovor potvrdan?
- Ako nepristran procjenitelj za $H(F)$ postoji, možemo li ga naći? Ako ih postoji nekoliko, koji je "najbolji"?

U sljedećim potpoglavljima, proučavajući U-statistike, odgovorit ćemo na navedena pitanja.

3.3 U-statistika

Primijetimo kako se neki od funkcionala $H(F)$ iz Primjera 3.1. mogu zapisati kao očekivanje. Takve funkcionale od sada ćemo označavati sa $\theta(F)$. Funkcional $\theta(F) = E_F[X]$ je sam po sebi već očekivanje. Međutim, funkcional $\theta(F) = F(c) = P_F(X \leq c)$, $c \in \mathbb{R}$ konstanta, također možemo zapisati kao $\theta(F) = E_F[\phi(X)]$, gdje je funkcija $\phi : \mathbb{R} \to \{0, 1\}$ definirana s

$$\phi(x) = \begin{cases} 0, & x > c; \\ 1, & x \leq c. \end{cases} \tag{3.1}$$

Kako bismo generalizirali ovu ideju, promotrimo izmjerivu funkciju $\phi : \mathbb{R}^a \to \mathbb{R}$ za neki $a \geq 1$ te definirajmo

$$\theta(F) = E_F[\phi(X_1, \ldots, X_a)], \tag{3.2}$$

gdje su $X_1, \ldots, X_a$ nezavisne i jednako distribuirane slučajne varijable s funkcijom distribucije $F \in \mathcal{F}$.
Uočimo kako bez smanjenja općenitosti u jednadžbi (3.2) možemo pretpostaviti da je funkcija ϕ simetrična[1] funkcija.
Naime, ako funkcija ϕ nije simetrična, zbog toga što su slučajne varijable $X_1, \ldots, X_a$ nezavisne i jednako distribuirane, vrijedi da je

$$E_F[\phi(X_1, \ldots, X_a)] = E_F[\phi(X_{\pi(1)}, \ldots, X_{\pi(a)})], \tag{3.3}$$

za bilo koju permutaciju $\pi : \{1, \ldots, a\} \to \{1, \ldots, a\}$. Budući da postoji $a!$ takvih permutacija, promotrimo funkciju $\phi^* : \mathbb{R}^a \to \mathbb{R}$,

$$\phi^*(x_1, \ldots, x_a) := \frac{1}{a!} \sum_{svi\, \pi} \phi(x_{\pi(1)}, \ldots, x_{\pi(a)}).$$

Funkcija ϕ^* je očito simetrična te zbog jednadžbe (3.3) vrijedi da je $E_F[\phi(X_1, \ldots, X_a)]$
$= E_F[\phi^*(X_1, \ldots, X_a)]$.
Statistički funkcional definiran jednadžbom (3.2) nazivamo *funkcional očekivanja*, a simetrična funkcija ϕ je *funkcija jezgre* funkcionala $\theta(F)$.

Razmotrimo sada procjenu funkcionala $\theta(F)$ na temelju slučajnog uzorka n nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, \ldots, X_n$ s funkcijom distribucije $F \in \mathcal{F}$. Pretpostavljamo da je $a \leq n$.
Pogledamo li jednadžbu (3.2), očito je da je statistika $\phi(X_1, \ldots, X_a)$, prema Definiciji 2.6., nepristran procjenitelj za $\theta(F)$, kao uostalom i svaka statistika $\phi(X_{\pi(1)}, \ldots, X_{\pi(a)})$. Međutim, statistika $\phi(X_1, \ldots, X_a)$ temelji se na samo a podataka od njih ukupno n s kojima

[1]Funkcija $f : \mathbb{R}^n \to \mathbb{R}, \quad f(x_1, \ldots, x_n)$ je simetrična ako poprima istu vrijednost za svaku permutaciju π svojih varijabli.

raspolažemo. Intuitivno, bolji procjenitelj bila bi statistika koja je kompozicija izmjerive simetrične funkcije $U : \mathbb{R}^n \to \mathbb{R}$ i svih n slučajnih varijabli $X_1, \ldots, X_n$. Takav procjenitelj je upravo U-statistika.

Definicija 3.2. *Neka je $a \in \mathbb{N}$ i $\phi(x_1, \ldots, x_a)$ funkcija jezgre funkcionala očekivanja $\theta(F)$. Pripadna U-statistika funkcionala $\theta(F)$ dana je s*

$$U_n := U(X_1, \ldots, X_n) := \frac{1}{\binom{n}{a}} \sum_{1 \le i_1 \le \cdots \le i_a \le n} \cdots \sum \phi(X_{i_1}, \ldots, X_{i_a}), \tag{3.4}$$

gdje su $X_1, \ldots, X_n$ nezavisne i jednako distribuirane slučajne varijable s funkcijom distribucije $F \in \mathcal{F}$ i $a \le n$.

Zbog toga što je ϕ simetrična, u jednadžbi (3.4) sumiramo po svih $\binom{n}{a}$ a-kombinacija n-članog skupa $\{1, \ldots, n\}$ te vidimo kako je i U simetrična funkcija. Primijetimo kako je U_n prosjek $\binom{n}{a}$ članova od kojih svaki ima očekivanje $\theta(F) = E_F[\phi(X_1, \ldots, X_a)]$ pa je stoga U_n nepristran procjenitelj za $\theta(F)$. Čak štoviše, "U" u nazivu "U-statistika" dolazi od engleske riječi *"unbiased"* što znači nepristran.

Budući da je U_n nepristran procjenitelj, zanima nas je li on i nepristran procjenitelj minimalne varijance za parametar $\theta(F)$.
Kako je U simetrična funkcija, vrijedi da je

$$U_n = U(X_1, \ldots, X_n) = U(X_{(1)}, \ldots, X_{(n)}),$$

gdje je $(X_{(1)}, \ldots, X_{(n)})$ uređajna statistika dana u Primjeru 2.1. Dakle, statistika U_n je funkcija uređajne statistike. Pretpostavili smo kako je $\mathcal{F}$ familija svih neprekidnih funkcija distribucije F. Nakon Teorema 2.1 i Definicije 2.5. pokazano je kako je tada uređajna statistika potpuna dovoljna statistika za svaki $F \in \mathcal{F}$. Pretpostavimo li i kako je varijanca statistike U_n konačna, tada je, prema Teoremu 2.3., statistika U_n nepristran procjenitelj najmanje varijance među svim nepristranim procjeniteljima konačne varijance za $\theta(F)$ te je ona gotovo sigurno jedinstvena.

U sljedećim potpoglavljima bavit ćemo se varijancom i asimptotskom distribucijom U-statistike, no prvo pogledajmo nekoliko primjera.
U svim primjerima pretpostavljamo kako su $X_1, \ldots, X_n$ nezavisne i jednako distribuirane slučajne varijable s funkcijom distribucije $F \in \mathcal{F}$.

Primjer 3.2. *Neka je $a = 1$ i $\phi : \mathbb{R} \to \mathbb{R}$ funkcija jezgre. Tada je $\theta(F) = E_F[\phi(X_1)]$ te je pripadna U-statistika dana s*

$$U_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i). \tag{3.5}$$

a. Za $\phi(x_1) = x_1$, pripadna U-statistika za funkcional $\theta(F) = E_F[X_1]$ je aritmetička sredina

$$U_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

b. Za ϕ kao u jednadžbi (3.1), pripadna U-statistika za funkcional $\theta(F) = P_F(X_1 \leq c)$, $c \in \mathbb{R}$ konstanta, je

$$U_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(X_i<c)}.$$

Dakle, očekivanje i vrijednost funkcije distribucije F u realnoj konstanti c možemo zapisati kao funkcionale očekivanja, a njihove procjenitelje kao U-statistike. Pokažimo sada kako je to slučaj i za varijancu čija je pripadna U-statistika korigirana uzoračka varijanca.

Primjer 3.3. *Neka je $a = 2$ i $\phi : \mathbb{R}^2 \to \mathbb{R}$ funkcija jezgre. Tada je $\theta(F) = E_F[\phi(X_1, X_2)]$ te je pripadna U-statistika dana s*

$$U_n = \frac{1}{\binom{n}{2}}\sum_{i<j} \phi(X_i, X_j). \tag{3.6}$$

Za $\phi(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$ dobijemo U-statistiku

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \qquad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}((X_i - \bar{X}_n)^2 + (X_j - \bar{X}_n)^2) \\ &= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}((X_i - \bar{X}_n) - (X_j - \bar{X}_n))^2 \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(X_i - X_j)^2 \\ &= \frac{1}{\binom{n}{2}}\sum_{i<j}\frac{1}{2}(X_i - X_j)^2 \end{aligned} \tag{3.7}$$

koja pripada funkcionalu očekivanja

$$\begin{aligned} \theta(F) = E_F[s_n^2] &= \frac{1}{2}E_F[(X_1 - X_2)^2] \\ &= \frac{1}{2}E_F[((X_1 - E_F[X_1]) - (X_2 - E_F[X_2]))^2] \\ &= \frac{1}{2}E_F[(X_1 - E_F[X_1])^2 - (X_2 - E_F[X_2])^2] \\ &= E_F[(X_1 - E_F[X_1])^2] = \mathrm{Var}_F(X_1). \end{aligned} \tag{3.8}$$

Varijanca se također može prikazati kao funkcional očekivanja, a korigirana uzoračka varijanca s_n^2 je pripadna U-statistika.
Pogledajmo još neke U-statistike u slučaju kada je $a = 2$.

a. *Za $\phi(x_1, x_2) = |x_2 - x_1|$ pripadna U-statistika za funkcional $\theta(F) = E_F[|X_2 - X_1|]$ je*

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} |X_j - X_i|,$$

poznata je pod imenom Ginijeva prosječna razlika.

b. *Za*

$$\phi(x_1, x_2) = \mathbf{I}_{(x_1+x_2>0)} = \begin{cases} 0, & x_1 + x_2 \leq 0; \\ 1, & x_1 + x_2 > 0. \end{cases}$$

pripadna U-statistika za funkcional $\theta(F) = E_F[\phi(X_1, X_2)] = P(X_1 + X_2 > 0)$ je

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{I}_{(X_i+X_j>0)}. \tag{3.9}$$

Ova statistika je usko povezana s test-statistikom u Wilcoxonovom testu ranga i predznaka.

3.3.1 Varijanca U-statistike

Neka je $\phi(x_1, \ldots, x_a)$ funkcija jezgre funkcionala očekivanja $\theta(F) = E_F[\phi(X_1, \ldots, X_a)]$. Pretpostavljamo da su slučajne varijable $X_1, \ldots, X_n$ nezavisne i jednako distribuirane s distribucijom $F \in \mathcal{F}$. U_n je U-statistika iz Definicije 3.2. Za $1 \leq i \leq a$ pretpostavljamo da su vrijednosti varijabli $X_1, \ldots, X_i$ konstantne, tj. $X_1 = x_1, \ldots, X_i = x_i$.
Za $i = 1, \ldots, a$ definiramo funkciju $\phi_i : \mathbb{R}^i \to \mathbb{R}$ s

$$\begin{aligned} \phi_i(x_1, \ldots, x_i) &= E_F[\phi(x_1, \ldots, x_i, X_{i+1}, \ldots, X_a)] \\ &= E_F[\phi(X_1, \ldots, X_a)|X_1 = x_1, \ldots, X_i = x_i]. \end{aligned} \tag{3.10}$$

Ekvivalentno možemo definirati

$$\phi_i(X_1, \ldots, X_i) = E_F[\phi(X_1, \ldots, X_a)|X_1, \ldots, X_i].$$

Dobivamo kako je za svaki $i = 1, \ldots, a$

$$E_F[\phi_i(X_1, \ldots, X_i)] = E_F\Big[E_F[\phi(X_1, \ldots, X_a)|X_1, \ldots, X_i]\Big] = E_F[\phi(X_1, \ldots, X_a)] = \theta(F). \tag{3.11}$$

Nadalje, za $i = 1, \ldots, a$ definiramo

$$\sigma_i^2 = \mathrm{Var}_F\Big(\phi_i(X_1, \ldots, X_i)\Big). \tag{3.12}$$

Sljedeća lema povezuje kovarijancu sa σ_i^2.

Lema 3.1. *Neka je* $\sigma_i^2 = \mathrm{Var}_F\left(\phi_i(X_1,\ldots,X_i)\right),\ i=1,\ldots,a,$
tada

$$\mathrm{Cov}_F\left(\phi(X_1,\ldots,X_i,X_{i+1},\ldots,X_a),\phi(X_1,\ldots,X_i,X_{i+1}^{'},\ldots,X_a^{'})\right)=\sigma_i^2, \tag{3.13}$$

gdje su $X_1,\ldots,X_i,\ X_{i+1},\ldots,X_a,\ X_{i+1}^{'},\ldots,X_a^{'}\subseteq\{X_1,\ldots,X_n\}$.

Dokaz:
Budući da vektori $(X_1,\ldots,X_i,X_{i+1},\ldots,X_a)$ i $(X_1,\ldots,X_i,X_{i+1}^{'},\ldots,X_a^{'})$ imaju i istih slučajnih varijabli i da su $X_{i+1},\ldots,X_a$ i $X_{i+1}^{'},\ldots,X_a^{'}$ međusobno nezavisne, uvjetno na $X_1,\ldots,X_i$ vektori su i nezavisni te vrijedi

$$\begin{aligned}
&\mathrm{Cov}_F\left(\phi(X_1,\ldots,X_i,X_{i+1},\ldots,X_a),\phi(X_1,\ldots,X_i,X_{i+1}^{'},\ldots,X_a^{'})\right)\\
&=E_F\Big[\Big(\phi(X_1,\ldots,X_i,X_{i+1},\ldots,X_a)-\theta(F)\Big)\Big(\phi(X_1,\ldots,X_i,X_{i+1}^{'},\ldots,X_a^{'})-\theta(F)\Big)\Big]\\
&=E_F\Big[E_F\Big[\Big(\phi(X_1,\ldots,X_i,X_{i+1},\ldots,X_a)-\theta(F)\Big)\Big(\phi(X_1,\ldots,X_i,X_{i+1}^{'},\ldots,X_a^{'})-\theta(F)\Big)|X_1,\ldots X_i\Big]\Big]\\
&=E_F\Big[\Big(\phi_i(X_1,\ldots,X_i)-\theta(F)\Big)\Big(\phi(X_1,\ldots,X_i)-\theta(F)\Big)\Big]\\
&=E_F\Big[\Big(\phi_i(X_1,\ldots,X_i)-\theta(F)\Big)^2\Big]\\
&=\mathrm{Var}_F\left(\phi_i(X_1,\ldots,X_i)\right)=\sigma_i^2.
\end{aligned} \tag{3.14}$$

□

Teorem 3.1. *Varijanca U-statistike dane Definicijom 3.2. iznosi*

$$\mathrm{Var}_F(U_n)=\frac{1}{\binom{n}{a}}\sum_{i=1}^{a}\binom{a}{i}\binom{n-a}{a-i}\sigma_i^2, \tag{3.15}$$

gdje je σ_i^2 *dan formulom* (3.13) *te* $X_1,\ldots,X_i,\ X_{i+1},\ldots,X_a,\ X_{i+1}^{'},\ldots,X_a^{'}\subseteq\{X_1,\ldots,X_n\}$.
Ako su $\sigma_i^2<\infty$, *za sve* $i=1,\ldots a$, *tada*

$$\mathrm{Var}(U_n)\sim\frac{a^2\sigma_1^2}{n},$$

za $n\in\mathbb{N}$ *velik broj.*

Dokaz:
Kako bismo izračunali varijancu statistike U_n započnimo s

$$\begin{aligned}
\mathrm{Var}_F(U_n)&=\mathrm{Var}_F\left(\binom{n}{a}^{-1}\sum_{1\le i_1\le\cdots\le i_a\le n}\cdots\sum\phi(X_{i_1},\ldots,X_{i_a})\right)\\
&=\binom{n}{a}^{-2}\sum_{1\le i_1\le\cdots\le i_a\le n}\cdots\sum\ \sum_{1\le j_1\le\cdots\le j_a\le n}\cdots\sum\mathrm{Cov}_F\left(\phi(X_{i_1},\ldots,X_{i_a}),\phi(X_{j_1},\ldots,X_{j_a})\right).
\end{aligned} \tag{3.16}$$

Kako bismo odredili varijancu, iz sume u (3.16) izdvajamo one $(X_{i_1}, \ldots, X_{i_a})$ i $(X_{j_1}, \ldots, X_{j_a})$ koji imaju točno $i = 1, \ldots, a$ zajedničkih varijabli (kao u Lemi 3.1.). Broj takvih vektora je točno $\binom{n}{a}\binom{a}{i}\binom{n-a}{a-i}$, jer postoji $\binom{n}{a}$ načina da odaberemo skup $i_1, \ldots i_a$ i tada $\binom{a}{i}$ načina da iz njega odaberemo podskup koji sadrži i elemenata te na kraju postoji $\binom{n-a}{a-i}$ načina da odaberemo preostalih $a - i$ elemenata skupa $j_1, \ldots j_a$ između preostalih $n - a$ brojeva. Primijetimo da, ako je $i = 0$, tj. ako vektori $(X_{i_1}, \ldots, X_{i_a})$ i $(X_{j_1}, \ldots, X_{j_a})$ nemaju zajedničkih varijabli, tada su oni nezavisni i $\mathrm{Cov}_F\left(\phi(X_{i_1}, \ldots, X_{i_a}), \phi(X_{j_1}, \ldots, X_{j_a})\right) = 0$. Stoga,

$$\begin{aligned} \mathrm{Var}_F(U_n) &= \binom{n}{a}^{-2} \sum_{i=1}^{a} \binom{n}{a}\binom{a}{i}\binom{n-a}{a-i}\sigma_i^2 \\ &= \binom{n}{a}^{-1} \sum_{i=1}^{a} \binom{a}{i}\binom{n-a}{a-i}\sigma_i^2. \end{aligned} \tag{3.17}$$

Kako je

$$\binom{n-a}{k} = \frac{1}{k!}(n-a)(n-a-1)\cdots(n-a-k+1) \sim \frac{n^k}{k!},$$

vidimo kako su, za velike n, u sumi (3.15) članovi koji odgovaraju $i \geq 2$ manjeg reda od prvog člana sume koji odgovara $i = 1$. U prvom članu, koeficijent uz σ_1^2 iznosi $a\binom{n-a}{a-1}\binom{n}{a}^{-1} \sim \frac{an^{a-1}}{(a-1)!}\frac{a!}{n^a} = \frac{a^2}{n}$. Dakle, $\mathrm{Var}(U_n) \sim \frac{a^2\sigma_1^2}{n}$, za veliki n.

□

3.3.2 Asimptotska normalnost U-statistike

Teorem 3.1 pokazuje kako varijanca od $\sqrt{n}U_n$ teži prema $a^2\sigma_1^2$ kada $n \to \infty$. Zanima nas je li $\sqrt{n}(U_n - \theta(F))$ asimptotski normalna s asimptotskom varijancom $a^2\sigma_1^2$.
U Primjeru 3.2., za $a = 1$, jednadžbom (3.5) dana je U-statisika koja je prosjek n nezavisnih i jednako distribuiranih slučajnih varijabli konačne varijance, asimptotska normalnost te U-statistike slijedi direktno iz Centralnog graničnog teorema, Teorem 2.6.
Primijetimo kako općenito U-statistika nije suma nezavisnih slučajnih varijabli te stoga nije moguće direktno primijeniti Slabi zakon velikih brojeva (Teorem 2.7) i Centralni granični teorem (Teorem 2.6) pri utvrđivanju asimptotske distribucije U-statistike.

Teorem 3.2. *Ako je* $0 < \sigma_1^2 < \infty$*, tada, kada* $n \to \infty$*,*

$$\sqrt{n}(U_n - \theta(F)) \xrightarrow{D} N(0, a^2\sigma_1^2); \tag{3.18}$$

ako vrijedi i da je

$$\sigma_i^2 < \infty, \quad \forall i = 2, \ldots, a, \tag{3.19}$$

tada je i

$$\frac{U_n - \theta(F)}{\sqrt{\mathrm{Var}_F(U_n)}} \xrightarrow{D} N(0,1). \tag{3.20}$$

Primjedba 3.1. *Budući da asimptotska varijanca u* (3.18) *sadržava samo* σ_1^2*, čini se neobičnim da u drugom dijelu Teorema 3.2 nije dovoljan osnovni uvjet da je* $\sigma_1^2 < \infty$*. Međutim, može doći do razlike između asimptotske varijance* σ_1^2 *i limesa prave varijance U-statistike. Naime, iz* (3.15) *je vidljivo kako prava varijanca U-statistike ne sadrži samo* σ_1^2*, već i preostale varijance* $\sigma_2^2, \ldots, \sigma_a^2$*. Ako je bilo koja od tih varijanci beskonačna te je istodobno* $\sigma_1^2 < \infty$*, imamo slučaj u kojem je prava varijanca od* $\sqrt{n}U_n$ *beskonačna pa je samim time i njen limes beskonačan. Istodobno, asimptotska varijanca je konačna.*

Teorem 3.2 dokazat ćemo u nekoliko koraka. Osnovna ideja dokaza sastoji se od toga da pokažemo kako $U_n - \theta(F)$ ima istu asimptotsku distribuciju kao i

$$\tilde{U_n} = \sum_{j=1}^{n} E_F[U_n - \theta(F)|X_j]. \tag{3.21}$$

Asimptotska distribucija od $\tilde{U_n}$ slijedi iz Centralnog graničnog teorema, Teorem 2.6, budući da je $\tilde{U_n}$ suma n nezavisnih i jednako distribuiranih slučajnih varijabli.

Lema 3.2. *Za sve* $1 \leq j \leq n$*,*

$$E_F[U_n - \theta(F)|X_j] = \frac{a}{n}(\phi_1(X_j) - \theta(F)).$$

Dokaz:
Uočimo kako je

$$E_F[U_n - \theta(F)|X_j] = \frac{1}{\binom{n}{a}} \sum_{1 \leq i_1 \leq \cdots \leq i_a \leq n} \cdots \sum E_F[\phi(X_{i_1}, \ldots, X_{i_a}) - \theta(F)|X_j],$$

gdje $E_F[\phi(X_{i_1}, \ldots, X_{i_a}) - \theta(F)|X_j]$ iznosi $\phi_1(X_j) - \theta(F)$ uvijek kada je $j \in \{i_1, \ldots, i_a\}$ i 0 inače. Broj načina na koji možemo odabrati $i_1, \ldots, i_a$ koje sadrže j je $\binom{n-1}{a-1}$. Dobivamo,

$$E_F[U_n - \theta(F)|X_j] = \frac{\binom{n-1}{a-1}}{\binom{n}{a}}(\phi_1(X_j) - \theta(F)) = \frac{a}{n}(\phi_1(X_j) - \theta(F)).$$

□

Uočimo kako su slučajne varijable $\frac{a}{n}(\phi_1(X_j) - \theta(F))$ također nezavisne i jednako distribuirane za svaki $j = 1, \ldots, n$.

Lema 3.3. *Ako je* $\sigma_1^2 < \infty$ *i* $\tilde{U_n}$ *definiran kao u jednadžbi* (3.21)*, tada*

$$\sqrt{n}\tilde{U_n} \xrightarrow{D} N(0, a^2\sigma_1^2), \quad n \to \infty.$$

Dokaz:
Primijetimo kako slučajna varijabla $a\phi_1(X_j)$ ima očekivanje

$$E_F[a\phi_1(X_j)] = aE_F[\phi_1(X_j)] \overset{(3.11)}{=} a\theta(F),$$

i varijancu

$$\mathrm{Var}_F(a\phi_1(X_j)) = a^2 \mathrm{Var}_F(\phi_1(X_j)) \overset{(3.12)}{=} a^2\sigma_1^2. \tag{3.22}$$

Prema Lemi 3.2., vrijedi da je

$$\begin{aligned} \tilde{U_n} = \sum_{j=1}^{n} E_F[U_n - \theta(F)|X_j] &= \frac{1}{n}\sum_{j=1}^{n} a(\phi_1(X_j) - \theta(F)) \\ &= \frac{1}{n}\sum_{j=1}^{n}(a\phi_1(X_j)) - a\theta(F). \end{aligned} \tag{3.23}$$

Prema Centralnom graničnom teoremu, Teorem 2.6,

$$\frac{\sqrt{n}\tilde{U_n}}{a^2\sigma_1^2} = \sqrt{n}\frac{\frac{1}{n}\sum_{j=1}^{n}(a\phi_1(X_j)) - a\theta(F)}{a^2\sigma_1^2} \xrightarrow{D} N(0,1), \quad n \to \infty,$$

tj. $\sqrt{n}\tilde{U_n} \xrightarrow{D} N(0, a^2\sigma_1^2), \quad n \to \infty.$

□

Lema 3.4.

$$E_F\Big[\tilde{U_n}(U_n - \theta(F))\Big] = E_F\Big[\tilde{U_n}^2\Big].$$

Dokaz:
Prema jednadžbi (3.21) i Lemi 3.2., očekivanje

$$E_F\Big[\tilde{U_n}\Big] = E_F\Big[\sum_{j=1}^{n} E_F[U_n - \theta(F)|X_j]\Big] = \sum_{j=1}^{n} E_F[U_n - \theta(F)] = \sum_{j=1}^{n}(\theta(F) - \theta(F)) = 0.$$

Samim time,

$$E_F\Big[\tilde{U_n}^2\Big] = \mathrm{Var}_F(\tilde{U_n}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_F(a\phi_1(X_j)) \overset{(3.22)}{=} \frac{a^2\sigma_1^2}{n}. \tag{3.24}$$

Nadalje,

$$\begin{aligned} E_F\Big[\tilde{U_n}(U_n - \theta(F))\Big] &= \frac{a}{n}\sum_{j=1}^{n} E_F\Big[(\phi_1(X_j) - \theta(F))(U_n - \theta(F))\Big] \\ &= \frac{a}{n}\sum_{j=1}^{n} E_F\Big[E_F\Big[(\phi_1(X_j) - \theta(F))(U_n - \theta(F))|X_j\Big]\Big] \\ &= \frac{a^2}{n^2}\sum_{j=1}^{n} E_F\Big[(\phi_1(X_j) - \theta(F))^2\Big] \\ &= \frac{a^2}{n^2}\sum_{j=1}^{n}\mathrm{Var}_F\Big(\phi_1(X_j) - \theta(F)\Big) \\ &= \frac{a^2\sigma_1^2}{n}. \end{aligned} \tag{3.25}$$

□

Lema 3.5. *Ako je* $\sigma_1^2 < \infty$*, tada*

$$\sqrt{n}\Big(U_n - \theta(F) - \tilde{U_n}\Big) \xrightarrow{D} 0, \quad n \to \infty.$$

Dokaz:
Prema Teoremu 2.4, konvergencija u srednjem reda 2 povlači konvergenciju po vjerojatnosti koja pak povlači konvergenciju po distribuciji. Stoga je dovoljno pokazati da

$$E_F\left[\Big(\sqrt{n}\Big(U_n - \theta(F) - \tilde{U_n}\Big)\Big)^2\right] \to 0, \quad n \to \infty.$$

Prema Lemi 3.4.,

$$\begin{aligned} E_F\left[\Big(\sqrt{n}\Big(U_n - \theta(F) - \tilde{U_n}\Big)\Big)^2\right] &= nE_F\left[\Big((U_n - \theta(F)) - \tilde{U_n}\Big)^2\right] \\ &= n\left(E_F\Big[\Big(U_n - \theta(F)\Big)^2\Big] - 2E_F\Big[\tilde{U_n}\Big(U_n - \theta(F)\Big)\Big] + E_F\Big[\tilde{U_n}^2\Big]\right) \\ &= n\left(E_F\Big[\Big(U_n - \theta(F)\Big)^2\Big] - E_F\Big[\tilde{U_n}^2\Big]\right) \\ &= n\,\mathrm{Var}_F(U_n) - nE_F\Big[\tilde{U_n}^2\Big]. \end{aligned} \tag{3.26}$$

Prema formuli (3.24) je $nE_F\Big[\tilde{U_n}^2\Big] = a^2\sigma_1^2$. Isto tako, prema Teoremu 3.1, $n\,\mathrm{Var}_F(U_n) = a^2\sigma_1^2, \quad n \to \infty$. Ovime je lema dokazana.

□

Dokaz Teorema 3.2:
Naposljetku, budući da je $\sqrt{n}(U_n - \theta(F)) = \sqrt{n}\tilde{U_n} + \sqrt{n}(U_n - \theta(F) - \tilde{U_n})$, Leme 3.3. i 3.5. zajedno za Slutskyjevim teoremom (Teorem 2.5) povlače tvrdnju Teorema 3.2,

$$\sqrt{n}(U_n - \theta(F)) \xrightarrow{D} N(0, a^2\sigma_1^2).$$

□

4 Wilcoxonov test ranga i predznaka

Wilcoxonov test ranga i predznaka razvio je 1945. godine statističar F. Wilcoxon te je to jedan od prvih neparametarskih statističkih testova. Smatramo ga neparametarskom alternativom t-testa za jedan uzorak u slučajevima kada pretpostavka o normalnosti slučajnog uzorka kojim modeliramo podatke nije ispunjena ili je uzorak premal.

4.1 Pretpostavke i hipoteze Wilcoxonovog testa ranga i predznaka

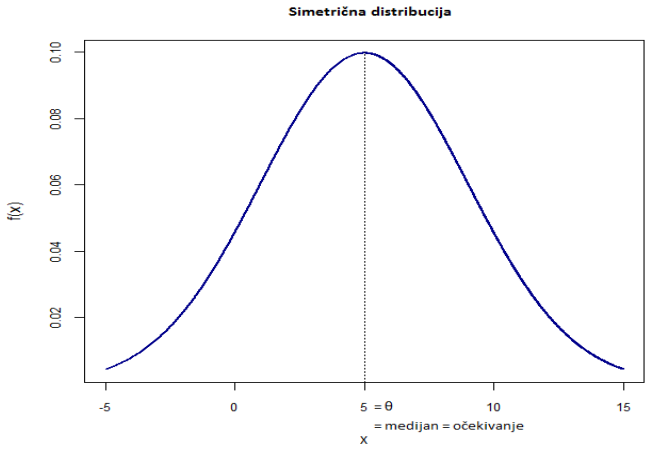
Pretpostavimo kako su $Z_1, \ldots, Z_n$ nezavisne i jednako distribuirane slučajne varijable koje dolaze iz neke neprekidne i simetrične distribucije. Dakle, pretpostavljamo:

- nezavisnost slučajnih varijabli
- jednaku distribuiranost slučajnih varijabli
- neprekidnost distribucije
- simetričnost distribucije

Kako se ovdje radi o neparametarskom statističkom testu, nemamo drugih pretpostavki o distribuciji, pogotovo ne o njenom obliku. Dovoljno je da je distribucija iz koje dolaze podaci simetrična i neprekidna.
Pogledajmo sada pobliže simetričnost distribucije. Za vjerojatnosnu distribuciju kažemo da je *simetrična* ako postoji vrijednost θ takva da vrijedi

$$f(\theta - x) = f(\theta + x),\ \forall x \in \mathbb{R},$$

gdje je f funkcija gustoće te distribucije. Vrijednost θ nazivamo *centar simetrije*.



Slika 4.1: Funkcija gustoće simetrične distribucije

Slika 4.1 prikazuje graf funkcije gustoće neke neprekidne simetrične distribucije sa centrom simetrije θ. Primijetimo kako je normalna distribucija također simetrična te u Wilcoxonovom testu ranga i predznaka pretpostavku o normalnoj distribuciji uzorka iz t-testa oslabljujemo na pretpostavku o simetričnoj distribuciji uzorka. Isto tako, primijetimo kako je centar simetrije θ također i očekivanje i medijan simetrične distribucije (ako oni postoje).
Uparavo je θ parametar o kojem želimo zaključivati u Wilcoxonovom testu ranga i predznaka te su hipoteze Wilcoxonovog testa ranga i predznaka o centru simetrije distribucije dane s:

$$\begin{aligned} &\mathcal{H}_0 : \theta = 0; \\ &\mathcal{H}_1 : \theta > 0. \end{aligned} \tag{4.1}$$

Prije svega, zanima nas koliko je restriktivna pretpostavka o simetričnosti distribucije te u kojim situacijama su pretpostavke testa zadovoljene.
Znamo kako mnoge vjerojatnosne distribucije nisu simetrične. Wilcoxonov test ranga i predznaka je originalno razvijen za testiranje hipoteza o razlikama, tj. efektu tretmana, među vezanim parovima podataka.

Postavljamo pitanje je li neki tretman B bolji od drugog tretmana A, tj. postoji li efekt tretmana. Na primjer: je li tretman lijekom koji se istražuje (B) bolji od placebo tretmana (A)?
Većinom je najučinkovitije uspoređivati tretmane na vezanim parovima podataka. Vezane parove podataka dobivamo na primjer primjenom oba tretmana na istoj jedinki ili pak primjenom različitih tretmana na jednojajčanim blizancima.

Razmotrimo sada slučaj kada smo tretmane A i B primijenili na n vezanih parova $(X_i, Y_i),\ i = 1, \ldots, n$ tako da u svakom paru prvi član X_i prima tretman A, a drugi član Y_i tretman B. Pretpostavljamo da su parovi $(X_i, Y_i),\ i = 1, \ldots, n$ nezavisni i jednako distribuirani s nepoznatom neprekidnom dvodimenzionalnom distribucijom. Dakle, pretpostavljamo da su slučajni vektori (X_i, Y_i) i (X_j, Y_j) nezavisni za sve $i, j = 1, \ldots, n$.
Definiramo slučajne varijable $Z_i = Y_i - X_i,\ i = 1, \ldots, n$ te vidimo kako su i one nezavisne i jednako distribuirane s nekom neprekidnom jednodimenzionalnom distribucijom F.
Primijetimo, ako ne postoji razlika u tretmanu, tj. ako su X_i i Y_i nezavisne i jednako distribuirane za svaki $i = 1, \ldots, n$, tada slučajna varijabla $Z_i = Y_i - X_i$ ima istu distribuciju kao $-Z_i = X_i - Y_i$, tj. distribucija od $Z_i,\ i = 1, \ldots, n$ je simetrična oko $\theta = 0$ (ishodišta). Slučajne varijable $Z_1, \ldots Z_n$ zadovoljavaju pretpostavke Wilcoxonovog testa ranga i predznaka.

4.2 Wilcoxonova test-statistika

Neka je $Z_1', \ldots, Z_n'$ permutacija $Z_1, \ldots, Z_n$ tako da vrijedi

$$|Z_1'| \leq |Z_2'| \leq \cdots \leq |Z_n'|.$$

Primjetimo kako su $i = 1, \ldots, n$ rangovi apsolutnih vrijednosti od $Z_1, \ldots, Z_n$ poredanih u rastućem poretku. Budući da je distribucija od Z_i neprekidna, vrijedi da je $P(|Z_i| = |Z_j|) = 0$, za sve $i, j = 1, \ldots, n,\ i \neq j$ te nemamo dva ista ranga.
Test-statistika Wilcoxonovog testa ranga i predznaka dana je s

$$W_n = \sum_{i=1}^{n} i \cdot \mathbf{I}_{(Z_i'>0)}. \tag{4.2}$$

Ova test-statistika, osim predznaka od Z_i, tj. Z_i' koristi i podatak o njegovom rangu.
Raspon vrijednosti w koji W_n može poprimiti kreće se od 0 (u slučaju kada niti jedan Z_i' nije pozitivan) do $\frac{n(n+1)}{2}$ (u slučaju kada su svi Z_i' pozitivni). U uvjetima hipoteze $\mathcal{H}_0$, zbog pretpostavke o simetričnosti, vjerojatnost da je Z_i' pozitivan je jednaka vjerojatnosti da je on negativan, tj. očekujemo kako će otprilike pola Z_i' biti pozitivno i pola negativno.

Pogledajmo sada egzaktnu distribuciju statistike W_n.
W_n je diskretna slučajna varijabla te je njena egzaktna nul-distribucija, tj. egzaktna distribucija u uvjetima hipoteze $\mathcal{H}_0$ dana kroz vjerojatnosti da W_n postigne bilo koju od mogućih vrijednosti $w = 0, \ldots, \frac{n(n+1)}{2}$. Neka je $F_W(w|n)$ broj svih mogućih kombinacija rangova i predznaka čija je suma w, n je veličina uzorka.

Primjer 4.1. *(Egzaktna nul-distribucija od W_3)*
Pretpostavimo da imamo $n = 3$ parova podataka. Tada imamo $2^3 = 8$ mogućih kombinacija rangova i predznaka $\pm1, \pm2, \pm3$ od kojih se svaka realizira s istom vjerojatnosti. Statistika W_3 poprima vrijednosti $w = 0, \ldots, 3 \cdot (3+1)/2 = 6$. Ako su svi predznaci pozitivni, $W_3 = 6$ te ni na koji drugi način ne možemo dobiti tu vrijednost statistike W_3. Stoga je $F_W(6|3) = 1$. Vrijednost $w = 5$ možemo postići samo uz kombinaciju $-1, +2, +3$ pa je $F_W(5|3) = 1$. Slično, postoji samo jedna kombinacija za vrijednosti $w = 4, 2, 1, 0$ te je $F_W(4|3) = F_W(2|3) = F_W(1|3) = F_W(0|3) = 1$, dok se $w = 3$ postiže za kombinacije $+1, +2, -3$ i $-1, -2, +3$ i $F_W(3|3) = 2$.
U konačnici, egzaktna distribucija od W_3 dana je s

$$W_3 \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{2}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{pmatrix} \tag{4.3}$$

Za veće uzorke $F_W(w|n)$ možemo dobiti pomoću rekurzivne formule

$$F_W(w|n) = F_W(w|n-1) - F_W(w-n|n-1).$$

Budući da postoji 2^n mogućih kombinacija rangova i predznaka od kojih se svaka realizira s istom vjerojatnosti, vrijedi da je

$$P(W_n \leq w|n) = \frac{\sum_{s=0}^{w} F_W(s|n)}{2^n}.$$

Općenito, vidimo kako je egzaktna distribucija od W_n simetrična.
Na razini značajnosti α, kritično područje egzaktnog Wilcoxonovog testa ranga i predznaka dano je s

$$\mathcal{C}_r = \Big\{W_n \geq c_\alpha\Big\}, \tag{4.4}$$

gdje je c_α kritična vrijednost egzaktne distribucije test-statistike W_n na razini značajnosti α.

Međutim, s povećanjem veličine uzorka n, računanje egzaktne distribucije statistike W_n postaje nepraktično te nas zanima i njena asimptotska distribucija.
U uvjetima hipoteze $\mathcal{H}_0$, $\mathbf{I}_{(Z_i'>0)},\ i = 1, \ldots, n$, su n nezavisnih jednako distribuiranih slučajnih varijabli s Bernoulijevom distribucijom

$$\mathbf{I}_{(Z_i'>0)} \sim \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix} \tag{4.5}$$

Stoga je

$$\begin{aligned} E[W_n] &= \sum_{i=1}^{n} \frac{i}{2} = \frac{n(n+1)}{4}; \\ \mathrm{Var}(W_n) &= \sum_{i=1}^{n} \frac{i^2}{4} = \frac{n(n+1)(2n+1)}{24}. \end{aligned} \tag{4.6}$$

Može se pokazati da

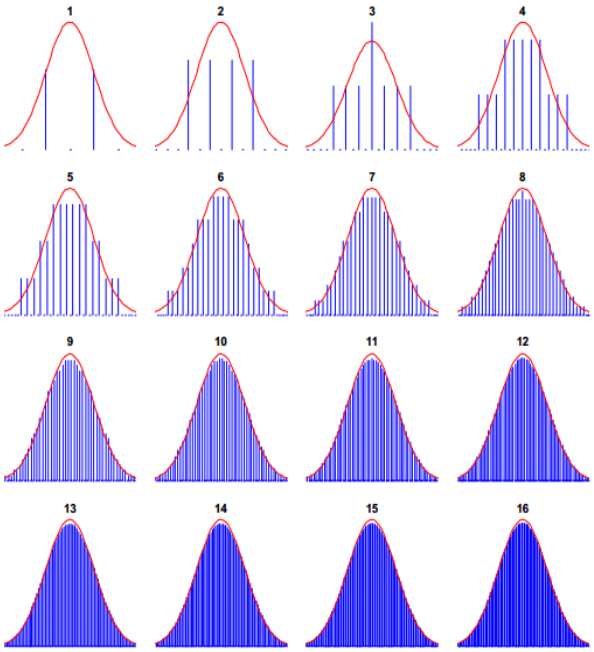
$$\frac{W_n - E[W_n]}{\sqrt{\mathrm{Var}(W_n)}} \xrightarrow{D} N(0,1), \quad n \to \infty. \tag{4.7}$$

Za dokaz pogledati ([4] Poglavlje 2., str. 102., Teorem 2.7.3.).

Stoga, na asimptotskoj razini značajnosti α, kritično područje Wilcoxonovog testa ranga i predznaka, asimptotski gledano, dano je s

$$\begin{aligned} \mathcal{C}_r &= \left\{\frac{W_n - E[W_n]}{\sqrt{\mathrm{Var}(W_n)}} \geq c_\alpha\right\} \\ &= \left\{W_n \geq c_\alpha\sqrt{\mathrm{Var}(W_n)} + E[W_n]\right\} \\ &= \left\{W_n \geq \frac{n(n+1)}{4} + c_\alpha\frac{\tau(0)}{\sqrt{n}}\right\}, \quad \tau(0) = n\sqrt{(n+1)(2n+1)/24}, \end{aligned} \tag{4.8}$$

gdje je c_α kritična vrijednost standardne normalne distribucije na asimptotskoj razini značajnosti α.
Zanima nas koliko se brzo distribucija statistike W_n približava normalnoj distribuciji. Wilcoxon je sugerirao kako se normalna aproksimacija može koristiti za dosta male uzorke, već od $n = 6$. Pokušajmo to pokazati grafički.



Slika 4.2: Egzaktna distribucija od W_n i asimptotska normalna distribucija

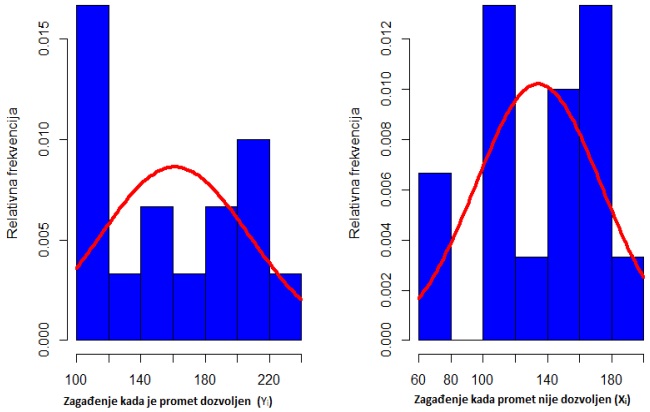
Slika 4.2 prikazuje funkciju gustoće egzaktne distribucije statistike W_n za uzorke veličine $n = 1, \ldots, 16$ i funkciju gustoće asimptotske normalne distribucije iste statistike. Vidimo kako se obje distribucije počinju poklapati relativno brzo te se zaista normalna aproksimacija može koristiti i za male uzorke.
Pogledajmo sada Wilcoxonov test ranga i predznaka na konkretnom primjeru.

Primjer 4.2. *Gradske vlasti žele ispitati hoće li se opća razina zagađenja zraka u gradu smanjiti ako zabranimo prometovanje automobila u gradu na jedan dan.*
Kako bismo ispitali ovu tvrdnju, nasumično odabiremo 15 ulica u gradu. U svakoj ulici dva puta mjerimo razinu zagađenja zraka. Prvo mjerenje radimo u 20:00h na dan kada je prometovanje automobila zabranjeno, dok drugi put mjerimo u 20:00h na dan kada je dopuštemo prometovanje automobila. Sljedećom tablicom dani su rezultati mjerenja:

y_i	214	159	169	202	103	119	200	109	132	142	194	104	219	119	234
x_i	159	135	141	101	102	168	62	167	174	159	66	118	181	171	112

Tablica 4.1: Podaci o razini zagađenja zraka

Tablicom 4.1. dani su parovi podataka (x_i, y_i), $i = 1, \ldots, 15$ *u kojima* x_i *predstavlja mjerenje razine zagađenja zraka u ulici i na dan kada je prometovanje automobila zabranjeno, dok su* y_i *mjerenja razine zagađenja zraka u ulici i na dan kada je prometovanje automobila dopušteno. Očito je kako raspolažemo uzorkom od* $n = 15$ *vezanih parova podataka.*
Svaki par (x_i, y_i) *modeliramo slučajnim vektorima* (X_i, Y_i), $i = 1, \ldots, 15$. *Smatramo kako razina zagađenja u ulici i ne ovisi o razini zagađenja u ulici j, kada* $i, j = 1, \ldots, 15$ *te su parovi* (X_i, Y_i) *nezavisni i jednako distribuirani nekom neprekidnom distribucijom.*



Slika 4.3: Histogrami X_i i Y_i, $i = 1, \ldots, 15$.

Slika 4.3 prikazuje histograme podataka distribucije slučajnih varijabli X_i *i* Y_i, $i = 1, \ldots, 15$. *Crvena linija predstavlja gustoću normalne distribucije čiji su očekivanje i varijanca procijenjeni na temelju podataka. Pogledamo li histograme, ne čini se da* X_i *i* Y_i *dolaze iz normalne distribucije. Isto tako, veličina uzorka* $n = 15$ *je premala da bismo smatrali kako normalnost distribucije vrijedi asimptotski. Stoga, nad ovim slučajnim uzorkom primjenjujemo Wilcoxonov test ranga i predznaka. Pripadne hipoteze su:*

$$\begin{aligned} \mathcal{H}_0 &: \textit{Opća razina zagađenja zraka je ista sa i bez prometa automobila;} \\ \mathcal{H}_1 &: \textit{Opća razina zagađenja zraka je veća kada ima automobilskog prometa.} \end{aligned} \tag{4.9}$$

Znamo kako se u Wilcoxonovom testu ranga i predznaka hipoteze (4.9) *svode na hipoteze:*

$$\begin{aligned} \mathcal{H}_0 &: \theta = 0; \\ \mathcal{H}_1 &: \theta > 0, \end{aligned} \tag{4.10}$$

gdje je θ centar simetrije funcije gustoće nezavisnih i jednako distribuiranih slučajnih varijabli $Z_i = Y_i - X_i$, $i = 1, \ldots, 15$, koje imaju neprekidnu i simetričnu distribuciju.
Testiranje hipoteza provedeno je u programskom paketu R.
Prvo smo proveli test u odnosu na egzaktnu distribuciju test-statistike W_n. Dobivena vrijednost test-statistike iznosi $W_n = 40$. Na razini značajnosti $\alpha = 0.05$ kritična vrijednost testa c_α iznosi 95, dok je pripadna p-vrijednost $p = 0.8738$. Kako je $p > \alpha$, ne možemo odbaciti nul-hipotezu da se opća razina zagađenja zraka neće smanjiti ako zabranimo automobilski promet na jedan dan.
Ukoliko isti test provedemo u odnosu na asimptotsku normalnu distribuciju test-statistike W_n, vrijednost test-statistike ponovno iznosi $W_n = 40$. Međutim, na asimptotskoj razini značajnosti $\alpha = 0.05$ kritična vrijednost testa c_α iznosi 88.96, dok je pripadna p-vrijednost $p = 0.8779$. I dalje je $p > \alpha$ te ne možemo odbaciti nul- hipotezu da se opća razina zagađenja zraka neće smanjiti ako zabranimo automobilski promet na jedan dan.
Primijetimo kako su kritične i p-vrijednosti u egzaktnoj i asimptotskoj verziji testa vrlo slične te je asimptotska distribucija zaista primjenjiva na relativno malim uzorcima. Ipak, egzaktni test je nešto konzervativniji.

4.3 Wilcoxonova test-statistika i U-statistika

Prikažimo sada test-statistiku W_n u Wilcoxonom testu ranga i predznaka pomoću U-statistika iz Poglavlja 3.

Test-statistika Wilcoxonovog testa ranga i predznaka dana je s

$$W_n = \sum_{i=1}^{n} i \cdot \mathbf{I}_{(Z_i'>0)}. \tag{4.11}$$

Budući da je $|Z_i'| \leq |Z_j'|$ za $i \leq j$, $i, j = 1, \ldots, n$, vrijedi da je $Z_i' + Z_j' > 0$ ako i samo ako je $Z_j' > 0$.
Stoga je

$$\sum_{i=1}^{j} \mathbf{I}_{(Z_i'+Z_j'>0)} = j\mathbf{I}_{(Z_j'>0)}$$

pa test-statistiku (4.11) možemo zapisati u obliku:

$$\begin{aligned} W_n &= \sum_{j=1}^{n} j\mathbf{I}_{(Z_j'>0)} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{j} \mathbf{I}_{(Z_i'+Z_j'>0)} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{j} \mathbf{I}_{(Z_i+Z_j>0)} \\ &= \sum_{i<j} \mathbf{I}_{(Z_i+Z_j>0)} + \sum_{i=1}^{n} \mathbf{I}_{(Z_i>0)} \\ &= \binom{n}{2} \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{I}_{(Z_i+Z_j>0)} + n\frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(Z_i>0)} \\ &= \binom{n}{2} U_n^{(1)} + nU_n^{(2)} \end{aligned} \tag{4.12}$$

Primijetimo kako je $U_n^{(1)}$ U-statistika iz Primjera 3.3. dana jednadžbom (3.9). Statistika $\binom{n}{2}U_n^{(1)}$ je broj svih $i < j$ za koje je $Z_i + Z_j > 0$. S druge stane, iz (4.12) možemo vidjeti kako je W_n broj svih $i \leq j$ za koje je $Z_i + Z_j > 0$.
Za velike n vrijedi da je

$$\binom{n}{2} = \frac{n^2 - n}{2} \sim \frac{n^2}{2}$$

te u (4.12) prvi član $\binom{n}{2}U_n^{(1)}$ dominira nad drugim. Možemo zaključiti kako se, asimptotski, W_n ponaša kao $\frac{n^2}{2}U_n^{(1)}$ te je kao test-statistiku Wilcoxonovog testa ranga i predznaka asimptotski dovoljno gledati U-statistiku $U_n^{(1)}$ i njenu asimptotsku distribuciju.

Naime, jednadžbu (4.12) možemo ekvivalentno zapisati u obliku

$$\binom{n}{2}^{-1} W_n = U_n^{(1)} + \frac{2}{n-1} U_n^{(2)} \tag{4.13}$$

Statistika $U_n^{(1)}$ je nepristran procjenitelj za $\theta' = P(Z_1 + Z_2 > 0)$. Ako pogledamo očekivanje obje strane u jednadžbi (4.13) u uvjetima hipoteze $\mathcal{H}_0 : \theta = 0$, vrijedi da je

$$\begin{aligned} E\left[\binom{n}{2}^{-1} W_n\right] &= \binom{n}{2}^{-1} \frac{n(n+1)}{4} = \frac{1}{2}; \\ E\left[U_n^{(1)} + \frac{2}{n-1} U_n^{(2)}\right] &= P(Z_1 + Z_2 > 0) + \frac{2}{n-1} P(Z_1 > 0) \\ &= \theta' + \frac{1}{n-1} \end{aligned} \tag{4.14}$$

Slijedi kako je

$$\theta' = \frac{1}{2} - \frac{1}{n-1} \to \frac{1}{2},\ n \to \infty.$$

Stoga možemo nul-hipotezu $\mathcal{H}_0 : \theta = 0$ izraziti u terminima θ' kao

$$\mathcal{H}_0 : \theta' = P(Z_1 + Z_2 > 0) = \frac{1}{2}$$

Promotrimo sada U-statistiku $U_n^{(1)}$ danu jednadžbom (3.9). Ona je procjenitelj za parametar $\theta' = P(Z_1 + Z_2 > 0)$.
Prema Lemi 3.1. znamo kako je

$$\begin{aligned} \sigma_1^2 &= \mathrm{Cov}(\phi(Z_1, Z_2), \phi(Z_1, Z_2')) \\ &= P(Z_1 + Z_2 > 0, Z_1 + Z_2' > 0) - \theta'^2, \end{aligned} \tag{4.15}$$

gdje su Z_1, Z_2, Z_2' nezavisne i jednako distribuirane.
U uvjetima hipoteze $\mathcal{H}_0$ znamo da je distribucija F od $Z_i,\ i = 1, \ldots, n$ simetrična oko 0, tj. da Z_i i $-Z_i$ imaju istu distribuciju.
Stoga je

$$P(Z_1 + Z_2 > 0) = P(Z_1 > -Z_2) = P(Z_1 > Z_2).$$

Kako je F neprekidna te je stoga $P(Z_2 = Z_1) = 0$, vrijedi da je

$$P(Z_1 + Z_2 > 0) = \frac{1}{2}.$$

Analogno,

$$P(Z_1 + Z_2 > 0, Z_1 + Z_2' > 0) = P(Z_1 > Z_2, Z_1 > Z_2')$$

je vjerojatnost da je Z_1 najveća od tri nezavisne i jednako distribuirane slučajne varijable Z_1, Z_2, Z_2'.
Za neprekidnu F imamo

$$P(Z_1 + Z_2 > 0, Z_1 + Z_2' > 0) = \frac{1}{3}.$$

Prema (4.15) slijedi da je

$$\sigma_1^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

Iz Teorema 3.2 znamo da

$$\sqrt{n}\left(U_n^{(1)} - \frac{1}{2}\right) \xrightarrow{D} N\left(0, \frac{1}{3}\right),$$

tj.

$$\frac{U_n^{(1)} - \frac{1}{2}}{\sqrt{\text{Var}\left(U_n^{(1)}\right)}} \xrightarrow{D} N(0,1).$$

Vidimo kako

$$\frac{U_n^{(1)} - E[U_n^{(1)}]}{\sqrt{\text{Var}\left(U_n^{(1)}\right)}} \; i \; \frac{W_n - E[W_n]}{\sqrt{\text{Var}(W_n)}}$$

imaju istu asimptotsku distribuciju.

Literatura

[1] P. BARTLETT, *Theoretical Statistics*, materijali s predavanja, Berkeley, 2013. http://www.stat.berkeley.edu/ bartlett/courses/2013spring-stat210b/, 13.2.2016.

[2] T. S. FERGUSON, *U-statistics*, materijali s predavanja, UCLA, 2005. https://www.math.ucla.edu/ tom/Stat200C/, 12.2.2016.

[3] D. R. HUNTER, *Asymptotic Tools*, materijali s predavanja, Penn State, 2006. http://sites.stat.psu.edu/ dhunter/asymp/fall2006/lectures/, 12.2.2016.

[4] E. L. LEHMANN, *Elements of large sample theory*, Springer, New York, 1999.

[5] E. L. LEHMANN, G. CASELLA, *Theory of Point Estimation, Second Edition*, Springer, New York, 1998.

[6] E. L. LEHMANN, J. P. ROMANO, *Testing Statistical Hypotheses, Third Edition*, Springer, New York, 2005.

[7] N. SARAPA, *Teorija Vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[8] S. B. VARDEMAN, *The Standard presentation of the Lehmann-Scheffe Theorem*, materijali s predavanja, ISU, 2005. http://www.public.iastate.edu/ vardeman/stat543/Handouts/, 11.2.2016.

Sažetak

Cilj ovog diplomskog rada bilo je upoznavanje s pojmom U-statistike kao nepristranog procjenitelja statističkog funkcionala očekivanja. Poseban naglasak stavljen je na primjenu U-statistika i njihovih svojstava. U radu su navedeni primjeri U-statistika koje se najčešće koriste pri procjeni parametara koje možemo zapisati pomoću funkcionala očekivanja. Također, dokazani su važni teoremi o varijanci i asimptotskoj varijanci U-statistika kao i teorem o njihovoj asimptotskoj normalnosti. U svrhu demonstriranje primjene U-statistika, obrađen je Wilcoxonov test ranga i predznaka, teorijski i primjerom. Budući da se pokazalo kako asimptotska distribucija test-statistike Wilcoxonovog testa ranga i predznaka odgovara asimptotskoj distribuciji određene U-statistike, iskoristili smo ranije dokazana svojstva U-statistika kako bismo pokazali asimptotsku normalnost test-statistike Wilcoxonovog testa ranga i predznaka.

Ključne riječi

Statistika, Statistički funkcional, U-statistika, Wilcoxonov test ranga i predznaka

U-statistics and application

Summary

The aim of this thesis was to introduce the concept of U-statistics as unbiased estimators of statistical expectation functionals. Special emphasis was put on the application and properties of U-statistics. Further in the thesis, several examples of U-statistics which are used in the assessment of parameters that can be represented by using expectation functionals were given. Additionally, important theorems regarding variance, asymptotic variance as well as asymptotic normality of U-statistics have been proven. In order to demonstrate the application of the u-statistics, Wilcoxon signed rank test was explained in detail, both theoretically and by example. Since asymptotic distribution of Wilcoxon signed rank test-statistics corresponds to the asymptotic distribution of a specific U-statistics, properties of U-statistics were used to show asymptotic normality of Wilcoxon signed rank test-statistics.

Key words

Statistics, Statistical functional, U-statistics, Wilcoxon signed rank test

Životopis

Rođena sam 27. studenog 1991. godine u Zagrebu. Osnovnoškolsko obrazovanje završila sam 2006. godine te sam iste godine upisala II. jezičnu gimnaziju u Osijeku. Tijekom srednje škole sudjelovala sam na županijskim natjecanjima iz matematike, informatike te latinskog jezika. Preddiplomski studij matematike upisujem 2010. godine na Odjelu za matematiku u Osijeku te isti završavam 2013. godine završnim radom na temu "Algoritmi kriptografije javnog ključa" pod mentorstvom izv.prof.dr.sc. Ivana Matića. Akademsko obrazovanje nastavljam na Odjelu za matematiku u Osijeku i upisujem diplomski studij matematike, smjer Financijska matematika i statistika. Tijekom završne godine diplomskog studija obavila sam višemjesečnu stručnu studentsku praksu u tvrtki Farmeron kao podatkovni analitičar i statističar. Također sam tijekom iste godine obavila i dvotjednu stručnu studentsku praksu u Uredu za validaciju Privredene banke Zagreb.