Originalni naučni rad
Original scientific paper
UDK: 631.15:005.51
DOI: 10.7251/AGRSR1304617M

University of Banjaluka, Faculty of Agriculture
Agro-knowledge Journal

Zamke deskriptivnog i inferencijalnog statističkog pristupa u biološkim i poljoprivrednim naukama

Nikola Mićić[1, 2], Borut Bosančić[1, 2]

[1]*Poljoprivredni fakultet Univerziteta u Banjoj Luci*
[2]*Institut za genetičke resurse Univerziteta u Banjoj Luci*

Sažetak

Deskriptivna statistika skup podataka tretira kao dati skup, tj. konačan i prebrojiv statistički skup i tako ga i interpretira, dok inferencijalna statistika kao prvi korak u analitičkom pristupu, saglasno cilju istraživanja, mora da definiše elementarna određenja osnovnog skupa (pojmovno, prostorno i vremenski), a potom i uzoraka koji se odnose na planirano istraživanje, odnosno, njihovu reprezentativnost u oceni parametara osnovnog skupa. Takođe, u inferencijalnoj statistici izbor logičko–matematičke argumentacije u oceni parametara osnovnog skupa mora da odredi i kolikoj greški će biti izloženi zaključci na osnovu kojih se procenjuje verovatnoća postavljenih hipoteza o osnovnom skupu. Tako, dok deskriptivna statistika konstatuje postojeće stanje u datom konačnom i prebrojivom skupu podataka, inferencijalna statistika, na osnovu eksperimentalnih, instrumentalnih i logičko-matematičkih metoda, analizira varijacije podataka u uzorcima i procenjuje objašnjene, neobjašnjene i dozvoljene varijacije posmatranog obeležja, kao mere verovatnoće ispoljavanja posmatranog svojstva u osnovnom skupu. Dakle, u deskriptivnoj statistici koristi se matematička aritmetička sredina, dok u inferencijalnoj statistici aritmetička sredina u stvari predstavlja centralnu tendenciju kao pouzdanu verovatnoću pojavljivanja ili ispoljavanja posmatranog obeležja u osnovnom skupu. Tako je u inferencijalnoj statistici reprezentativnost uzoraka u stvari reprezentativnost centralnih tendencija uzoraka, koja se argumentuje dozvoljenim varijacijama posmatranih vrednosti obeležja, odnosno, dozvoljenim intervalom relativnog varijabiliteta (5 % < Vk < 30 %). Naime, uzorci čiji su koeficijenti varijacije manji od 5 % "suviše su dobri" (odnose se na skup istovetnih statističkih jedinica), a uzorci sa koeficijentima varijacije većim od 30 % moraju se razložiti na poduzorke sa dozvoljenim varijabilitetom za centralnu tendenciju i osnovnom pretpostavkom za analizu strukture podskupova posmatranog obeležja u osnovnom skupu. U ovom radu obrađeno je pitanje biometričke analize uzoraka sa nedozvoljenim relativnim varijabilitetom podataka u argumentaciji centralne tendencije.

Ključne reči: uzorci, centralna tendencija, interval homogenizacije, biometrika

Uvod

U biološkim i poljoprivrednim naukama posle 2007. godine i usvajanja inovirane FOS klasifikacije naučnih oblasti sa kojom su primenjena i eksperimentalna statistika klasifikovane kao posebna uža naučna oblast u svih šest naučnih oblasti, pa tako i u poljoprivrednim naukama, izvršena je generalna podela statističkih metoda, prema određenju statističkog skupa na kome se vrši posmatranje, interpretaciji i oceni parametara statističkog skupa, kao i logičko–matematičkoj argumentaciji verovatnoće u testiranju hipoteza, na dva osnovna pristupa: deskriptivnu i inferencijalnu statistiku.

Deskriptivna biometrika se dijeli na statistiku lociranja i disperzije (Sokal i Rohlf 1995; Rao 2007). Ni jedna od ovih statistika nije pouzdano merilo parametara osnovnog skupa, ali njihova kohezija i interakcija nas dovodi do relativno pouzdanih procena traženih parametara. Poznavanje lokacione i disperzione statistike predstavlja osnovu svakog rada u agronomskim istraživanjima. Najčešće pogreške koje se mogu susresti u deskriptivnoj biometrici su pogreške u uočavanju pravilnosti u podacima, interpretacija odstupanja od pretpostavljenog, uočavanje neuobičajenih vrijednosti i u samom unošenju podataka. Greške ukoliko nastanu na ovom nivou su praktično nepopravljive u daljem radu na inferencijalnom delu istraživanja, čime se dakle u velikoj meri ugrožava pouzdanost procene parametara i samim tim pouzdanost i validnost donošenja zaključaka.

U ovom radu prevashodno smo se posvetili kvantitativnim karakteristikama kao najznačajnijim za poljoprivredna istraživanja. Sa takvim podacima pre svega potrebno je grafički predstaviti uzorak i distribuciju frekvencija u uzorku, što je osnova pregleda podataka i analize u poljoprivrednim i biološkim istraživanjima (Kohler i sar. 2012). Pored normalnosti rasporeda i homogenosti varijansi, najjednostavnije, čak i važnije, je pregledati statistiku disperzije. Naime, postoje brojna istraživanja bioloških procesa u poljoprivrednoj proizvodnji čije varijacije u skupu istovrsnih jedinica posmatranja nije moguće argumentovati na nivo centralnih tendencija, odnosno, biološki istovrsnih entiteta u eksperimentalno ujednačenim uslovima čija se obeležja bez obzira na broj ponavljanja ispoljavaju sa varijacijama iznad 30%. Potrebno je napomenuti da iz uzorka kod kojeg je koeficijent varijacije veći od 30% nije moguće u standardnoj postavci ogleda otkriti značajnu razliku čak i kada se stvarne sredine ispitivanih osnovnih skupova razlikuju i preko 20%, kao i to da je standardna devijacija sadržana u koeficijentu varijacije mera izbora, upravo zbog svoje sadržajnosti. Značajno je u ovom kontekstu postaviti i kritike testiranja hipoteza u klasičnom statističkom smislu, tj. sa *a priori* određenim verovatnoćama pogreške. Jasno je da sa dovoljno velikim uzorkom veoma male razlike postaju visoko značajne (Quinn i Keough 2002), sem u onim slučajevim kada su osnovni skup, tj. uzorci strukturisani u podskupove. Stoga je najbolje unapred odrediti veličinu efekta tretmana koji se testira, a koji je cilj istraživanja, te u skladu s tim odrediti veličinu uzorka, ili velike uzorke sa nedozvoljenim varijacijama testirati na prisustvo podskupova koje je moguće homogenizovati.

Za određeni uzorak, izračunata aritmetička sredina uvek predstavlja istinitu meru lokacije, a istinitu meru disperzije uvek predstavlja tačno izračunata standardna

devijacija (Sokal i Rohlf 1995). Ipak, u poljoprivrednom istraživanju najčešće nas interesuje osnovni skup iz koga je potekao uzorak, tj. procena parametara osnovnog skupa.

Varijabilitet podataka je notorno prihvaćena mera u tumačenju preciznosti u kvantitativnim poljoprivrednim istraživanjima, pre svega prinosa, veličine ili mase ciljnih organa. Međutim, veoma je malo gotovih algoritama ili softverskih alatki koje omogućavaju istraživaču da jednostavno poveže preciznost dobijenih rezultata sa njihovim tumačenjem. U ovom radu razrađen je model analitičkog pristupa argumentaciji centralne tendencije uzorka na osnovu analize varijabiliteta, odnosno, identifikacije podskupova posmatranog obeležja.

Koeficijent varijacije se do sada pokazao superiornim merilom varijabiliteta u biološkim i poljoprivrednim istraživanjima (Mićić i Bosančić, 2012), ali i u drugim oblastima (Reed i sar. 2002; Weber i Sharoni 2004). Dormann i Kuhn (2012) stavljaju koeficijent varijacije na prvo mesto u proceni biometričkih parametara. U ovom radu razrađeni su slučajevi uzorkovanja iz konzistentnog osnovnog skupa gde se svi uticaji uzimaju u obzir i s druge strane uzorkovanje iz osnovnog skupa koji je pod uticajem različitih faktora koji su ostali neprimećeni. Uz to precizno se razjašnjava kakve su procedure i moguće pogreške u deskriptivnim i inferencijalnim računanjima koje proizilaze iz ispravnog ili pogrešnog tumačenja podataka, pre svega, proizišlog iz logičko–matematičke argumentacije centralne tendencije uzoraka.

Materijal i metode

Hipotetički ogled modeliran je tako da logičko–matematička argumentacija bude dosledno primenjena i očita. Modelirana su tri tipa uzorka iz osnovnih skupova sa različitom zastupljenošću posmatranog svojstva (ψ) kao analizirane karakteristike biometričkih jedinica posmatranja (Ω). Osnovni skupovi predstavljaju biološki entiteti (individue – Ω) sa različitom strukturom vrednosti svojstva ψ, u uzorcima označenim sa simbolima A, B, C, D i E, u identičnim datim uslovima. Primenjeni modeli treba da pokažu kako se različiti uticaji na biometričke jedinice posmatranja, koji se u uzorcima ili određenim skupu opažaju kao varijacije posmatranog svojstva, različito tumače u zavisnosti od deskriptivnog ili inferencijalnog statističkog pristupa. Naime, deskriptivne statističke metode varijabilitet podataka u posmatranom skupu tretiraju kao dato svojstvo, dok biometričke analize u domenu inferencijalne statistike moraju da se baziraju na argumentaciji varijacija posmatranog svojstva na nivou reprezentativnosti centralne tendencije uzoraka, jer u suprotnom se dolazi do potpuno nelogičnih i pogrešnih opažanja i procena, pa time i do pogrešnih zaključaka.

Rezultati istraživanja sa diskusijom

Rezultat istraživanja u ovom radu predstavljaju modeli logičko–matematičke analize uzoraka sa nedozvoljenim varijabilitetom vrednosti posmatranog obeležja sa kojim se argumentuje centralna tendencija uzoraka.

Evidentiranje ili analiza nekog svojstva u osnovnom skupu na osnovu izračunatih srednjih vrednosti, kao elementarnu pretpostavku koja proizlazi iz definicije osnovnog skupa, podrazumeva da sve jedinice posmatranja (Ω) u osnovnom skupu moraju da imaju to posmatrano svojstvo (ψ).

Zamka analitičkog posmatranja aritmetičke sredine u domenu deskriptivne ili inferencijalne statistike stoji u definiciji njene reprezentativnosti. Naime, u deskriptivnoj statistici aritmetička sredina konstatuje stanje u datom skupu podataka tako da samo srednja vrednost prezentuje posmatranu pojavu, ali u inferencijalnoj statistici aritmetička sredina po definiciji treba da predstavlja centralnu tendenciju, što znači da aritmetička sredina u inferencijalnoj statistici ne konstatuje samo dato stanje, već predstavlja procenu verovatnoće pojavljivanja posmatranog svojstva (ψ) u osnovnom skupu. Dakle, u inferencijalnoj statistici aritmetičku sredinu nužno mora da prate i ocene varijabiliteta ispitivanog obeležja u uzorku, čime se u stvari i određuje reprezentativnost aritmetičke sredine da predstavlja ocenu posmatranog svojstva u osnovnom skupu.

Zamka u koju se upada ako se ne konkretizuje pristup biometričkoj analizi u okviru deskriptivne ili inferencijalne statističke metode dokazana je na sledećem primeru:

Statistika: *Statistics*	Deskriptivna *Descriptive*		Inferencijalna * *Inferential* *		
	$\overline{X}$	$\Delta\overline{X}$	$\overline{X} \pm S_{\overline{X}}$	V_K	
Uzorak I (Ω_I) *Sample I (Ω_I)*	38,56	13,59	38,56 ± 4,42	52,3 %	$t_{\overline{X}_I - \overline{X}_{II}} = 1{,}941^{nz} \Rightarrow \overline{X}_I = \overline{X}_{II}$
Uzorak II (Ω_{II}) *Sample II (Ω_{II})*	52,15		52,15 ± 6,16	53,6 %	

** Uzorci sa varijacijama izvan intervala dozvoljenih varijacija za reprezentativnost centralne tendencije: 5 % < V_K < 30 %*

**Samples with variations outside of the inteval of allowed variations for representativeness of the central tendency 5 % < V_K < 30 %*

Na prikazanom modelu vidimo da u deskriptivnoj statistici zaključujemo da je prosečna razlika između uzoraka $\Delta\overline{X}_{\Omega I - \Omega II} = 13{,}59$ mernih jedinica svojstva ψ, a u inferencijalnoj statistici zaključujemo da je ova razlika u osnovnom skupu slučajna, odnosno, da je $\overline{X}_I = \overline{X}_{II}$. Dakle, ako iz iskustva znamo, ili u definisanom modelu tvrdimo ako je $\Delta\overline{X} \geq 10$, da to predstavlja značajan efekat svojstva ψ koje ispoljava Ω u datim uslovima, zaključak obadve statističke analize može da bude jednako pogrešan. Evidentno je da zamku za pogrešno zaključivanje u deskriptivnoj statistici predstavlja činjenica da varijacije vrednosti obeležja ψ nisu argumentovane (nepoznato je da li u posmatranom skupu postoje podskupovi sa indikativnim razlikama u ispoljavanju analiziranog svojstva), a u inferencijalnoj statistici problem predstavljaju nedozvoljene varijacije za reprezentativnost centralne tendencije u uzorcima, odnosno, velika varijacija osporava matematičku argumentaciju značajne razlike između podskupova u osnovnom skupu.

Interpretacija modela analitičkog pristupa uzorcima u kojima biometričke jedinice posmatranja (Ω) imaju vrednosti svojstva ψ sa varijacijama izvan opsega centralne tendencije, biće izvedena na uzorcima sa sledećom strukturom ispoljavanja posmatranog svojstva ψ:

I.) Ω sa prisutnim vrednostima svojstva ψ na nivou očekivanih varijacija, i Ω bez prisustva svojstva ψ (ψ = 0);

II.) svako Ω poseduje svojstvo ψ u rasponu koji se atributivno determiniše kao: ψ malih apsolutnih vrednosti, ψ očekivanih (prosečnih) apsolutnih vrednosti, i ψ velikih apsolutnih vrednosti.

Uzorci u kojima se nalaze biometričke jedinice posmatranja sa i bez ispoljenog svojstva

Ovde se otvara pitanje kako u analizi prosečne vrednosti određenog svojstva (ψ) posmatrati istovrsne statističke jedinice (Ω) koje imaju merljivu vrednost ispitivanog svojstva i one statističke jedinice kod kojih to svojstvo nije prisutno (ψ = 0). Ovde dilema nije u tome da se statističke jedinice posmatranja sa i bez određenog svojstva posmatraju kao atributivna obeležja koja se argumentuju neparametarskim statističkim testovima. Ovo pitanje je otvoreno kao kardinalan primer za razumevanje pristupa onim slučajevima kada sve istovrsne statističke jedinice posmatranja u osnovnom skupu, odnosno uzorku, imaju analizirano svojstvo koje je različito distribuirano u podskupovima. Naime, ako je posmatrano svojstvo ψ masa nekog metaboličkog produkta biometričkih jedinica posmatranja Ω, i ako se na proizvodnoj površini od 100 m^2 nalazi 30 entiteta Ω, te ako u ovom slučaju ukupna produkcija ψ iznosi 390 *kg*, onda u deskriptivnom statističkom pristupu možemo tvrditi da je prosečna produkcija ψ po jednom Ω jednaka 13 *kg*. Ako znamo da je svih 30 jedinica Ω imalo isti proizvodni tretman, te da su troškovi proizvodnje sumirani za celu proizvodnu površinu, odnosno, da su troškovi ravnomerno raspoređeni na sve proizvodnje jedinice Ω, onda se tvrdnja da je prosečni prinos ψ jednak 13 kg, može prihvatiti kao korektan u domenu deskriptivnog statističkog pristupa. Model ove analize dat je u uzorku A, tabela 1.

Tab. 1. Uzorak A sa sledećim utvrđenim vrednostima ψ (n = 30)
Sample A with the following established values of ψ (n = 30)

20,3	19,6	16,7	18,9	18,5	17,2	18,9	0,0	20,1	0,0
17,5	0,0	17,9	18,0	20,6	0,0	19,6	0,0	18,6	17,8
18,9	0,0	19,1	0,0	16,4	18,7	19,3	17,4	0,0	0,0

Sličnu situaciju imamo i u uzorku B, gde se na proizvodnoj površini od 100 m^2, takođe nalazi 30 entiteta Ω, ali njihova ukupna produkcija ψ, sada iznosi 525 *kg*, pa u deskriptivnom statističkom pristupu možemo tvrditi da je prosečna produkcija ψ po jednom Ω jednaka 17,5 *kg*. Model ove analize dat je u uzorku B, tabela 2.

Tab. 2. Uzorak B sa sledećim utvrđenim vrednostima ψ (n = 30)
Sample B with the following established values of ψ (n = 30)

27,8	24,2	23,7	27,2	21,6	25,3	28,6	0,0	21,8	0,0
29,5	0,0	23,9	24,8	23,1	0,0	29,7	0,0	21,6	24,8
25,8	0,0	25,6	0,0	23,3	24,9	24,3	23,5	0,0	0,0

Ako se izračuna prosečna produkcija ψ, kg /Ω u uzorcima A i B sa cele proizvodne površine 100 m^2, odnosno, za n = 30, dobijaju se sledeći rezultati:

Uzorak A / *Sample A*	Uzorak B / *Sample B*
$\overline{X} = 13{,}00$	$\overline{X} = 17{,}50$
$\sigma_x = 8{,}709$	$\sigma_x = 11{,}822$
$S_{\overline{X}} = 1{,}590$	$S_{\overline{X}} = 2{,}158$
$V_{K_A} = 66{,}99\%$	$V_{K_B} = 67{,}55\%$

Razlika u ukupnoj produkciji ψ sa proizvodne površine 100 m^2, između A i B iznosi 135 kg, a na prosečnom nivou ta razlika je $\Delta\overline{X}_{A-B} = 4{,}5\,kg$. Međutim, testiranje statističke značajnosti ove razlike kaže da je ispoljena razlika statistički slučajna ($t_{\overline{X}_A - \overline{X}_B} = 1{,}678^{nz}$). Jasno je da ovaj statistički zaključak nije relevantan jer znamo da u uzorcima imamo statističke jedinice Ω, koje nemaju produkciju svojstva ψ, a to nam govore i koeficijenti varijacije čije vrednosti su iznad dozvoljenih za argumentovanje centralne tendencije uzoraka. Ako sada izračunamo prosečnu produkciju ψ, samo za one biometričke jedinice Ω koje imaju produkciju svojstva ψ, dobijaju se sledeći rezultati:

Uzorak A / *Sample A*	Uzorak B / *Sample B*
$\overline{X} = 18{,}57$	$\overline{X} = 25{,}00$
$\sigma_x = 1{,}155$	$\sigma_x = 2{,}401$
$S_{\overline{X}} = 0{,}252$	$S_{\overline{X}} = 0{,}524$
$V_k = 6{,}22\%$	$V_k = 9{,}60\%$

Analiza prosečne produkcije ψ, samo sa statističkih jedinica Ω koje su ispoljile svojstvo ψ, pokazuje da su dobijene srednje vrednosti argumentovane kao centralne tendencije jer su im koeficijenti varijacije u intervalu dozvoljenih varijacija (5 % < Vk < 30 %). Takođe, razlika između prosečne produkcije uzoraka A i B ($\Delta\overline{X}_{A-B} = 6{,}43\,kg$) statistički je visoko značajna ($t_{\overline{X}_A - \overline{X}_B} = 11{,}059^{**}$).

Ova analiza, kao inferencijalni statistički pristup otvara sledeća pitanja, čime i sam analitički postupak ima sve atribute naučno relevantnog:

- produktivnost ψ u posmatranom skupu ispoljilo je 70 % statističkih jedinica Ω, dok 30 % statističkih jedinica Ω ne ispoljava svojstvo ψ, što u konačnom znači

da u posmatranom skupu podataka imamo dva podskupa: podskup Ω sa svojstvom ψ, i podskup Ω bez svojstva ψ;

- eksperimentalnim metodama neophodno je ustanoviti uzroke odsustva ψ kod 30 % statističkih jedinica Ω (otvara se pitanje potpune ili parcijalne alternativnosti, kao i mogućeg metodološkog postupka pomoću koga će sve statističke jedinice Ω uspeti da realizuju svojstvo ψ);
- izvedeni analitički postupak otvara pitanje povećanja ukupne produktivnosti za 30 % ako se utvrde razlozi zbog kojih se svojstvo ψ ne realizuje kod, u ovom uzorku evidentiranih statističkih jedinica Ω (Ω – bez ispoljenog svojstva ψ).

Modelirani primer uzoraka A i B, predstavlja kardinalan primer postajanja dva podskupa u posmatranim uzorcima, međutim, ova analiza je mnogo složenija kada sve statističke jedinice Ω imaju svojstvo ψ ali ono varira preko gornje dozvoljene granice za argumentovanje centralne tendencije uzoraka, tj. preko 30 %.

Uzorci u kojima sve biometričke jedinice posmatranja ispoljavaju analizirano svojstvo ali sa nedozvoljenim varijacijama za argumentovanje centralne tendencije

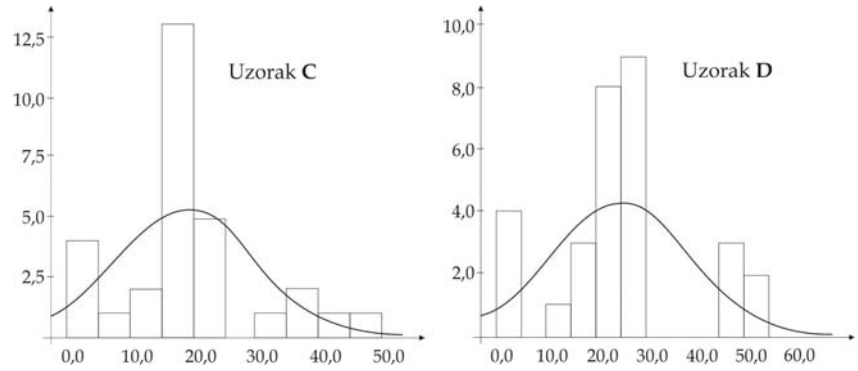
Ovde se otvara pitanje kako u analizi prosečne vrednosti određenog svojstva (ψ) posmatrati istovrsne statističke jedinice (Ω) koje imaju merljive vrednost ispitivanog svojstva ali je njihova varijacija iznad dozvoljenog varijabiliteta kojom se može argumentovati centralna tendencija uzorka, odnosno, centralna tendencija osnovnog skupa. Naime, u biološkim i poljoprivrednim istraživanjima relativno često je prisutna situacija da povećavanje broja statističkih jedinaca posmatranja u uzorcima ne dovodi do smanjenja varijacije analiziranog svojstva. Konačno, ako u uzorcima imamo kritičan broj statističkih jedinica posmatranja Ω, i ako je varijacija analiziranog svojstva ψ iznad dozvoljene varijacije za argumentaciju centralne tendencije, onda je jasno da u uzorku imamo pojavu podskupova sa indikativnim razlikama u ispoljavanju svojstva ψ. Model ove analize dat je u uzorcima C i D, tabela 4 i 5.

Tab. 4. Uzorak C sa sledećim utvrđenim vrednostima ψ (n = 30)
Sample C with the following established values of ψ (n = 30)

20,3	21,6	16,7	23,9	14,6	17,3	16,9	0,7	21,1	39,3
16,2	1,3	15,6	23,2	8,3	42,6	19,1	1,1	16,2	16,6
18,9	0,9	12,3	46,5	16,2	18,7	18,3	15,9	35,2	32,3

Tab. 5. Uzorak D sa sledećim utvrđenim vrednostima ψ (n = 30)
Sample D with the following established values of ψ (n = 30)

26,8	24,2	21,7	29,2	21,1	26,9	28,6	0,5	17,3	52,6
29,8	0,9	19,2	22,6	10,1	45,4	29,7	1,2	19,7	28,2
26,1	1,3	22,3	48,5	23,8	26,2	22,5	21,3	49,6	51,7



Sl. 1. *Grafički prikaz distribucije uzoraka C i D.*
Graphical illustration of samples C and D distributions

U tabelama 4 i 5 vidimo da sve statističke jedinice posmatranja Ω imaju definisano svojstvo ψ, a njihove prosečne vrednosti sa pripadajućim statističkim pokazateljima dati su u tabeli 6.

Tab. 6. Prosečna produkcija ψ, statističkih jedinica Ω, u uzorcima C i D
Average production of ψ,statistical units Ω, in samples C and D

Uzorak C / *Sample C*	Uzorak D / *Sample D*
$\overline{X} = 18{,}92\,kg$	$\overline{X} = 24{,}96\,kg$
$\sigma_x = 11{,}388$	$\sigma_x = 14{,}210$
$V_k = 60{,}18\%$	$V_k = 56{,}92\%$
$S_{\overline{X}} = 2{,}079$	$S_{\overline{X}} = 2{,}594$

Iz tabele 6. vidimo da oba uzorka imaju visoke koeficijente varijacije (60,18 % i 56,92 %) što znači da izračunate aritmetičke sredine ne predstavljaju centralne tendencije ovih uzoraka, odnosno, da je u posmatranim skupovima prisutno dejstvo eksperimentalno neopaženih faktora. Uzorci sa utvrđenim varijacijama ne argumentuju centralne tendencije, što u konačnom ima za posledicu da na osnovu istih nije moguće izvesti ni testiranja značajnosti postavljenih hipoteza. U datom primeru razlika između uzoraka C i D ($\Delta\overline{X}_{C-D} = 6{,}04\,kg$) statistički je slučajna ($t_{\overline{X}_C - \overline{X}_D} = 1{,}817^{nz}$) iako se iz iskustva zna da je ova razlika indikativna sa aspekta opšte produktivnosti svojstva ψ.

Ako pretpostavimo da se u posmatranom konačnom i prebrojivom skupu, odnosno, uzorku, visoke varijacije javljaju kao posledica dejstva eksperimentalno neopaženih faktora, onda možemo i da pretpostavimo da ukupni efekat ovih faktora ima za posledicu grupisanje statističkih jedinica posmatranja Ω u podskupove sa različitim centralnim tendencijama ispoljavanja svojstva ψ. Sada se otvara pitanje analitičkog pristupa u stratifikaciji ovakvog skupa ili uzoraka na podskupove sa argumentovanim centralnim tendencijama.

Matematičko–statistički metod determinisanja podskupova u konačnim i prebrojivim skupovima, odnosno, uzorcima, mora se bazirati na određivanju intervala varijacije kojim se uređena statistička serija segmentira na podskupove saglasno dozvoljenim varijacijama za argumentaciju centralne tendencije. Određivanje intervala varijacije mora se zasnivati na proceni standardne devijacije, odnosno, proceni one standardne devijacije koja bi se izračunala na osnovu statističkih jedinica koje se grupišu oko aritmetičke sredine sa zadanom varijacijom, a zatim i njenim stavljanjem u odnos sa utvrđenom aritmetičkom sredinom, tj. $\overline{X} \pm 3\sigma_x$. Za izračunatu aritmetičku sredinu iz date serije moguće je izvesti ovu matematizaciju preko unapred određenog koeficijenta varijacije. Npr. ako je Vk = 25 % $\Rightarrow$ $25 = \frac{\sigma_x}{\overline{X}} \cdot 100$. Dakle kako nam je $\overline{X}$ poznato, interval varijacije, koji sada možemo zvati i interval homogenizacije (*Ih*) u kome treba da se nalaze vrednosti statističkih jedinca, date uređenje statističke serije, koje variraju oko $\overline{X}$ sa varijacijom od 25 % nalaziće se u intervalu $\overline{X} \pm 0{,}75 \cdot \overline{X}$. Takođe, za varijaciju od 20 % interval homogenizacije je $\overline{X} \pm 0{,}6 \cdot \overline{X}$, za varijaciju od 15 % interval homogenizacije je $\overline{X} \pm 0{,}45 \cdot \overline{X}$, a za varijaciju od 10 % interval homogenizacije je $\overline{X} \pm 0{,}3 \cdot \overline{X}$.

Proverimo izvedenu matematizaciju na uzorku C: $\overline{X} \pm 0{,}6 \cdot 18{,}92 \Rightarrow$ *Ih*: 7,56 – 30,27.

0,7	0,9	1,1	1,3	8,3	12,3	14,6	15,6	15,9	16,2
16,2	16,2	16,6	16,7	16,9	17,3	18,3	18,7	18,9	19,1
20,3	21,1	21,6	23,2	23,9	32,3	35,2	39,3	42,6	46,5

Pregledom apsolutnih vrednosti u ovako podeljenom uzorku vidimo tri podskupa podataka. Centralne tendencije ovih podskupova prikazane su u tabeli 7.

Tab. 7. Centralne tendencije svojstva ψ, kod statističkih jedinica Ω u podskupovima uzorka C.
Central tendencies of characteristic ψ, of statistical units Ω, in the subsets of sample C

Podskup niske produkcije ψ *Subset of low production ψ*	Podskup prosečne produkcije ψ *Subset of average production ψ*	Podskup visoke produkcije ψ *Subset of high production ψ*
$n = 4$	$n = 21$	$n = 5$
$\overline{X} = 1{,}00\,kg$	$\overline{X} = 17{,}51\,kg$	$\overline{X} = 39{,}18 kg$
$\sigma_x = 0{,}258$	$\sigma_x = 3{,}541$	$\sigma_x = 5{,}667$
$V_k = 25{,}82\%$	$V_k = 20{,}21\%$	$V_k = 14{,}46\%$
$S_{\overline{X}} = 0{,}129$	$S_{\overline{X}} = 0{,}773$	$S_{\overline{X}} = 2{,}534$

Primenom iste matematizacije na uzorku D, dobijaju se takođe tri podskupa podataka čije su centralne tendencije prikazane u tabeli 8.

Tab. 8. Centralne tendencije svojstva ψ, kod statističkih jedinica Ω u podskupovima uzorka D.
Central tendencies of characteristic ψ, of statistical units Ω, in the subsets of sample D

Podskup niske produkcije ψ *Subset of low production ψ*	Podskup prosečne produkcije ψ *Subset of average production ψ*	Podskup visoke produkcije ψ *Subset of high production ψ*
$n = 4$	$n = 21$	$n = 5$
$\overline{X} = 0{,}975\ kg$	$\overline{X} = 23{,}68\ kg$	$\overline{X} = 49{,}56\ kg$
$\sigma_x = 0{,}359$	$\sigma_x = 4{,}812$	$\sigma_x = 2{,}839$
$V_k = 36{,}86\%$	$V_k = 20{,}32\%$	$V_k = 5{,}73\%$
$S_{\overline{X}} = 0{,}179$	$S_{\overline{X}} = 1{,}050$	$S_{\overline{X}} = 1{,}270$

Ako sada testiramo značajnost razlika centralnih tendencija svojstva ψ, kod istovrsnih statističkih jedinica Ω u podskupovima uzoraka C i D, dolazimo do sledećih zaključaka:

Podskupovi uzoraka C i D *Subsets of samples C and D*	$\Delta\overline{X}_{C-D}$ *(kg)*	$t_{\overline{X}_C - \overline{X}_D}$
Podskup niske produkcije ψ *Subset of low production ψ*	0,025	$0{,}113^{nz}$
Podskup prosečne produkcije ψ *Subset of average production ψ*	6,170	4,732**
Podskup visoke produkcije ψ *Subset of high production ψ*	10,38	3,662**

U konačnoj matematazaciji uzoraka C i D vidimo da statistička analiza ova dva uzorka za sve eksperimentalno utvrđene vrednosti ψ (n = 30), prikazana u tabeli 6. daje sasvim pogrešan zaključak, odnosno, pri ispoljenim ukupnim varijacijama većim od 30 % nije moguće argumentovati centralnu tendenciju u oba ova skupa podataka. Međutim, razdvajanjem vrednosti obeležja u podskupove sa unapred zadatom varijacijom podataka oko izračunate aritmetičke sredine možemo zaključiti sledeće:

- u posmatranim uzorcima evidentirana su tri podskupa sa argumentovanim centralnim tendencijama;
- eksperimentalna i logičko–matematička analiza strukture ovih podskupova treba da dâ odgovor o uzrocima ovakvog grupisanja analiziranih statističkih jedinica posmatranja;
- centralne tendencije podskupova prosečne produkcije ψ i podskupova visoke produkcije ψ između uzoraka C i D statistički je visoko značajna.

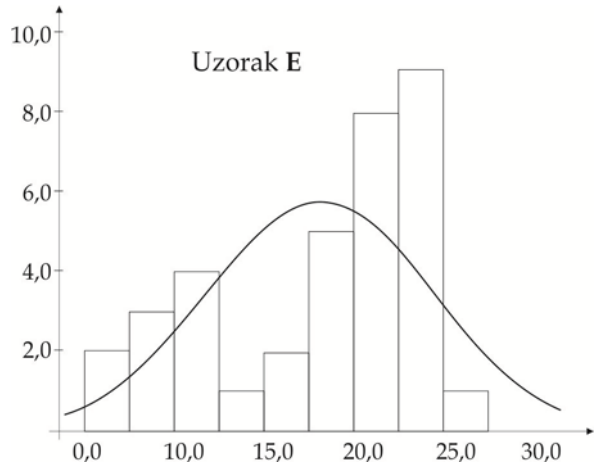
Ovako protumačeni rezultati istraživanja mogu da pruže realniju argumentaciju odnosa u posmatranim skupovima, ali i da otvore nova relevantna pitanja za dalja istraživanja.

Neophodno je naglasiti da grupisanje statističkih jedinica posmatranja u podskupove nužno nameće i pitanje diskusije izdvojenih podskupova. Naime, u praksi je prisutan i takav pristup, da se u slučaju visokih varijacija u uzorcima sukcesivno uklanjaju najmanje i najveće vrednosti što je neprihvatljivo jer se u konačnom dobija pogrešan rezultat. Procenu intervala homogenizacije u okviru koga se grupišu vrednosti obeležja saglasno centralnim tendencijama, treba izvesti na više nivoa homogenizacije (0,3; 0,45; 0,60; 0,75) kako bi se što bolje odredila granica podskupova. Na ovaj način određeni podskupovi moraju se posebno argumentovati, ali i utvrditi njiva struktura u uzorku, odnosno u osnovnom skupu.

Na sledećem statističkom skupu modeliran je primer homogenizacije statističkih jedinica posmatranja u podskupove kod uzoraka sa varijacijama na granici argumentacije centralne tendencije, uzorak E, tabela 9.

Tab. 9. Uzorak E sa sledećim utvrđenim vrednostima ψ (n = 35)
Sample E with following values ψ (n = 35)

6,0	6,3	7,8	7,9	8,5	10,5	11,0
11,6	12,3	12,6	15,4	17,3	18,4	18,6
18,8	18,8	18,9	20,2	20,6	21,2	21,7
21,8	22,0	22,3	22,3	22,6	23,1	23,1
24,0	24,0	24,3	24,6	24,7	24,9	25,0



Sl. 2. *Grafički prikaz distribucije uzorka E.*
Graphical illustration of sample E distribution

Statistički parametri uzorka E: $\overline{X} = 18{,}09$; $\sigma_x = 6{,}106$; $V_k = 33{,}75\%$ i $S_{\overline{X}} = 1{,}032$. Evidentno je da varijacija od 33,75 % ne argumentuje centralnu tendenciju uzorka E.

U cilju analize mogućeg prisustva podskupova u uzorku E, izvedena je homogenizacija na nivou varijacije od 20 % ($\overline{X} \pm 0{,}6 \cdot \overline{X}$) i 15 % ($\overline{X} \pm 0{,}45 \cdot \overline{X}$):

Interval homogenizacije: *Interval of homogenization* $\overline{X} \pm 0{,}6 \cdot \overline{X}$	Interval homogenizacije: *Interval of homogenization* $\overline{X} \pm 0{,}45 \cdot \overline{X}$
Ih: 7,8 – 25,0	*Ih*: 10,5 – 25,0
$n = 33$	$n = 30$
$\overline{X} = 18{,}81$	$\overline{X} = 19{,}88$
$\sigma_x = 5{,}492$	$\sigma_x = 4{,}487$
$V_k = 29{,}19\%$	$V_k = 22{,}56\%$
$S_{\overline{X}} = 0{,}956$	$S_{\overline{X}} = 0{,}819$

Rezultati homogenizacije na nivou $\overline{X} \pm 0{,}6 \cdot \overline{X}$, pokazuju da je neophodno isključiti prve dve vrednosti obeležja iz serije (6,0 i 6,3), kako bi se centralna tendencija argumentovala na nivou *Vk* < 30 %, tj. *Vk* = 29,19 % . Međutim, jasno je da ove dve isključene vrednosti ne dokazuju prisustvo podskupova u uzorku E. Zato je urađena homogenizacija na nivou $\overline{X} \pm 0{,}45 \cdot \overline{X}$, i ovaj interval homogenizacije sada pokazuje da prvih pet vrednosti ψ u uzorku E (6,0; 6,3; 7,8; 7,9 i 8,5) predstavlja podskup niskog ispoljavanja svojstva ψ, u odnosu na argumentovanu centralnu tendenciju prosečnog ispoljavanja svojstva ψ.

Diskusija dobijenih rezultata analize prosečne vrednosti svojstva ψ u uzorku E, izvodi se na sledeći način:

- Analizirano svojstvo ψ kod statističkih jedinica Ω u uzorku E pojavljuje se u dva podskupa: 1) statističke jedinice Ω sa niskim ispoljavanjem svojstva ψ ($\overline{X} = 7{,}3 \pm 0{,}49$,*Vk* = 14,91 %); i 2) statističke jedinice Ω sa očekivanim ispoljavanjem svojstva ψ ($\overline{X} = 19{,}88 \pm 0{,}82$,*Vk* = 22,56 %);
- Podskup statističkih jedinica Ω sa niskim prosečnim sadržajem ψ ($\overline{X} = 7{,}3$) u uzorku E čini 14,28 % statističkih jedinica Ω.

Testiranjem značajnosti razlike aritmetičkih sredina uzorka E između izračunate i korigovanih sredina, ispoljena razlika je statistički slučajna. Dakle, u uzorku E, 85,71 % statističkih jedinica posmatranja Ω ima vrednost ispitivanog svojstva ψ na nivou očekivanog, odnosno centralna tendencija argumentovano predstavlja prosečan sadržaj ψ. Takođe, izdvajanje podskupa statističkih jedinica Ω sa niskim sadržajem svojstva ψ, nije uticalo na značajnu promenu prosečne vrednosti očekivanog sadržaja ψ u argumentaciji centralne tendencije, ali je otvorilo pitanje naknadnog istraživanja uzroka za ovu pojavu.

Zaključak

Zamka analitičkog posmatranja aritmetičke sredine u domenu deskriptivne ili inferencijalne statistike stoji u definiciji njene reprezentativnosti. Predloženi metod analize deskriptivnih podataka pomoću koeficijenta varijacije i stratifikacije intervalom homogenizacije, omogućuje izdvajanje homogenih poskupova u uzorcima, odnosno, adekvatnu argumentaciju centralnih tendencija u onim statističkim skupovima gde je varijabilitet podataka iznad dozvoljene varijacije za argumentaciju centralne tendencije. Uz prilagođenost primenjenim poljoprivrednim istraživanjima predložena matematičko-logička utemeljenost opisanog metoda predstavlja njegovu glavnu prednost u odnosu na grafičke i druge slične rasprostranjene metode. Preciznom deskriptivnom statistikom, u biometričkom smislu, dolazimo do ispravno procenjenih parametara i preciznih biometričkih zaključaka. Zamke koje su predstavljene u ovom radu tiču se pre svega preciznosti deskriptivne statistike kako se mora shvatiti u biometrici. Tačno izračunata deskriptivna statistika predstavlja samo polaznu osnovu za precizan biometrički pristup. Tačno izračunate mere deskriptivne statistike neispitane dodatnim biometričkim metodama predstavljaju sumnjivu i često netačnu osnovu za donošenje zaključaka i inferencijalnu statistiku uopšte, te stoga i jesu nazvane zamkama. Metoda opisana u ovom radu predstavlja jednu dodatnu jednostavnu matematičko-logičku alatku i algoritam za izbegavanje tih zamki u primenjenim poljoprivrednim i biološkim istraživanjima.

Literatura

Dormann, C.F. & Kuhn, I. (2012). *Angewandte Statistik fur die biologischen Wissenschaften, Second Edition*. Helmholtz Zentrum fur Umweltforschung, UFZ.

Kohler, W., Schachtel, G. & Voleske, P. (2012). *Biostatistik, Eine Einführung für Biologen und Agrarwissenschaftler, 5th Ed*. Berlin-Heidelberg: Springer.

Mićić, N. & Bosančić, B. (2012) Variability and Variation Coefficients in Biological and Agricultural Experimental Research. *Agro-knowledge Journal 13*(3), 331-341.

Rao, N.G. (2007). *Statistics for Agricultural Sciences, 2nd Ed.* Hyderabad: BS Publications.

Reed, G.F., Lynn, F. & Meade, D.B. (2002). Use of Coefficient of Variation in Assessing Variability of Quantitative Assays. *Clinical and Diagnostic Laboratory Immunology, 9*(6), 1235–1239.

Sokal, R.R. & Rohlf, J.F. (1995). *Biometry, the Principles and Practice of Statistics in Biological Research, 3rd Ed*. New York: W. H. Freeman and Company.

Quinn, G.P. & Keough, M.J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge, UK: Cambridge University Press.

Weber, E.U. & Sharoni, S. (2004). Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review, 111*(2), 430-445.

Pitfalls of Descriptive and Inferential Statistical Approach in Biological and Agricultural Sciences

Nikola Mićić[1, 2], Borut Bosančić[1, 2]

[1]*Faculty of Agriculture, University of Banja Luka, Bosnia and Herzegovina*
[2]*Genetic Resources Institute,University of Banja Luka, Bosnia and Herzegovina*

Abstract

Descriptive statistics observes a data set as the given set, i.e. final and countable statistical set and interprets it in that manner, while on the other hand inferential statistics as a first step in the analytical approach, in accordance with the research aim, has to define elementary determinations of the basic set (term, space and time) and then also samples related to the planned research, i.e. their representativeness in assessing the parameters of the basic set. Furthermore, in the inferential statistics the choice of the logical-mathematical argumentation in the assessment of parameters of the basic set has to determine also how large is the error the conclusions will be exposed to on the basis of which the probability of the defined hypothesis is assessed for the basic set. Hence while descriptive statistics establishes the present state in the given final and countable data set, the inferential statistics on the basis of experimental, instrumental and logical-mathematical methods analyses variations of data in the samples and assesses the explained, unexplained and allowed variations of observed characteristic, as a measure of occurrence probability for the observed character in the given set. Consequently, in the descriptive statistics used is the mathematical arithmetic mean, while in the inferential statistics arithmetic mean actually represents central tendency as a reliable probability for occurrence or exhibit of the observed characteristic in the basic set. Therefore, in the inferential statistics the representativeness of the samples is in reality the representativeness of the central tendencies of the samples, which is supported with argument of allowable variations of the observed values of the characteristic, i.e. allowable relative variability interval (5%<CV<30%). Namely, samples where the coefficient of variation are lower that 5% "are too good" (are related to the set of same statistical units), and samples with coefficient of variation larger than 30% have to be decomposed into subsamples with allowable variability for the central tendency and basic assumption for the analysis of the structure of the subsets of the observed characteristic in the basic set. In this paper elaborated is the issue of biometrical analysis of samples with unallowable relative variability of data in argumentation of the central tendency.

Key words: samples, central tendency, interval of homogenization, biometry

Nikola Mićić
E-mail address:
nikmicic@yahoo.com