POGLAVJE 4

Vzorčenje

V prvem poglavju smo obravnavali količine, kot so povprečje, mere razpršenosti in odstotki enot z dano lastnostjo v populaciji. Pogosto se znajdemo v situaciji, ko nimamo podatkov za celotno populacijo, ker bi bilo zbiranje preprosto prezamudno ali predrago. V takih primerih si pomagamo z vzorčenjem. Iz populacije izberemo samo manjše število enot in na podlagi podatkov za te enote poskušamo oceniti količine, ki nas zanimajo, za celotno populacijo.

Z vzorčnimi podatki se srečujemo vsak dan. Primer so predvolilne ankete, ki poskušajo na podlagi odgovorov nekaj tisoč volivcev napovedati izid volitev. Drugi primer je ocenjevanje rasti življenjskih stroškov v Sloveniji, tretji revizorsko "vzorčenje" vknjižb v podjetjih, seveda pa je primerov še več. Ocene, ki jih dobimo na podlagi izbranega vzorca, zajamejo samo del celotne populacije, zato se moramo seveda vprašati, koliko se na nanje lahko zanesemo. Odgovor na to vprašanje ni vedno preprost in je odvisen tako od velikosti vzorca kot tudi od načina izbire enot v vzorec. V tem poglavju se bomo lotili vprašanja, kako je treba izbirati vzorec in kaj lahko potem trdimo o natančnosti ocen. Vpeljali bomo pojem vzorčnega načrta in standardne napake ocene, ki smo jo izračunali na podlagi vzorca.

4.1 Uvodni primeri

4.1.1 Plebiscit 1990

Plebiscit o neodvisnosti Slovenije decembra 1990 je bil prelomni dogodek pri osamosvajanju. Ko je takrat jeseni naraščala napetost, so mnogi nestrpno pričakovali rezultate predplebiscitnih anket. Ena od odmevnejših je bila SJM 90 (Slovensko javno mnenje 90), ki so jo izpeljali na Fakulteti za družbene vede na Univerzi v Ljubljani.

Za kaj pravzaprav gre pri taki anketi? Pred plebiscitom je nemogoče ugotoviti mnenje vsakega volivca, zato se je treba zateči k izbiranju manjšega števila enot iz populacije volivcev. Izbrani skupini v statističnem žargonu pravimo *vzorec*. Anketa SJM 90 je bila zasnovana na izbiri vzorca velikosti 2074 volivcev, od katerih se jih je 1306 nedvoumno izreklo tako za samostojnost kot za odcepitev Slovenije. Natančni rezultati ankete so v spodnji tabeli [1].

		Samostojnost		
		Da	Ne	Drugo
Odcepitev	Da	1306	11	34
	Ne	183	125	63
	Drugo	110	12	230

Tabela 4.1: Tabela razultatov SJM90

Podatki iz vzorca kažejo, da se je velika večina volivcev iz vzorca izrekla za neodvisnost Slovenije. Iz tega bi sklepali, da se bo tudi večina vseh volilcev v Sloveniji odločila za neodvisnost. Ta sklep je sedaj, ko je rezultat plebiscita že zdavnaj znan, očiten. Poskusimo pa se postaviti v čas jeseni 1990. Vzorec je zajel samo neznaten delež celotnega volilnega telesa, komajda nekaj več kot 0,1% vseh upravičencev. Skeptiki bi se gotovo vprašali, ali lahko na podlagi tako neznatnega vzorca zanesljivo sklepamo o

[1] Vir: SJM90

volji celotne populacije volivcev. Lahko si zamislimo, da bi se v vzorcu znašlo ali preveč privržencev neodvisnosti ali preveč nasprotnikov. V takem primeru bi bila napoved izida seveda napačna. Ali nam statistika lahko pomaga, da presodimo zanesljivost napovedi na podlagi vzorčnih podatkov? Odgovor je pritrdilen.

Prvi korak pri razmišljanju o zanesljivosti ocen na podlagi vzorca mora biti opis načina izbire vzorca ali, kot se temu pravi v statistiki, vzorčni načrt. Pri ustrezno izpeljanih anketah je postopek izbire vzorca natanko predpisan. Poglejmo si način vzorčenja pri anketah SJM, ki je bil uporabljen tudi za anketo o plebiscitu.

Okvir vzorčenja pri anketah SJM je centralni register prebivalstva, ki ga vzdržuje Statistični urad RS. Gre preprosto za primerno urejen seznam vseh prebivalcev Slovenije, iz katerega zlahka dobimo tudi seznam vseh volivcev. Ta seznam je za potrebe vzorčenja po geografskem ključu razdeljen na manjše dele po 4200 volivcev. Tem skupinam pravimo *primarne vzorčne enote*. Vsaka od teh manjših skupin po 4200 volivcev je nadalje razdeljena na skupine po 100, spet po takem ključu, da volivci v podskupini živijo v skupnosti, kot je naselje ali vas. Tem manjšim skupinam pravimo *sekundarne vzorčne enote*. Izbira vzorca SJM poteka v treh korakih:

1. Na prvem koraku anketarji naključno izberejo 140 primarnih vzorčnih enot, torej 140 skupin po 4200 volivcev.

2. Na drugem koraku so v vsaki na prvem koraku izbrani primarni vzorčni enoti izbrane tri sekundarne vzorčne enote, torej tri skupine po 100 volivcev.

3. Na tretjem koraku nato anketarji v vsaki izbrani sekundarni vzorčni enoti naključno izberejo 5 volivcev.

Če izračunamo, je na koncu v vzorec izbranih $140 \cdot 3 \cdot 5 = 2100$ volivcev. Nato je seveda treba z izbranimi volivci stopiti v stik in jih povprašati po njihovem mnenju. Izvajalci ankete na terenu imajo napotek, da od vsake izbrane osebe poskusijo dobiti odgovor. Če prvi poskus ni uspešen, anketarji poskušajo znova do največ petkrat, če je to potrebno. Če so vsi poskusi neuspešni, anketarji neznani odgovor obravnavajo kot manjkajoči podatek. Torej, način vzorčenja je vnaprej določen in anketarji se ga morajo držati. Drug pomemben dejavnik pri vzorčenju za SJM je, da je izbira enot na vsakem koraku naključna. Ni vnaprej določeno, katere skupine ali podskupine ali

nazadnje volivci bodo izbrani, temveč so vse enote izbrane kot na "loteriji". Gotovo je na tem mestu utemeljeno vprašanje, zakaj je treba uvesti tovrstno naključno izbiro. Razlog je v odpravljanju vsakršne pristranosti pri izbiri vzorca. Nekako tako, kot da bi pred jemanjem vzorca kapljice tekočine iz steklenice le-to dobro pretresli. Vzorec iz dobro pretresene steklenice mnogo bolje pokaže njeno vsebino, kot pa če steklenice ne bi pretresli.

Kot bomo videli v nadaljevanju, način vzorčenja bistveno vpliva na zanesljivost ocen. Na podlagi opisa vzorčenja lahko domnevamo, da so bili rezultati ankete SJM 90 precej zanesljivi, saj je bila "steklenica zelo dobro pretresena". V naslednjem razdelku si bomo ogledali, kako natančneje opišemo, do kolikšne mere lahko verjamemo, da vzorčni rezultati zanesljivo odražajo stanje v celotni populaciji.

4.1.2 Indeks cen življenjskih potrebščin

Statistični urad Republike Slovenije vsak mesec objavi indeks cen življenjskih potrebščin, ki je eden od osnovnih indikatorjev inflacije. Osnovi za izračun tega indeksa sta:

1. Seznam najvažnejših predmetov in storitev gospodinjske porabe in podatki o višini stroškov za posamezne postavke. Osnova za določitev obsega te porabe je anketa o porabi v gospodinjstvih, ki jo izvaja Statistični urad RS.

2. Povprečna porast cen na drobno za posamezne postavke s seznama.

Formula, ki jo Statistični urad RS uporablja za izračun tega indeksa, je

$$I = \frac{\sum_{i=1}^{m} (p_{1i}/p_{0i}) \cdot w_{0i}}{\sum_{i=1}^{m} w_{0i}} \cdot 100\% \,.$$

Pri tem je kvocient p_{1i}/p_{0i} porast cene na drobno za predmet ali storitev i s seznama, utež w_{0i} pa je povprečni delež stroškov, ki je namenjen za dani predmet ali storitev glede na celotne izdatke v gospodinjstvu. Za primer navedimo, da gospodinjstva v Sloveniji za hrano v povprečju namenijo 23,1%[2] celotne mesečne porabe.

[2] Vir: Statistični letopis 1996, str. 234

Zgornji formuli ne bomo posvetili posebne pozornosti. Omenimo le, da so uteži potrebne zato, ker morajo imeti predmeti ali storitve s seznama, za katere gospodinjstva namenijo večji delež izdatkov, pri določanju rasti življenjskih stroškov večjo težo. Podražitev hrane ima na življenjsko raven večji vpliv kot, recimo, zvišanje cen frizerskih storitev.

Če želimo zgornjo formulo uporabiti, moramo priti na dan z dejanskimi številkami. Med drugim moramo za posamezne predmete ali storitve s seznama ugotoviti povprečne deleže porabe v gospodinjstvih. Ker bi bilo težko spremljati porabo v vseh gospodinjstvih, si Statistični urad RS pomaga z vzorčenjem. Gospodinjstva v Sloveniji so popisana in urejena v popisne okoliše. Za potrebe izbire vzorca so popisni okoliši razdeljeni v 6 podskupin glede na lokacijo in tip. Te podskupine v statistiki imenujejo *stratumi*.

Vzorčenje poteka v dveh korakih:

1. Najprej v vsakem stratumu naključno izberemo popisne okoliše.

2. Na drugem koraku v vsakem izbranem popisnem okolišu naključno izberemo 5 gospodinjstev.

V vzorec je zajetih 3270 gospodinjstev. Število popisnih okolišev, izbranih na prvem koraku, je tako, da v vsakem stratumu na koncu postopka izbire zajamemo 0,5% gospodinjstev. Statistični urad RS izbranim gospodinjstvom razdeli vprašalnike za vodenje evidence o porabi. Anketarji urada zberejo izpolnjene evidenčne vprašalnike in na podlagi tako zbranih podatkov določijo deleže porabe za posamezne predmete ali storitve. Prej omenjenih 23,1% stroškov za hrano je ena od tako dobljenih ocen.

Seveda se moramo vprašati, do kolikšne mere lahko ocenam na podlagi vzorca zaupamo. Odgovor je odvisen od tega, kako dobro je bil vzorčni načrt premišljen in ali je sámo vzorčenje bilo izvedeno z nadzorom. Element naključnosti je bistvena sestavina vzorčenja, saj nam izkušnje iz preteklosti in teoretična razmišljanja kažejo, da tako najbolje dosežemo nepristranost vzorčnih ocen, poleg tega pa edino strog vzorčni načrt omogoča presojo o tem, kako zanesljiva je ocena. Brez naključne izbire vzorca taka presoja ni možna. Velikost vzorca je izbrana tako, da so dobljeni vzorčni odstotki dovolj zanesljivi za uporabo v formuli za izračun indeksa cen življenjskih potrebščin.

Rezultat zgornje formule za december 1996 je bil 1,3%.

4.1.3 TIMSS v Sloveniji

Naslednji primer, kjer zbiranje podatkov poteka z vzorčenjem, so raziskave uspešnosti šolskih sistemov. Učenci slovenskih osnovnih in srednjih šol so vključeni v mednarodne primerjalne raziskave znanja matematike, naravoslovja, računalništva, bralne pismenosti in drugih področij znanja. V Sloveniji izvaja te primerjalne raziskave Pedagoški inštitut v Ljubljani. Za opis postopka vzorčenja v teh raziskavah si izberimo zadnjo izmed raziskav TIMSS, ki je potekala od leta 1991 in je bila končana leta 1998. To raziskavo smo že srečali v prejšnjih poglavjih.

V jeziku prvega poglavja spadajo v populacijo, ki nas v tem primeru zanima, vsi učenci sedmih in osmih razredov slovenskih šol. V letu 1995, ko je potekalo dejansko vzorčenje, je bila ta populacija velika $N = 54.965$ učencev[3]. Iz praktičnih razlogov je bilo izključenih 310 učencev (večinoma učenci iz šol za slepe in slabovidne ter podobnih), tako da je bila na koncu populacija, iz katere je bil izbran vzorec, velika $N = 54.655$.

Zanima nas znanje matematike in naravoslovja pri učencih iz opisane populacije. Znanje je bilo merjeno z nalogami, ki so bile sestavljene posebej za TIMSS. Če se spet vrnemo k jeziku prvega poglavja, sta spremenljivki raven znanja matematike in raven znanja naravoslovja posameznega učenca. Ta ravni izrazimo kot število na posebnih lestvicah, ki so premišljene tako, da omogočajo smiselno primerjavo med državami, ki so bile vključene v raziskavo. Za nas je v tem trenutku pomembno to, da vsaki enoti v populaciji pripadata neki vrednosti, ki pomenita znanje matematike oziroma naravoslovja.

Preden se lotimo opisa vzorčenja, se moramo seveda vprašati, zakaj je vzorčenje sploh potrebno. Populacija je dokaj velika, zato si ni težko predstavljati, koliko dela bi bilo z izvedbo preizkušanja znanja, ki bi zajela vse učence v populaciji, da o vrtoglavih stroških tako velikega projekta sploh ne govorimo. Take raziskave za celotno populacijo učencev ni mogoče izpeljati iz povsem praktičnih razlogov. Rešitev je v vzorčenju. Iz celotne populacije izberemo manjši vzorec učencev in povprečno raven znanja slovenskih sedmošolcev in osmošolcev ocenimo na podlagi znanja učencev iz tega vzorca. V raziskavi TIMSS je bilo v vzorec zajetih 5927 učencev.

[3] Vir: Pedagoški inštitut Ljubljana

© Visoka šola za podjetništvo Piran

Samo izbiranje vzorca je bilo mednarodno usklajeno in vnaprej predpisano. Vzorčenje ni potekalo z neposrednim izbiranjem učencev, temveč posredno v dveh korakih. Na prvem koraku so sodelavci Pedagoškega inštituta med 455 osnovnimi šolami v Sloveniji naključno izbrali 150 šol, in sicer tako, da so imele večje šole nekaj več verjetnosti, da bodo izbrane v vzorec. Tukaj besedo "naključno" uporabljamo še v nekoliko ohlapnem pomenu, kasneje pa bomo ta pojem tudi natančneje opredelili. Ko so bile šole izbrane, je bil naslednji korak izbiranje razreda. Izbira je bila spet naključna, in sicer taka, da je bil izbran po en 7. in en 8. razred na vsaki šoli, ki je bila izbrana na prvem koraku. V končnem vzorcu so bili zajeti vsi učenci iz izbranih razredov, torej zgoraj omenjenih 5927 učencev. Takemu načinu izbiranja vzorca statistiki pravijo "vzorčenje v skupinicah", ker izbiramo celotne skupinice enot, v tem primeru razrede. Vsi izbrani učenci so reševali izbrane naloge in na podlagi njihovih odgovorov je bila ocenjena povprečna raven znanja matematike oziroma naravoslovja vseh slovenskih sedmošolcev in osmošolcev.

Tudi tukaj si moramo zastaviti vprašanje o zanesljivosti dobljenih ocen. Raziskava je zajela samo okrog 10% vseh učencev iz opisane populacije in na podlagi njihovih rezultatov je bila potem ocenjena raven znanja za celotno populacijo. Do kolikšne mere je to utemeljeno? Kako opisati zanesljivost ocen? Na to vprašanje bomo odgovorili v razdelkih, ki sledijo.

4.1.4 Predsedniške volitve v ZDA leta 1936

Kot zadnji primer vzorčenja si oglejmo znamenito anketo, kjer so šle stvari pri napovedovanju izida volitev zelo hudo narobe. Pred predsedniškimi volitvami v ZDA leta 1936 je prestižna revija *Literary Digest* na podlagi obsežne ankete napovedala zmagovalca. Danes vemo, da je bil na omenjenih volitvah s precejšnjo večino izvoljen Franklin D. Roosevelt, revija pa je takrat napovedala, da bo zmagal njegov republikanski tekmec Alfred Landon. V tabeli 4.2 so vsebovane napovedi revije *Literary Digest* in dejanski rezultati. Jasno je, da anketa ni mogla zajeti vseh volivcev v ZDA leta 1936, zato je bilo treba izbrati vzorec. Vzorec, ki so ga izbrali anketarji revije *Literary Digest*, je bil največji, kar so jih sploh kdaj izbrali, in je vseboval 2,4 milijona oseb. Kljub tako velikemu vzorcu pa se je napoved razlikovala od dejanskega rezultata za skoraj 20%!

	Napoved revije *Literary Digest*	Dejanski rezultat
F. D. Roosevelt	43%	62%
A. Landon	57%	38%

Tabela 4.2: Napovedi in rezultati predsedniških volitev v ZDA leta 1936

Zanimivo je seveda vprašanje, kaj je povzročilo tako zelo zgrešeno napoved. Če želimo dobiti odgovor na to vprašanje, si moramo ogledati, kako je bil vzorec izbran. Izvor za izbiro volivcev v vzorec so bili seznami naročnikov revije *Literary Digest*, telefonski imeniki, seznami članov elitnih klubov in podobno. Vprašalnike so poslali izbranim osebam po pošti, in sicer kar 10 milijonom, odgovorilo pa je samo 2,4 milijona naslovnikov, kar je že razlog za previdnost. Če pomislimo, da je bilo leto 1936 najbolj črno leto velike depresije in je bilo v ZDA 11 milijonov brezposelnih, nam takoj pade v oči, da od le-teh velika večina ni imela telefona in torej njihova imena niso bila v telefonskem imeniku ali na seznamu članov kakšnega elitnega kluba. Če še pomislimo, da republikansko stranko v ZDA po pravilu podpirajo bogatejši sloji, je eden od razlogov za slabe napovedi na dlani. Anketa je bila že vnaprej načrtovana tako, da so bili v vzorec zajeti tisti, ki so tudi med depresijo imeli kaj pod palcem. Velikost vzorca seveda ni pomagala, ker se je samo ponavljala ena in ista napaka. V vzorcu se je vedno znova znašlo več republikanskih volivcev kot pa podpornikov predsednika Roosevelta. V statističnem žargonu bi lahko rekli, da je bil vzorčni načrt pri tej anketi slab, z drugimi besedami, način izbire vzorca je bil slabo premišljen. Še bolj primerna izjava bi bila, da vzorčnega načrta sploh ni bilo. Revija *Literary Digest* se je kmalu po volitvah 1936 znašla v stečaju.

Kot zanimivost lahko povemo še to, da je v istem času mladi statistik George Gallup na podlagi vzorca velikosti 5000 oseb napovedal zmago F. D. Roosevelta s 56% glasov. Še bolj zanimivo je to, da je George Gallup napovedal tudi napoved revije *Literary Digest*, še preden jo je ta objavila. Izbral je vzorec velikosti 3000 izmed tistih, ki so prejeli vprašalnike, in na podlagi izbranega vzorca napovedal, da bo napoved 44% za Roosevelta. Res ne bi bilo treba izbirati vzorca velikosti 2,4 milijona!

Pogosto je nemogoče zbrati podatke o vrednostih spremenljivke za vse enote v populaciji. Zato iz populacije izberemo vzorec, ki zajema le njen manjši del. Iz vrednosti spremenljivke za enote v vzorcu potem ocenimo želene količine, na primer povprečno vrednost spremenljivke ali odstotek enot z določeno lastnostjo, za celotno populacijo. Če nas zanima povprečna vrednost spremenljivke za celotno populacijo, kot oceno za to količino vzamemo povprečno vrednost spremenljivke za enote v vzorcu. Če nas zanima odstotek enot v populaciji z neko lastnostjo, za oceno tega odstotka vzamemo odstotek enot v vzorcu, ki imajo izbrano lastnost.

Vzorčni načrt je vnaprej predpisan postopek izbiranja vzorca iz vnaprej določene in natančno opredeljene populacije. Če je izbira enot ali skupin enot naključna, potem takemu vzorčenju pravimo *verjetnostno vzorčenje*. Z vpeljavo naključnosti se najbolje izognemo pristranosti.

4.2 Enostavno slučajno vzorčenje

4.2.1 Pojem enostavnega slučajnega vzorca

Iz primerov v prejšnjem razdelku, posebej iz zadnjega, je razvidno, da je način izbire vzorca zelo pomemben. V tem razdelku si bomo ogledali najpreprostejši vzorčni načrt: *enostavno slučajno vzorčenje*. Čeprav se ta tip vzorčenja v dejanskih anketah redko uporablja, je sestavni del mnogih, tudi bolj zapletenih vzorčnih načrtov, poleg tega pa je pri njem najbolj razviden pojem standardne napake. Za okvir razmišljanja vzemimo populacijo velikosti N, iz katere želimo izbrati vzorec velikosti n. Zamislimo

si, da imamo za vsako od N enot listek, ki ga damo v škatlo, škatlo dobro pretresemo in naključno izberemo n listkov. S tem dosežemo, da je vsaka enota iz populacije zajeta v vzorec z enako verjetnostjo.

> ☞ Pri enostavnem slučajnem vzorčenju je verjetnost, da je enota zajeta v vzorec, enaka za vse enote. Še več, vsi možni vzorci velikosti n iz populacije velikosti N so pri enostavnem slučajnem vzorčenju enako verjetni.

Vseh možnih izbranih vzorcev je za večje populacije seveda zelo veliko. Poglejmo si preprost primer, ko je $N = 6$ in $n = 3$. Izbiramo torej vzorec velikosti 3 iz populacije velikosti 6. Vsi možni vzorci so naslednji:

{1, 2, 3}	{1, 2, 4}	{1, 2, 5}	{1, 2, 6}	{1, 3, 4}
{1, 3, 5}	{1, 3, 6}	{1, 4, 5}	{1, 4, 6}	{1, 5, 6}
{2, 3, 4}	{2, 3, 5}	{2, 3, 6}	{2, 4, 5}	{2, 4, 6}
{2, 5, 6}	{3, 4, 5}	{3, 4, 6}	{3, 5, 6}	{4, 5, 6}

Različnih možnih izbir vzorca velikosti 3 je 20. Za večje populacije dobimo mnogo večja števila možnih vzorcev. Za občutek poglejmo, na koliko načinov lahko izberemo vzorec velikosti n iz populacije velikosti N za različne vrednosti. Iz matematike si sposodimo formulo, da je število teh načinov enako

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}\,.$$

Z uporabo te formule dobimo, da je za:

© Visoka šola za podjetništvo Piran

$$N = 10 \qquad n = 3 \qquad \binom{10}{3} = 120$$

$$N = 20 \qquad n = 5 \qquad \binom{20}{5} = 15504$$

$$N = 100 \qquad n = 10 \qquad \binom{100}{10} = 17310309456440$$

$$N = 1.500.000 \quad n = 1000 \quad \binom{1.500.000}{1000} \approx 2,19 \cdot 10^{3608}$$

V zadnjem primeru, ki bi ustrezal približno temu, da bi izbirali vzorec velikosti 1000 iz populacije volilnih upravičencev v Sloveniji, je število vseh možnih vzorcev že nepredstavljivo veliko.

4.2.2 Vzorčna porazdelitev

V tem razdelku si zastavljamo vprašanje o zanesljivosti vzorčnih ocen. Na podlagi enot, ki jih izberemo v vzorec po nekem vzorčnem načrtu, izračunamo oceno za povprečje ali odstotek, ki nas zanima. Seveda ne moremo pričakovati, da bo ocena povprečja ali odstotka na podlagi podatkov iz vzorca točno enaka povprečju ali odstotku za celotno populacijo, saj je vzorec le neznaten del celotne populacije. Kot smo videli, je možnih vzorcev res veliko in pri dejanski izbiri bomo dobili pač enega od možnih vzorcev. Lahko si zamislimo, da bi pozabili na izbrani vzorec in izbirali novega še enkrat. Po vsej verjetnosti bi novi vzorec vseboval druge enote kot prejšnji in s tem bi se od prejšnje verjetno razlikovala tudi ocena povprečja ali odstotka. Seveda bi lahko ta postopek vsaj miselno ponavljali: vedno znova bi izbirali vzorec in vsakič izračunali nekoliko drugačno oceno. V realnosti si anketarji tega seveda ne morejo privoščiti, tukaj pa si bomo pomagali z računalniško simulacijo.

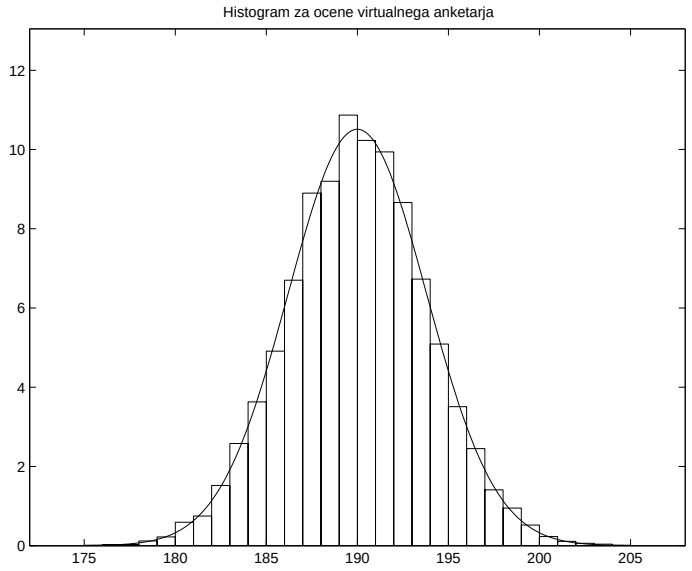
Za računalniško simulacijo si izberimo konkreten primer. Populacija naj bo velikosti $N = 1.000.000$, izbirali pa bomo vzorec velikosti $n = 1000$. V populaciji naj bo

povprečje spremenljivke enako 190 standardni odklon pa 120. Predstavljajmo si na primer, da gre za populacijo vseh zaposlenih, pri čemer je spremenljivka bruto plača. Nepoučenemu simuliranemu anketarju ali, kot se temu danes pravi, virtualnemu anketarju naročimo, naj izbere vzorec velikosti $n = 1000$. Potem ko anketar izbere prvi vzorec, lahko igrico seveda ponavljamo in simuliranega anketarja naženemo, da postopek izbiranja enostavnega slučajnega vzorca ponovi. Naš nevedni virtualni anketar potem vsakič pridno oceni povprečno plačo, tako da ugotovi izračuna povprčje spremenljivke za enote v trenutno izbranem vzorcu. Kaj si lahko od virtualnega anketarja in njegovih ocen obetamo? Iz veliko ponavljanj vzorčenja nam bo uspelo razbrati kaj o zanesljivosti ocen. Najprej si oglejmo nekaj prvih vzorčnih ocen, ki jih je dobil naš virtualni anketar.

188,36	183,68	190,48	191,09	185,65	194,52	194,51	189,86
191,24	190,66	189,29	192,75	187,77	198,28	189,48	190,43
194,05	190,22	189,64	186,84	191,12	184,93	192,71	196,16
187,37	193,26	194,76	183,95	184,53	192,17	188,48	192,62

Tabela 4.3: Ocene, ki jih je dobival virtualni anketar.

Kot vidimo, so anketarjeve ocene v splošnem blizu pravemu povprečju 190, le nekajkrat znaša razlika več kot 6. Razliki med oceno in pravim povprečjem pravimo *napaka vzorčne ocene*. To ime je upravičeno, saj pové, za koliko ocena na podlagi vzorca zgreši dejansko povprečje v populaciji. Iz samih številk je težko razbrati zakonitost v obnašanju napake, zato si pomagamo s histogrami, ki jih poznamo iz prvega poglavja. Slika 4.1 prikazuje histogram za zelo veliko ocen, ki jih je izračunal računalniški anketar, torej histogram za zelo veliko številk s seznama, katerega začetek je v tabeli 4.3. Na histogramu je razvidno, kakšne ocene za odstotek v populaciji lahko pričakujemo, ko izbiramo enostavni slučajni vzorec. Ploščina dela histograma nad intervalom med 185 in 195 nam recimo pove, v kolikšnem odstotku izbir vzorcev je marljivi računalniški anketar s svojo oceno zadel med 185 in 195, torej v kolikšnem odstotku primerov je bila napaka vzorčne ocene manjša od 5.



Sl. 4.1: Histogram za veliko število simuliranih vzorčnih ocen.

Zdaj že lahko začnemo presojati zanesljivost vzorčnih ocen na podlagi enostavnega slučajnega vzorca. S histograma vidimo, da je verjetnost napake, večje od 10, torej verjetnost da bi bila vzorčna ocena, ki bi jo izračunal anketar, manjša od 180 ali večja od 200, zelo majhna. To kaže ploščina histograma levo od vrednosti 180 in desno od vrednosti 200, ki je zelo majhna. S precejšnjo gotovostjo lahko trdimo, da bo napaka ocene iskanega odstotka na podlagi izbranega vzorca manjša od 10. To nam torej dá prvi občutek o zanesljivosti vzorčnih ocen.

Statistično ime za zgornji histogram je *vzorčna porazdelitev*. Porazdelitev zato, ker nam pove, kako so porazdeljene vzorčne ocene pri zelo velikem številu ponavljanj izbiranja vzorca. Oblika vzorčne porazdelitve je ključnega pomena pri presoji o zanesljivosti vzorčnih ocen. Kot je razvidno s slike 4.1 se vzorčni porazdelitvi dobro prilega ena od normalnih krivulj. Iz prvega poglavja vemo, da sta za opis histograma, ki se prilega normalni krivulji, potrebna le povprečje in standardni odklon. Kakšni sta ti dve količini tukaj, bomo povedali v naslednjem razdelku. Vzorčna porazdelitev bo pri enostavnem slučajnem vzorčenju vedno podobna eni od normalnih krivulj ne

glede na to, ali ocenjujemo povprečje neke spremenljivke v populaciji ali odstotek enot v populaciji z določeno lastnostjo. Trditev velja tudi za drugačne tipe vzorčenja, ne samo za enostavno slučajno vzorčenje. Več o tem nekoliko kasneje.

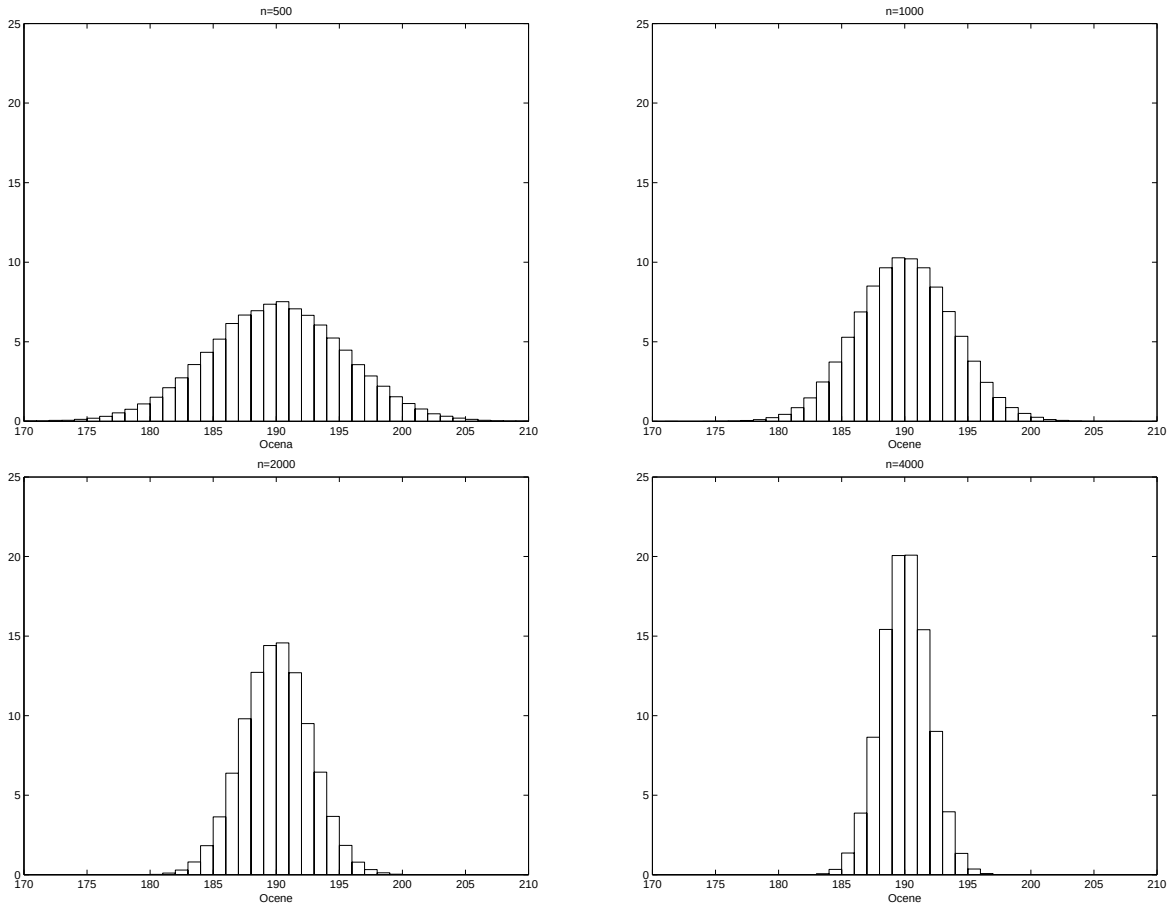
> ☞ Histogram za vzorčne ocene se za enostavno slučajno vzorčenje dobro prilega normalni krivulji s primernima parametroma. To velja ne glede na to, ali ocenjujemo populacijsko povprečje ali odstotek enot v populaciji z neko lastnostjo. To dejstvo je ključnega pomena za raziskovanje zanesljivosti vzorčnih ocen.

4.2.3 Standardna napaka ocene

Kako nam oblika vzorčne porazdelitve pomaga pri raziskovanju zanesljivosti vzorčnih ocen? Ugotovili smo, da se je naš virtualni anketar le malokrat zmotil za več kot za 10, pa vendar se je v mnogo ponavljanjih to nekajkrat zgodilo. Tako ne moremo reči, da se anketar nikoli ne bo zmotil za več kot za 10. Rečemo lahko le, da je verjetnost napake, ki bi bila večja od 10, zelo majhna. Torej lahko govorimo le o tem, s kolikšno verjetnostjo bomo z vzorčno oceno "zadeli" blizu prave, vendar neznane, količine v populaciji. Za enostavno slučajno vzorčenje, ki ga je izvajal virtualni anketar, bomo v nadaljevanju izračunali, da je verjetnost pomote pri ocenjevanju, ki bi bila večja od 10, manjša od 1%. V stotih ponovitvah izbire vzorca in računanja ocene bi torej lahko pričakovali, da se bomo le enkrat zmotili za več kot za 10.

Pri ugotavljanju napake vzorčne ocene nam pomaga mera razpršenosti ocen v vzorčni porazdelitvi. Mere razpršenosti podatkov smo obravnavali v prvem poglavju in jih lahko uporabimo tudi za vzorčne porazdelitve, kot na primer na sliki 4.1. Za histograme, ki so podobni normalni krivulji, je posebej primerna mera razpršenosti standardni odklon. Toda kako izračunati standardni odklon za vzorčno porazdelitev? Za zdaj lahko rečemo le, da bo za večje vzorce standardni odklon manjši. To bi pričakovali že po zdravi pameti, saj bo z večjim vzorcem ocena verjetno bolj zanesljiva. Da bi dobili nekaj občutka, si predstavljajmo, da bi virtualni anketar ponovil veliko število izbiranj vzorcev, ki bi bili različnih velikosti. Izberimo si po vrsti velikosti $n = 500$,

$n = 1000$, $n = 2000$ in $n = 4000$. Na sliki 4.2 so vzorčne porazdelitve za izbrane velikosti vzorca. Virtualni anketar je izbiral iz populacije velikosti $N = 1.000.000$ vzorec dane velikosti n in ocenjeval povprečje. Po pričakovanju je za večje velikosti n porazdelitev ožja, torej z večjim vzorcem lahko pričakujemo večjo zanesljivost ocen.



Sl. 4.2: Vzorčne porazdelitve za $n = 500$, $n = 1000$, $n = 2000$ in $n = 4000$.

Standardni odklon vzorčne porazdelitve imenujemo *standardna napaka*. Označili

jo bomo s SE, kar pride iz angleškega izraza *standard error*. Ta količina je v jeziku statistike merilo za natančnost vzorčne ocene. Kako to merilo uporabljamo, bomo videli v primerih. Izračun standardnega odklona vzorčne porazdelitve ni vedno enostaven. Obravnavali bomo primera, ko na podlagi vzorca ocenjujemo populacijsko povprečje ali odstotek. Formuli za standardni napaki teh dveh ocen si bomo sposodili iz teoretične statistike.

☞ Če ocenjujemo povprečje vrednosti spremenljivk za celotno populacijo na podlagi enostavnega slučajnega vzorca, je standardna napaka

$$SE = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\,.$$

Pri tem je N velikost populacije, n velikost vzorca in σ standardni odklon vrednosti spremenljivke za celotno populacijo.

Kvadratni koren $\sqrt{\frac{N-n}{N-1}}$ imenujemo popravni faktor. Razlogi za tako ime so skriti v matematični izpeljavi formule, ker moramo "popraviti" dejstvo, da vsako enoto v vzorec izberemo kvečjemu enkrat. Na srečo je v večini primerov popravni faktor tako blizu 1, da ga lahko zanemarimo. Kot grobo pravilo velja, da popravni faktor lahko zanemarimo, če je velikost vzorca manjša od 10% velikosti populacije, kar brž vidimo, če izračunamo to količino za nekaj primerov N in n.

$$N = 10 \quad n = 3 \quad \sqrt{\frac{N-n}{N-1}} = 0,88$$

$$N = 20 \quad n = 5 \quad \sqrt{\frac{N-n}{N-1}} = 0,89$$

© Visoka šola za podjetništvo Piran

$N = 100 \qquad n = 10 \qquad \sqrt{\frac{N-n}{N-1}} = 0,95$

$N = 1.500.000 \quad n = 1000 \quad \sqrt{\frac{N-n}{N-1}} = 0,9997$

Primer: V razdelku 4.2.2 smo spremljali, kaj se je dogajalo z vzorčnimi ocenami virtualnega anketarja, ko je ponavljal izbiranje vzorca. Histogram ocen, ki jih je anketar izračunal iz posameznih vzorcev, smo narisali na sliki 4.1. Zdaj lahko povemo nekaj več. Vemo, da je bilo povprečje 190 in standardni odklon 120. Po formuli v okvirčku izračunamo ($n = 1000$ in $N = 1.000.000$)

$$SE = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = 3,79\,.$$

Standardna napaka je torej 3,79. Ko že poznamo SE, lahko začnemo govoriti tudi o verjetnostih, da za ocene dobimo takšne ali drugačne vrednosti. Iz prvega poglavja vemo, da je ploščina med $190 - 3,79$ in $190 + 3,79$ pod normalno krivuljo, ki se tesno prilega vzorčni porazdelitvi, enaka približno 68%. Ravno v takem odstotku primerov vzorcev je virtualni anketar tudi dejansko dobil oceno, ki se je od 190 razlikovala manj kot za $3,79$. Nadalje iz tabele za normalno porazdelitev razberemo, da je ploščina med $-1,96$ in $1,96$ enaka 95% vse ploščine pod normalno krivuljo. Lahko torej trdimo, da se bo anketarjeva ocena razlikovala od prave vrednosti za manj kot $1,96 \cdot SE = 7,44$ z verjetnostjo 95%. S podobnim razmislekom bi ugotovili, da se z verjetnostjo samo 1% lahko zmotimo za več kot $9,71$, ker iz tabele za normalno porazdelitev razberemo, da moramo standardno napako množiti s približno 2,56.

Primer: Med uvodnimi primeri smo obravnavali raziskavo TIMSS. V Sloveniji je bilo v vzorec izbranih $n = 5927$ učencev, ki so reševali na prezkusu znanja reševali naloge iz matematike in naravoslovja. Povprečje njihovih dosežkov pri naravoslovju je bilo $547,8$. Povprečje znanja naravoslovja vseh učencev v Sloveniji ocenimo s $547,8$. Zanima nas, kako zanesljiva utegne biti ta ocena. Zamislimo si spet za trenutek, da je

bilo vzorčenje enostavno slučajno. Za izračun standardne napake potrebujemo standardni odklon vzorčnih podatkov, ki ga bomo označili s $\hat{\sigma}$. S pomočjo računalniškega programa dobimo $\hat{\sigma} = 84,2$ in nato po zgornji formuli

$$SE = \frac{84,2}{\sqrt{5927}}\sqrt{\frac{53655-5927}{53655-1}} = 1,03\,.$$

Kot vidimo, je SE dokaj majhna, torej smo z vzorcem že zelo natančno določili povprečje. Z verjetnostjo le 1% smo se zmotili za več kot $2,55 \cdot 1,03 = 2,6$ točke, napaka, večja od 4 točk, pa je praktično nemogoča.

Čeprav smo si razmišljanje poenostavili s privzetkom, da je bilo vzorčenje enostavno slučajno, velja podoben razmislek tudi brez te predpostavke, le da je treba za izračun standardne napake uporabiti matematično bolj zahtevne formule. Interpretacija standardne napake pa je tudi v tem primeru enaka. Kot zanimivost naj povemo, da je bila standardna napaka za povprečje znanja naravoslovja ocenjena s približno 2,5. Standardna napaka je večja zaradi drugačnega vzorčnega načrta, kot je enostavno slučajno vzorčenje.

Pozoren bralec je morda že opazil, da za izračun SE potrebujemo pravi standardni odklon v populaciji. Kako torej ravnamo v bolj realistični situaciji, ko nimamo na voljo populacijskega standardnega odklona, ki smo ga v formuli za SE označili s σ? Kaj storiti? Tudi σ lahko ocenimo na podlagi izbranega vzorca tako, da izračunamo standardni odklon vrednost spremenljivke za izbrane enote kot v zgornjem primeru. Ta vzorčni standardni odklon označimo z $\hat{\sigma}$. Strešica nad črko σ v statistiki vedno označuje ocene količin na podlagi vzorca. V zgornjem primeru je bila $\hat{\sigma} = 84,2$. Oznako izgovarjamo kot "sigma-strešica".

Obravnavajmo še primer, ko na podlagi vzorca ocenjujemo odstotek enot v populaciji z dano lastnostjo. Standardno napako ocenimo po formuli v spodnjem okvirčku. Formule sledijo na podlagi izračuna standardne napake vsote v tretjem poglavju.

Če ocenjujemo povprečje vrednosti spremenljivke za celotno populacijo na podlagi enostavnega slučajnega vzorca, potem za standardno napako vzamemo

$$SE = \frac{\hat{\sigma}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\,.$$

Pri tem je N velikost populacije, n velikost vzorca in $\hat{\sigma}$ standardni odklon vrednosti spremenljivke v vzorcu.

Če ocenjujemo odstotek enot v populaciji z dano lastnostjo na podlagi enostavnega slučajnega vzorca, je

$$SE = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot 100\%\,.$$

Pri tem je N velikost populacije, n velikost vzorca in p dejanski delež enot z dano lastnostjo v celotni populaciji, torej $p \cdot 100\%$ dejanski odstotek.

Primer: Velikost vzorca, ki so ga pred plebiscitom leta 1990 izbrali anketarji SJM 90, je bila 2074. Privzemimo za trenutek, da je bilo vzorčenje enostavno slučajno, čeprav vemo, da je bil v resnici vzorčni načrt bolj zapleten. Med anketiranci se jih je za odcepitev in samostojnost izreklo 1306, kar je 63%. Odstotek volivcev, ki so podpirali tako samostojnost kot odcepitev, bi torej ocenili s 63%. Pretirani skeptiki bi lahko še vedno trdili, da so te vrednosti le vzorčne in niso nujno pravilne za vso populacijo volivcev. To je seveda res. Vendar ni vse izgubljeno, saj s pomočjo standardne napake ugotovimo verjetnost, da je vzorčna ocena napačna za toliko in toliko. Za izračun standardne napake bi morali poznati p, tega pa šele ocenjujemo! Rešitev je v tem, da

namesto p uporabimo kar ocenjeni delež na podlagi vzorčnih podatkov, kot da bi bil pravi. Standarno napako torej računamo v tem primeru po formuli

$$SE = \frac{\sqrt{0,63 \cdot 0,37}}{\sqrt{2074}} \cdot 100\% = 1\% \,.$$

Popravni faktor smo izpustili, ker je zelo blizu 1. Statistiki znajo pokazati, da se s tem ne pregrešimo preveč. Tako izračunani SE je večinoma zelo blizu pravi vrednosti standardne napake in nam omogoči zanesljivo presojo o natančnosti ocene.

Iz tabele za normalno porazdelitev lahko zdaj razberemo, da je verjetnost, da se pri ocenjevanju odstotka na podlagi vzorca v našem primeru zmotimo za več kot 4%, praktično 0. Če bi kdo še dvomil o uspehu plebiscita, bi tako moral verjeti, da se lahko zgodijo praktično nemogoči dogodki.

Kot zadnjo pripombo k temu primeru dodajmo še, da vzorec pri SJM 90 ni bil enostavni slučajni, vendar razmislek s standardnimi napakami kljub temu velja, pri čemer moramo za izračun SE uporabiti spremenjene, matematično zahtevnejše formule. Dejanska standardna napaka je bila 1,5%, torej je bila vzorčna ocena dovolj zanesljiva za presojo o tem, ali bi za samostojnost in odcepitev glasovalo več kot 50% volivcev ali ne. Dvom o uspehu plebiscita je bil v tem primeru izključen.

Tudi v tem primeru smo v formuli za SE uporabili kar ocenjeni delež $\hat{p}$ namesto neznanega deleža p. S strešico v oznaki $\hat{p}$ statistiki povedo, da gre za vrednost, ki je bila ocenjena na podlagi vzorca. Dejansko tudi $\hat{p}$ izgovarjamo kot p-strešica.

Povzemimo še osnovno ugotovitev o standardni napaki, ko ocenjujemo odstotek enot v populaciji z določno lastnostjo, recimo odstotek volivcev, ki podpirajo neko stranko.

Če ocenjujemo odstotek enot v populaciji z dano lastnostjo na podlagi enostavnega slučajnega vzorca, za standardno napako vzamemo

$$SE = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot 100\% \,.$$

Pri tem je N velikost populacije, n velikost vzorca in $\hat{p}$ ocenjeni delež enot z dano lastnostjo v celotni populaciji, torej $\hat{p} \cdot 100\%$ ocenjeni odstotek.

Primer: Od 13. stoletja je težo zlatnikov iz Kraljevske kovnice Anglije kontrolirala posebna komisija s proceduro, ki se je imenovala *The Trial of the Pyx*. Naključno so izbrali 1000 gvinej (zlatnikov), od katerih naj bi po predpisih vsak tehtal 128 zrn. Teh 1000 zlatnikov so stehtali in skupna teža se od 128.000 zrn ni smela razlikovati za več kot 640 zrn. Če se to ni zgodilo, je bil mojster kovnice, ali kot bi danes rekli, guverner centralne banke, hudo kaznovan.

Recimo, da je bilo leta 1799 mogoče s takratno tehnologijo kovati zlatnike z natančnostjo 128/200 zrna. Standardni odklon teže vseh kovancev, ki so jih nakovali v kovnici, je bil torej omenjena količina. Recimo, da je mojster kovnice pošten in kuje kovance, ki so v povprečju težki 128 zrn. Kraljevska komisija je izbrala enostavni slučajni vzorec 1000 gvinej iz populacije vseh gvinej, nakovanih leta 1799, in jih stehtala. Kolikšna je verjetnost, da bo mojster kovnice obtožen nepoštenosti?

Omejitev 640 zrn, ki jo je postavila kraljevska komisija, lahko opišemo tudi s povprečji. Povprečje 1000 vzorčenih gvinej se od 128 zrn ni smelo razlikovati za več kot 0,64 zrna. Iz podatkov za tehnologijo kovanja lahko izračunamo standardno napako za oceno povprečne teže zlatnika:

$$SE = \frac{0,64}{\sqrt{1000}} = 0,02 \,.$$

Kot vidimo, je omejitev 0,64 enako 32 standardnih napak! Če je bil mojster pošten, so bile meje tako varne, da je bila obtožba po krivem praktično nemogoča.

Predpostavimo, da bi bil mojster kovnice nepošten in bi koval zlatnike, ki v povprečju tehtajo 127 zrn in imajo standardni odklon 128/200 zrna. Kolikšna je zdaj verjetnost, da bo mojster kovnice obtožen nepoštenosti? S standardnim odklonom, ki ga predpostavljamo za skovane zlatnike, smo lahko popolnoma prepričani, da bo kraljevska komisija ujela nepoštenega mojstra. Verjetnost, da bi bila ocena povprečne teže zlatnika v tem primeru znotraj predpisanih mej, je seveda praktično 0, saj bi se ocena od 127 zrn morala razlikovati kar za približno 18 standardnih napak, kar pa je nemogoče!

Primer: Leta 1965 je ameriško vrhovno sodišče zavrnilo pritožbo obsojenca A. Swaina, ki ga je sodišče v Talladega County, Alabama, obsodilo na smrt zaradi posilstva bele ženske (registracija: 380 US 202, 13 L ed 2nd 759, 85 S Ct 824). A. Swain je bil črnec.

Odvetniki A. Swaina so se pritožili na vrhovno sodišče, češ da je bila izbira porote pristranska. Pri kazenskih postopkih sodišče najprej izbere "panel" 100 potencialnih porotnikov, izmed katerih po dolgotrajnem in zapletenem postopku izberejo 12 porotnikov. Tukaj nas zanima samo prvotni panel. Pritožba na vrhovno sodišče je namreč vsebovala podatek, da nihče od še živečih v Talladega County ne pomni, da bi bil črnec porotnik v kakšnem kazenskem ali civilnem procesu, vključno s procesom, o katerem govorimo. Med 100 potencialnimi porotniki je bilo samo 8 črncev, čeprav naj bi bila izbira "slepa", v Talladega County pa je črnega prebivalstva 26%. Pričakovali bi torej, vsaj pri slepi izbiri, da bi bilo med potencialnimi porotniki približno 26 črncev.

Vrhovno sodišče je pritožbo zavrnilo z naslednjo statistično utemeljitvijo: poroto izberejo na sodišču med 100 "slučajno izbranimi" moškimi nad 21 let. Med kandidati za porotnike je bilo 8 črncev, kar ne kaže na namerno pristrano izbiro potencialnih porotnikov. Kaj naj si mislimo?

Oglejmo si utemeljitev nekoliko podrobneje. Privzemimo, da "slepa" izbira pomeni enostavno slučajno vzorčenje. Panel 100 porotnikov je bil torej enostavni slučajni vzorec velikosti $n = 100$ iz populacije vseh moških nad 21 let, ki jih je bilo približno $N = 16.000$. Ker poznamo delež $p = 0,26$, lahko brž izračunamo SE:

$$SE = \frac{\sqrt{p(1-p)}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \cdot 100\% = 4,37\%\,.$$

© Visoka šola za podjetništvo Piran

Če si predstavljamo, da bi na podlagi enostavnega slučajnega vzorca velikosti 100 ocenjevali odstotek črnih moških v populaciji odraslih moških v Talladega County, bi se torej z zelo majhno verjetnostjo zmotili za več kot $2,55 \cdot 4,37\% = 11,1\%$. Torej je praktično nemogoče, da bi pri enostavnem slučajnem vzorčenju v vzorec dobili samo 8 črncev. Vzorčna ocena odstotka bi bila le preveč napačna, da bi lahko sploh verjeli predpostavki, da je šlo za "slepo" izbiro. Utemeljitev vrhovnega sodišča je bila torej, vsaj statistično, na zelo trhlih nogah.

4.3 Intervali zaupanja

Kot smo videli v prejšnjem razdelku, ne moremo pričakovati, da bi bila ocena odstotka ali povprečja na podlagi enostavnega slučajnega vzorca povsem točna, je le približek. Standardna napaka nam pove, kolikšna utegne biti razlika med dejanskim odstotkom ali povprečjem in oceno tega odstotka ali povprečja, ki jo dobimo na podlagi vzorca.

Intervali zaupanja so le drugačen način izražanja o standardnih napakah. Ideja je preprosta: oceno, ki jo dobimo na podlagi vzorca, "napihnemo" v interval, in sicer tako, da bo ta interval pokril pravo vrednost, ki jo ocenjujemo z vnaprej predpisano verjetnostjo. Če bi želeli, da bi interval zaupanja z gotovostjo pokrival pravo vrednost, bi moral biti zelo širok. To ni praktično, zato se sprijaznimo z možnostjo, da interval zaupanja ne bo pokrival prave vrednosti, predpišemo pa *verjetnost*, s katero mora interval zaupanja pokrivati pravo vrednost. Oglejmo si primer.

Primer: Vrnimo se še enkrat k virtualnemu anketarju, ki nas spremlja že ves čas. Predstavljajmo si, da anketar vsakič, ko izbere vzorec in oceni iskani odstotek, "napihne" oceno v interval tako, da gre od dobljene ocene na levo in na desno za $1,96 \cdot SE$, kjer SE izračuna po formuli iz prejšnjega razdelka. Če torej dobi za odstotek oceno 31,2%, je $SE = 1,47\%$ in je torej spodnja meja intervala $31,2\% - 2,87\%$, zgornja meja pa $31,2\% + 2,87\%$. Zakaj smo izbrali ravno faktor 1,96? Na sliki 4.3 si oglejmo 100 intervalov zaupanja, ki jih je na opisani način dobil virtualni anketar. Nekateri izmed teh 100 intervalov pokrivajo pravo vrednost odstotka, ki je 30%, nekateri pa ne. Če preštejemo, ugotovimo, da interval zaupanja ne pokriva vrednosti 30% v 5 primerih od 100. Če bi postopek še nadaljevali, bi ugotovili, da pri velikem številu ponavljanj

ravno 5% intervalov zaupanja ne pokrije vrednosti 30%. Zdaj si lahko pojasnimo, zakaj smo izbrali faktor 1,96 in z njim množili SE. Interval zaupanja pokrije vrednost 30% natanko tedaj, ko se pri ocenjevanju nismo zmotili več kot za $1,96 \cdot SE$, to pa se zgodi, kot vemo iz prejšnjega razdelka, z verjetnostjo 95%. Če bi torej SE množili s faktorjem 2,55, bi dobili intervale zaupanja, ki bi pravo vrednost pokrili z verjetnostjo 99%.

Verjetnost, da interval zaupanja ne pokrije prave vrednosti, si, kot rečeno, lahko izberemo vnaprej. Udomačen izraz za to izbrano verjetnost je *stopnja tveganja*, ki jo pogosto označimo z grško črko α, in govorimo o intervalih zaupanja pri dani stopnji tveganja. Pogosto si namesto verjetnosti, da interval zaupanja *ne* pokriva prave vrednosti, izberemo verjetnost, da *jo* pokriva. Potem govorimo o *stopnji zaupanja* namesto o stopnji tveganja.

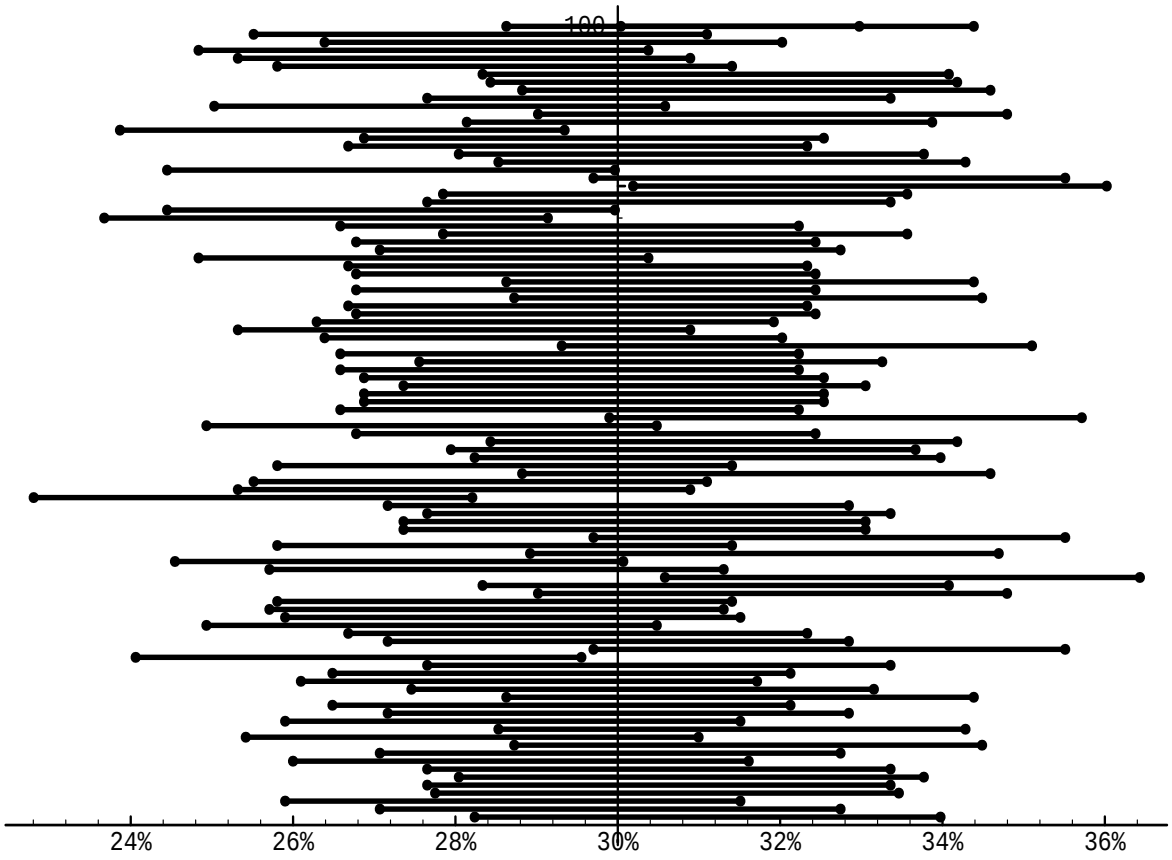
☞ Pri dani stopnji tveganja α dobimo spodnjo mejo intervala zaupanja tako, da od dobljene ocene (odstotka ali povprečja) odštejemo $z_\alpha \cdot SE$, kjer je z_α tako število, da je ploščina pod standardno normalno krivuljo med $-z_\alpha$ in z_α enaka $1-\alpha$. Večinoma bo $\alpha = 5\%$ in torej $z_\alpha = 1,96$, ali $\alpha = 1\%$ in $z_\alpha = 2,55$. Zgornjo mejo intervala zaupanja dobimo tako, da produkt $z_\alpha \cdot SE$ prištejemo. Smisel intervalov zaupanja je na drugačen način opisati zanesljivost vzorčnih ocen.

Opomba: Kot v tretjem poglavju tudi tukaj uporabljamo črko z namesto s za standardne enote.

Primer: V časopisju večkrat zasledimo rezultate anket, ki navajajo tudi natančnost ocene, vendar mnogokrat pomanjkljivo. Primer takega nepopolnega poročanja je naslednji:

```
V nedavni anketi so ocenili, da je povprečna doba šolanja
12,4 leta.  Ocena je natančna na ±0,7 leta.
```


© Visoka šola za podjetništvo Piran



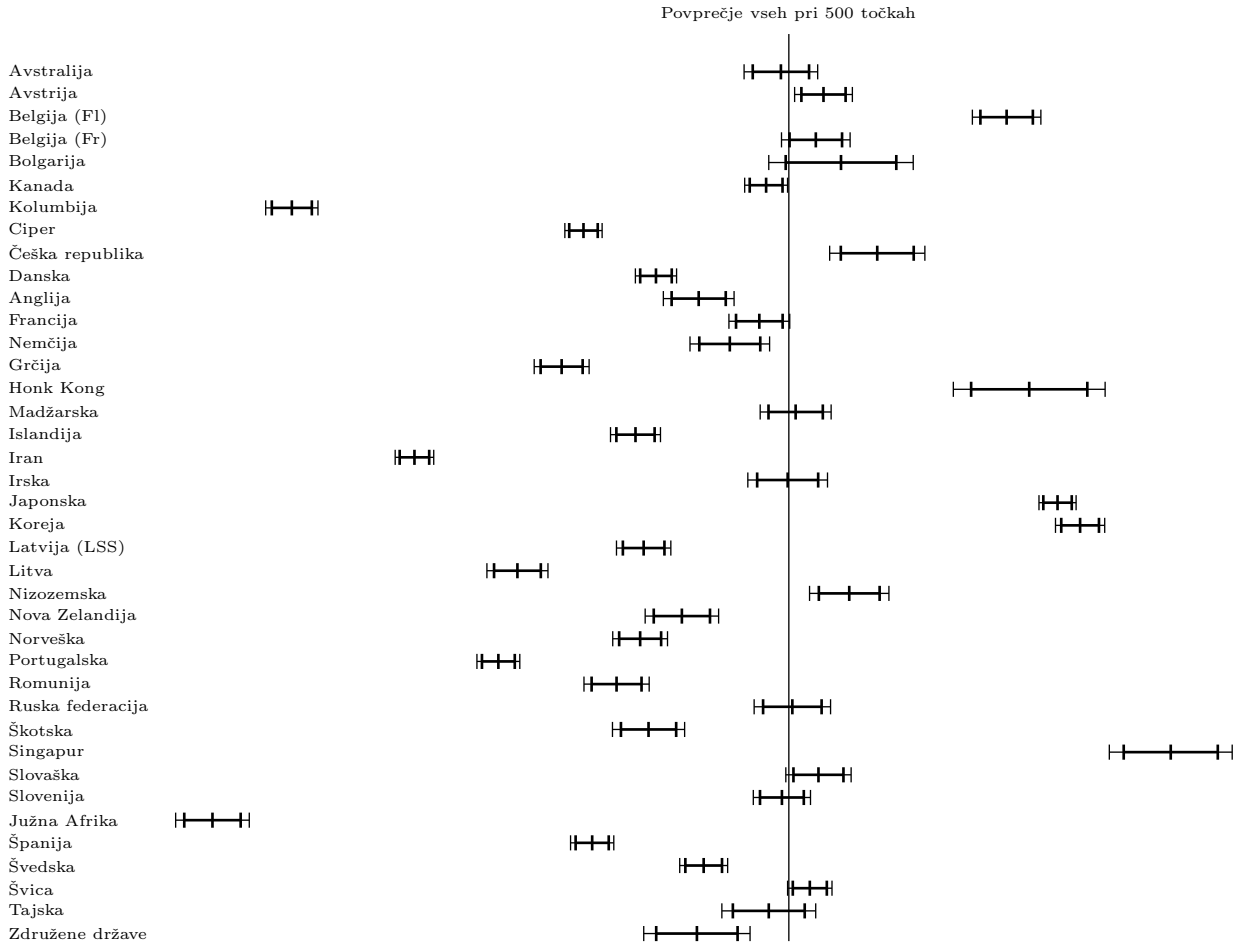
Sl. 4.3: 100 intervalov zaupanja virtualnega anketarja

Kaj pravzaprav pomeni $\pm 0,7$ leta? Naj to razumemo kot standardno napako? Ali naj razumemo, da je interval od 11,7 do 13,1 leta interval zaupanja pri stopnji tveganja $\alpha = 5\%$? Pravilna formulacija bi seveda morala vsebovati bolj natančen opis, kaj navedene količine pomenijo. Pripomnimo, da pri drugačnih vrstah vzorčenja SE dobimo na drugačen način, kar smo omenili že v prejšnjih primerih, intervale zaupanja pa še vedno izračunamo tako kot zgoraj in tudi interpretacija še vedno drži.

Primer: Že večkrat smo govorili o raziskavi TIMSS. Na sliki 4.4 [4] sta za 39 držav, ki so sodelovale, narisana interval zaupanja za oceno povprečja dosežkov pri matematiki za 8. razred pri stopnjah tveganja $\alpha = 0,05$ (krajši) in $\alpha = 0,01$ (daljši). Kot je videti s slike, nam intervali zaupanja takoj posredujejo informacijo, s kakšno zanesljivostjo so bila ocenjena povprečja v posameznih državah. Vemo pa še več! Vemo, da krajši intervali z verjetnostjo 95% pokrivajo pravo povprečje, tisti daljši pa z verjetnostjo 99%. Poleg tega nam intervali zaupanja omogočajo, da že na prvi pogled vidimo, ali lahko rečemo, da je bila neka država res dejansko boljša od druge. Zaradi svoje vizualne privlačnosti in preglednosti so intevali zaupanja postali standardno sredstvo podajanja zanesljivosti ocen pri vzorčenju.

[4] Vir: TIMSS mednarodno poročilo.

© Visoka šola za podjetništvo Piran



Sl. 4.4: Intervali zaupanja za ocene povprečja dosežkov pri matematiki za posamezne države pri $\alpha = 0,05$ in $\alpha = 0,01$.

1. Rezultati 1255 srednješolcev na sprejemnem izpitu za visoko šolo na EF leta 1991 so bili približno normalno porazdeljeni s povprečjem 25,5 točke in standardnim odklonom 12,1 točke.

   a. Približno kolikšen odstotek srednješolcev je doseglo od 26 do 30 točk?

   b. Recimo, da bi ocenjevali povprečje v tej populaciji z $N = 1255$ enotami na podlagi enostavnega slučajnega vzorca velikosti $n = 100$. Kolikšna bi bila verjetnost, da bi bila ocena višja od 25,15? Utemeljite odgovor!

Rešitev:

a. *Dani meji za število točk spremenimo v standardne enote, da lahko izračunamo ploščino pod standardno normalno krivuljo.*

$$\frac{26-25,5}{12,1} = 0,04 \qquad \textit{Ploščina na levo je } 51,6\%.$$
$$\frac{30-25,5}{12,1} = 0,37 \qquad \textit{Ploščina na levo je } 64,5\%.$$

Razlika teh dveh ploščin, $12,9\%$*, nam da odstotek srednješolcev, ki so imeli na sprejemnem izpitu od* 26 *do* 30 *točk.*

b. *Vse možne ocene za povprečje v tej populaciji na podlagi vzorca s* 100 *enotami se porazdeljujejo po vzorčni porazdelitvi, ki se prilega normalni krivulji, s povprečjem* $\mu = 25,5$ *in standardnim odklonom*

$$SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{12,1}{\sqrt{100}}\sqrt{\frac{1255-100}{1255-1}} = 1,16\,.$$

Verjetnost, da bo ocena višja od $25,15$*, je enaka ploščini desno od te vrednosti. V standardnih enotah je ta vrednost* $-0,30$ *in ploščina na desno od te standardne enote je* $100\% - 38,2\% = 61,8\%$.

 © Visoka šola za podjetništvo Piran

2. Oblika porazdelitve bruto osebnega dohodka zaposlenih v Sloveniji za leto 1995 je prikazana na sliki 1.2.

   a. Recimo, da bi takrat izbrali enostavni slučajni vzorec velikosti 225 zaposlenih in ugotovili, da je vzorčno povprečje 108.000 SIT in vzorčna ocena standardnega odklona 15.000 SIT. Ali bi lahko trdili, da 16% zaposlenih zasluži več od 123.000 SIT? Utemeljite odgovor!

   b. Ali bi lahko pri istih podatkih kot v točki a. trdili, da je imelo 68% zaposlenih bruto osebni dohodek med 107.000 SIT in 109.000 SIT? Utemeljite odgovor! Popravni faktor zanemarite!

   c. Kolikšna je približno verjetnost, da bo vzorčna ocena bruto osebnega dohodka na podlagi vzorca velikosti $n = 225$ previsoka za več kot 1000 SIT? Popravni faktor zanemarite!

   Rešitev:

   a. Ne. Porazdelitev bruto osebnega dohodka ni podobna normalni porazdelitvi, zato tega ne moremo trditi. Odstotek zaposlenih z osebnimi dohodki, višjimi od 123.000 *SIT, bi ugotovili, če bi lahko izračunali odstotek ploščine na histogramu na desno od vrednosti* 123.000 *SIT.*

   b. Ne. Utemeljitev je enaka kot v a.

   c. Histogram za vzorčne ocene povprečnega bruto osebnega dohodka se se prilega normalni porazdelitvi s parametroma $\mu = 108.000$ *SIT in* $SE = \frac{\sigma}{\sqrt{n}} = 1000$ *SIT. Verjetnost, da bo vzorčna ocena previsoka za več kot* 1000 *SIT, je enaka ploščini pod normalno krivuljo na desno od vrednosti* 109.000 *SIT. Ko te vrednosti spremenimo v standardne enote, dobimo, da je ta ploščina enaka* 16%.

3. V časopisju večkrat zasledimo rezultate anket, ki navajajo tudi zanesljivost ocene, mnogokrat pomanjkljivo. Primer takega nepopolnega poročanja je naslednji:

`V nedavni anketi so ocenili, da je povprečna doba šolanja 12 let. Ocena je natančna na ±0,7 leta.`

Privzemite, da so oceno dobili na osnovi enostavnega slučajnega vzorca velikosti 1500, in odgovorite na naslednja vprašanja:

a. Ali naj zgornjo izjavo razumemo tako, da je tisto pravo povprečje med 11,3 leta in 12,7 leta?
b. Ali v stavku o natančnosti ocene kaj manjka?
c. Kakšna bi bila pravilna formulacija navedbe zanesljivosti ocene?
d. Glede na vaš odgovor v točki c., kolikšna bi torej bila verjetnost, da se pri ocenjevanju povprečne dobe šolanja zmotimo za manj kot za 0,7 leta?

Rešitev:

a. Ne. Pravo povprečje je še vedno neznano, lahko pa govorimo o verjetnostih, da določeni intervali pokrivajo pravo povprečje.
b. Da. Manjka opis, kaj natančnost na $\pm 0,7$ *leta pomeni; ali je to standardna napaka ali kakšna druga količina, s katero lahko merimo napako vzorčne ocene.*
c. Možnosti za pravilno formulacijo je več. Lahko rečemo, da je standardna napaka ocene $0,7$ *leta.*
d. Če vzamemo, da je $SE = 0,7$*, je iskana verjetnost* 68%.

4. Recimo, da je nekdo 10-krat izbral enostavni slučajni vzorec velikosti $n = 400$ iz neke populacije, v kateri je 38% dobrih in 62% slabih enot. Vsakič je ocenil delež p dobrih enot z vzorčnim deležem, torej, če je bilo, recimo, v vzorcu velikosti $n = 400$ dobrih enot 156, je ocenil delež p z vzorčnim deležem $\hat{p} = 0,39$. Velikost populacije je, recimo, $N = 1.500.000$. Za katero od spodnjih zaporedij ocen za p mislite, da je tisto, ki ga je ta "nekdo" dobil? Podajte kratko utemeljitev!

 © Visoka šola za podjetništvo Piran

a. 0,375, 0,3925, 0,3675, 0,3625, 0,3875, 0,375, 0,37, 0,385, 0,36, 0,39

b. 0,3696, 0,3400, 0,3830, 0,3869, 0,3525, 0,4086, 0,4085, 0,3791, 0,3879, 0,3842

c. 0,4075, 0,375, 0,4375, 0,4025, 0,47, 0,415, 0,4025, 0,4275, 0,405, 0,4475

Rešitev: Najprej izračunamo standardno napako.

$$SE = \sqrt{\frac{p(1-p)}{n}} \cdot 100\% = 2,4\%\,.$$

Zaporedje c, odpade, ker je 9 *ocen od* 10 *večjih od pravega povprečja, nekatere celo za* 3 SE. *Med drugima dvema možnostima se odločimo glede na standardni odklon, ki bi moral biti blizu tistemu teoretičnemu, torej* SE. 10 *ocen iz b. ima standardni odklon* $2,06\%$, 10 *ocen iz a. pa* $41,11\%$. *Odgovor je torej b.*

5. Kot populacijo izberemo vse davčne formularje v RS leta 1991, spremenljivka pa je znesek, ki po formularju pripada državi (lahko tudi negativen). Spremenljivka NI normalno porazdeljena. Recimo, da izberemo enostavni slučajni vzorec 400 formularjev. Katera od spodnjih dveh izjav je pravilna?

   a. Verjetnost, da se vzorčno povprečje razlikuje od pravega povprečja za več kot 100 SIT, izračunamo s pomočjo SE in normalne krivulje.

   b. Verjetnosti, da se vzorčno povprečje razlikuje od pravega povprečja za več kot 100 SIT, ne moremo izračunati s pomočjo SE in s pomočjo normalne krivulje, ker spremenljivka ni normalno porazdeljena.

   Obkrožite eno od možnosti in utemeljite izbiro.

   Rešitev: Pravilna je izjava a. Vzorčna povprečja na podlagi vseh možnih vzorcev so porazdeljena po normalni porazdelitvi s parametroma μ *(pravo povprečje v populaciji) in* $SE = \frac{\sigma}{\sqrt{n}}$, *kjer je* σ *standardni odklon zneskov v populaciji.*

6. Na podlagi enostavnega slučajnega vzorca velikosti $n = 346$ ocenjujemo odstotek prebivalstva, ki je naročen na neki časopis. Od oseb v vzorcu je 248 naročnikov.

   a. Pri stopnji tveganja $\alpha = 0,05$ poiščite zgornjo in spodnjo mejo zaupanja za odstotek naročnikov.

   b. Ali lahko z gotovostjo trdite, da je dejanski odstotek naročnikov večji od 66,9%. Obrazložite odgovor!

   c. Kolikšen bi moral biti vzorec, da bi lahko pri stopnji tveganja $\alpha = 0,32$ trdili, da je napaka, ki smo jo zagrešili pri ocenjevanju, manjša od 2,4%?

   Rešitev:

   a. *Meji zaupanja pri stopnji tveganja* $\alpha = 0,05$ *dobimo po formuli* $\hat{p} - 1,96 \cdot SE$ *in* $\hat{p} + 1,96 \cdot SE$*, kjer je* $\hat{p} = \frac{248}{346} = 72\%$ *in* $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot 100\% = 2,4\%$*. Meji zaupanja za pravi odstotek naročnikov v populaciji sta torej* $72\% - 1,96 \cdot 2,4\%$ *in* $72\% + 1,96 \cdot 2,4\%$*. Pri* $\alpha = 0,05$ *ta interval z verjetnostjo 95% pokriva pravi odstotek naročnikov.*

   b. *Ne. Verjetnost, da je dejanski odstotek naročnikov večji od* $66,9\%$*, je manjša od 1.*

   c. *Ravno pri vzorcu* $n = 346$ *je* $SE = 2,4\%$*, torej je pri stopnji tveganja* $\alpha = 0,32$ *napaka vzorčne ocene manjša od* $2,4\%$*.*

7. Iz pošiljke 4000 elektronskih komponent izberemo enostavni slučajni vzorec 225 komponent, da bi ugotovili, ali prenesejo napetost 100V, kot trdi proizvajalec. Pri preizkušanju je 205 komponent preneslo predpisano napetost. Bi pri $\alpha = 0,05$ verjeli, da je odstotek dobrih komponent večji od 90%?

   Rešitev: Izjavi bomo verjeli, če interval zaupanja pri $\alpha = 0,05$ *ne bo pokrival vrednosti* 90%*. Ocena za odstotek dobrih komponent je* $\hat{p} = \frac{205}{225} = 91\%$ *in standardna napaka te ocene (brez popravnega faktorja) je* $SE = 1,9\%$*. Spodnja meja*

© Visoka šola za podjetništvo Piran

intervala zaupanja $91\% - 1,96 \cdot 1,9\%$ *je manjša od 90%, zato zgornji izjavi ne moremo verjeti.*

8. Iz populacije vseh zaposlenih je statistična organizacija izbrala enostavni slučajni vzorec. V vzorcu je bilo 27% ljudi, ki niso še nikoli zamenjali delovnega mesta.

   a. Zgornja in spodnja meja zaupanja pri stopnji tveganja $\alpha = 0,05$ sta bili 27,76% in 26,24%. Kolikšen je bil vzorec? Zanemarite popravni faktor!

   b. Kako velik bi po vašem mnenju moral biti vzorec, da bi bila spodnja meja zaupanja 24%? Zanemarite popravni faktor!

   Rešitev:

   a. Zgornjo in spodnjo mejo zaupanja pri $\alpha = 0,05$ *dobimo po formuli* $\hat{p} + 1,96 \cdot SE$ *in* $\hat{p} - 1,96 \cdot SE$*, torej je* $SE = 0,39\%$*. Iz formule za* SE *lahko izračunamo velikost vzorca in dobimo* $n = 12.959$*.*

   b. SE *je v tem primeru* $1,53\%$ *in* $n = 842$*.*

9. Recimo, da izberemo enostavni slučajni vzorec velikosti $n = 400$ iz populacije velikosti $N = 1.500.000$ in ocenimo povprečje s 37,45, za meji zaupanja pa pri $\alpha = 0,05$ dobimo 34,25 in 40,65. Verjetnost, da je celoten interval zaupanja levo od prave vrednosti parametra, je

   a. 5%.

   b. 2,5%.

   c. 10%.

   d. 1%.

   e. Nobeden od zgornjih odstotkov.

Podajte kratko utemeljitev!

Rešitev: Pravilen je odgovor b. Celoten interval zaupanja bo levo od pravega povprečja, če bo vzorčna ocena padla "dovolj daleč" na levo stran v vzorčni porazdelitvi. Vzorčna ocena bo dovolj daleč, če bo ploščina levo od nje pod vzorčnim histogramom manjša od $2,5\%$.

10. Kot del zdravstvene raziskave so izbrali 900 študentov neke univerze. Uporabili so enostavno slučajno vzorčenje. Povprečje višin izbranih študentov je bilo 174 cm in standardni odklon 12 cm. Ko so narisali histogram za te podatke, se je izkazalo, da se tesno prilega normalni porazdelitvi. Povprečno višino študentov na univerzi so ocenili iz vzorca kot 174 cm s standardno napako 0,40 cm. Razmislite o pravilnosti naslednjih sklepov:

    a. Približno 68% študentov te univerze je visokih med 173,60 cm in 174,40 cm.

    b. Približni interval zaupanja za povprečno višino pri stopnji tveganja $\alpha = 0,05$ ima spodnjo mejo 173,22 cm in zgornjo mejo 174,78 cm.

    c. Če nekdo izbere enostavni slučajni vzorec 900 študentov in gre na levo in na desno od izračunanega vzorčnega povprečja za standardno napako 0,40 cm, potem je verjetnost, da bo zadel pravo vrednost povprečja, enaka 68%.

    d. Približno 68% študentov v vzorcu je bilo visokih med 163 cm in 182 cm.

Rešitev:

a. Sklep ni pravilen. Porazdelitev višin ima standardni odklon 12 *cm in ne* $0,40$ *cm.*

b. Pravilno. Ocene za povprečno višino so porazdeljene po vzorčni porazdelitvi.

c. Pravilno.

d. Sklep ni pravilen. 68% *študentov zajamemo z mejama* 162 *cm in* 186 *cm.*

© **Visoka šola za podjetništvo Piran**

11. Na sprejemnem izpitu za visoko šolo na EF leta 1991 je bilo možnih največ 100 točk. Povprečje za vse kandidate je bilo 58,68, standardni odklon pa 17,14. Kandidatov je bilo $N = 1259$. Recimo, da bi vzeli iz populacije kandidatov enostavni slučajni vzorec velikosti 225. Ali lahko trdimo, da bo vzorčno povprečje z verjetnostjo približno 84% manjše od 59,72? Odgovor utemeljite!

Rešitev: To bomo lahko trdili, če je ploščina pod vzorčno porazdelitvijo z $\mu = 58,68$ *in* $SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 1,04$ *na levo od* $59,72$ *enaka* 84%. *Vrednost* $59,72$ *spremenimo v standardne enote*

$$\frac{59,72 - 58,68}{1,04} = 1\,.$$

Ploščina na levo od te vrednosti je 84%, *torej je trditev pravilna.*

12. Pri velikih serijah enakih izdelkov lahko kontroliramo kvaliteto z enostavnim slučajnim vzorčenjem. Privzemite, da so serije velikosti $N = 10.000$. Recimo, da izberemo enostavni slučajni vzorec velikosti $n = 200$ in ugotovimo, da je v vzorcu neuporabnih 8% izdelkov. Proizvajalec trdi, da je v celotni seriji neuporabnih izdelkov le 5%. Bi zavrnili njegovo trditev?

Rešitev: Proizvajalčevo trditev bi zavrnili, če pri neki izbrani stopnji tveganja interval zaupanja okrog ocene 8$*nebipokrivalvrednosti*5%4. *To bi pomenilo, da z izbrano stopnjo tveganja zavrnemo proizvajalčevo trditev, kljub temu da ima proizvajalec prav.*

Izberimo si stopnjo tveganja $\alpha = 0,05$. *Standardna napaka ocene (brez popravnega faktorja) je* $SE = 1,9\%$ *in spodnja meja zaupanja je* $8\% - 1,96 \cdot 1,9\% = 4,2\%$. *Pri tej stopnji tveganja ne moremo zavrniti proizvajalčeve trditve. Tvegati bi morali več, če bi hoteli trditev zavrniti.*

13. Recimo, da delež volivcev, naklonjenih neki stranki, ocenimo na podlagi enostavnega slučajnega vzorca velikosti $n = 900$ in izračunamo, da je interval zaupanja pri stopnji tveganja $\alpha = 0,05$ enak $z_\alpha \cdot SE = 2,4\%$. Interval zaupanja pri stopnji tveganja $\alpha = 0,01$ bi bil

a. 2-krat daljši.

b. 5-krat daljši.

c. 1,45-krat daljši.

d. Krajši.

e. Nobena od zgornjih možnosti.

Utemeljite v enem stavku!

Rešitev: Odgovor je e. Interval zaupanja bi bil $1,3$*-krat daljši, kar je razmerje med* $2,55$ *in* $1,96$.

 © **Visoka šola za podjetništvo Piran**