STATISTIKA

Matjaž Blejec Marko Lovrečič-Saražin

Mihael Perman Mojca Štraus

Visoka šola za podjetništvo Piran

Kazalo

Kazalo slik in grafov

Kazalo tabel

Predgovor

Pričujoči učbenik je nastal iz gradiv, ki so jih avtorji pripravili za predmet Poslovna statistika na Visoki šoli za podjetništvo. Pogosto je slišati, da je statistika dolgočasen predmet, pri katerem je treba le prekladati številke, risati nezanimive grafe in preštevati. Pa vendar nas statistično izrazoslovje spremlja v vsakdanjem življenju. Ko odpremo časopis, beremo o inflaciji, o rezultatih te in one javnomnenjske raziskave, o napovedih gospodarske rasti in še bi lahko naštevali. Pri sestavljanju učbenika nas je vodila želja, da bi študentom približali osnovne pojme in postopke statistike, ne da bi morali poseči po zahtevnih matematičnih sredstvih. Razlage zato temeljijo na primerih in grafičnih prikazih, ki zahtevajo le nekaj več kot srednješolsko znanje matematike. Upamo, da bo učbenik dobro izhodišče za ekonomske in marketinške predmete, ki se naslanjajo na statistično izrazoslovje in razmišljanje.

Učbenik je razdeljen na štiri poglavja. Vsako se začne z nekaj primeri iz vsakdanjega življenja. Vpeljava statističnih pojmov se potem naslanja na izbrane primere. Sledijo primeri uporabe, ki smo jih poskušali izbrati tako, da bi bili čim bolj življenjski. Vsako poglavje zaključujejo naloge z rešitvami. V prilogi sta še primer pisnega izpita za samostojen preizkus znanja in tabela za normalno porazdelitev.

Sedanja oblika učbenika je v veliki meri rezultat odziva študentov. Radi bi se zahvalili predvsem prvi generaciji študentov, ki so se prebijali skozi zgodnje verzije gradiv. Zahvala gre tudi recenzentoma prof. dr. Anuški Ferligoj in prof. dr. Lovrencu Pfajfarju za podrobno branje in številne vsebinske in oblikovne pripombe. Ravno tako gre zahvala g. Janezu Juvanu za pozoren jezikovni pregled. K večji preglednosti so prispevali tudi predavatelji in asistenti pri predmetu. Posebej bi se zahvalili Poloni Grešak, Petri Grošelj in Gregorju Šegi za sezname nedoslednosti in napak. Nenazadnje bi se

želeli zahvaliti tudi vodstvu Visoke šole za podjetništvo, ki nas je ves čas vspodbujalo.

Ljubljana, 15. januar 2003 Avtorji

POGLAVJE 1

Opisne statistike

Statistika je veda, ki na podlagi zbranih podatkov odgovarja na vprašanja, ki si jih zastavljamo. V ekonomiji, na primer, želimo na podlagi danih podatkov oceniti gibanje ekonomskih kazalcev in razbrati trende razvoja. Marketinške raziskovalce zanima odnos potencialnih kupcev do novih produktov, v medicini se statistika uporablja za presojo, ali je neka nova terapija uspešna ali ne, in še bi lahko naštevali. Neposredno iz podatkov je odgovor na zastavljeno vprašanje pogosto težko razviden, zato je treba podatke predstaviti na čim preglednejši način oziroma v podatkih vsebovano informacijo nekako povzeti. V tem poglavju so predstavljene metode povzemanja podatkov, kot so na primer povprečje, kvantili in mere razpršenosti. Drugi, morda še uporabnejši način povzemanja podatkov so grafične metode. Primerno izbran graf nam lahko v trenutku predstavi podatke. Od grafičnih metod predstavljanja podatkov si bomo ogledali histograme, ki so eden od zelo razširjenih načinov predstavitve podatkov.

1.1 Populacije in spremenljivke

Vsak dan prihajamo v stik s podatki statistične narave. V časopisju srečujemo rezultate anket ali podatke o porastu cen življenjskih potrebščin, govorimo o inflaciji, spremljamo gibanje deviznih tečajev ali podatke o gospodarskih gibanjih. Statistika se pri tem pojavlja kot orodje, s katerim lahko iz množice podatkov, ki so nam na voljo, izluščimo bolj strnjeno informacijo. Hkrati je poznavanje osnov statistike koristno tudi za bolj kritično presojo o dejanskem pomenu množice številk, ki nas obdajajo. Pričujoča gradiva bodo poskušala osvetliti osnovne statistične pojme na primerih uporabe teh pojmov pri obdelavi in interpretaciji podatkov.

Okvir razmišljanja v statistiki so *populacije*. Primeri populacij, ki jih bomo obravnavali, so volilni upravičenci v Sloveniji, vsi zaposleni, učenci določene starosti in podobno. Pri tem moramo besedo "populacija" razumeti širše, kot bi sklepali iz latinske besede populus, ki pomeni ljudstvo. Poleg populacij, sestavljenih iz ljudi, nas bodo zanimale tudi populacije, Kot so vsa gospodinjstva v Sloveniji, podjetja in tudi izdelki, narejeni v danem časovnem obdobju. Ne gre torej izključno za populacijo ljudi, čeprav bo to pogosto res, temveč s to besedo označimo skupek vseh ljudi, skupin ljudi ali predmetov, ki jih obravnavamo. Sestavne dele populacije bomo imenovali *enote*, kar so zopet lahko ljudje, skupine ljudi ali predmeti.

Primer: Ko govorimo o povprečni plači v Sloveniji, se moramo vprašati, kaj ta količina pravzaprav pomeni. Prvi korak pri odgovoru je, da opišemo populacijo, ki jo obravnavamo. Povedati moramo, čigave plače upoštevamo, ko računamo povprečje. Odgovor v tem primeru je, da kot populacijo obravnavamo vse redno ali začasno zaposlene v Sloveniji. Enote populacije so torej zaposleni posamezniki.

Primer: Slovenija je bila vključena v mednarodno primerjalno raziskavo znanja matematike in naravoslovja z naslovom Third International Mathematics and Science Study. (TIMSS) . Pod drobnogledom so bili tudi učenci sedmih in osmih razredov osnovne šole, ki so reševali precej obsežne delovne zvezke nalog, na podlagi katerih je bilo potem ocenjeno njihovo znanje. Populacijo v tem primeru sestavljajo učenci v danih razredih v času izvedbe raziskave. Kaj so tukaj enote, je na dlani.

PRIMER: Enote so lahko tudi podjetja ene ali več držav in govorimo o populaciji podjetij. Kasneje se bomo srečali s podatki o nekaterih slovenskih podjetjih in na podlagi le-teh skušali sklepati o nekaterih ekonomskih zakonitostih. Na prvi pogled je morda nekoliko nenavadno, da bi govorili o "populaciji" podjetij, vendar gre za ustaljen način izražanja v statistiki, ki ga bomo privzeli tudi tukaj.

PRIMER: Pri kontroli kvalitete izdelkov je pogosto treba preverjati veliko število izdelkov. Tudi tukaj govorimo o populaciji vseh izdelkov, narejenih v nekem časovnem obdobju.

V statistiki nas bodo zanimale populacije, ki so sestavljene iz enot. Pri tem ne gre samo za populacije v dobesednem smislu, temveč je ta pojem splošnejši in zajema tudi populacije, kot so vsa gospodinjstva, podjetja ali tudi izdelki, narejeni v danem obdobju.

V statistiki nas ne zanimajo populacije kot take, ampak nas zanimajo podatki o enotah v teh populacijah. Ti podatki so lahko *številski*, kot je recimo višina plače, ali *opisni*, kot je ime politične stranke, za katero se volivec odloči. V statistiki imenujemo lastnost enot, ki nas zanima, *spremenljivka*. Vsaki enoti v dani populaciji pripada neka vrednost spremenljivke.

PRIMER: Če nadaljujemo primer o povprečni plači v Sloveniji, moramo potem, ko smo se dogovorili, kaj je populacija in kaj so enote, povedati tudi, kaj je spremenljivka. V primeru plač je očitno, da je spremenljivka višina plače posameznega zaposlenega. Ni treba posebej razlagati, da se ta "spremenljivka" res spreminja od enote do enote.

PRIMER: V enem od zgornjih primerov smo govorili o učencih 7. in 8. razredov. Vrednost spremenljivke, ki je pripadala posameznemu učencu oziroma enoti, je bila doseženo število točk na preizkusu znanja. Ker je preizkus reševalo več kot 300.000

učencev iz vseh sodelujočih držav, seznam njihovih dosežkov ne dá jasne slike o znanju učencev posameznih držav. Vloga statistike je ravno v tem, da iz takšne množice podatkov izlušči uporabno informacijo.

PRIMER: Oglejmo si še primer volitev v Sloveniji. Populacija so slovenski volivci in vsak volivec voli eno politično stranko. V tem primeru ne moremo reči, da vsaki enoti pripada neka številska vrednost, saj gre za izrekanje o strankah. Lahko pa še vedno govorimo o spremenljivki: volivci in volivke se odločajo za to ali ono stranko po svoji presoji, in tudi tukaj se odločitev spreminja od enote do enote. Zato bomo tudi v teh primerih govorili o spremenljivki, katere vrednosti pa so opisne. V vsakdanjem življenju odstotku enot, za katere ima spremenljivka vrednost "STRANKA X", pravimo odstotek volivcev, ki so se odločili za stranko X.

Obravnavano lastnost enot populacije v statističnem žargonu imenujemo spremenljivka. Vsaki enoti v populaciji pripada vrednost spremenljivke, ki je lahko številska ali opisna. Večinoma nas ne bodo zanimale vrednosti spremenljivke za posamezne enote, temveč na primeren način strnjena informacija o spremenljivki za celotno populacijo.

1.2 HISTOGRAMI

V množici podatkov je pogosto težko prepoznati zakonitosti ali si predstavljati obseg vrednosti spremenljivke za dano populacijo. Zato nas pri statistiki zanima *porazdelitev* vrednosti spremenljivke, ali skrajšano, porazdelitev spremenljivke. Ker so vrednosti spremenljivke lahko različne, želimo predstaviti, kolikšen odstotek enot ima dane vrednosti. Vprašamo se lahko na primer, kolikšen odstotek zaposlenih v Sloveniji ima plače med 70.000 SIT in 90.000 SIT. Skupni opis vrednosti spremenljivke in odstotkov enot, za katere ima spremenljivka dane vrednosti, imenujemo porazdelitev. Porazdelitve

spremenljivk si bomo predočili z grafikoni. Za zdaj se bomo omejili le na spremenljivke, katerih vrednosti so številske.

Porazdelitev spremenljivke s številskimi vrednostmi je opis, ki za poljubni dve mejni vrednosti poda odstotek enot, ki imajo vrednosti spremenljivke med tema mejama. Porazdelitve si najbolje ponazorimo s primerno izbranimi grafikoni.

Grafični prikaz porazdelitve spremenljivke, ki si ga bomo ogledali, je *histogram*. Začeli bomo kar s primerom.

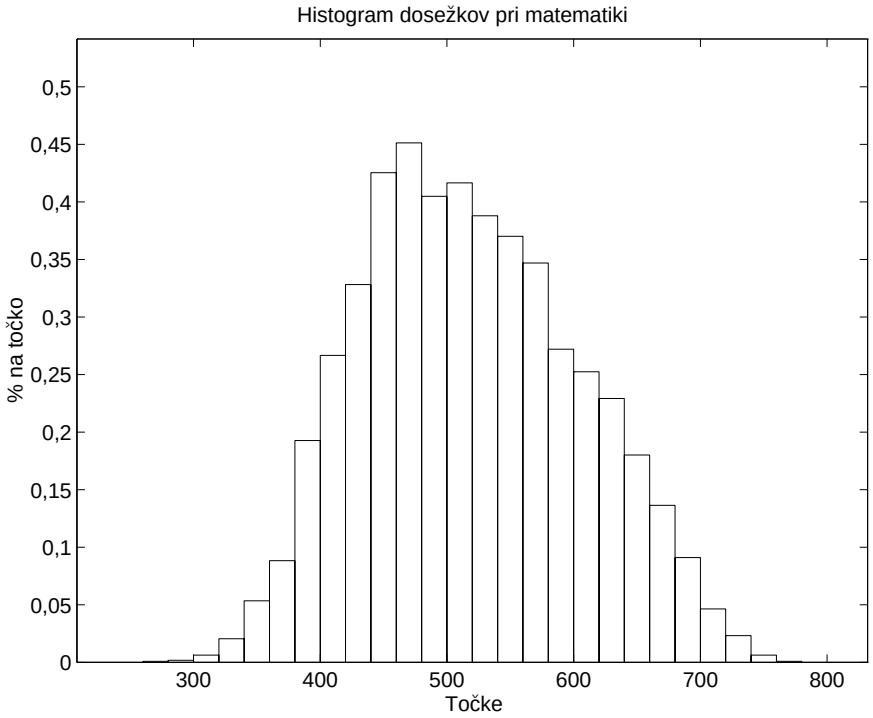
PRIMER: V Sloveniji je v že omenjeni raziskavi TIMSS sodelovalo 5606 učencev sedmih in osmih razredov. Njihovi rezultati na preizkusu znanja so bili izračunani po posebnem postopku in so številke med približno 250 in 800. Seznam dosežkov iz matematike za 24 učencev je v spodnji tabeli.[1]

423,7	398,1	311,9	414,1	401,5	478,8	453,5	433,2
512,5	625,4	520,2	468,7	543,8	440,1	490,6	545,6
533,8	575,4	463,4	632,6	391,9	522,0	432,3	613,6

Tabela 1.1: Primer rezultatov na preizkusu znanja iz matematike.

Po pričakovanju je iz samih številk težko razbrati porazdelitev dosežkov, zato si pomagamo s histogramom. Na sliki 1.1 je histogram za dosežke vseh 5606 slovenskih učencev na preizkusu znanja iz matematike. Ideja histograma je v tem, da odstotke

[1] Vir: Pedagoški inštitut, Raziskava TIMSS 1995.



Sl. 1.1: Histogram dosežkov na preizkusu iz matematike za 5606 učencev.

predstavimo s *ploščinami* narisanih pravokotnikov v histogramu. Odstotek učencev, ki so dosegli od 500 do 600 točk, je enak odstotku skupne ploščine pravokotnikov histograma med 500 in 600 glede na celotno ploščino histograma. Umestno je vprašanje, zakaj smo za predstavitev odstotkov izbrali ravno ploščino. Odgovor je, da tako najlaže predstavimo vrednosti spremenljivke skupaj z odstotki enot, ki so imele vrednost spremenljivke med danimi mejami. Pogosto lahko že iz geometrijske oblike histograma sklepamo o nekaterih značilnostih populacije. V naslednjih poglavjih bomo videli, da je predstavitev porazdelitve spremenljivk s histogrami zelo primerna tudi za obravnavanje nekaterih bolj teoretičnih vprašanj v statistiki.

Kaj lahko rečemo o porazdelitvi dosežkov učencev na podlagi histograma 1.1? Kot prvo, da je največja "gneča" okrog rezultata 480. Bolj strokovno se izrazimo, če

rečemo, da je na tem intervalu *gostota* največja. Odstotek učencev z dosežkom med 480 in 500 točkami je občutno večji kot odstotek učencev z dosežki med 300 in 320 točkami. Interval med 480 in 500 točkami je gosteje "naseljen", ker se v njem gnete več učencev kot na intervalu med 300 in 320 točkami, ki je enako dolg.

Enota na navpični osi histograma je na prvi pogled morda nekoliko nenavadna. Rekli smo, da so v histogramih odstotki enot z vrednostmi spremenljivke med danima mejama predstavljeni s ploščino pravokotnikov ali dela histograma med tema mejama. Zato mora biti enota na navpični osi izbrana tako, da iz produkta z enoto na vodoravni osi za ploščino zares dobimo odstotke.

Poskusimo oceniti odstotek učencev, ki so dosegli od 480 do 500 točk, na podlagi histograma. Ta odstotek je enak ploščini pravokotnika nad tem intervalom, ta pa je približno 20 točk $\cdot$ $0,4\%$ na točko $= 8\%$. Intervale na vodoravni osi merimo v tem primeru s točkami in te se morajo, vsaj formalno, pokrajšati.

Posvetimo se še geometrijski obliki histograma 1.1. Rezultati testa so približno simetrično porazdeljeni okrog sredine histograma. Odstotek učencev z dosežki nad 500 točk je le nekoliko večji kot odstotek učencev z dosežkom pod 500 točk.

Histogram grafično ponazori porazdelitev vrednosti neke spremenljivke. Pomembno je dejstvo, da je odstotek enot, za katere je vrednost spremenljivke med danima mejama, enak ploščini pravokotnikov ali dela histograma med tema mejama. Enote na navpični osi so izbrane tako, da je celotna ploščina histograma 100%. Teh enot pogosto niti ne navajamo.

PRIMER: Oglejmo si histogram za bruto plače v Sloveniji v letu 1995.[2] Histogram je narisan na podlagi podatkov iz tabele na naslednji strani. Naj omenimo, da so lahko pravokotniki v histogramu različno široki. Pomembno je le to, da odstotke predstavlja ploščina. Kaj bi lahko še rekli o histogramu na sliki 1.2? Rečemo lahko, da plače

[2]Vir: Urad za statistiko RS, Statistični letopis 1996

niso enakomerno porazdeljene, temveč je večina ploščine pomaknjena proti levi strani histograma, torej proti nižjim plačam. Pri višjih plačah je histogram občutno nižji. V visokih plačilnih razredih se ne "gnete" toliko posameznikov oziroma je tam gostota nižja.

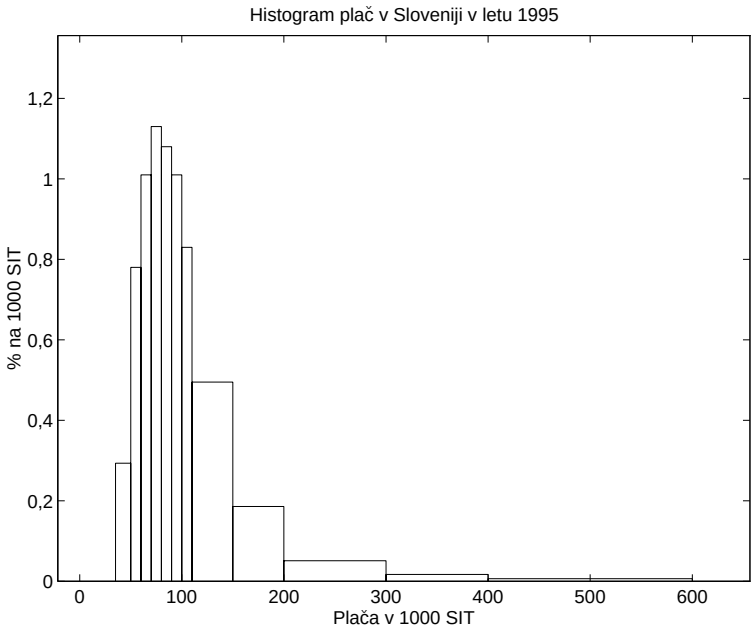
Plačilni razred (v 1000 SIT)	%
35 – 50	4,4
50 – 60	7,8
60 – 70	10,1
70 – 80	11,3
80 – 90	10,8
90 – 100	10,1
100 – 110	8,3
110 – 150	19,8
150 – 200	9,3
200 – 400	6,8
400 – 600	1,3

Tabela 1.2: Tabela plačilnih razredov za RS v letu 1996.

Poskusimo oceniti, kolikšen odstotek zaposlenih je prejemal plače med 150.000 SIT in 200.000 SIT. Gostota na tem intervalu je približno 0,186% na tisoč tolarjev in ploščino dobimo tako, da to številko pomnožimo z osnovnico pravokotnika nad omenjenim intervalom, torej 0,186% na tisoč tolarjev $\cdot$ 50.000 tolarjev $= 9,3\%$. Postavimo si še nekoliko težje vprašanje. Za katero višino plače lahko trdimo, da je 50% zaposlenih zaslužilo več, 50% pa manj? Vprašanje lahko prevedemo na vprašanje o ploščinah. Histogram moramo navpično prerezati tako, da bo na levi polovica ploščine in na desni polovica ploščine. Podatki iz zgornje tabele pokažejo, da je bila bruto plača za 44,4% zaposlenih nižja od 90.000 SIT, za 54,5% pa nižja od 100.000 SIT. To pomeni, da bo višina plače, ki jo iščemo, med 90.000 in 100.000 SIT. Od pravokotnika nad tem intervalom moramo sedaj z leve strani odrezati košček s ploščino 5,6%. Gostota na

tem intervalu je 1,01% na 1000 SIT, torej mora biti osnovnica odrezanega koščka 5,6% deljeno z 1,01% na 1000 SIT, kar je $5,54 \cdot 1000$ SIT. Iskana višina plače je 95.540 SIT.

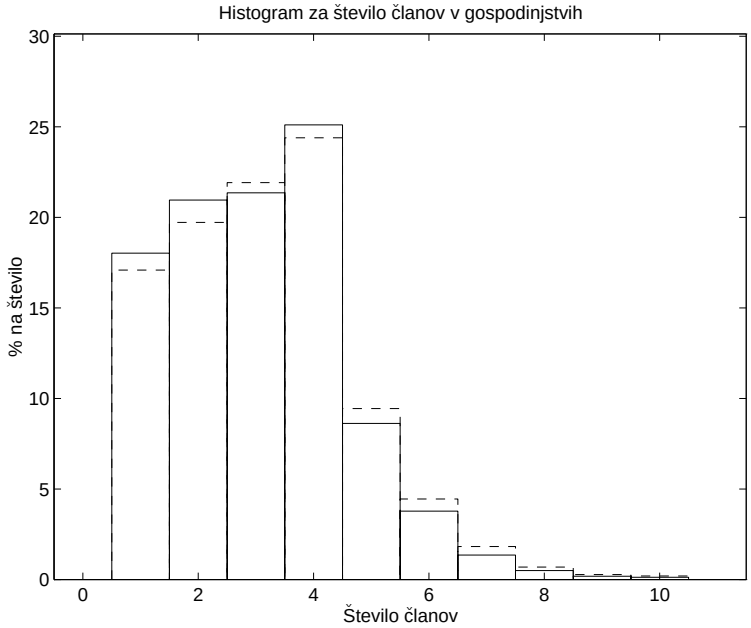
Posebna vrsta podatkov so celoštevilski podatki. Na primer, za populacijo gospodinjstev nas lahko zanima število vseh članov v gospodinjstvu, število otrok ali morda število avtomobilov.



Sl. 1.2: Histogram za plače v Sloveniji.

Skupno vsem tem spremenljivkam je, da so njihove vrednosti lahko samo cela števila. Ne moremo imeti gospodinjstva z 1,2 člana, ali morda 0,99 avtomobila. Tudi pri takih spremenljivkah si za grafično predstavitev porazdelitve vrednosti pomagamo s histogrami. Da se izognemo težavam pri interpretaciji, se dogovorimo, da bodo pravokotniki v histogramu, ki predstavljajo odstotek enot z dano vrednostjo celoštevilske spremenljivke, postavljeni točno nad tistim številom.

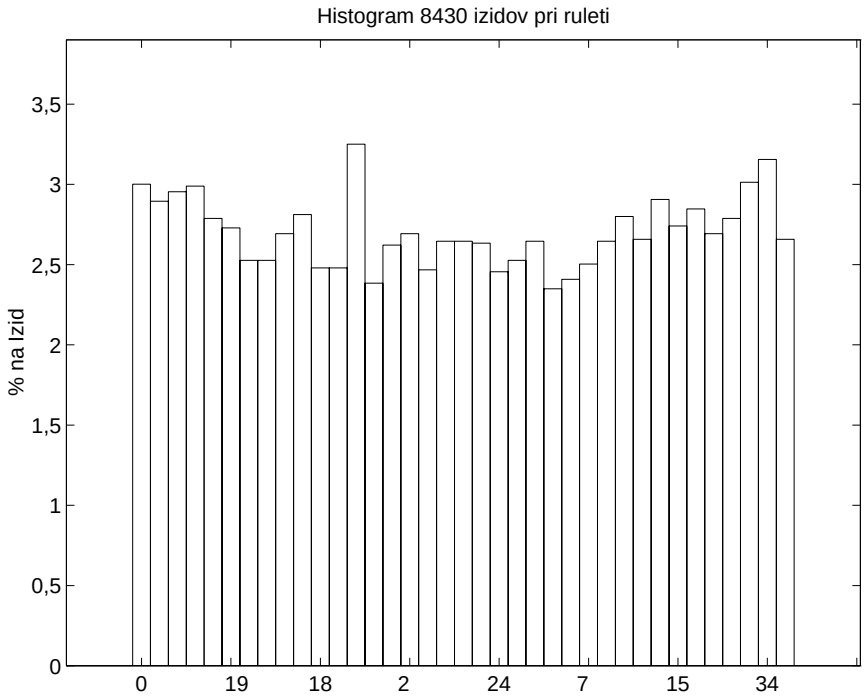
Primer: Populacijo naj sestavljajo slovenska gospodinjstva, spremenljivka pa bodi število članov. Podatke lahko dobimo v Statističnem letopisu RS, ki ga izdaja Statistični urad RS. Kot vedno, tudi tukaj ploščina predstavlja odstotke enot z danimi vrednostmi spremenljivke, le da jih je v tem primeru še laže razbrati. Odstotki so namreč enaki višini stolpca nad vsakim številom. Tako vidimo, da je leta 1991 imelo približno 18% gospodinjstev enega člana in 21% gospodinjstev dva člana. Če bi nas zanimalo, koliko odstotkov gospodinjstev je imelo dva člana ali manj, bi bil odgovor torej 39%.



Sl. 1.3: Histogram števila članov v gospodinjstvih (1981 črtkano, 1991 polno).

Primer: Oglejmo si še nekoliko bolj nenavaden primer, kjer bomo s histogramom zelo dobro videli, kakšno informacijo vsebujejo podatki, ki jih imamo na voljo. Poglejmo si francosko ruleto, ki je v navadi v slovenskih igralnicah. Vsakič, ko krupje zavrti cilinder rulete in vanj spusti kroglico, dobimo neki izid, ki je številka med 0 in 36. Re-

cimo, da si zabeležimo veliko število izidov in se vprašamo, ali obstaja izid ali skupina izidov, ki bi bili bolj verjetni kot preostali.



Sl. 1.4: Histogram izidov pri ruleti.

Vprašanje je zanimivo tako za igralnico kot za tiste, ki prisegajo, da so iznašli "pravi" sistem, s katerim v igri dobivajo. Odgovor na to vprašanje je vse prej kot preprost, prvi korak pa je, da si podatke predočimo grafično. Histogram 1.4, ki ima 37 stolpcev, je grafična predstavitev porazdelitve izidov za 8430 iger na cilindru v eni od slovenskih igralnic. Stolpec nad dano številko predstavlja, kolikokrat se je ta številka pojavila, merjeno v odstotkih. Kot primer navedimo, da se je 0 pojavila 253-krat, kar je v 3% vseh iger. Če privzamemo, da se vsak izid pojavi z enako verjetnostjo, bi pričakovali, da se bo v velikem številu iger vsak izid pojavil v 2,7% iger. Odstotek 2,7% dobimo, če 100% razdelimo na 37 delov, kolikor je možnih izidov.

Preden se lotimo ugibanja o tem, kaj naj si mislimo o cilindru, pripomnimo nekaj o histogramu samem. Za boljše razumevanje povejmo, da so stolpci v enakem vrstnem redu kot na obodu cilindra, če gremo v smeri, nasprotni urinemu kazalcu.[3] Tako je na petem mestu številka 19, ker je pač peta od 0 po obodu. Gre za histogram, v katerem ploščine stolpcev predstavljajo odstotke, te pa je zelo lahko odčitati, ker je širina vseh stolpcev enaka 1 in so zato njihove ploščine enake višinam, vsaj, če si za trenutek odmislimo enote. V statističnem jeziku bi v tem primeru govorili o "populaciji" vseh iger, vrednost spremenljivke, ki bi pripadala vsaki igri, pa bi bila izid.

Kaj bi lahko sklepali o verjetnostih posameznih izidov na podlagi histograma 1.4? Morda, da se bo kroglica z večjo verjetnostjo ustavila na eni od številk blizu 0 kot pa na številkah, ki so na cilindru nasproti 0. Poleg tega je zelo izrazit tudi stolpec nad 33, kar pomeni, da se je številka 33 pojavila večkrat kot druge številke v okolici. Podrobnejša analiza izidov na histogramu je pokazala, da je bilo na tej ruleti dejansko mogoče staviti v korist igralca, zato je igralnica zgornji cilinder odstranila iz uporabe.

1.3 Povprečje in standardni odklon

Ena od nalog statistike je, da informacijo, vsebovano v množici podatkov, povzame na pregleden način. V prejšnjem razdelku smo videli, da so histogrami pripravno orodje za prikaz oblike porazdelitve. Pogosto pa potrebujemo tudi povzetke v obliki številk. Najpogostejši povzetek podatkov je *povprečje*. Vsi vemo, da je povprečje seznama števil preprosto vsota števil deljena s številom teh števil. V statistiki obstajajo tudi bolj zapletene vrste povprečij, vendar se bomo tukaj omejili le na omenjeno obliko.

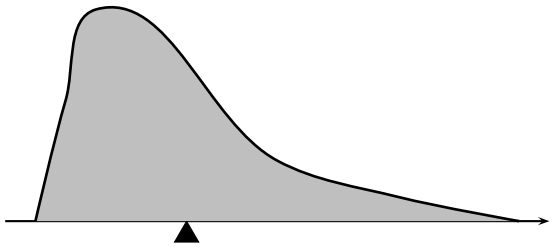
V prvem razdelku smo govorili o populacijah in spremenljivkah. Vsaki enoti v populaciji pripada neka vrednost spremenljivke. Omejimo se na primer, ko so vrednosti spremenljivke številske, in si zamislimo, da bi vrednosti spremenljivke za enote iz populacije zapisali v seznam. Povprečje števil v tem seznamu imenujemo *povprečno vrednost spremenljivke* ali kar *povprečje spremenljivke*.

[3] Vrstni red številk na cilindru francoske rulete od 0 v smeri, nasprotni urinemu kazalcu: 0, 23, 6, 35, 4, 19, 10, 31, 16, 27, 18, 14, 33, 12, 25, 2, 21, 8, 29, 3, 24, 5, 28, 17, 20, 7, 36, 11, 32, 30, 15, 26, 1, 22, 9, 34, 13.

Povprečna vrednost spremenljivke je povprečje vrednosti spremenljivke za posamezne enote v populaciji. Povprečno vrednost spremenljivke pogosto označimo z grško črko μ.

PRIMER: Ko govorimo o povprečni plači v Sloveniji, mislimo na populacijo vseh redno ali začasno zaposlenih, vrednost spremenljivke pa je za vsakega zaposlenega, torej za vsako enoto, enaka višini bruto plače. Povprečje precej dolgega seznama višin plač je potem povprečna vrednost spremenljivke. Za leto 1995 je bilo to povprečje 112.105 SIT[4]. Če bi populacijo spremenili, bi za povprečje bruto plače verjetno dobili drugačen rezultat. Tako je bila za populacijo vseh zaposlenih v proizvodnji obutve in galanterije povprečna bruto plača le 65.135 SIT, medtem ko je bilo najvišje povprečje bruto plače v populaciji zaposlenih v sektorju uprave z javnimi skladi, in sicer 212.764 SIT.

Vprašajmo se še, kako bi lahko "na oko" ocenili povprečje vrednosti, če imamo na razpolago samo njihov histogram. Zamislimo si, da bi imeli histogram izrezan iz kartona.



Sl. 1.5: Povprečje je tam, kjer je histogram uravnotežen.

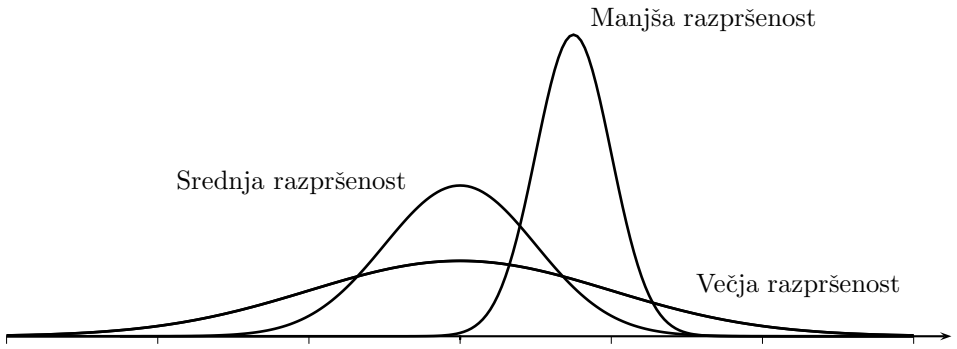
Uravnotežen bi bil takrat, ko bi podporno točko postavili natanko pod povprečje. Na sliki 1.5 je shematično prikazan histogram, podporna točka pa je prikazana kot majhen trikotnik. Poskusite se prepričati, da bi za primer histograma na sliki 1.2

[4]Vir: Urad za statistiko RS, Statistični letopis 1996

povprečna plača res ustrezala tisti, ki bi jo "na oko" ocenili na podlagi oblike histograma.

Povprečje je pomemben način povzemanja informacije iz podatkov. S to količino v strnjeni obliki z enim samim številom povemo nekaj o porazdelitvi spremenljivke. Pogosto pa nas ne bo zanimalo samo povprečje, ampak bomo s števili želeli povzeti tudi informacijo o obliki porazdelitve ali vsaj, ali so vrednosti spremenljivke bolj strnjene okrog povprečja ali bolj razpršene. Kot prvi korak v tej smeri se vprašajmo, kako bi v enem številu strnili informacijo o tem, koliko so podatki ali vrednosti spremenljivke razpršeni. Za občutek si oglejmo shematično prikazane histograme na sliki 1.6.

Takoj se vidi, za kateri histogram lahko rečemo, da ponazarja porazdelitev bolj razpršenih vrednosti. To je tisti z oznako "večja razpršenost". Želeli pa bi to razpršenost povzeti tudi s številom. Količino, ki jo statistiki uporabljajo v ta namen, imenujemo *standardni odklon*. Kako izračunamo standardni odklon za seznam števil, si oglejmo na primeru.



Sl. 1.6: Histogrami za različno razpršene vrednosti.

Primer: Naj bodo za neko populacijo desetih enot dane vrednosti spremenljivke 59, 37, 33, 48, 43, 55, 53, 57, 72, 43. Najprej izračunamo povprečje, ki je v tem primeru 50. Vsem vrednostim nato odštejemo 50 in dobimo odklone posameznih vrednosti od povprečja, ki so 9, -13, -17, -2, -7, 5, 3, 7, 22, -7. Čim večji so ti odkloni po

absolutni vrednosti, tem večja je razpršenost vrednosti. Kako natančno merimo vpliv teh odklonov na razpršenost podatkov? Možnosti je seveda več. Ideja standardnega odklona je v tem, da predstavlja primerno mero za povprečno velikost teh odklonov, ki jo izberemo takole: odklone najprej kvadriramo, da se znebimo negativnih predznakov, in dobimo 189, 169, 289, 4, 49, 25, 9, 49, 484, 49. Povprečje teh kvadratov je 120,8. Standardni odklon je kvadratni koren iz tega povprečja, torej $\sqrt{120,8} = 10,99$.

Iz postopka v zgornjem primeru je jasno, kako izračunamo standardni odklon za poljubno število vrednosti. Pri velikem številu vrednosti seveda uporabimo kakšen modernejši način računanja.[5] Če izračunamo standardni odklon za vrednosti spremenljivke za vse enote v populaciji, potem govorimo o *standardnem odklonu vrednosti spremenljivke* ali kar o *standardnem odklonu spremenljivke*. Kot opombo morda povejmo, da pri standardnem odklonu ne gre za nekaj, kar bi bilo samo po sebi standardno, gre le za udomačeno statistično izrazoslovje. V razdelku 1.5 bomo srečali histograme, katerih obliko lahko popolnoma opišemo samo s povprečjem in standardnim odklonom. V splošnem sta povprečje in standardni odklon le številska povzetka, ki informacijo v vrednostih spremenljivke strneta, pri tem pa se je nekaj izgubi.

Standardni odklon spremenljivke je najpogosteje uporabljena mera razpršenosti njenih vrednosti. Postopek za izračun je naslednji: najprej izračunamo povprečje vrednosti in ga odštejemo vsaki vrednosti. Tako dobljene razlike kvadriramo in izračunamo njihovo povprečje. Standardni odklon je kvadratni koren tega zadnjega povprečja. Pogosto bomo standardni odklon označili z grško črko σ.

[5] Žepna računala pogosto imajo funkcijo, ki izračuna standardni odklon za dane vrednosti. Večinoma pa ta računala vsoto kvadratov delijo z 1 manj, kot je členov. Razlogi so matematične narave in jih ne bomo obravnavali.

1.4 Kvantili

Histogram na sliki 1.2 prikazuje porazdelitev bruto plač v Sloveniji. Iz histograma smo ugotovili, da je bila bruto plača v letu 1995 za 50% zaposlenih nižja od 95.540 SIT, povprečje pa je bilo 112.105 SIT. Lahko bi se vprašali, katera od teh dveh številk je boljši povzetek podatkov o plačah, ki so nam na voljo. Ne ena ne druga! Obe količini imata svoje prednosti in svoje slabosti. Če poznamo povprečno plačo in število zaposlenih, lahko ocenimo celotno količino vseh izplačanih plač, česar pa ne moremo narediti, če poznamo samo številko 95.540 SIT. Slaba stran povprečja pa je, da je preveč "občutljivo" na visoke bruto plače. Majhen odstotek zelo visokih plač je dovolj, da povprečje naraste in tako dá na videz nekoliko nerealno sliko o dejanskih prejemkih. Druga količina je v takem primeru boljša, ker predstavi, s kolikšnimi dohodki mora shajati večina zaposlenih. Številka 95.540 SIT je primer količine, ki jo v statistiki imenujemo *kvantil*.

Kvantili so še eden od načinov povzemanja podatkov. O kvantilu govorimo, ko iščemo vrednost, za katero je določen odstotek vrednosti spremenljivke pod njo in preostalo nad njo. 25. kvantil seznama vrednosti spremenljivke je vrednost, za katero je 25% vrednosti spremenljivke pod njo, 75% vrednosti pa nad njo. Povsem na enak način bi lahko govorili o 40. kvantilu ali o 99. kvantilu.

Kvantili so številski povzetki informacije, ki je vsebovana v podatkih ali porazdelitvah. Za poljuben x definiramo x-ti kvantil kot vrednost, za katero je x% danih vrednosti pod njo. Pri spremenljivkah govorimo o x-tem kvantilu porazdelitve te spremenljivke. Kvantili so drugačna vrsta povzemanja podatkov kot povprečje in imajo svoje prednosti in svoje slabosti.

Za izbrane vrednosti odstotkov imajo ustrezni kvantili posebna imena. Tako rečemo 50. kvantilu *mediana*, kar prihaja iz latinske besede "medianus" za sredino. Mediana je torej vrednost "na sredi"; pod njo je polovica in nad njo polovica vrednosti. Posebni

imeni imata še 25. kvantil, ki mu statistiki pravijo *prvi kvartil*, in 75. kvantil, ki ga imenujemo *tretji kvartil*. Kvartil prihaja iz latinskega "quartus" za četrtino. Zakaj, je seveda jasno.

PRIMER: V razdelku 1.2 smo govorili o raziskavi TIMSS . V mednarodnem poročilu so za vsako sodelujočo državo navedeni tudi 90. kvantili. Tako je bil 90. kvantil za dosežke pri matematiki za slovenske sedmošolce in osmošolce enak 638,5 točke. Če bi za populacijo namesto slovenskih učencev vzeli učence vseh sodelujočih držav skupaj, bi za 90. kvantil dosežkov dobili 628,2 točke.

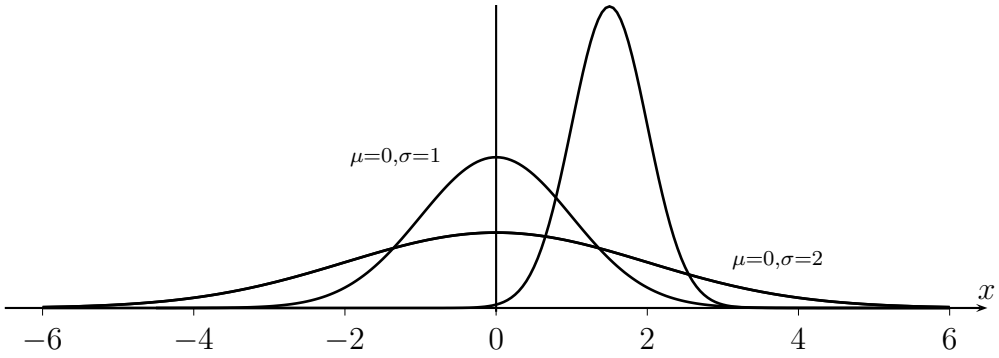
PRIMER: Ameriške univerze imajo navado, da poleg rezultatov na izpitih za vsakega študenta objavijo še kvantil njihovega rezultata. Za nekoga recimo lahko rečejo, da je bil "na 68. kvantilu", kar bi pomenilo, da je bil boljši od 68% študentov na izpitu.

Vprašajmo se še, kako "na oko" oceniti kvantile s histograma. Odgovor je v načelu preprost: za 68. kvantil moramo histogram navpično prerezati tako, da ostane 68% ploščine na levi.

1.5 NORMALNA PORAZDELITEV

Histogrami za nekatere spremenljivke, kot so telesna višina, inteligenčni kvocient, ali histogrami dosežkov na testih, ki jih mora opraviti veliko število ljudi, se tesno prilegajo krivuljam, ki jih matematiki imenujejo *normalne krivulje*. Tako krivuljo opišemo kot graf funkcije, ki je definirana v okvirčku na naslednji strani.

Normalne krivulje za vrednosti parametrov $\mu = 0, \sigma = 1$, $\mu = 1,5, \sigma = 0,5$ in $\mu = 0, \sigma = 2$ so na sliki 1.7. Kot vidimo, parameter μ pove, kje je sredina normalne krivulje, parameter σ pa pove, koliko je krivulja "raztegnjena" v vodoravni smeri. Za nas bo pomembna interpretacija parametra σ, ki pravi, da bi bil σ standardni odklon za histogram, ki bi se tesno prilegal taki normalni krivulji. Spomnimo se še na to, da je za vsak histogram enota na navpični osi izbrana tako, da je celotna ploščina histograma enaka 100%. To velja tudi pri normalnih krivuljah.



Sl. 1.7: Primeri normalnih krivulj.

Matematični opis normalne krivulje je dan s formulo

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

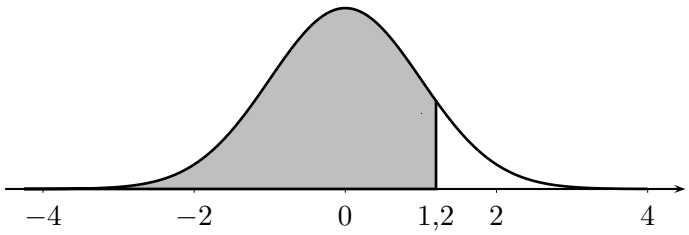
kjer sta μ in σ dani števili ali, kot tudi rečemo, dana parametra. Številu μ pravimo pričakovana vrednost, številu σ pa standardni odklon. Razlogi za táko poimenovanje so podani v besedilu.

Kakšen je pomen normalne krivulje? V statistiko je ta pojem vpeljal belgijski statistik Adolphe Quetelet okoli leta 1870 kot neke vrste "idealni" histogram, ki bi opisoval "naravno" porazdelitev vrednosti spremenljivk. Iz prejšnjega razdelka vemo, da nekateri histogrami niso niti malo podobni normalnim, seveda pa zato niso nenormalni ali nenaravni. Rečemo lahko samo, da je mogoče histograme, ki se tesno prilegajo normalni krivulji, zelo učinkovito opisati s samo dvema številoma: povprečjem μ in s standardnim odklonom σ. Ti dve števili sta, kot bomo videli na primerih, dovolj, da izračunamo poljuben kvantil ali katero drugo količino v histogramu.

Uporabnost normalnih krivulj bomo spoznali v poglavju o vzorčenju, kjer bomo

videli, da se dobro prilegajo posebni vrsti histogramov. Normalne krivulje imajo zato v statistični teoriji pomembno vlogo.

Pri histogramih nas je pogosto zanimala ploščina med danima vrednostma. Kako ravnamo, ko imamo opraviti z normalnimi krivuljami? Normalne krivulje so popolnoma opisane s parametroma μ in σ, zato lahko izračunamo poljubno ploščino, takoj ko ju poznamo. Na srečo je res še več. Če znamo izračunati ploščino med danima mejama pod normalno krivuljo s parametroma $\mu = 0$ in $\sigma = 1$, potem bomo znali izračunati ploščino tudi za normalno krivuljo z drugačnima parametroma. Zaradi tega je normalna krivulja s parametroma $\mu = 0$ in $\sigma = 1$ pomembna in ima posebno ime: *standardna normalna krivulja*.

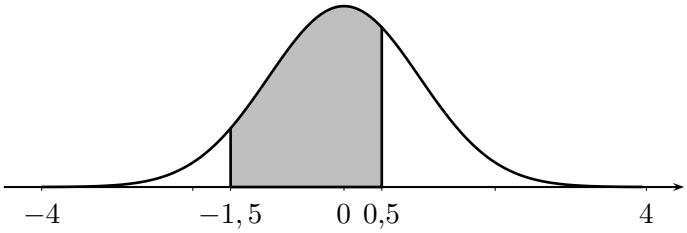


Sl. 1.8: Ploščina pod normalno krivuljo levo od $x = 1,2$.

Za računanje ploščin iz porazdelitve, ki jo ta krivulja opisuje, je na koncu pričujočih gradiv dodana posebna tabela. Podatki v tabeli za posamezne meje povedo, kolikšen je odstotek ploščine pod krivuljo levo od dane meje. Kot primer si oglejmo, kako iz tabele odčitamo ploščino na sliki 1.8. V stolpcu, označenem z x, poiščemo 1,2 in zraven te vrednosti odčitamo odstotek ploščine na levo pod normalno krivuljo. Ta odstotek je enak 88,49.

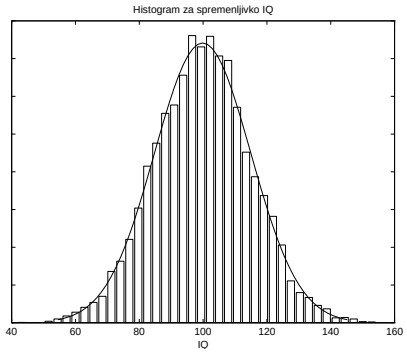
Za računanje ploščin drugačnih področij pod normalno krivuljo si pomagamo z nekaj iznajdljivosti. Recimo, da želimo vedeti odstotek ploščine pod normalno krivuljo med $-1,5$ in $0,5$. S slike 1.9 brž razberemo, da moramo najti odstotek, ki pripada vrednosti $x = 0,5$, in od njega odšteti odstotek, ki pripada $x = -1,5$. Iz tabele

preberemo, da sta odstotka enaka 69,1% in 6,7%. Ploščina osenčenega dela na sliki 1.9 je torej $69,1\% - 6,7\% = 62,4\%$.



Sl. 1.9: Ploščina pod standardno normalno krivuljo med $x = -1,5$ in $x = 0,5$.

PRIMER: Na sliki 1.10 je histogram za inteligenčni kvocient ljudi v določeni skupini in starosti, ki se mu tesno prilega normalna krivulja s sredino v $\mu = 100$ in standardnim odklonom $\sigma = 15$. Zanima nas, kolikšen odstotek posameznikov ima IQ večji od 125. Vemo, da je ta odstotek enak odstotku ploščine histograma desno od 125, tega pa lahko nadomestimo z odstotkom ploščine pod prilegajočo se normalno krivuljo. Tabele za standardno normalno krivuljo ne moremo uporabiti neposredno, ampak moramo najprej 125 pretvoriti v *standardne enote*.



Sl. 1.10: Histogram za inteligenčni kvocient s prilegajočo se normalno krivuljo.

Postopek je preprost: dani vrednosti odštejemo povprečje vseh vrednosti v populaciji in razliko delimo s standardnim odklonom. V našem primeru dobimo, da je 125 v standardnih enotah enako $(125 - 100)/15 = 1,67$. Iz tabele za to vrednost dobimo, da je ploščina pod standardno normalno krivuljo levo od te vrednosti približno 95% (za x smo vzeli 1,65, ki je najbliže 1,67), kar je odstotek ljudi z IQ, nižjim od 125. Torej ima 5% ljudi IQ višji od 125.

Postavimo si še nekoliko težje vprašanje. Kolikšen je 99. kvantil za histogram na sliki 1.10? Iščemo vrednost IQ, pod katero je 99% posameznikov, nad njo pa samo 1% izbrancev. Iz tabele za standardno normalno porazdelitev razberemo, da je levo od $x = 2,35$ približno 99% vse ploščine. Torej je 2,35 vrednost, ki jo iščemo, izražena v standardnih enotah. Pretvorimo jo v prvotne enote za IQ: 2,35 moramo pomnožiti s standardnim odklonom in prišteti povprečje. Dobimo $2,35 \cdot 15 + 100 = 35 + 100 = 135$. Zdaj lahko odgovorimo, da je 99. kvantil na histogramu vrednost IQ = 135.

Za izračun ploščin pod normalno krivuljo s parametroma μ in σ lahko uporabljamo tabelo za standardno normalno porazdelitev, če meje, med katerimi nas zanima ploščina, pretvorimo v standardne enote. Vrednost x pretvorimo v standardne enote po formuli

$$s = \frac{x - \mu}{\sigma}\,.$$

Obratno pretvorimo vrednost v standardnih enotah s v originalne enote po formuli

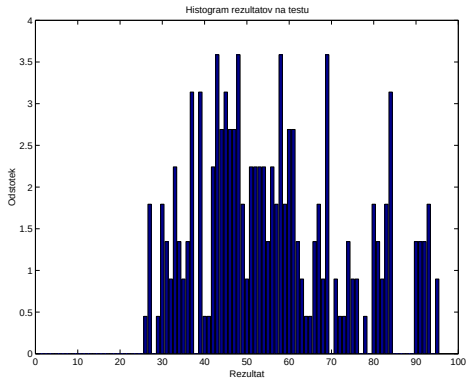
$$x = s \cdot \sigma + \mu\,.$$

1. Kandidati za službe v mestni upravi Chicaga, ZDA, so izbrani na podlagi rezultatov na izpitu, ki je namenjen oceni njihove kvalifikacije. V spodnji tabeli so rezultati 223 kandidatov na izpitu 23. marca 1966. Na voljo je bilo 15 delovnih mest. Na podlagi spodnjih podatkov so bili člani izpitne komisije obtoženi goljufije. Narišite histogram za podatke in povejte razlog za obtožbo.[6]

26	27	27	27	27	29	30	30	30	30	31	31	31	32	32
33	33	33	33	33	34	34	34	35	35	36	36	36	37	37
37	37	37	37	37	39	39	39	39	39	39	39	40	41	42
42	42	42	42	43	43	43	43	43	43	43	43	44	44	44
44	44	44	45	45	45	45	45	45	45	46	46	46	46	46
46	47	47	47	47	47	47	48	48	48	48	48	48	48	48
49	49	49	49	50	50	51	51	51	51	51	52	52	52	52
52	53	53	53	53	53	54	54	54	54	54	55	55	55	56
56	56	56	56	57	57	57	57	58	58	58	58	58	58	58
58	59	59	59	59	60	60	60	60	60	60	61	61	61	61
61	61	62	62	62	63	63	64	65	66	66	66	67	67	67
67	68	68	69	69	69	69	69	69	69	69	71	71	72	73
74	74	74	75	75	76	76	78	80	80	80	80	81	81	81
82	82	83	83	83	83	84	84	84	84	84	84	84	90	90
90	91	91	91	92	92	92	93	93	93	93	95	95		

Rešitev: Razlog za obtožbo goljufije je bilo najverjetneje to, da je natanko 15 *kandidatov doseglo* 90 *točk ali več, vsi drugi pa manj kot* 84*. To kaže, da ocenjevanje ni bilo pošteno, temveč, da je bilo v prid vnaprej izbranim kandidatom.*

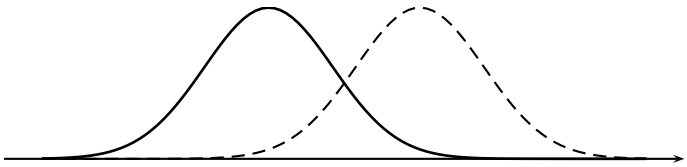
[6]Vir: Freedman, Pisani, Purves, Adhikari, STATISTICS, 2nd Ed., W. W. Norton & Company, 1991



2. Slika 1.3 prikazuje histogram za število članov v gospodinjstvih za leti 1981 in 1991. Na podlagi histograma poskusite ugotoviti, ali število članov v gospodinjstvih upada ali narašča.

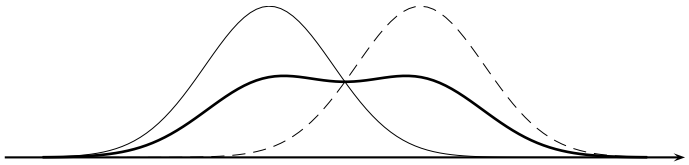
   Rešitev: Dejstvo, da je histogram za leto 1981 višji pri višjih vrednostih spremenljivke, histogram za leto 1991 pa pri nižjih vrednostih, kaže na to, da število članov gospodinjstev v povprečju upada.

3. Na spodnji sliki sta shematično prikazana histograma za telesno višino moških med 25. in 30. letom (črtkano) in žensk (polno) v istem starostnem razredu.



Privzemite, da je v populaciji polovica žensk in polovica moških. Kakšna je po vašem mnenju oblika histograma za telesno višino populacije, ki jo dobimo tako, da združimo populaciji moških in žensk? Narišite ta histogram shematično in utemeljite odgovor!

Rešitev: Višina novega histograma je tako povsod preprosto povprečje višin posameznih histogramov. Približna oblika je na spodnji sliki.



4. Povprečna neto plača v podjetju je bila v danem letu 72.738 SIT. Povprečna neto plača za moške v tem podjetju je bila 80.200 SIT, za ženske pa 71.100 SIT. Ugotovite odstotek moških v podjetju.

   Rešitev: Recimo, da je odstotek moških v podjetju enak x, torej je odstotek žensk $100\% - x$. *Zveza med povprečno plačo v celotnem podjetju in povprečnima plačama moških in žensk posebej je*

   $$x \cdot 80.200 + (100\% - x) \cdot 71.100 = 72.738\,.$$

   Iz te zveze dobimo $x = 18\%$.

5. Predpostavite, da imate dane vrednosti spremenljivke za vsako enoto v populaciji. Kaj se zgodi s povprečjem in standardnim odklonom te spremenljivke:

a. če vrednosti spremenljivke za vsako enoto prištejemo isto število?
b. če vrednost spremenljivke za vsako enoto pomnožimo z istim številom?
c. če vrednosti spremenljivke za vsako enoto odštejemo povprečje?
d. če vrednost spremenljivke za vsako enoto delimo s standardnim odklonom?

Rešitev:

a. Povprečje se poveča za isto število, standardni odklon se ne spremeni.
b. Povprečje in standardni odklon se pomnožita z istim številom.
c. Novo povprečje je 0*, standardni odklon se ne spremeni.*
d. Novo povprečje je staro povprečje, deljeno s standardnim odklonom, nov standardni odklon je 1*.*

6. Janez Novak je na maturi iz fizike dosegel 84 točk, iz matematike pa 90 točk. Rezultati na maturi iz fizike so bili porazdeljeni približno normalno s povprečjem 76 in standardnim odklonom 10, rezultati na maturi iz matematike pa so bili prav tako približno normalno porazdeljeni s povprečjem 82 točk in standardnim odklonom 16. Pri katerem predmetu se je Janez Novak odrezal bolje? Premislite, kako bi primerjali razultata iz matematike in fizike. Utemeljite vašo izbiro.

Rešitev: Izračunamo kvantila Janezovih dosežkov pri matematiki in fiziki:

$$\begin{aligned} \textit{Fizika:} \quad & \frac{84-76}{10} = 0,8 \\ \textit{Matematika:} \quad & \frac{90-82}{16} = 0,5 \end{aligned}$$

Pri fiziki je bil Janez Novak boljši od 79 *odstotkov kolegov, medtem ko je bil pri matematiki boljši samo od* 69 *odstotkov kolegov. Boljši je bil torej pri fiziki.*

7. Na sistematskem pregledu 11-letnih dečkov v nekem mestu je zdravnik ugotovil, da je njihova višina približno normalno porazdeljena s povprečjem 146 cm in standardnim odklonom 8 cm.

   a. Približno kolikšen odstotek pregledanih dečkov je visokih med 138 cm in 154 cm? Kolikšen odstotek pa jih je visokih med 130 cm in 162 cm?

   b. Če bi zdravnik moral uganiti višino naključno izbranega dečka, preden ga vidi, kaj bi mu svetovali? Za koliko bi pričakovali, da se bo zdravnik zmotil? Za 2 cm, za 4 cm ali za 8 cm?

Rešitev:

a. Izračunamo ploščino pod normalno krivuljo. Meji za višino prej spremenimo v standardne enote.

$$\begin{aligned} \frac{138-146}{8} &= -1 \\ \frac{154-146}{8} &= 1 \end{aligned}$$

Ploščina pod standardno normalno krivuljo med tema dvema mejama je 68%*, torej je bilo* 68% *dečkov visokih med* 138 *cm in* 154 *cm.*

Tudi za drugi primer meji spremenimo v standardne enote

$$\begin{aligned} \frac{130-146}{8} &= -2 \\ \frac{162-146}{8} &= 2 \end{aligned}$$

Ploščina pod standardno normalno krivuljo med tema dvema mejama je $95,5\%$.

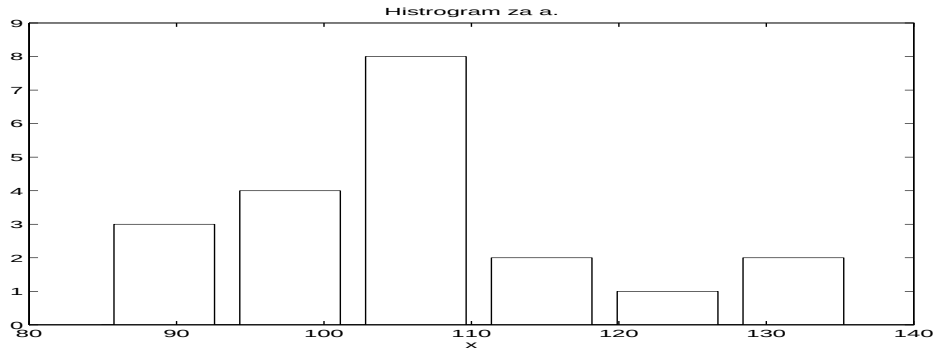
b. *Zdravnik bi lahko ugibal, da bo deček visok, kot je povprečje, torej* 146 *cm. Pričakovana napaka, ki bi jo zdravnik pri tem naredil, bi bila standardni odklon, torej* 8 *cm.*

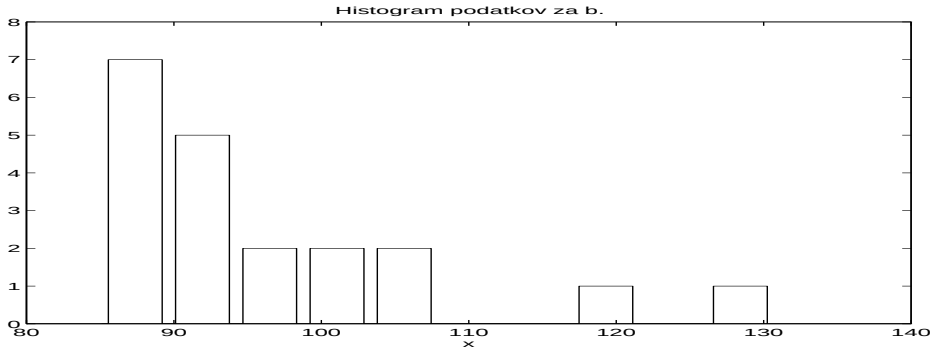
8. Iz treh populacij naključno izberemo 20 enot. Vrednosti spremenljivke za izbrane enote iz posameznih populacij so naslednje:

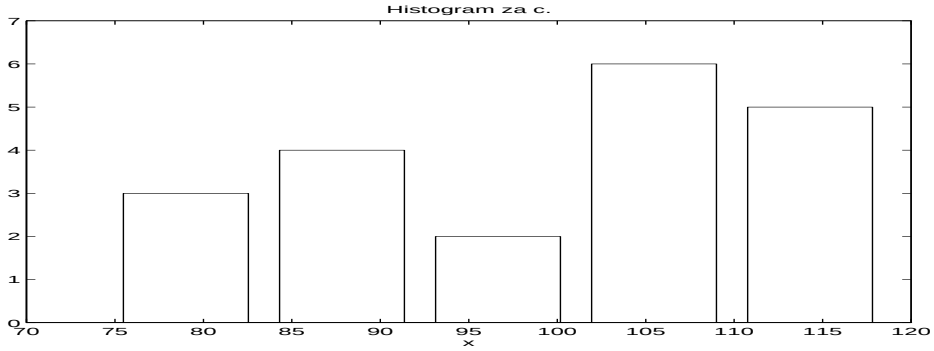
   - Populacija 1: 103,8 86,7 89,7 107,4 127,9 123,9 109,9 95,4 97,3 98,6 97,9 84,9 107,8 108,5 102,0 108,3 113,2 136,1 107,4 114,5
   - Populacija 2: 94,6 120,7 91,3 98,1 90,3 86,4 89,0 105,0 130,7 89,5 101,7 91,8 89,1 85,1 100,1 105,9 85,2 89,8 89,2 91,4
   - Populacija 3: 97,7 86,1 102,9 113,6 107,8 118,7 80,4 90,8 109,0 98,2 86,8 115,0 112,9 108,3 87,7 108,3 79,3 74,6 108,7 112,1

   Za katero populacijo bi rekli, da je spremenljivka normalno porazdeljena? Odgovor utemeljite!

   Rešitev: Narišemo tri histograme in pogledamo, kateri bi bil najbolj podoben normalni porazdelitvi.







Za presojo potrebujemo še nekaj količin. Strnimo jih v tabelo

	Povprečje	Std. odklon	Znotraj 1σ	Znotraj 2σ
a.	106,06	12,8	14(70%)	19(95%)
b.	96,26	11,6	18(80%)	18(80%)
c.	99,95	13,2	13(65%)	20(100%)

Primer b. lahko takoj izključimo zaradi oblike histograma. Poleg tega bi znotraj 1σ *pričakovali okrog* 68% *enot znotraj* 2σ *pa okrog* 95% *enot. Pri ostalih*

dveh histogramih je znotraj 1σ *in* 2σ *približno pravi odsotek enot, vendar je tako ujemanje odstotkov kot tudi oblika histograma bolj združljiva z normalnim histogramom za histogram a.*

POGLAVJE 2

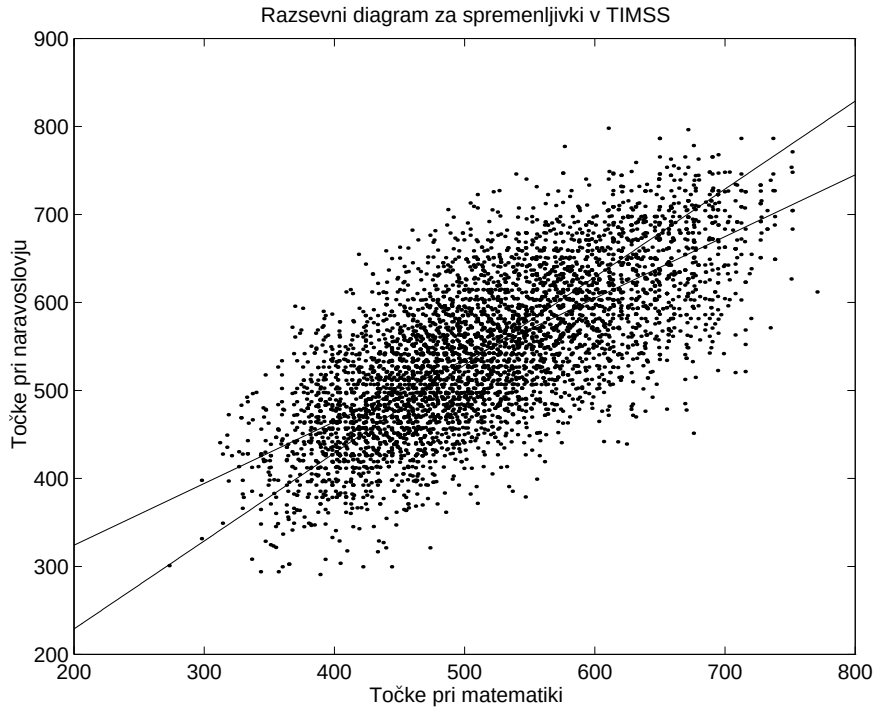
Korelacija in regresija

Različne spremenljivke niso med sabo neodvisne količine, temveč med njimi pogosto obstaja neka zveza. Tako na primer obstaja zveza med vloženim delom in dodano vrednostjo, med kupno močjo prebivalstva in bruto dohodkom na prebivalca, primerov pa je seveda še več. Statistika je orodje, s pomočjo katerega lahko te zveze opišemo in raziskujemo. Če namreč poznamo eno od količin, ne moremo vedno natančno napovedati druge, lahko pa povemo, katera vrednost druge količine je najverjetnejša. V tem poglavju bomo najprej vpeljali korelacijski koeficient, ki meri povezanost dveh količin, potem pa se bomo ukvarjali s pojmom regresije in regresijske premice.

2.1 Uvodni primeri

2.1.1 TIMSS v Sloveniji

Ali lahko pričakujemo, da sta dosežka učenca pri matematiki in pri naravoslovju povezana? V vsakdanjem življenju bi rekli, da so otroci, ki so nadarjeni za matematiko, običajno boljši tudi pri naravoslovnih predmetih. V prvem poglavju smo omenili raziskavo TIMSS, v kateri je bil eden od namenov oceniti znanje sedmošolcev in osmošolcev pri matematiki in pri naravoslovnih predmetih. Za 5606 slovenskih učencev imamo zbrane njihove dosežke na obeh preizkusih znanja. Najprej želimo predstaviti podatke s primernim grafom.



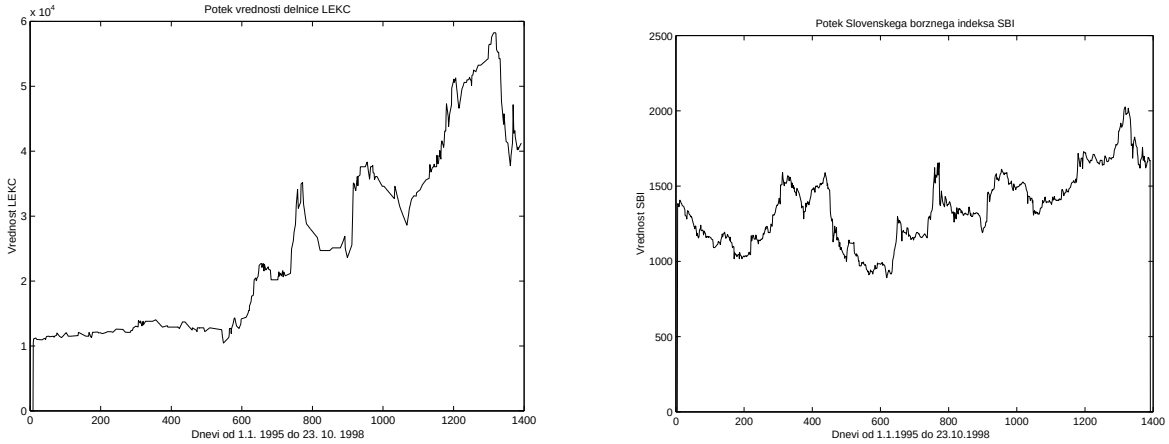
Sl. 2.1: Razsevni grafikon za dosežke pri matematiki in pri naravoslovju.

Za raziskovanje povezave med matematičnim in naravoslovnim dosežkom bomo izbrali *razsevni grafikon*. Na razsevnem grafikonu na sliki 2.1 vsaka točka predstavlja enega učenca. Koordinata na osi x je dosežek učenca pri matematiki, koordinata na osi y pa dosežek na preizkusu iz naravoslovja.

Kaj bi lahko rekli na podlagi razsevnega grafikona? Na prvi pogled lahko ugotovimo, da "oblak" točk teži navzgor, ko se pomikamo po osi x proti desni. To bi pomenilo, da so v splošnem učenci z boljšimi rezultati na preizkusu iz matematike boljši tudi pri preizkusu iz naravoslovja. Na grafikonu lahko vidimo, da povezava med spremenljivkama ni popolna. S tem želimo reči, da na podlagi dosežka iz matematike ni mogoče povsem natančno napovedati dosežka pri naravoslovju. Med 5606 učenci so tudi taki, ki so bili na preizkusu iz matematike slabši od povprečja, hkrati pa boljši od povprečja na preizkusu iz naravoslovja. Ne moremo torej reči, da poznavanje ene spremenljivke natančno določa tudi drugo. Lahko pa trdimo, da so učenci z boljšimi dosežki pri matematiki *v povprečju* boljši tudi pri naravoslovju. Povprečje dosežkov pri naravoslovju za vse učence je bilo 547,8 točke, povprečje pri matematiki pa 518,9 točke. Vzemimo na primer samo učence, katerih dosežek na preizkusu iz matematike je bil med 550 in 560 točkami, torej tiste z zelo dobrimi rezultati. Njihovo povprečje na preizkusu iz naravoslovja je bilo 574,13 točke, kar je občutno višje od celotnega povprečja pri naravoslovju. Kljub temu je med njimi tudi učenec, ki je pri naravoslovju dosegel le 399,2 točke, kar gotovo ni navdušujoč rezultat. Izjava o tem, da so učenci, ki so bili boljši pri matematiki, boljši tudi pri naravoslovju, velja torej le za povprečja.

V tem poglavju bomo uvedli pojma korelacijskega koeficienta in regresijske premice, s katerima lahko natančneje opišemo povezavo med spremenljivkama. S korelacijskim koeficientom in regresijsko premico opisujemo povezave, ki so podobnega tipa kot v zgornjem primeru, torej povezave, ki napovedujejo povprečje ene spremenljivke na podlagi vrednosti druge spremenljivke.

Položnejša premica na sliki 2.1 je *regresijska premica*, druga premica pa je tako imenovana *simetrala*, glede na katero je "oblak" točk v razsevnem grafikonu simetričen. Kako ti premici določimo, bomo opisali v kasnejših razdelkih tega poglavja.



Sl. 2.2: Gibanje cen delnice LEKC in gibanje slovenskega borznega indeksa SBI za obdobje 1. 1. 1995 do 23. 10. 1998.

2.1.2 Gibanje cen vrednostnih papirjev

Na sliki 2.2 sta predstavljena gibanje cen delnice LEKC in gibanje slovenskega borznega indeksa SBI v obdobju od 1. januarja 1995 do 23. oktobra 1998. Vprašamo se lahko, ali je delnica LEKC "sledila" gibanju celotnega trga, ki ga povzema indeks SBI. Tukaj imamo opraviti samo z eno delnico, na tržišču, posebej na razvitih borzah, pa je delnic lahko zelo veliko. Borzne posrednike in ponudnike vzajemnih skladov, torej naborov velikega števila delnic, zanima, do kolikšne mere kaka delnica sledi trgu. Ta informacija je uporabna pri sestavljanju nabora delnic, tako imenovanega portfelja, ki bi bil optimalen v smislu, da je pri želeni donosnosti čim manj tvegan.

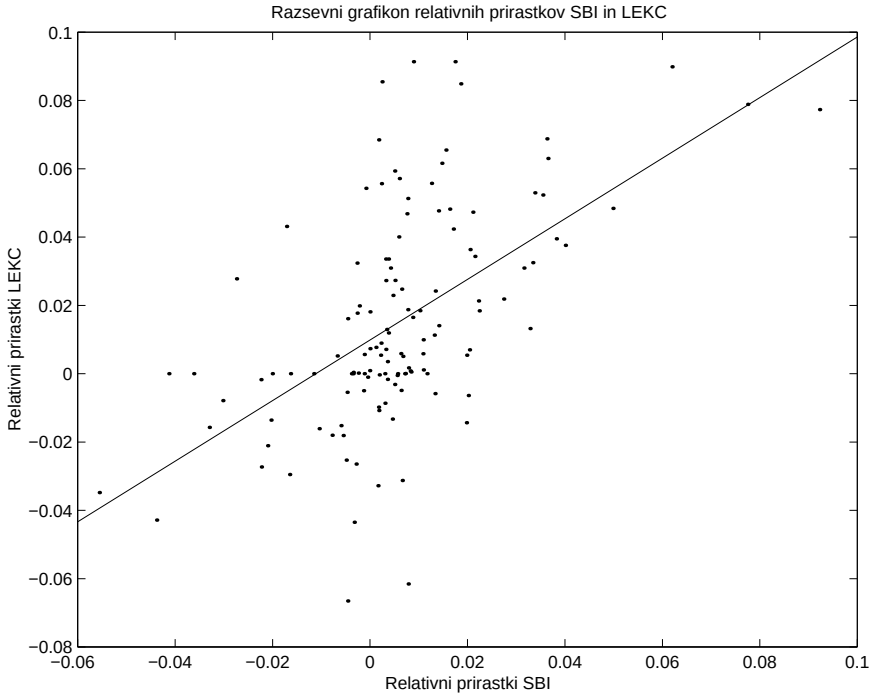
V kolikšni meri delnica sledi trgu, poskušamo opisati tako, da primerjamo relativne dnevne prirastke delnice s prirastki indeksa. Relativni dnevni prirastek je preprosto delež, za katerega se je spremenila cena delnice v danem dnevu. Kot primer vzemimo podatka, da je bila vrednost delnice LEKC dne 9. 1. 1995 10.500 SIT, dne 10. 1. 1995 pa 11.100 SIT. Relativna dnevna sprememba R dne 10. 1. 1995 je potem

$$R = \frac{11.100 - 10.500}{10.500} = 0,057\,.$$

Če označimo ceno delnice LEKC na dan i s S_i, potem je relativni prirastek na dan i

$$R_i = \frac{S_i - S_{i-1}}{S_{i-1}} .$$

Podobno lahko izračunamo relativne dnevne prirastke za katerokoli drugo delnico in tudi za indeks SBI. Zdaj lahko primerjamo relativne dnevne prirastke za SBI in za delnico LEKC ter na podlagi tega sklepamo o "moči" povezave. Na sliki 2.3 je razsevni grafikon za relativne dnevne prirastke SBI in relativne dnevne prirastke LEKC. Narisana premica je, kot v prvem uvodnem primeru, regresijska premica. Njen pomen bomo pojasnili kasneje.



Sl. 2.3: Razsevni grafikon za relativne dnevne prirastke SBI in LEKC.

Kaj bi lahko sklepali iz tega razsevnega grafikona? Vsekakor obstaja povezava, saj

že na prvi pogled lahko rečemo, da so večji prirastki za SBI povezani z večjimi prirastki za LEKC. Bolj natančna mera za moč te povezave pa je korelacijski koeficient, ki ga bomo obravnavali v naslednjem razdelku.

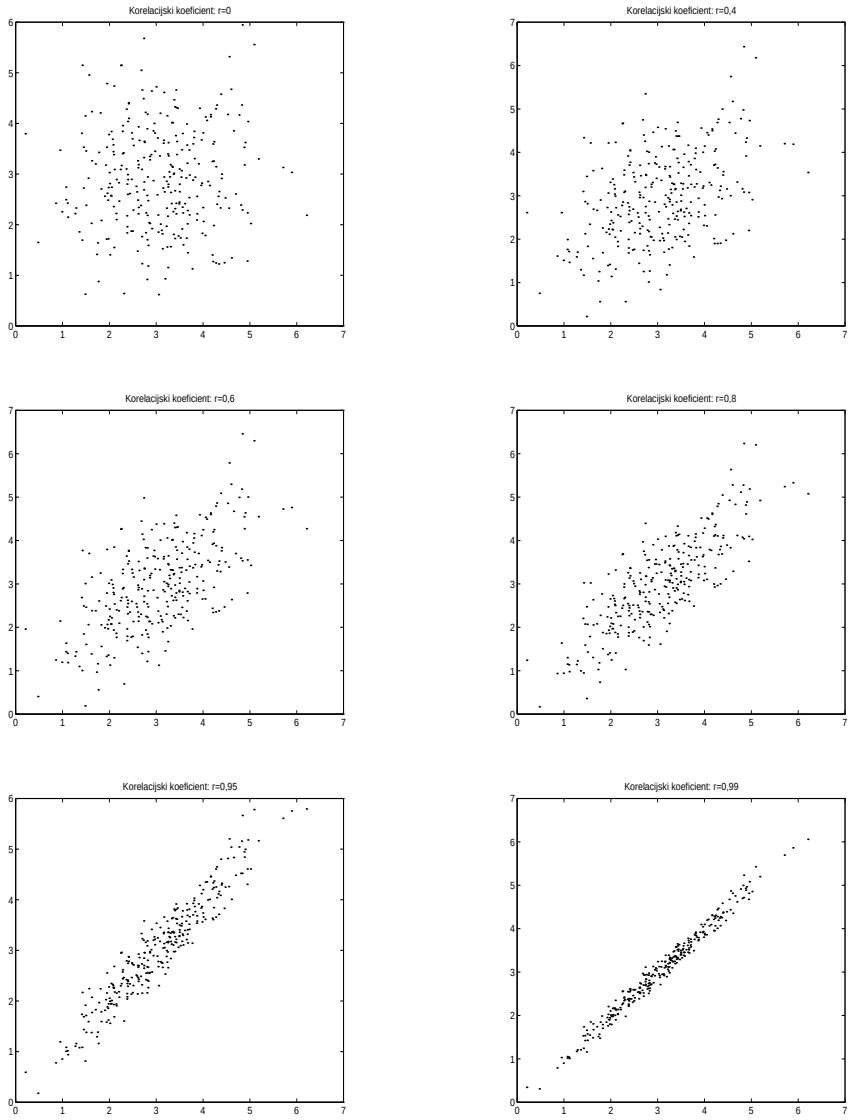
2.2 Korelacijski koeficient

V uvodnih primerih smo postavili vprašanje, kako bi s številom povzeli povezanost dveh spremenljivk. Imamo populacijo in dve spremenljivki, torej za vsako enoto dve vrednosti. Za lažje izražanje bomo prvo od obeh spremenljivk označili z X, drugo pa z Y. Ti dve spremenljivki sta lahko bolj ali manj povezani. Poznavanje vrednosti ene spremenljivke nam lahko nekaj pove o vrednostih druge. Od mere za povezanost bi želeli, da bi na primerni lestvici izmerila tako moč kot smer povezanosti med spremenljivkama. Smer povezanosti je pozitivna, če so večje vrednosti spremenljivke X povezane z v povprečju večjimi vrednostmi spremenljivke Y. Smer povezanosti je negativna, če so večje vrednosti spremenljivke X povezane z v povprečju manjšimi vrednostmi spremenljivke Y.

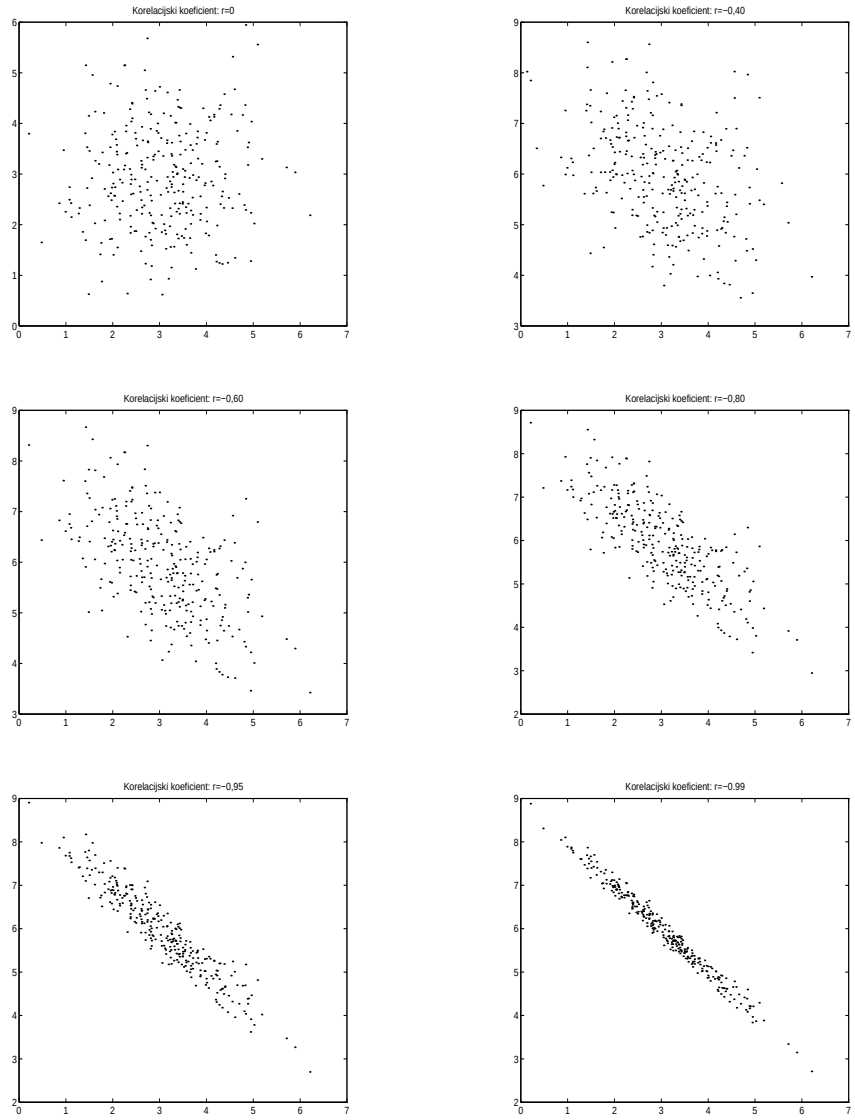
Odgovor na zastavljeno vprašanje je *korelacijski koeficient*, ki ga je vpeljal angleški statistik sir Francis Galton (1822–1911). Ta mera povezanosti je vedno na intervalu od -1 do 1. Absolutna vrednost korelacijskega koeficienta je povzetek moči, predznak pa smeri povezave. Preden se lotimo računanja korelacijskega koeficienta, si oglejmo na slikah 2.4 in 2.5 nekaj razsevnih grafikonov in pripadajočih korelacijskih koeficientov.

V skladu z zahtevami, katerim naj bi ustrezala vsaka mera povezanosti, vidimo, da je pri večjem korelacijskem koeficientu oblak točk bolj zgoščen, torej lahko natančneje napovemo vrednost ene spremenljivke na podlagi vrednosti druge.

Lotiti se moramo še izračuna korelacijskega koeficienta. Kot podatke imamo dane pare vrednosti spremenljivk za posamezne enote, po eno vrednost za X in eno vrednost za Y. Koraki, ki jih moramo pri računanju narediti, so naslednji.



Sl. 2.4: Razsevni grafikoni s pripadajočimi pozitivnimi korelacijskimi koeficienti.



Sl. 2.5: Razsevni grafikoni s pripadajočimi negativnimi korelacijskimi koeficienti.

- Izračunamo povprečji $\bar{x}$ in $\bar{y}$ za vrednosti spremenljivk X in Y.
- Izračunamo standardna odklona σ_x in σ_y za vrednosti spremenljivk X in Y.
- Vrednosti spremenljivke X pretvorimo v standardne enote. Prav tako vrednosti Y pretvorimo v standardne enote.
- Korelacijski koeficient je povprečje produktov vrednosti X v standardnih enotah in pripadajočih vrednosti Y v standardnih enotah.

PRIMER: Oglejmo si zgornji postopek na dejanskih podatkih. Recimo, da imamo telesne višine za 5 naključno izbranih očetov in sinov. Podatki v centimetrih so v spodnji tabeli.

Očetje	173	175	184	166	172
Sinovi	171	181	176	170	182

Najprej moramo dane podatke za vsako spremenljivko pretvoriti v standardne enote. Označimo velikost očetov z X in velikost sinov z Y. Povprečje telesnih višin očetov označimo z $\bar{x}$ in njihov standardni odklon s σ_x. Privesek x pri tej zadnji oznaki pomeni, da imamo v mislih standardni odklon za vrednosti spremenljivke X. Podobno označimo povprečje vrednosti spremenljivke Y z $\bar{y}$ in njihov standardni odklon s σ_y. Iz danih podatkov dobimo naslednje vrednosti:

$$\bar{x} = 174 \quad \sigma_x = 5,83$$
$$\bar{y} = 176 \quad \sigma_y = 4,94$$

Sedaj lahko telesne višine očetov in sinov pretvorimo v standardne enote, kot je opisano v prvem poglavju. Dobimo

Očetje	$-0,1715$	$0,1715$	$1,715344$	$-1,3722$	$-0,3431$
Sinovi	$-1,0121$	$1,01214$	0	$-1,2146$	$1,2146$

Korelacijski koeficient je povprečje petih produktov, ki jih dobimo, če množimo standardizirane telesne višine očetov in sinov. Produkti so

$$0{,}1736 \quad 0{,}1736 \quad 0 \quad 1{,}6668 \quad -0,4167\ ,$$

njihovo povprečje pa je 0,32. Za korelacijski koeficient bomo uporabljali oznako r. V zgornjem primeru je potem $r = 0,32$.

Korelacijski koeficient je mera linearne povezanosti med dvema spremenljivkama. Linearne zato, ker meri "zgoščenost" razsevnega grafikona okoli premice. Njegova vrednost je vedno med -1 in 1, pri čemer absolutna vrednost koeficienta meri moč povezanosti in predznak smer. Čim bliže je korelacijski koeficient -1 ali 1, tem bolj zanesljivo lahko iz vrednosti ene spremenljivke napovemo vrednost druge spremenljivke. Če je povezava med spremenljivkama linearna in je korelacijski koeficient blizu 0, nam poznavanje vrednosti ene spremenljivke ne pomaga napovedati vrednosti druge.

PRIMER: Drugi uvodni primer je govoril o tem, do kolikšne mere je gibanje cene delnice LEKC povezano z gibanjem celotnega trga, kar povzema slovenski borzni indeks SBI.

Delnica	Kor. koef.	Delnica	Kor. koef.
BTC	0.53	NBS8	−0.09
DAD	−0.04	PETG	0.50
DRPG	0.56	PFNP	−0.28
LEKA	0.81	RARG	0.52
LEKC	0.56	SKBR	0.62

Tabela 2.1: Korelacijski koeficienti med relativnimi prirastki SBI in relativnimi prirastki delnic.

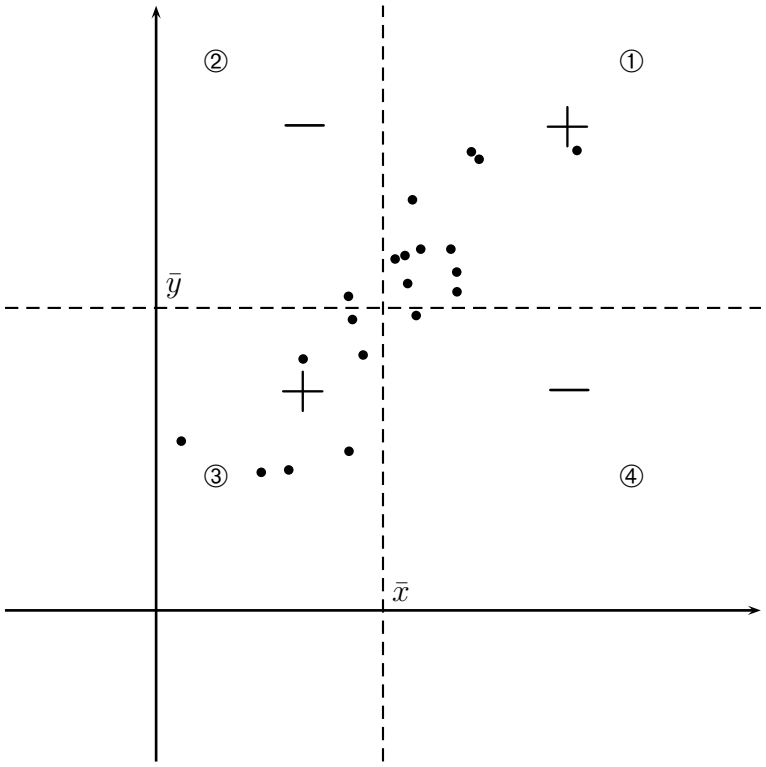
Razsevni grafikon za podatke je na sliki 2.3. Odgovor na zastavljeno vprašanje o moči povezave je, kot smo omenili že v uvodu, ravno korelacijski koeficient. Iz

podatkov lahko izračunamo, da je ta koeficient $r = 0,56$. Za primerjavo tabela 2.1 vsebuje še nekaj drugih korelacijskih koeficientov med relativnimi dnevnimi prirastki SBI in relativnimi prirastki posameznih delnic.

Poskusimo si ponazoriti idejo v ozadju korelacijskega koeficienta. Na sliki 2.6 je hipotetični razsevni grafikon. Črtkani črti se sekata v točki, katere koordinata x je povprečje vrednosti spremenljivke X, koordinata y pa povprečje vrednosti spremenljivke Y. Ko vrednosti spremenljivk X in Y pretvarjamo v standardne enote, v delu, označenem z ①, dobimo pozitivne standardne enote preprosto zato, ker so te vrednosti nad povprečjem. V delu ravnine, označenem s ③, dobimo za standardne enote obeh spremenljivk negativne vrednosti, vendar so produkti teh vrednosti med sabo spet pozitivni. V delih ravnine, ki sta označena z ② in ④, pa so produkti vrednosti standardnih enot negativni, ker je za točke v teh delih vrednost X v standardnih enotah vedno drugače predznačena kot vrednost Y. Če torej oblak točk sili navzgor, pozitivni produkti prevladajo in dobimo pozitiven korelacijski koeficient. To velja še posebej, če je oblak točk zelo ozek, ker je tedaj točk, ki prispevajo negativne produkte v formuli za korelacijski koeficient, zelo malo v primerjavi s tistimi, ki prispevajo pozitivne produkte. Bralec se bo zlahka prepričal, da zgodba velja z negativnim predznakom, če oblak točk sili navzdol.

Opozoriti moramo, da korelacijski koeficient meri le linearno povezanost. Z drugimi besedami, korelacijski koeficient je dobra mera povezanosti le tedaj, ko je razsevni grafikon ovalne oblike. Pogosto imamo razsevni grafikon za majhno število točk in moramo sami presoditi, ali je privzetek o ovalnosti smiseln. Kot primer, da je predpostavka o ovalnosti razsevnega grafikona zares potrebna, si oglejmo razsevni grafikon na sliki 2.7. Očitno sta spremenljivki povezani, saj lahko vrednost koordinate y neke točke precej zanesljivo napovemo na podlagi vrednosti koordinate x, vendar pa je korelacijski koeficient le 0,006, kar bi kazalo na zelo majhno povezanost. Razlog je ta, da razsevni grafikon ni ovalne oblike in korelacijski koeficient v tem primeru ni pravi način za raziskovanje povezanosti spremenljivk.

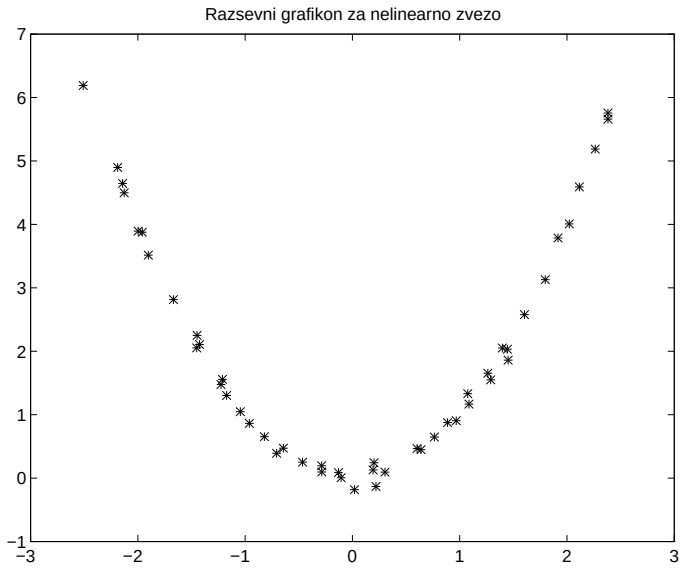
Korelacijski koeficient meri linearno povezanost med spremenljivkama.



Sl. 2.6: Pomen korelacijskega koeficienta.

Opozorimo še na eno dejstvo. Korelacijski koeficient meri povezanost med spremenljivkama, ne trdi pa, da je naraščanje vrednosti ene spremenljivke tudi vzrok za naraščanje ali padanje vrednosti druge. Kot primer si oglejmo naslednji preprost razmislek.

PRIMER: Korelacijski koeficient med višino plače in delovno dobo za populacijo moških med 25. in 35. letom je $-0,3$. Ali to pomeni, da daljša delovna doba povzroča manjšo plačo? Ne, razlog je verjetno ta, da je krajša delovna doba za moške v dani starosti povezana z daljšim trajanjem izobraževanja, to pa ima po drugi strani za posledico v povprečju višje plače.



Sl. 2.7: Primer nelinearne povezanosti.

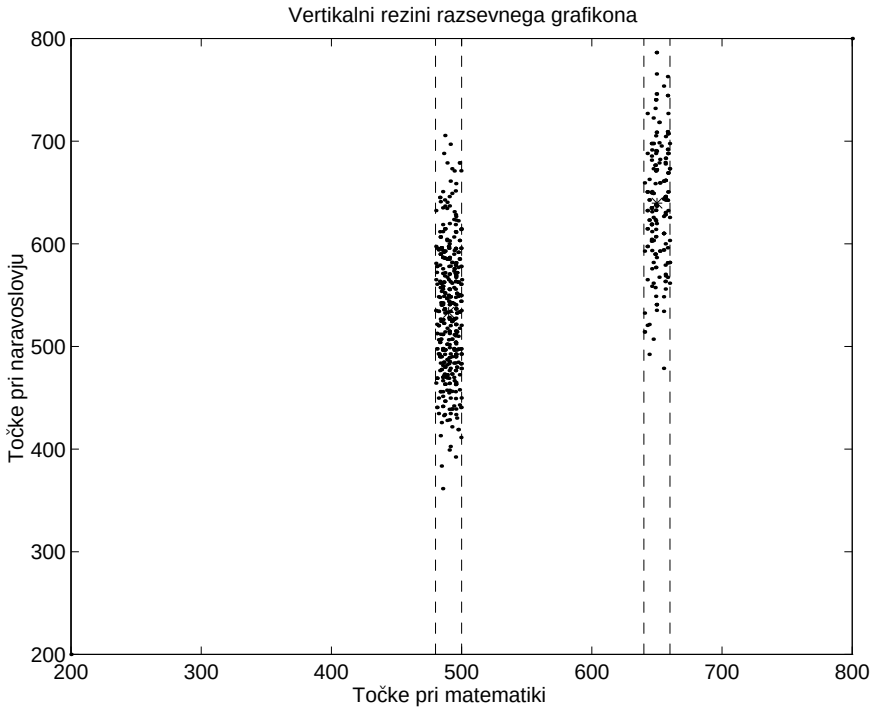
Plača in delovna doba sta sicer povezani, vendar daljša delovna doba ni vzrok za v povprečju manjše plače, temveč je verjetno prej posledica krajše dobe izobraževanja.

Korelacijski koeficient pove, da sta spremenljivki povezani, ne pove pa, da je naraščanje vrednosti ene spremenljivke tudi vzrok za naraščanje ali padanje vrednosti druge spremenljivke. Povezanost ne pomeni, da obstaja tudi vzročna povezava.

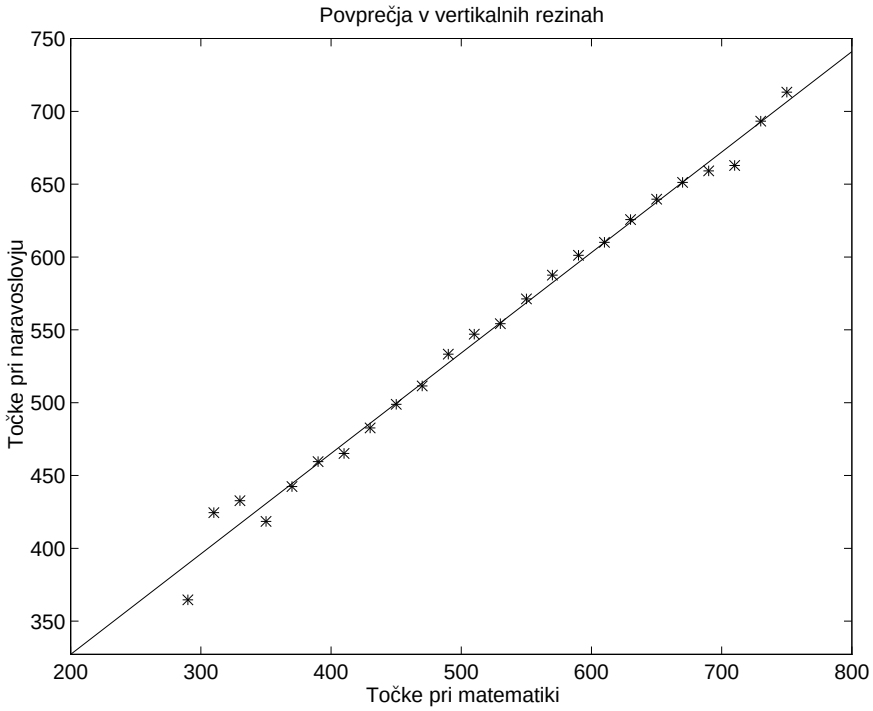
2.3 Regresijska premica

Iz podatkov TIMSS na sliki 2.1 smo ugotovili, da imajo učenci z boljšim dosežkom pri matematiki v povprečju tudi boljši dosežek pri naravoslovju. Na sliki 2.8 je del razsevnega grafikona za učence, ki so imeli dosežek pri matematiki med 480 in 500 ali med 640 in 660. Z zvezdico sta označeni povprečji dosežkov pri naravoslovju za učence, ki "padejo" v navpično rezino razsevnega grafikona. Po pričakovanju je pri učencih z dosežkom med 640 in 660 na preizkusu iz matematike povprečje na preizkusu znanja iz naravoslovja višje kot pri učencih z matematičnim dosežkom med 480 in 500.

Isti razmislek bi lahko naredili za poljubne rezine. Če razsevni grafikon razrežemo v navpične rezine, široke po 20 točk, in za vsako rezino izračunamo povprečje dosežkov



Sl. 2.8: Vertikalni rezini razsevnega grafikona.



Sl. 2.9: Povprečja pri naravoslovju po podskupinah.

pri naravoslovju, dobimo sliko 2.9. Zvezdice na sliki označujejo povprečja dosežkov iz naravoslovja za vsako rezino po 20 točk. Tako recimo zvezdica nad 490 ponazarja povprečje dosežkov iz naravoslovja za učence, ki so pri matematiki dosegli med 480 in 500 točk.

Kot je razvidno s slike 2.9, ležijo povprečja skoraj natanko na premici. To premico, ki povezuje povprečja, bomo imenovali *regresijska premica*. Beseda regresija ima latinski koren, ki pomeni vračanje na nekaj prejšnjega. Tukaj se "vračamo" od spremenljivke Y k spremenljivki X in poskušamo pojasniti spreminjanje povprečij spremenljivke Y po navpišnih rezinah s spremenljivko X.

Regresijska premica povezuje povprečja po posameznih navpičnih rezinah. Rezino v razsevnem grafikonu razumemo kot množico tistih enot, za katere je vrednost spremenljivke X določena. Grafično so to tiste točke na razsevnem grafikonu, ki ležijo točno nad dano vrednostjo spremenljivke X.

V splošnem povprečja spremenljivke Y po navpičnih rezinah ne ležijo točno na premici. Zavedati se moramo, da so v mnogih primerih podatki za razsevni grafikon podani le za del populacije, recimo za vzorec. Tako kot vzorčne ocene niso vedno enake dejanskemu povprečju, tudi vzorčna povprečja po navpičnih rezinah ne ležijo natančno na premici. Kljub temu nam privzetek, da dejanska povprečja ležijo na premici, pomaga dobro oceniti naklon regresijske premice, ta količina pa je pogosto pomembna za razumevanje strukture podatkov ali za napovedovanje vrednosti spremenljivke Y, če poznamo vrednost spremenljivke X.

Poglejmo sedaj, kako regresijsko premico izračunamo. Premico v ravnini opišemo z enačbo

$$y = \alpha + \beta x\,,$$

kjer je α presečišče z osjo y, β pa je naklon premice. Naklon pove, kolikšen je prirastek premice v smeri osi y, če se v smeri osi x premaknemo za 1 v desno.

Kako izračunamo naklon regresijske premice? Tega vprašanja se lotimo na primeru podatkov TIMSS. S slike 2.1 vidimo, da je povečanje spremenljivke X povezano s povečanjem povprečja spremenljivke Y v posameznih rezinah. Ugotovili smo tudi, da moč te povezave merimo s korelacijskim koeficientom. Če se vrednosti spremenljivke X povečajo za 1 standardno enoto, označili smo jo s σ_x, ali lahko pričakujemo povečanje povprečja Y za σ_y? Z drugimi besedami, če si ogledamo navpični rezini, ki sta za σ_x narazen, ali lahko pričakujemo, da se povprečji spremenljivke Y v teh dveh rezinah razlikujeta za σ_y? Odgovor je ne! Če bi to držalo, bi povprečja po rezinah ležala na simetrali, ki je vrisana kot bolj strma premica na razsevnem grafikonu 2.1. Tukaj nastopi korelacijski koeficient. Povečanje vrednosti spremenljivke X za σ_x je povezano s povečanjem povprečja spremenljivke Y za $r \cdot \sigma_y$, kjer je r korelacijski koeficient.

Temu razmisleku lahko dodamo še opazko, da bodo učenci, ki so na preizkusu iz matematike dosegli povprečen rezultat, tudi na preizkusu iz naravoslovja v povprečju dosegli povprečen rezultat. Za regresijsko premico to pomeni, da gre skozi točko $(\bar{x}, \bar{y})$, kjer $\bar{x}$ označuje povprečje vrednosti za spremenljivko X in $\bar{y}$ povprečje vrednosti spremenljivke Y. Oznaki $\bar{x}$ in $\bar{y}$ sta standardni pri regresiji in ju bomo uporabljali za povprečji namesto oznake μ iz prvega poglavja. Iz povedanega sledi, da lahko ocenimo naklon regresijske premice in njeno presečišče z osjo y po formulah

$$\hat{\beta} = r\frac{\sigma_y}{\sigma_x} \qquad \text{in} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}\,.$$

V zgornji formuli govorimo o $\hat{\beta}$ in $\hat{\alpha}$. Uporabljali bomo oznako $\hat{}$, da bomo ločili ocene od pravih vrednosti. Postavimo se namreč na stališče, da so enote, na osnovi katerih smo narisali razsevni grafikon, le vzorec celotne populacije. Vzorčenje bomo podrobneje obravnavali v 4. poglavju. Če imamo na voljo le vzorec, potem lahko govorimo samo o ocenah količin, ki jih želimo izračunati. Uvedene oznake nas spominjajo na to dejstvo. Večinoma imamo v razsevnem grafikonu le vzorčne podatke in skušamo oceniti naklon regresijske premice za celotno populacijo. Zgornje formule nam iz podatkov *ocenijo* naklon in presečišče regresijske premice z osjo y.

Regresijska premica gre skozi točko $(\bar{x}, \bar{y})$ v ravnini, prirastek σ_x v vrednosti spremenljivke X pa je povezan s prirastkom povprečja vrednosti spremenljivke Y za $r \cdot \sigma_y$. Tukaj sta σ_x in σ_y standardna odklona vrednosti spremenljivke X in vrednosti spremenljivke Y, $\bar{x}$ in $\bar{y}$ povprečji in r korelacijski koeficient. Iz tega dobimo še oceni

$$\hat{\beta} = r\frac{\sigma_y}{\sigma_x} \qquad \text{in} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

za naklon regresijske premice in njeno presečišče z osjo y.

PRIMER: Oglejmo si še enkrat podatke TIMSS . Z X smo označili dosežek učenca na

preizkusu iz matematike, z Y pa njegov dosežek pri naravoslovju. V spodnji tabeli so navedene vse količine, ki jih potrebujemo za izračun $\hat{\beta}$ in $\hat{\alpha}$.

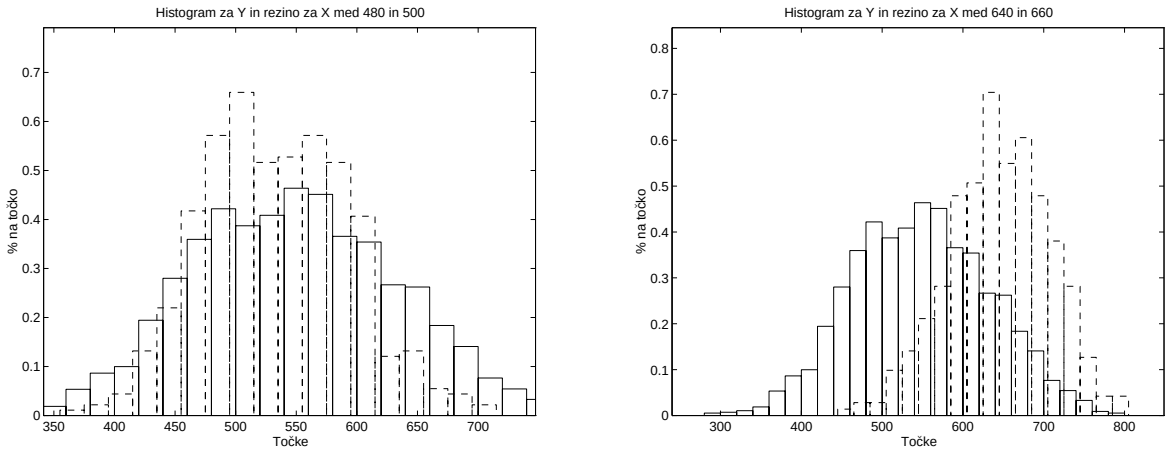
$\bar{x} = 518,9$ točke $\quad \sigma_x = 85,7$ točke
$\bar{y} = 547,8$ točke $\quad \sigma_y = 84,3$ točke
$r = 0,71$

Prirastek za 85,7 točke pri matematiki je torej povezan s prirastkom za 0,71 · 84,3 točke pri naravoslovju. Izračunajmo še ocene za naklon regresijske premice in presečišče z osjo y, pa dobimo

$$\hat{\beta} = 0,71 \cdot \frac{84,3}{85,7} = 0,70 \qquad \text{in} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 547,8 - 0,70 \cdot 518,9 = 183,69\,.$$

Enačbo regresijske premice zapišemo kot $y = 0,70 \cdot x + 183,69$. Iz ocene $\hat{\beta}$ naklona regresijske premice lahko razberemo, da tisti, ki so na preizkusu iz matematike dosegli eno točko več, na preizkusu iz naravoslovja dosežejo 0,70 točke več. Povezava torej ni taka, da bi točka na enem preizkusu prispevala prirastek za točko na drugem preizkusu.

Posvetimo se še vprašanju, kaj lahko povemo o standardnem odklonu vrednosti spremenljivke Y v posameznih rezinah razsevnega grafikona. Pričakovali bi, da je leta manjši, saj imajo enote v rezini zelo podobne vrednosti spremenljivke X in bi zato pričakovali, da so bolj skupaj tudi vrednosti spremenljivke Y za te enote. Oglejmo si ta razmislek na podatkih raziskave TIMSS . Na sliki 2.10 sta s polno črto narisana histograma za vse vrednosti spremenljivke Y. Na prvem histogramu je dodatno s črtkano črto narisan histogram vrednosti spremenljivke Y za učence iz rezine matematičnih dosežkov med 480 in 500 točkami, na drugem pa je črtkano narisan histogram vrednosti Y za učence iz rezine z dosežkom med 640 in 660 točkami pri matematiki. S slike lahko razberemo, da so histogrami, ki pripadajo rezini učencev z večjim dosežkom pri matematiki, pomaknjeni bolj na desno, kot bi pričakovali, saj vemo, da je tudi njihovo povprečje višje. Opazimo pa še več. Histogrami za rezine po 20 točk imajo tudi manjši standardni odklon za vrednosti spremenljivke Y, kot smo pričakovali. To vidimo po tem, da so črtkani histogrami ožji in višji.



Sl. 2.10: Histogrami za dosežke pri naravoslovju za vse učence in za učence iz posameznih rezin.

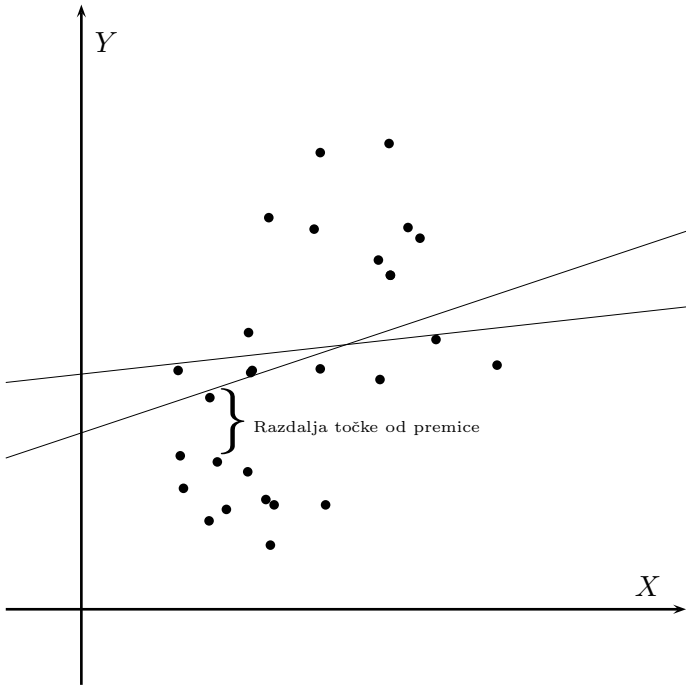
Standardni odklon vrednosti spremenljivke Y v rezinah lahko izračunamo iz že znanih količin. Če je standardni odklon v vsaki rezini enak, potem je to vprašanje gotovo smiselno. Takim razsevnim grafikonom, pri katerih je standardni odklon enak v vsaki rezini, pravimo *homoshedastični*, tistim, pri katerih se standardni odklon od rezine do rezine spreminja, pa *heteroshedastični*. Za homoshedastične razsevne grafikone je značilno, da so ovalne oblike, tako kot grafikon na sliki 2.1. Zanje je mogoče izraziti standardni odklon posameznih rezin na zelo preprost način.

Zgornji obrazec je povezana s še eno interpretacijo regresijske premice. Vsaka točka v razsevnem grafikonu je v navpični smeri od regresijske premice oddaljena za neko razdaljo. Standardni odklon teh razdalj je ravno količina $\sqrt{1-r^2} \cdot \sigma_y$. Res je še več. Recimo, da bi v razsevni grafikon postavili poljubno premico, za vsako točko izračunali razdaljo od premice, to razdaljo kvadrirali in sešteli vse kvadrate. Vsota teh kvadratov bi bila za nekatere premice večja in za nekatere manjša. Premica, pri kateri je ta vsota najmanjša, je regresijska premica, pri kateri je standardni odklon razdalj točk od premice ravno RMS. Zato pogosto srečamo izraz, da naklon regresijske premice ocenimo po *metodi najmanjših kvadratov*. Dokaz te trditve bomo prepustili matematikom in verjeli na besedo.

Za homoshedastične razsevne grafikone velja, da izračunamo standardni odklon znotraj posameznih rezin po formuli

$$RMS = \sqrt{1 - r^2} \cdot \sigma_y \,,$$

kjer je r korelacijski koeficient in σ_y standardni odklon vrednosti spremenljivke Y. Količino RMS imenujemo v statističnem žargonu koren srednjih kvadratov. Izraz prihaja iz engleškega *root mean square*.



Sl. 2.11: Metoda najmanjših kvadratov.

Za regresijsko premico velja, da je vsota kvadratov odklonov točk od premice najmanjša med vsemi možnimi premicami v razsevnem grafikonu. Zato pogosto rečemo, da naklon β in presečišče α z osjo y ocenimo po *metodi najmanjših kvadratov*.

Za homoshedastične razsevne grafikone pogosto privzamemo, da so porazdelitve spremenljivke Y po navpičnih rezinah približno normalne. Kot vemo iz prvega poglavja, normalno porazdelitev poznamo, če poznamo dve količini: povprečje μ in standardni odklon σ. Za spremenljivko Y v rezinah homoshedastičnega razsevnega grafikona povprečje izračunamo z regresijsko premico in standardni odklon s formulo $RMS = \sqrt{1-r^2} \cdot \sigma_y$. Če se histogrami za posamezne rezine prilegajo normalni krivulji, lahko računamo odstotke znotraj rezin s pomočjo normalne krivulje. Za homoshedastične razsevne grafikone dobimo dobre ocene odstotkov.

Primer: Na sliki 2.10 je na levem grafikonu črtkano narisan histogram dosežkov pri naravoslovju za učence, ki so pri matematiki dosegli med 640 in 660 točkami. Kolikšen odstotek teh učencev, bila sta 202, je tudi na preizkusu iz naravoslovja dosegel nadpovprečne rezultate? Nadpovprečni rezultat je tukaj rezultat nad povprečjem pri naravoslovju, ki je bilo $\bar{y} = 547,8$ točke. Najprej ocenimo povprečni rezultat pri naravoslovju za ta 202 učenca. Ker imamo rezino podano med 640 in 660 točkami pri matematiki, bomo kot približek za vrednost spremenljivke X vzeli 650 točk. Dobimo

$$\hat{\beta} \cdot 650 + \hat{\alpha} = 0,70 * 650 + 183,69 = 639,9\,.$$

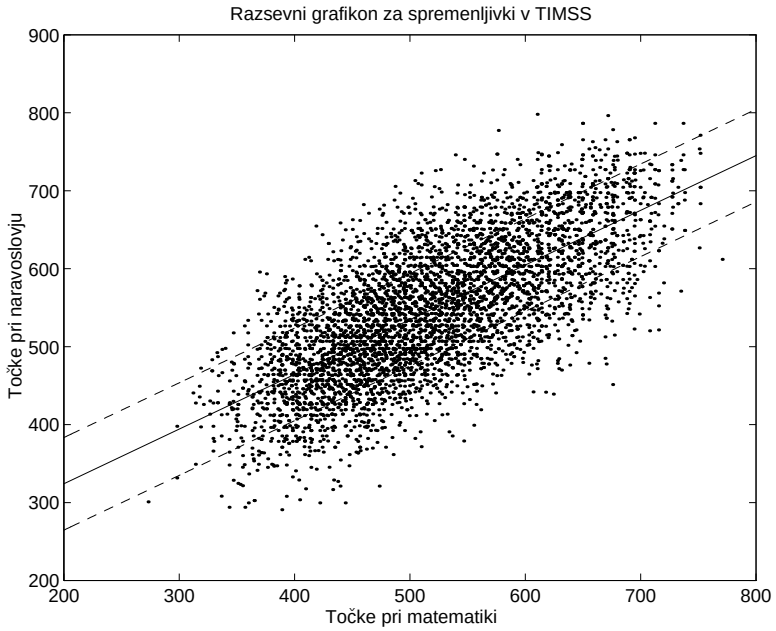
Omenimo, da je ocena presenetljivo dobra, saj je dejansko povprečje za 202 učenca, o katerih teče beseda, 639,7 točke. Standardni odklon dosežkov pri naravoslovju dobimo po formuli

$$RMS = \sqrt{1-r^2} \cdot \sigma_y = \sqrt{1-0,71^2} \cdot 84,3 = 59,4\,.$$

Tudi tukaj je ujemanje presenetljivo dobro, saj je dejanski standardni odklon vrednosti spremenljivke Y za 202 učenca v rezini 59,2. Zdaj lahko uporabimo normalno krivuljo za oceno želenega odstotka. Povprečje 547,8 pretvorimo v standardne enote

$(547,8 - 639,9)/59,4 = -1,55$. Odstotek, ki ga iščemo, je enak odstotku ploščine pod standardno normalno krivuljo desno od zgornje vrednosti. Z uporabo normalne tabele ocenimo ta odstotek s 93,9%. Dejansko je od 202 učencev v obravnavani rezini na preizkusu iz naravoslovja nadpovprečni rezultat doseglo 189 učencev, kar je 93,5%. Tudi tukaj je ujemanje ocen z dejanskimi rezultati zelo dobro.

Podobno kot v zgornjem primeru bi lahko obravnavali poljubno rezino v razsevnem grafikonu in ugotovili, da je v vsaki navpični rezini 68% točk takih, da se vrednost spremenljivke Y od povprečja v rezini razlikuje manj kot za RMS. Ker to velja za vsako rezino, velja tudi za celoten razsevni grafikon. Na sliki 2.12 sta črtkano narisani vzporednici z regresijsko premico, ki sta od nje v navpični smeri oddaljeni točno za RMS. Po prejšnjem razmisleku bi moralo biti v pasu med tema premicama 68% vseh točk v razsevnem grafikonu. Dejansko število teh točk je 3828, kar je 68,3%.



Sl. 2.12: 68% točk je znotraj enega RMS od regresijske premice.

2.4 Uporaba regresijske premice

Ideja regresijske premice je uporabna tako v ekonomiji, zavarovalništvu, tehniki, financah kot tudi v družboslovnih vedah. V veji ekonomije, ki se imenuje *ekonometrija*, so metode linearne regresije osnovno orodje za ocenjevanje ekonomskih kazalcev. Ob koncu poglavja si bomo v tem razdelku ogledali dva primera uporabe regresijske premice.

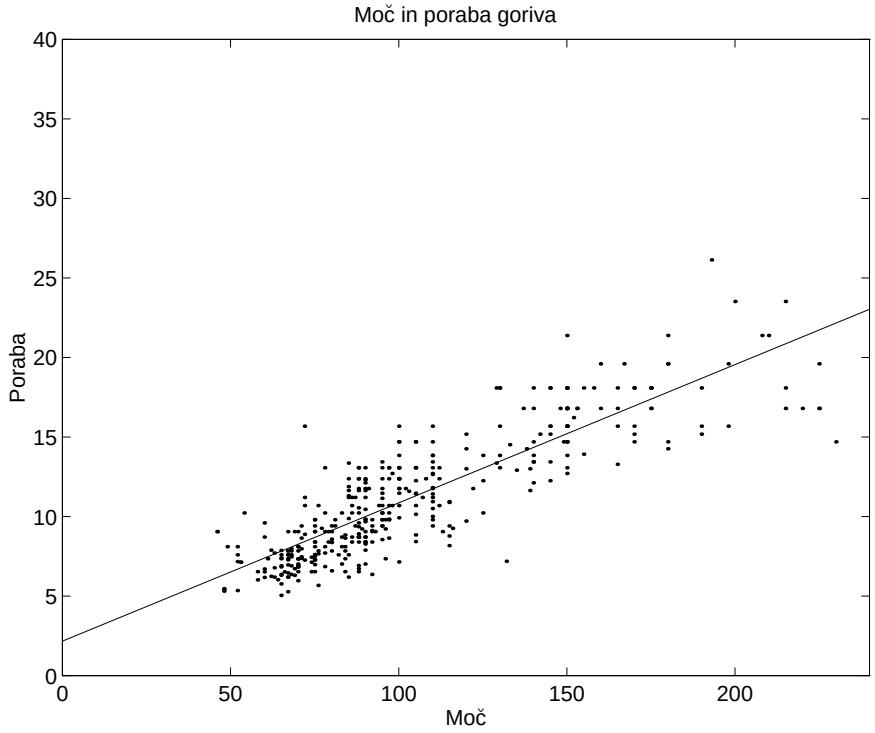
Ko raziskujemo povezavo med dvema spremenljivkama, si običajno eno od spremenljivk izberemo kot *neodvisno ali pojasnjevalno spremenljivko* in z regresijsko premico poskušamo napovedati povprečje druge spremenljivke, ki jo zato poimenujemo odvisna spremenljivka. Katero spremenljivko bomo izbrali za neodvisno, je odvisno od vprašanja, ki si ga zastavljamo, in od vsebinske interpretacije podatkov. Pri določanju neodvisne spremenljivke ne gre za to, da bi bila izbrana spremenljivka neodvisna od vseh drugih spremenljivk, ampak gre le za ustaljen način izražanja. Drugo spremenljivko imenujemo *odvisna spremenljivka*, ker nas zanima "odvisnost" te spremenljivke od izbrane neodvisne spremenljivke. Standardna terminologija je nekoliko nesrečno izbrana, saj smo v razdelku o korelacijskem koeficientu videli, da pogosto ne moremo govoriti o odvisnosti ene spremenljivke od druge, temveč lahko govorimo le o povezanosti. Običajno bomo neodvisno spremenljivko označili z X in odvisno spremenljivko z Y.

Pri uporabi regresijske premice si eno od obeh spremenljivk izberemo za *neodvisno* in na podlagi njenih vrednosti ocenjujemo povprečje ali napovedujemo vrednost druge, *odvisne* spremenljivke. Pri določanju neodvisne in odvisne spremenljivke gre samo za ustaljen način izražanja. Neodvisno spremenljivko običajno označimo z X in odvisno spremenljivko z Y.

Primer: Pri podatkih TIMSS smo si za neodvisno spremenljivko izbrali dosežek na preizkusu iz matematike in za odvisno spremenljivko dosežek na preizkusu iz nar-

avoslovja. Lahko bi si izbrali za neodvisno spremenljivko tudi dosežek pri naravoslovju. Regresijska premica bi bila v tem primeru drugačna od regresijske premice, ki smo jo obravnavali v prejšnjem razdelku, sámo razmišljanje in ideja linerne regresije pa bi bila enaka.

PRIMER: Med močjo motorja pri avtomobilu in porabo goriva seveda obstaja zveza. Pričakovali bi, da močnejši motorji porabijo več goriva. Razsevni grafikon na sliki 2.13 prikazuje moč motorja, merjeno v kilovatih, in porabo goriva v litrih na 100 km za 392 ameriških, evropskih in japonskih avtomobilov. Najprej se moramo odločiti, katero od dveh spremenljivk si bomo izbrali za neodvisno.



Sl. 2.13: Razsevni grafikon za moč motorja in porabo goriva.

Na podlagi vsebine podatkov je bolj smiselno izbrati za neodvisno spremenljivko moč motorja in z regresijsko premico oceniti povprečje porabe goriva. Osnovni povzetki podatkov, ki jih potrebujemo za izračun naklona regresijske premice in RMS, so:

$$\begin{array}{ll} \bar{x} = 104,47 & \sigma_x = 38,49 \\ \bar{y} = 11,25 & \sigma_y = 3,91 \\ r = 0,85 & \end{array}$$

Iz teh osnovnih podatkov lahko ocenimo naklon regresijske premice in presečišče z osjo y z

$$\hat{\beta} = 0,087 \qquad \text{in} \qquad \hat{\alpha} = 2,17\,.$$

Izračunajmo še $RMS = \sqrt{1-0,85^2} \cdot 3,91 = 2,06$. Kaj lahko trdimo na podlagi izračunanih ocen? Splošno izjavo o tem, da močnejši avtomobili porabljajo več, lahko zdaj opremimo tudi s primernimi številkami. Povečanje moči avtomobilov za 1 kilovat v povprečju poveča porabo za 0,087 litra na 100 km. Je ta ocena uporabna? Vsekakor bi bila uporabna za naftne družbe, saj bi lahko na podlagi regresijske premice, histograma moči avtomobilov v prometu in podatkov o prevoženih kilometrih napovedali porabo. Za posameznika je naklon regresijske premice uporaben pri nakupu avtomobila, saj lahko približno ocenimo, ali se poraba goriva pri dani moči bistveno razlikuje od povprečja za avtomobile iste moči. Kolikšno odstopanje bi pomenilo bistveno odstopanje, ocenimo z RMS. Če bi neki avto moči 100 kW porabil 15 litrov goriva na 100 km, bi se morda zamislili. Regresijska premica v razsevnem grafikonu na sliki 2.13 napove za avtomobile z močjo 100 kW porabo v povprečju 10,86 litra na 100 km, RMS pa je, kot smo videli, 2,17. Avto enake moči s porabo 15 litrov na 100 km bi nekako spadal med avtomobile z izjemno visoko porabo, saj je v svoji kategoriji, ali kot bi rekli v jeziku razsevnih grafikonov, v svoji navpični rezini več kot dva standardna odklona nad povprečjem.

Pri uporabi statističnih metod se moramo vedno tudi vprašati, ali so zares primerne za podatke, ki jih obravnavamo. Za presojo o primernosti tega ali onega postopka v statistiki ni enotnega pravila. V našem primeru bi morali preveriti, ali lahko privzamemo, da povprečja spremenljivke Y po posameznih rezinah ležijo na premici.

Razsevni grafikon na sliki 2.13 gotovo kaže na linearno zvezo med spremenljivkama, torej je privzetek utemeljen. Naslednje vprašanje bi bilo, ali je razsevni grafikon homoshedastičen. Na to je nekoliko teže odgovoriti. V razsevnem grafikonu na sliki 2.13 ni nobene rezine, v kateri bi bil standardni odklon odvisne spremenljivke že na videz bistveno večji ali manjši kot v kakšni drugi navpični rezini. Metoda linearne regresije je seveda uporabna tudi za razsevne grafikone, ki niso homoshedastični, le da moramo v tem primeru uporabiti primerne formule za ocenjevanje naklona regresijske premice. Teh formul tukaj ne bomo obravnavali, ker segajo čez okvir pričujočih gradiv. Omenimo le, da dokler standardni odkloni za spremenljivko Y po navpičnih rezinah niso med sabo bistveno različni, lahko uporabimo formule, ki smo jih navedli.

PRIMER: V ekonomski teoriji se pogosto vprašamo, koliko k celotnemu produktu podjetja, ali tudi celotnega gospodarstva, prispevajo posamezni dejavniki, kot so vloženo delo, osnovna sredstva in drugi. Pri tem si pomagamo s produkcijsko funkcijo, ki nam pove, kako vloženo delo in osnovna sredstva vplivajo na celotni produkt. Poznavanje produkcijske funkcije v podjetju nam omogoča ocenjevanje prispevka dela in osnovnih sredstev k celotnemu produktu. Agregatne produkcijske funkcije pa nam omogočijo oceniti prispevkek dela in kapitala k celotnemu produktu gospodarstva v kaki državi. Zahtevnejši bralec bo bolj podrobno razlago našel v večini učbenikov osnov ekonomije[1].

Na konkretnih podatkih si bomo ogledali Cobb-Douglasovo produkcijsko funkcijo. Na voljo imamo podatke o dodani vrednosti, vloženem delu in osnovnih sredstvih v štirih slovenskih podjetjih za obdobje od 1975 do 1985. Dodana vrednost je primerno preračunana v denarne enote za leto 1985, kar omogoča primerjavo kljub takratni visoki inflaciji. Vloženo delo merimo na osnovi vloženih ur dela, ki so jih v proces produkcije vložili zaposleni v podjetju, vrednost osnovnih sredstev pa vzamemo za mero vloženega kapitala. V tabeli 2.2 so prikazani omenjeni podatki.

Produkcijska funkcija je opis odvisnosti med dodano vrednostjo na eni strani ter kapitalom in vloženim delom na drugi. Če označimo dodano vrednost s P, vloženo delo z L in kapital s C, potem Cobb-Douglasova produkcijska funkcija pravi, da med temi količinami velja zveza

$$P = KL^{\beta}C^{1-\beta}\,,$$

[1]Na primer Paul A. Samuelson & William D. Nordhaus, ECONOMICS, McGraw Hill, 1985.

Podjetje 1			Podjetje 2		
Dod. vred.	Ure	Osn. sred.	Dod. vred.	Ure	Osn. sred.
1842977,12	1893	1960560,67	2040090,61	1227	2258107,07
2032220,01	1978	3614022,42	1725235,50	1175	3589397,42
2548712,42	2073	3975380,79	1786674,77	1202	3541184,66
3340485,22	2180	4064767,37	1797138,92	1191	4831685,41
3780860,79	2418	5451114,83	2206903,41	1195	4844338,24
5275763,78	2452	5229302,62	2542427,00	1220	4730904,48
4425268,74	2641	7293819,03	2289878,18	1259	5009050,13
4586158,99	2664	8557207,19	3312290,52	1299	5750000,46
5120452,06	2706	8336195,39	3756506,63	1501	7128678,59
4777002,41	2815	10124830,11	4893422,24	1636	8028116,43
8530131,00	2989	10160107,00	5199309,02	1734	13412480,00
Podjetje 3			Podjetje 4		
Dod, vred,	Ure	Osn, sred,	Dod, vred,	Ure	Osn, sred,
1086079,52	356	2314749,47	2368005,33	1550	1337449,92
1047348,08	386	2064328,82	2742266,96	1455	1286679,27
1209027,81	416	4695394,68	2849316,78	1776	2137971,20
1411120,59	431	4698184,24	4128882,70	1754	2216218,20
1438907,65	437	4646628,02	3748530,74	1704	2128540,76
1154057,48	437	4293430,30	3775082,10	1726	2223972,95
1044886,70	442	4513585,95	3865823,55	1744	2240995,28
1132584,86	461	5109936,19	4220784,33	1746	2540259,85
1315067,60	468	4624410,03	4226701,07	1799	2361357,33
1557022,25	471	5061143,31	3397018,21	1854	2608158,72
2469025,00	507	5151507,00	3654183,00	1912	2744468,00

Tabela 2.2: Dodana vrednost, vloženo delo in osnovna sredstva.

kjer je β koeficient elastičnosti, K pa je neka konstanta. Koeficient elastičnosti nam pove, kako sta delo in kapital udeležena pri nastajanju dodane vrednosti.

Zgornja produkcijska funkcija je idealizacija. Če zberemo dejanske podatke, bomo lahko β le ocenili. Najprej bi želeli problem prevesti na ocenjevanje naklona regresijske premice. Tukaj na pomoč priskoči logaritem, ki zvezo prevede na

$$\log(P) = \log(K) + \beta \log(L) + (1 - \beta) \log(C).$$

To zvezo še nekoliko predelamo in dobimo

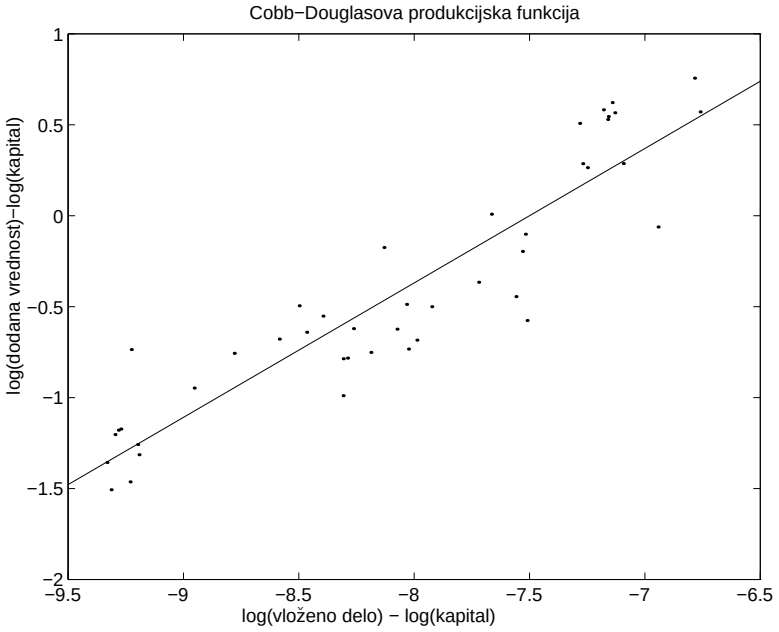
$$\log(P) - \log(C) = \log(K) + \beta(\log(L) - \log(C)),$$

kar lahko napišemo tudi kot

$$\log(P/C) = \log(K) + \beta \log(L/C).$$

Ekonomska teorija nam narekuje, da za odvisno spremenljivko izberemo logaritem kvocienta med dodano vrednostjo in kapitalom, torej $Y = \log(P/C)$, za neodvisno pa logaritem kvocienta med vloženim delom in kapitalom, torej $X = \log(L/C)$. Na podlagi podatkov za 4 slovenska podjetja za 11 let, ki so v tabeli 2.2, lahko izračunamo vrednosti za X in Y. Razsevni grafikon, ki ga dobimo, je na sliki 2.14.

Kaj lahko razberemo z razsevnega grafikona na sliki 2.14? Gotovo to, da sta večje vloženo delo in večji kapital povezana z večjo dodano vrednostjo. Kot običajno pri regresiji tudi tukaj ne moremo trditi, da je zveza med spremenljivkama enolična. Poznavanje vrednosti ene od spremenljivk še ne določa enolično tudi druge spremenljivke. Vse, kar lahko trdimo, je, da je več dela in več kapitala povezano z v povprečju večjo dodano vrednostjo, ni pa to seveda res za vsako posamezno podjetje za vsako leto. Kljub temu lahko trend naraščanja povzamemo z regresijsko premico.



Sl. 2.14: Razsevni grafikon za $\log(P/C)$ in $\log(L/C)$.

Na razsevnem grafikonu je vrisana regresijska premica. Podatke o X in Y strnemo v tabeli.

$$\bar{x} = -8,07 \quad \sigma_x = 0,80$$
$$\bar{y} = -0,42 \quad \sigma_y = 0,64$$
$$r = 0,92$$

Iz teh podatkov lahko ocenimo, da je $\hat{\beta} = 0,73$. Ta naklon je ocena koeficienta elastičnosti.

Kako naj interpretiramo uporabo regresije v tem primeru? Možni sta dve pravilni interpretaciji. Lahko si predstavljamo, da so zgornji podatki vzorec iz neke večje populacije podjetij, za katera velja, da z večanjem vloženega dela in vloženega kapitala

narašča tudi dodana vrednost. Vemo, da zveza ni povsem natančna in da obstaja le naraščajoči trend, ki ga povzamemo z regresijsko premico. Druga možna interpretacija je ta, da predpostavimo, da povezava med vloženim delom in dodano vrednostjo sledi zakonitosti, ki jo predlaga Cobb-Douglasova produkcijska funkcija, vendar pride zaradi slučajnih vplivov do odstopanj. V načelu bi morale vse točke v razsevnem grafikonu, če ne bi bilo slučajnih vplivov, ležati na regresijski premici. Ti slučajni vplivi so lahko posledica različnih načinov vodenja in upravljanja, sprememb na trgu, nihanja cen, konkurence in podobno. Še vedno pa verjamemo v osnovno zakonitost, le zaradi slučajnih vplivov jo vidimo nekoliko zabrisano. Vedno pa ima naklon regresijske premice isti pomen v ekonomski teoriji.

Ekonomska interpretacija naklona β je, da v obravnavanih podjetjih k dodani vrednosti delo v povprečju prispeva 73% in kapital 27%. V ZDA je ocena elastičnosti za veliko število podatkov približno 75%. Zanimivo se je potem vprašati, kako je razdeljen dobiček od dodane vrednosti med delom in kapitalom. Izkaže se, da se porazdelitev presenetljivo dobro ujema z odstotki, ki smo jih predvideli na podlagi produkcijske funkcije, torej 75% dobička dobi delo, 25% pa kapital.

PRIMER: Vrnimo se še enkrat k primeru 2 iz uvodnega dela. Na sliki 2.3 smo imeli razsevni grafikon za relativne dnevne prirastke SBI in delnice LEKC. Na grafikonu je tudi že vrisana regresijska premica. Iz podatkov, ki jih tukaj ne navajamo, lahko izračunamo, da je naklon regresijske premice $\hat{\beta} = 0,88$. Borzni posredniki imenujejo to količino dejansko *beta* in jo uporabljajo pri uvrščanju delnic v portfelj. Ekonomska interpretacija je taka, da 1% relativni prirastek SBI pomeni v povprečju 0,88% relativni prirastek LEKC. Podobno 1% padec SBI pomeni v povprečju 0,88% padec vrednosti delnice LEKC.

1. Zavarovalnice imajo veliko število podatkov o zahtevkih pri nezgodnem zavarovanju in dejansko izplačanih odškodninah. Recimo, da se omejimo na avtomobilsko zavarovanje v določenem obdobju. Korelacijski koeficient med višino zahtevkov in dejansko izplačanimi odškodninami je 0,99. Pojasnite, zakaj ta korelacijski koeficient ni enak 1?

   Rešitev: Zavarovalnice ne izplačajo vseh zahtevkov v višinah, kot so zahtevani. Nekatere zahtevke zavarovalnice zavrnejo, nekatere pa izplačajo v zmanjšanih zneskih.

2. Po popisu prebivalstva ZDA leta 1988 je bila za moške, starejše od 25 let, povprečna starost 47 let in povprečno število let izobrazbe 12,5 leta. Korelacijski koeficient med tema dvema spremenljivkama je bil $r = -0,27$. Na podlagi korelacijskega koeficienta bi torej sklepali, da s časom postanejo moški manj izobraženi.

   Komentirajte zgornjo izjavo v treh stavkih! Kako bi sicer pojasnili negativni predznak korelacijskega koeficienta?

   Rešitev: Negativni korelacijski koeficient kaže na to, da so starejši moški manj izobraženi v primerjavi z mlajšimi. Vzrok tej povezavi je, da se ljudje v preteklosti niso toliko izobraževali kot danes, in ne, da s starostjo postanejo manj izobraženi.

3. Za neki razsevni grafikon lahko rečemo, da je povečanje spremenljivke X za en standardni odklon σ_x povezano s povečanjem povprečja spremenljivke Y za σ_y. Kolikšen je korelacijski koeficient? Podajte kratko utemeljitev.

Rešitev: Korelacijski koeficient je 1. Pri manjšem korelacijskem koeficientu je povečanje spremenljivke Y v povprečju enako $r\sigma_y$.

4. Predpostavite, da imate dane vrednosti dveh spremenljivk za vse enote v populaciji. Kaj se zgodi s korelacijskim koeficientom:

   a. če vrednostim obeh spremenljivk za vsako enoto prištejemo isto število?

   b. če vrednostim ene spremenljivke za vsako enoto prištejemo isto število, drugo spremenljivko pa pustimo nespremenjeno?

   c. če vrednosti obeh spremenljivk za vsako enoto pomnožimo z istim številom?

   d. če vrednosti ene spremenljivke pomnožimo z istim številom, drugo spremenljivko pa pustimo nespremenjeno?

Rešitev: V vseh primerih se korelacijski koeficient ne spremeni, ker vrednosti spremenljivk v standardnih enotah pri navedenih operacijah ostanejo enake.

5. V spodnji tabeli so podatki za višine očetov (spremenljivka X) in višine njihovih najstarejših sinov (spremenljivka Y). Mere so podane v centimetrih.

Očetje	164	159	170	162	172	157	177	167	172	170	175	180
Sinovi	172	167	172	164	175	167	172	164	180	170	172	177

   a. Narišite razsevni grafikon za te podatke.

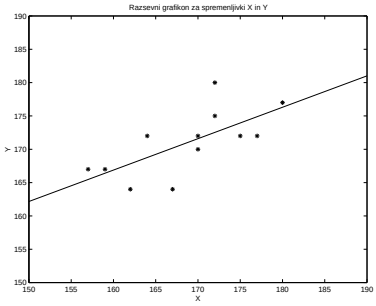
   b. Izračunajte korelacijski koeficient za spremenljivko X glede na spremenljivko Y.

   c. Izračunajte korelacijski koeficient za spremenljivko Y glede na spremenljivko X.

   d. Izračunajte enačbo regresijske premice za Y glede na X. Premico tudi narišite na razsevnem grafikonu.

Rešitev:

a. Razsevni grafikon je na spodnji sliki. Vrisana je tudi regresijska premica.



b., c. Odgovor na obe vprašanji je enak. Izračunati moramo naslednje količine:

$$\begin{aligned}
\bar{x} &= \frac{1}{12}\left(164 + 159 + \ldots + 180\right) = 168,75 \\
\bar{y} &= \frac{1}{12}\left(172 + 167 + \ldots + 177\right) = 171 \\
\sigma_x &= \sqrt{\frac{1}{12}\left((164 - 168,75)^2 + (159 - 168,75)^2 + \ldots + (180 - 168,75)^2\right)} \\
&= \sqrt{\frac{1}{12}\left(22,56 + 95,06 + \ldots + 126,56\right)} = 6,85 \\
\sigma_y &= \sqrt{\frac{1}{12}\left((172 - 171)^2 + (167 - 171)^2 + \ldots + (177 - 171)^2\right)} \\
&= \sqrt{\frac{1}{12}\left(1 + 16 + \ldots + 36\right)} = 4,73
\end{aligned}$$

Vrednosti spremenljivk X in Y spremenimo v standardne enote in jih medsebojno zmnožimo

spremenljivka X:	$-0,69$	$-1,42$	$\ldots$	$1,64$
spremenljivka Y:	$0,21$	$-0,84$	$\ldots$	$1,27$
produkti:	$-0,15$	$1,21$	$\ldots$	$2,09$

Korelacijski koeficient je povprečje izračunanih produktov

$$r = \frac{1}{12}\left(-0,15 + 1,21 + \ldots + 2,09\right) = 0,68$$

d. Izračunati moramo naklon regresijske premice in njeno presečišče z osjo y.

$$\begin{aligned}\hat{\beta} &= r \cdot \frac{\sigma_y}{\sigma_x} = 0,68 \cdot \frac{4,73}{6,85} = 0,47\\ \hat{\alpha} &= \bar{y} - \hat{\beta} \cdot \bar{x} = 171 - 0,47 \cdot 168,75 = 91,69\end{aligned}$$

Enačba regresijske premice je $y = 91,69 + 0,47x$.

6. V raziskavi spreminjanja inteligenčnega kvocienta s starostjo so veliko skupino posameznikov testirali v 18. letu in še enkrat v 35. letu. Dobili so naslednje rezultate:

18. leto:	povprečni IQ	100	std. odklon	15
35. leto:	povprečni IQ	100	std. odklon	15

Korelacijski koeficient med spremenljivkama je bil $r = 0,80$.

a. Ocenite povprečni IQ v 35. letu za vse posameznike, ki so imeli v 18. letu IQ 115.

b. Napovejte IQ v 35. letu za osebo, ki je v 18. letu imela IQ 115.

Rešitev: Odgovor je za oba primera enak, ker povprečje v rezini ocenjujemo na enak način, kot napovedujemo vrednost za neko osebo v isti rezini. IQ v 18. letu določimo za neodvisno spremenljivko in IQ v 35. letu za odvisno spremenljivko. Izračunati moramo regresijsko premico:

$$\begin{aligned}\hat{\beta} &= r\frac{\sigma_y}{\sigma_x} = 0,80\\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 100 - 0,80 \cdot 100 = 20\end{aligned}$$

Enačba regresijske premice je $y = 20 + 0,80x$. *Iz vrednosti neodvisne spremenljivke* $x = 115$ *lahko sedaj napovemo vrednost odvisne spremenljivke* $y = 20 + 0,80 \cdot 115 = 112$.

7. V neki zdravstveni raziskavi v populaciji moških med 18. in 74. letom so dobili naslednje rezultate:

povprečna višina	175 cm	std. odklon	8 cm
povprečna teža	84 kg	std. odklon	15 kg

Korelacijski koeficient med višino in težo moških je bil $r = 0,40$. Ocenite približno težo moških, ki so visoki

a. 174 cm

b. 167 cm

c. 61 cm

d. 0 cm

Komentirajte rezultate v točkah c. in d.

Rešitev: Ker ocenjujemo težo moških glede na njihovo višino, za neodvisno spremenljivko določimo višino in za odvisno spremenljivko težo. Za ocenjevanje bomo najprej potrebovali regresijsko premico.

$$\hat{\beta} = r \cdot \frac{\sigma_y}{\sigma_x} = 0,40 \cdot \frac{15}{8} = 0,75$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 84 - 0,75 \cdot 175 = -47,25$$

Enačba regresijske premice je $y = -47,25 + 0,75x$.

a. *V enačbo regresijske premice vstavimo za vrednost neodvisne spremenljivke* 174 *cm in dobimo oceno za odvisno spremenljivko* $y = -47,25+0,75 \cdot 174 = 83,25$. *Moški v raziskavi, ki so visoki* 174 *cm, so v povprečju težki* 83 *kilogramov.*

b. *Na enak način kot v primeru a. dobimo* $y = 78$.

c. $y = -1,5$.

d. $y = -47,25$.

Vrednosti odvisne spremenljivke v točkah c. in d. lahko izračunamo, vendar so nesmiselne, ker so dani podatki za višino moških zunaj območja razsevnega grafikona za podatke iz raziskave.

8. V neki zdravstveni raziskavi v populaciji moških med 18. in 74. letom so dobili naslednje rezultate:

povprečna višina	175 cm	std. odklon	8 cm
povprečna teža	84 kg	std. odklon	15 kg

Korelacijski koeficient med višino in težo moških je bil $r = 0,40$. Ocenite približno višino moških, ki so težki:

a. 95 kg

b. 80 kg

c. 40 kg

d. 0 kg

Komentirajte rezultate v točkah c. in d.

Rešitev: Ta naloga je podobna prejšnji, le da tokrat iz teže napovemo višino. Za neodvisno spremenljivko X torej določimo težo in za odvisno spremenljivko Y višino moških v raziskavi.

Za vse zahtevane napovedi bomo potrebovali regresijsko premico.

$$\begin{aligned} \hat{\beta} &= r\frac{\sigma_y}{\sigma_x} = 0,21 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 157,36 \end{aligned}$$

Enačba regresijske premice je $y = 157,36 + 0,21x$. Enačba je drugačna od tiste v prejšnji nalogi. Razlog je ta, da sta vlogi višine in teže zamenjani. Vse zahtevane napovedi izračunamo tako, da za neodvisno spremenljivko x vstavimo ustrezne dane vrednosti teže.

a. $y = 177,31$

b. $y = 174,16$

c. $y = 165,76$

d. $y = 157,36$

Odgovora v točkah c. in d. sicer lahko izračunamo, vendar vsaj odgovor v točki d. ni smiseln, ker vrednosti teže X niso iz opazovane populacije moških in zato nam tudi rezultat za višino iz regresijske premice ne pove veliko.

9. Korelacijski koeficient med številom let šolanja moža in številom let šolanja žene je $r = 0,5$. Recimo, da je povprečje za žene in može enako 12 let in standardni odklon 3 leta.

   a. Ocenite povprečno število let šolanja žensk, katerih možje imajo za seboj 18 let šolanja.

   b. Ocenite povprečno število let šolanja mož, katerih žene imajo za seboj 15 let šolanja.

   c. Iz zgornjega bi torej sledilo, da se dobro izobraženi moški poročijo v povprečju z manj izobraženimi ženskami, te pa se poročijo s še manj izobraženimi moškimi. Pojasnite, kje je napaka v razmisleku.

Rešitev:

a. Povprečno število let šolanja žen bomo ocenili z regresijsko premico.

$$\begin{aligned} \hat{\beta} &= r\frac{\sigma_y}{\sigma_x} = 0,5 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 6 \end{aligned}$$

Regresijska premica je $y = 6 + 0,5x$*. Za povprečno število let šolanja žensk, katerih možje imajo* 18 *let šolanja (spremenljivka* X*), tako dobimo* 15 *let (spremenljivka* Y*).*

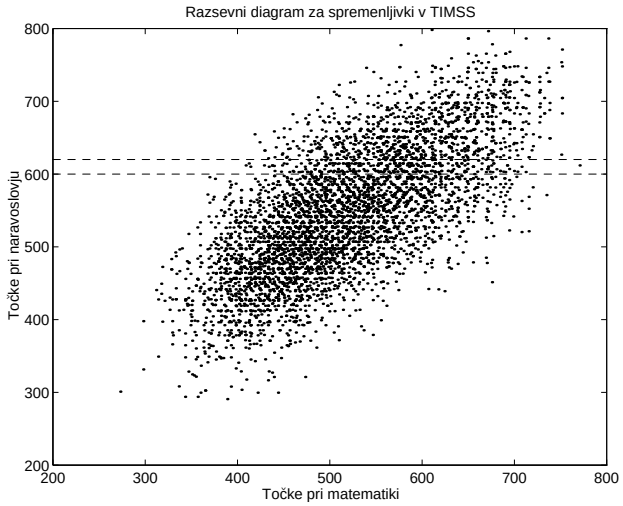
b. V tem primeru za neodvisno spremenljivko X *določimo število let šolanja žen in za odvisno spremenljivko* Y *število let šolanja mož. Ker so povprečja in standardni odkloni tako za može kot za žene enaki, je regresijska premica tudi v tem primeru* $y = 6 + 0,5x$*. Povprečno število let šolanja mož, katerih žene imajo* 15 *let šolanja, je tako* $13,5$ *leta.*

c. *Napaka v razmisleku je privzetek, da obravnavamo iste enote. Enote, ki jih obravnavamo v primeru a., so zakonski pari, v katerih imajo možje* 18 *let šolanja in žene poljubno mnogo let šolanja. V primeru b. pa obravnavamo zakonske pare, v katerih imajo žene* 15 *let šolanja in možje poljubno mnogo let šolanja.*

10. Obravnavali smo raziskavo TIMSS. Na spodnji sliki je še enkrat razsevni grafikon za dosežke pri matematiki in dosežke pri naravoslovju za 5606 slovenskih učencev. Potrebni podatki so v spodnji tabeli.

povprečje pri matematiki	518,9	std. odklon	85,7
povprečje pri naravoslovju	547,9	std. odklon	84,3

$r = 0,71$



Razsevni grafikon za podatke iz raziskave TIMSS

a. Vodoravna rezina prikazuje učence, ki so pri naravoslovju dosegli med 600 in 620 točkami. Ocenite povprečje pri matematiki za te učence.

b. Kolikšen je standardni odklon dosežkov pri matematiki za učence, ki so v vodoravni rezini?

c. Učenci, ki so imeli pri naravoslovju dosežek med 600 in 620 točkami, so pri matematiki v povprečju dosegli 564 točk. Po drugi strani so učenci, ki so pri matematiki dosegli 564 točk, pri naravoslovju v povprečju dosegli 579 točk. Torej je nekaj narobe, saj smo začeli z učenci, ki so imeli pri naravoslovju dosežek med 600 in 620 točkami, in nazadnje ugotovili, da je njihovo povprečje 579 točk. Preverite zgornje ocene in ugotovite, kje je napaka v razmisleku.

Rešitev:

a. *Ker ocenjujemo povprečje pri matematiki in je vrednost pri naravoslovju dana, za neodvisno spremenljivko X določimo dosežek pri naravoslovju in za odvisno spremenljivko Y dosežek pri matematiki. Izračunati moramo regresijsko premico.*

$$\begin{aligned} \hat{\beta} &= r\frac{\sigma_y}{\sigma_x} = 0,72 \\ \hat{\alpha} &= \bar{y} - \hat{\beta} \cdot \bar{x} = 124,4 \end{aligned}$$

Enačba regresijske premice je torej $y = 124,4 + 0,72x$. Za vrednost x vzamemo približek 610 točk pri naravoslovju in tako za povprečni dosežek pri matematiki dobimo $y = 564$ točk.

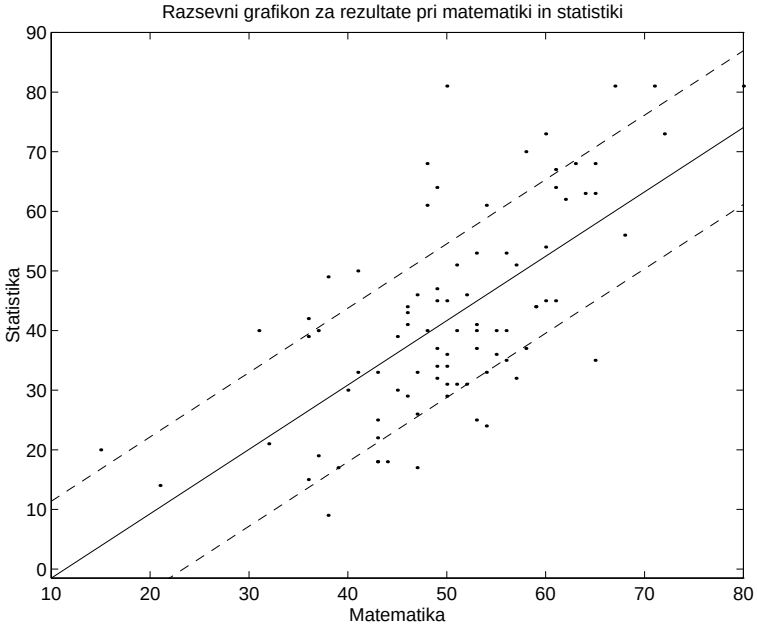
b. *Ker ima razsevni grafikon ovalno obliko, lahko predpostavimo, da je homoshedastičen. V tem primeru je standardni odklon vrednosti spremenljivke Y v posameznih rezinah dan s formulo $RMS = \sqrt{1-r^2}\sigma_y = 60,4$.*

c. *V razmisleku ni napake, ker gledamo dve različni rezini. Primerjajte razmislek z nalogo 9. V prvem primeru gledamo učence v vodoravni rezini, ki so pri naravoslovju dosegli rezultat med 600 in 620 točkami, in računamo njihov*

povprečni dosežek pri matematiki, v drugem primeru pa gledamo učence v navpični rezini, ki so dosegli pri matematiki 564 *točk, in računamo njihovo povprečje pri naravoslovju.*

11. Razsevni grafikon na spodnji sliki prikazuje rezultate na izpitih iz matematike in statistike za 88 študentov univerze v Hullu v Veliki Britaniji. Vrisane so tudi regresijska premica in premici, ki ju dobimo, če regresijski premici prištejemo oziroma odštejemo $\sqrt{1-r^2}\cdot\sigma_y$. Količine, ki jih potrebujete za izračun naklona regresijske premice in njenega presečišča z osjo y, so v spodnji tabeli.

Matematika: $\bar{x} = 50,6 \quad \sigma_x = 10,62$
Statistika: $\bar{y} = 42,3 \quad \sigma_y = 17,25$
$r = 0,66$



Razsevni grafikon za rezultate pri matematiki in statistiki.

a. Privzemite, da je histogram za rezultate iz statistike približno normalen. Približno kolikšen odstotek študentov je na izpitu iz statistike imelo med 29 in 54 točkami? (*Opomba: Dejansko* 37 *od* 88.)

b. Kako bi ocenili odstotek točk v razsevnem grafikonu, ki ležijo med črtkanima premicama? (*Opomba: Dejansko leži med premicama* 61 *točk.*)

c. Recimo, da je imel neki študent pri matematiki 62 točk. Napovejte rezultat na izpitu iz statistike za tega študenta. Za koliko bi pričakovali, da se bo napoved razlikovala od dejanskega rezultata?

Rešitev:

a. *V tem primeru nas zanima samo porazdelitev rezultata pri statistiki in ne povezava med spremenljivkama. Vrednosti* 29 *in* 54 *pretvorimo v standardne enote*

$$\frac{29-42,3}{17,25}=-0,77 \qquad \frac{54-42,3}{17,25}=0,68$$

in izračunamo ploščino pod normalno krivuljo med tema mejama, ki je 53%.

b. *Če privzamemo, da je razsevni grafikon homoshedastičen, je med premicama* 68 *odstotkov točk.*

c. *Rezultat pri statistiki napovemo z regresijsko premico.*

$$\begin{aligned}\hat{\beta} &= r\frac{\sigma_y}{\sigma_x}=1,07\\ \hat{\alpha} &= \bar{y}-\hat{\beta}\bar{x}=-11,8\end{aligned}$$

Enačba regresijske premice je $y=-11,8+1,07x$. *Povprečni rezultat pri statistiki za študente, ki so dosegli* 62 *točk pri matematiki, dobimo tako, da*

vstavimo $x = 62$ v zgornjo enačbo, kar nam da $y = 54,5$ točke. Pričakovali bi, da se bo napoved od dejanskega rezultata razlikovala za $RMS = \sqrt{1 - r^2} \cdot \sigma_y = 13$ točk.

12. V prvem letniku na fakulteti je bil korelacijski koeficient med sprejemnim izpitom in prvim matematičnim izpitom $r = 0,60$. Rezultati na obeh izpitih so bili približno normalno porazdeljeni. Napovejte, na katerem kvantilu bo rezultat matematičnega izpita študenta, ki je imel rezultat na sprejemnem izpitu boljši od:

    a. 50 odstotkov kolegov;

    b. 90 odstotkov kolegov;

    c. 30 odstotkov kolegov;

    d. neznanega števila kolegov.

Rešitev: Za neodvisno spremenljivko X določimo rezultat na sprejemnem izpitu in za odvisno spremenljivko Y rezultat na matematičnem izpitu. Napoved izračunamo z regresijsko premico. Podatki so v standardnih enotah, zato lahko zapišemo

$$\begin{array}{ll} \bar{x} = 0 & \sigma_x = 1 \\ \bar{y} = 0 & \sigma_y = 1 \\ r = 0,60 & \end{array}$$

Izračunati moramo

$$\begin{aligned} \hat{\beta} &= r \cdot \frac{\sigma_y}{\sigma_x} = 0,60 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 0 \end{aligned}$$

Ker so podatki v standardnih enotah, je enačba regresijske premice kar $y = rx$ oziroma, če uporabimo oznake za standardne enote iz prvega poglavja, $s_y = r s_x$.

a. *Iz podatka, da je bil študent na sprejemnem izpitu boljši od* 50 *odstotkov kolegov, iz normalne tabele razberemo standardno enoto za ta odstotek* $s_x = 0$. *Standardna enota za odvisno spremenljivko je potem* $s_y = 0$. *Iz normalne tabele razberemo, da je bil ta študent tudi na matematičnem izpitu boljši od* 50 *odstotkov kolegov.*

b. *Podobno kot v primeru a. iz normalne tabele preberemo, da je standardna enota za* 90. *kvantil* $s_x = 1,28$ *in jo vstavimo v regresijsko premico, da dobimo standardno enoto za spremenljivko* Y $s_y = r \cdot s_x = 0,60 \cdot 1,28 = 0,77$. *Iz normalne tabele odčitamo, da je to* 78. *kvantil. Študent je bil torej boljši od* 78 *odstotkov študentov.*

c. *Standardna enota za* 30. *kvantil v normalni porazdelitvi je* $s_x = -0,52$. *Iz regresijske premice dobimo* $s_y = -0,31$ *in to nam da* 38. *kvantil za spremenljivko* Y.

d. *Če za študenta ne poznamo rezultata na sprejemnem izpitu, je za njegov rezultat na matematičnem izpitu najboljši približek kar povprečje vseh rezultatov na matematičnem izpitu. Napovedali bi, da je bil ta študent na matematičnem izpitu boljši od* 50 *odstotkov svojih kolegov.*

13. V prvem letniku na fakulteti je bil korelacijski koeficient med sprejemnim izpitom in prvim matematičnim izpitom $r = 0,60$. Rezultati na obeh izpitih so bili približno normalno porazdeljeni. Za nekega študenta smo na podlagi njegovega rezultata na sprejemnem izpitu napovedali, da bo imel rezultat na matematičnem izpitu boljši od:

    a. 50 odstotkov kolegov;

    b. 90 odstotkov kolegov;

    c. 30 odstotkov kolegov.

    Na katerem kvantilu je bil rezultat tega študenta na sprejemnem izpitu?

Rešitev: Ta naloga je na videz podobna prejšnji, vendar bomo med računanjem opazili razliko. Ker naloga pravi, da smo matematični izpit napovedovali iz sprejemnega izpita, določimo za neodvisno spremenljivko X sprejemni izpit in za odvisno spremenljivko Y matematični izpit. Tu gre za odvisnost v smislu regresije, četudi je v tej nalogi neznana spremenljivka X in znana Y.

Podatki so v standardnih enotah, zato vemo že iz prejšnje naloge, da ima regresijska premica obliko $s_y = rs_x$. Vrednosti s_y odčitamo iz normalne tabele iz danih odstotkov za spremenljivko Y in iz regresijske premice za posamezne primere izračunamo s_x.

a. *Vrednost $s_y = 0$, iz česar dobimo $s_x = 0$, kar je 50. kvantil. Rezultat na sprejemnem izpitu bi moral biti na 50. kvantilu, če bi hoteli napovedati, da bo rezultat na matematičnem izpitu tudi na 50. kvantilu.*

b. *Vrednost $s_y = 1,28$, iz česar dobimo $s_x = 2,13$, kar je 98. kvantil. Rezultat na sprejemnem izpitu bi moral biti na 98. kvantilu, če bi hoteli napovedati, da bo rezultat na matematičnem izpitu na 90. kvantilu.*

c. *Vrednost $s_y = -0,52$, iz česar dobimo $s_x = -0,87$, kar je 19. kvantil.*

14. Angleški statistik Francis Galton je raziskoval podobnost med starši in otroki. Med drugim je zbral podatke za 1078 očetov in najstarejših sinov. Označimo z X velikost očeta in z Y velikost sina. Potrebne količine so naslednje:

$\bar{x} = 172,7$ cm $\quad \sigma_x = 6,9$ cm
$\bar{y} = 175,3$ cm $\quad \sigma_y = 6,9$ cm
$r = 0,5$

a. Kako visoki so v povprečju sinovi očetov, ki so visoki 190 cm?

b. Ocenite odstotek očetov, visokih 190 cm, katerih sinovi so višji od 190 cm. Privzemite, da je porazdelitev višin sinov za vsako navpično rezino v razsevnem grafikonu približno normalna.

Rešitev:

a. *Iskano povprečje izračunamo z regresijsko premico.*

$$\begin{aligned}\hat{\beta} &= r\frac{\sigma_y}{\sigma_x} = 0,5\\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 89\end{aligned}$$

Enačba regresijske premice je $y = 89 + 0,5x$. *Ko za* x *vstavimo vrednost* 190 *cm, dobimo* $y = 184$, *kar je povprečna višina sinov* 190 *cm visokih očetov.*

b. *Vsak* 190 *cm visok oče ima sina, ki je lahko višji ali nižji od* 190 *cm. Iščemo odstotek očetov, katerih sinovi so višji od* 190 *cm. To je pravzaprav odstotek sinov* 190 *cm visokih očetov, ki so sami višji od* 190 *cm. Ker privzamemo, da so višine sinov v rezini normalno porazdeljene, moramo izračunati ploščino pod normalno krivuljo desno od* 190 *cm.*

V primeru a. smo izračunali, da je povprečje normalne porazdelitve 184 *cm. Standardni odklon je* $RMS = \sqrt{1 - r^2}\sigma_y = 6$. *Meja* 190 *cm v standardnih enotah je* $s = 1$, *kar pomeni, da je 16 odstotkov ploščine pod normalno krivuljo na desno od te meje. Torej je 16 odstotkov sinov* 190 *cm visokih očetov višjih od* 190 *cm.*

15. Govorili smo že o Galtonovem primeru primerjave višine očetov in sinov. Oglejte si naslednji razmislek:

V povprečju so sinovi višjih očetov bliže povprečni višini za sinove, kot so bili njihovi očetje blizu povprečni višini očetov. Podobno so sinovi nižjih očetov zopet v povprečju bliže povprečni višini sinov, kot so očetje blizu povprečni višini očetov. Iz tega bi sklepali, da bo to "gibanje" višin sinov proti povprečju povzročilo manjši standardni odklon za višine sinov. Še posebej bi morda sklepali, da bi po več generacijah bili sinovi enako visoki.

Kot smo videli, je bil standardni odklon v Galtonovem primeru enak 6,9 cm tako za očete kot za sinove. Kje je napaka v zgornjem razmisleku?

Rešitev: Gibanje višin sinov proti povprečni višini je navidezni učinek, ker računamo povprečno višino sinov za vsako rezino za višino očetov. Ker korelacijski koeficient ni enak 1, so ta povprečja po rezinah bliže celotnemu povprečju, kot so višine očetov blizu celotnemu povprečju za očete. Na standardni odklon višin vseh sinov to ne vpliva, ker v rezinah računamo povprečja.

16. Leta 1981 je *Seattle School District* (uprava lokalnih šolskih oblasti) tožil zvezno državo Washington z namenom, da bi dobili več denarja za izobraževanje iz državnih sredstev. Tožniki so trdili, da je rezultat izobraževanja odvisen od višine sredstev, ki jih za to namenja država. Del dokaznega gradiva med sodnim procesom je bila regresijska premica! (Vir: DeGroot, Fienberg & Kadane, Statistics and the Law, Wiley Classic Library, 1994.)

    a. Odvisna spremenljivka Y je bila dosežek učenca na standardiziranem preizkusu iz angleškega jezika. Za neodvisno spremenljivko je bilo več možnosti. Tukaj si bomo ogledali primer, ko je bila neodvisna spremenljivka X število učencev na učitelja na šoli, v katero je hodil učenec, čigar dosežek so merili. Povprečje spremenljivke Y je bilo 60,14 točke s standardnim odklonom $8,63$ točke. Korelacijski koeficient med spremenljivkama je bil $-0,8$. Kolikšen odstotek učencev v šolah, kjer je bilo razmerje med učitelji in učenci enako povprečnemu razmerju za vse šole, je imelo dosežke med približno 52 in 69 točkami? Privzemite, da so porazdelitve po rezinah približno normalne.

    b. Standardni odklon spremenljivke X je bil 10, povprečje pa 28,35. Ocenite povprečni dosežek učencev v šolah, kjer je bilo razmerje med učenci in učitelji enako 35.

    c. Predstavljajte si, da ste sodnik v zgoraj opisanem procesu. Tožnik trdi, da bi z zahtevanimi sredstvi lahko zmanjšal število učencev na učitelja za 5 in s tem občutno izboljšal povprečni dosežek učencev. Bi dali tožniku prav?

Rešitev:

a. *Zahtevani odstotek bomo napovedali z regresijsko premico. Ker gledamo rezino okrog $\bar{x}$, je povprečje spremenljivke Y v tej rezini kar $\bar{y}$. Standardni odklon v rezini je $RMS = \sqrt{1-r^2}\sigma_y = 5,18$. Meje za dosežke na preizkusu pretvorimo v standardne enote:*

$$\frac{52-60,14}{5,18} = -1,57 \qquad \frac{69-60,14}{5,18} = 1,71$$

Ploščina pod normalno krivuljo med tema dvema mejama je 90 odstotkov. Torej je 90 odstotkov učencev v šolah, kjer je bilo razmerje med učitelji in učenci enako povprečnemu razmerju za vse šole, imelo dosežke med 52 in 69 točkami.

b. *V tem primeru moramo regresijsko premico dejansko izračunati.*

$$\begin{aligned} \hat{\beta} &= r\frac{\sigma_y}{\sigma_x} = -0,69 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 79,72 \end{aligned}$$

Enačba regresijske premice je $y = 79,72 - 0,69x$. Za x vstavimo vrednost 35 in dobimo povprečni dosežek $y = 55,6$.

c. *Povprečni dosežek učencev v šolah, ki imajo razmerje med učenci in učitelji za 5 manjše kot splošno povprečje, izračunamo z regresijsko premico in dobimo $y = 63,6$. Ta dosežek je za $3,6$ točke boljši od povprečnega dosežka vseh učencev, vendar ta izboljšava ni večja od RMS, kar pomeni, da ni bistvena.*

POGLAVJE 3

Verjetnost

V vsakdanjem življenju pogosto govorimo o verjetnostih dogodkov. Govorimo o verjetnosti, da bo naslednji dan deževalo, o verjetnostih prometnih nezgod, še najbolj pa je jezik verjetnosti povezan z igrami na srečo. Matematična disciplina, ki ji pravimo verjetnostni račun, se je razvila iz razmišljanja o igrah na srečo, dandanes pa je uporaba verjetnostnega računa široka in sega od kvantne fizike in genetike do upravljanja z vrednostnimi papirji in zavarovalništva. V tem poglavju bomo spoznali osnovne pojme verjetnostnega računa. To nam bo omogočilo boljše razumevanje pojma verjetnosti, saj ga v vsakdanjem življenju uporabljamo precej ohlapno. Naučili se bomo izračunati verjetnosti mnogih zanimivih dogodkov.

3.1 Uvodni primeri

3.1.1 Kockanje v 17. stoletju

V 17. stoletju je bila med italijanskimi kockarji popularna stava na skupno število pik na treh kockah. Možno je bilo staviti na dogodek, da bo skupno število pik enako neki številki. Če je bil seštevek pik po dejanskem metu kock enak tistemu, na katerega je igralec stavil, je le-ta stavo dobil, sicer pa jo je izgubil. Posebej pogosti sta bili stavi na vsoto 9 ali 10. Po razmišljanju takratnih kockarjev naj bi bili obe vsoti enako verjetni in torej stavi enakovredni. Njihov argument sta bila naslednja seznama možnih izidov, ki dajo vsoto 9 oziroma 10.

Vsota 9	Vsota 10
1 2 6	1 4 5
1 3 5	1 3 6
1 4 4	2 2 6
2 3 4	2 3 5
2 2 5	2 4 4
3 3 3	3 3 4

Na podlagi teh dveh seznamov so kockarji sklepali, da je verjetnost vsote 9 enaka verjetnosti vsote 10. Strastni kockarji pa so vendar opazili, da se 10 pojavlja nekaj pogosteje kot 9, kljub zgornji "teoretični" razlagi. To protislovje med kockarsko prakso in takratno teorijo je pripeljalo do precejšnje zmede. V svoji zmedenosti so se kockarji obrnili na slavnega sodobnika Galileja (1564–1642). Ta je problem rešil na zelo preprost način. Enostavno je napisal vse možnosti za izide pri metu treh kock in preštel trojčke z vsoto 9 in trojčke z vsoto 10. Na prvi kocki imamo 6 možnih števil pik, ravno tako na drugi in na tretji. To pomeni, da imamo $6 \cdot 6 \cdot 6 = 216$ možnih izidov, ki so zapisani v tabeli 3.1. V tabeli zlahka preštejemo trojčke z vsoto 9 in tiste z vsoto 10. Trojčki z vsoto 9 so označni z okvirčkom in zlahka preštejemo, da jih je 25. Trojčkov z vsoto 10 je 27.

1 1 1	1 1 2	1 1 3	1 1 4	1 1 5	1 1 6
1 2 1	1 2 2	1 2 3	1 2 4	1 2 5	1 2 6
1 3 1	1 3 2	1 3 3	1 3 4	1 3 5	1 3 6
1 4 1	1 4 2	1 4 3	1 4 4	1 4 5	1 4 6
1 5 1	1 5 2	1 5 3	1 5 4	1 5 5	1 5 6
1 6 1	1 6 2	1 6 3	1 6 4	1 6 5	1 6 6
2 1 1	2 1 2	2 1 3	2 1 4	2 1 5	2 1 6
2 2 1	2 2 2	2 2 3	2 2 4	2 2 5	2 2 6
2 3 1	2 3 2	2 3 3	2 3 4	2 3 5	2 3 6
2 4 1	2 4 2	2 4 3	2 4 4	2 4 5	2 4 6
2 5 1	2 5 2	2 5 3	2 5 4	2 5 5	2 5 6
2 6 1	2 6 2	2 6 3	2 6 4	2 6 5	2 6 6
3 1 1	3 1 2	3 1 3	3 1 4	3 1 5	3 1 6
3 2 1	3 2 2	3 2 3	3 2 4	3 2 5	3 2 6
3 3 1	3 3 2	3 3 3	3 3 4	3 3 5	3 3 6
3 4 1	3 4 2	3 4 3	3 4 4	3 4 5	3 4 6
3 5 1	3 5 2	3 5 3	3 5 4	3 5 5	3 5 6
3 6 1	3 6 2	3 6 3	3 6 4	3 6 5	3 6 6
4 1 1	4 1 2	4 1 3	4 1 4	4 1 5	4 1 6
4 2 1	4 2 2	4 2 3	4 2 4	4 2 5	4 2 6
4 3 1	4 3 2	4 3 3	4 3 4	4 3 5	4 3 6
4 4 1	4 4 2	4 4 3	4 4 4	4 4 5	4 4 6
4 5 1	4 5 2	4 5 3	4 5 4	4 5 5	4 5 6
4 6 1	4 6 2	4 6 3	4 6 4	4 6 5	4 6 6
5 1 1	5 1 2	5 1 3	5 1 4	5 1 5	5 1 6
5 2 1	5 2 2	5 2 3	5 2 4	5 2 5	5 2 6
5 3 1	5 3 2	5 3 3	5 3 4	5 3 5	5 3 6
5 4 1	5 4 2	5 4 3	5 4 4	5 4 5	5 4 6
5 5 1	5 5 2	5 5 3	5 5 4	5 5 5	5 5 6
5 6 1	5 6 2	5 6 3	5 6 4	5 6 5	5 6 6
6 1 1	6 1 2	6 1 3	6 1 4	6 1 5	6 1 6
6 2 1	6 2 2	6 2 3	6 2 4	6 2 5	6 2 6
6 3 1	6 3 2	6 3 3	6 3 4	6 3 5	6 3 6
6 4 1	6 4 2	6 4 3	6 4 4	6 4 5	6 4 6
6 5 1	6 5 2	6 5 3	6 5 4	6 5 5	6 5 6
6 6 1	6 6 2	6 6 3	6 6 4	6 6 5	6 6 6

Tabela 3.1: Galilejev seznam vseh možnih izidov pri metanju 3 kock.

Kje so se kockarji zmotili? Pri zapisovanju ugodnih izidov za posamezno stavo kockarji niso upoštevali različnih možnih vrstnih redov za pike na treh kockah. Ko ohlapno govorimo o verjetnostih, v resnici na tiho razmišljamo o nekem seznamu vseh možnosti, od katerih se bo ena zgodila. V tem primeru so vsi izidi v tabeli 3.1 enako verjetni in je zato potem verjetnost, da dobimo vsoto 10, enaka deležu ugodnih izidov med vsemi izidi. Za vsoto 10 je takih izidov 27, torej je verjetnost dogodka, da bo vsota pik 10, enaka 27/216. Za vsoto 9 pa je ugodnih izidov 25 in tako je verjetnost enaka 25/216. Vidimo, da je bila kockarska praksa boljša od kockarske "teorije".

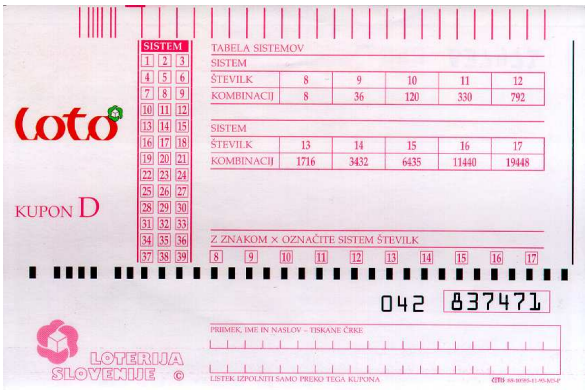
Če so vsi možni izidi nekega poskusa s slučajnim izidom enako verjetni, potem je verjetnost danega dogodka enaka deležu za dogodek ugodnih izidov.

3.1.2 Loterija Republike Slovenije

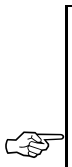
Na sliki 3.1 je eden od lističev Loterije Slovenije. Glede na to, koliko ste pripravljeni vložiti v igro, lahko med 39 številkami izberete od 8 do 17 številk. Za večje število izbranih številk mora biti vložek večji. Na koncu bo izžrebanih 7 številk, in če ste tudi vi s svojo izbiro od 8 do 17 številk pokrili teh 7 številk, ste dobitnik lahko tudi zelo velike vsote. Morda si boste morali to vsoto deliti še s kom, ki je tudi "zadel", ampak nas tukaj predvsem zanimajo verjetnosti. Če kupite listič, vas bo gotovo zanimalo, kolikšna je verjetnost polnega zadetka? Spomnimo se na Galileja in na vse možne izide. Kaj bi bili možni izidi tukaj? Izžrebanih bo 7 številk, torej so vsi možni izidi vsi možni nabori sedmih izmed 39 številk. Oglejmo si tri primere možnih izidov. Vseh možnih izbir 7 izmed 39 številk je zelo zelo veliko. Za izračun, koliko je teh možnih izbir, potrebujemo formulo za število možnih izbir k predmetov izmed n predmetov. V matematičnem jeziku temu rečemo, da računamo število možnih *kombinacij* k elementov izmed n elementov in to število označimo s C_n^k.

1	2	3	4	5	6	7
1	2	5	7	8	9	12
5	13	17	21	24	37	38

Tabela 3.2: Primeri možnih izidov pri žrebanju.



Sl. 3.1: Listič Loterije Slovenije D.



Če izbiramo k predmetov izmed n predmetov, jih lahko izberemo na

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

načinov. Pri tem je $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$, ali z besedami n-fakulteta. Oznaki $\binom{n}{k}$ pravimo binomski simbol.

Formulo za število kombinacij uporabimo za izračun, na koliko načinov lahko izbe-

remo 7 števil izmed 39 na loterijski kartici. Po formuli je

$$C_{39}^{7} = \binom{39}{7} = \frac{39!}{7! \cdot 32!} = \frac{39 \cdot 38 \cdots 33}{7 \cdot 6 \cdots 2 \cdot 1} = 15.380.937\,.$$

Vseh možnih takšnih izbir je torej kar 15.380.937.

Vrnimo se k vprašanju, kolikšna je verjetnost polnega zadetka. Recimo, da smo si že izbrali številke na lističu. Predstavljajte si, da zdaj vneto gledate žrebanje, katerega izid bo eden od vseh možnih sedmerčkov. Med vsemi možnostmi moramo prešteti tiste sedmerčke, ki bodo za nas ugodni. Ker so vsi izidi enako verjetni, bo potem verjetnost zadetka enaka deležu ugodnih izidov. Odgovor je odvisen tudi od tega, koliko številk na lističu smo obkrožili. Začnimo s primerom, ko smo obkrožili 17 številk. Vemo, da bo izžrebanih 7 številk, mi pa bomo zadeli, če bo teh 7 številk med tistimi 17, ki smo jih izbrali. Koliko je naborov po 7 številk, ki so vsebovani med 17 številkami, ki smo jih izbrali? Po formuli za kombinacije je takih sedmerčkov

$$C_{17}^{7} = \binom{17}{7} = 19.448\,.$$

Če torej na lističu izberemo 17 številk, je 19.448 izidov od 15.380.93 za nas ugodnih. Verjetnost zadetka v tem primeru je

$$\frac{\binom{17}{7}}{\binom{39}{7}} = 0,001264422\,.$$

Naše možnosti so približno 1:1000, tudi če smo pripravljeni precej globoko poseči v žep in plačati za 17 številk. Če pa obkrožimo manj kot 17 številk, se verjetnosti dobitka še zmanjšajo. V tabeli 3.3 so verjetnosti dobitkov za različna števila obkroženih številk na lističu.

k	Verjetnost
8	0.0000005201
9	0.0000023406
10	0.0000078019
11	0.0000214551
12	0.0000514923
13	0.0001115667
14	0.0002231334
15	0.0004183750
16	0.0007437778
17	0.0012644223

Tabela 3.3: Verjetnost glavnega dobitka, če na lističu obkrožimo k številk.

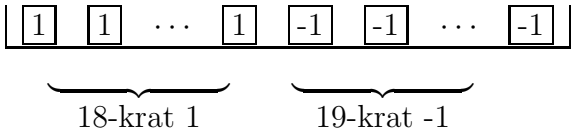
Kot vidimo iz tabele 3.3, so verjetnosti glavnega dobitka precej majhne. Verjetnost glavnega dobitka, če obkrožite samo 8 številk, je komaj 1 v 2 milijonih.

3.2 Verjetnostni modeli

V tem poglavju si bomo ogledali preprost način, kako si predstavimo verjetnosti za posamezne dogodke, ko je v igri naključnost. Predstavljajte si, da igrate ruleto in vsakič stavite na rdeče. Pravila te igre so taka, da vam tedaj, ko se kroglica ustavi na rdečem, vrnejo začetno stavo, povrhu pa dobite še enkrat toliko denarja. Na koncu ste torej za svojo vloženo stavo bogatejši. Če pa se kroglica ustavi na črnem ali zelenem izseku, ste stavo izgubili. Po vsaki igri ste ali bogatejši ali revnejši za svojo stavo. Predpostavimo, da vedno stavimo 1 enoto denarja[1]. Namesto vrtenja dejanske rulete si lahko predstavljamo, da na slepo izbiramo lističe iz škatlice na sliki 3.2. Na ruletnem cilindru je 18 rdečih, 18 črnih in en zelen izsek. Za našo stavo na rdeče je "dobrih" 18 izsekov in 19 "slabih". Čeprav je dejansko igranje rulete morda bolj zabavno, je naključno izbiranje lističa iz škatlice popolnoma enakovredno, vsaj kar se

[1]Ena enota denarja v Perli v Novi Gorici pomeni 2 EVRA

tiče verjetnosti dogodkov. Ko potegnemo listič, dobimo toliko denarja, kolikor piše na lističu, in igra je tako povsem enakovredna dejanski ruleti. Ta preprosta ideja nam bo prišla še posebej prav, če bomo govorili o večkratnem igranju rulete in stavah na rdeče. Predstavljali si bomo lahko, da izbiramo lističe iz škatlice večkrat, pri čemer izbrane lističe pred ponovnim izbiranjem vrnemo v škatlo. Celotni dobiček bo potem preprosto vsota številk na izbranih lističih.



Sl. 3.2: Predstavitev rulete s škatlico in lističi.

Podobno si lahko z izbiranjem lističev iz škatle predstavimo še mnogo iger na srečo in tudi drugih situacij, v katerih je v igri naključnost.

PRIMER: Recimo, da večkrat mečemo kovanec in nas zanima samo celotno število grbov. Kako bi si ta "poskus" predstavili z izbiranjem lističev iz škatle? Ena od možnosti je gotovo škatlica | 0 1 |, kjer listek 1 predstavlja padec grba, listek 0 pa padec številke. Če kovanec vržemo 100-krat in preštejemo grbe, bodo verjetnosti posameznih izidov povsem enake, kot če bi 100-krat izbrali listič iz škatlice in sešteli izbrana števila. Pri tem predpostavljamo, da lističe med posameznimi izbirami vračamo.

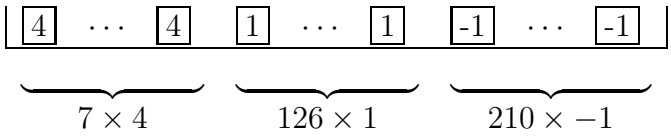
PRIMER: Oglejmo si še nekoliko bolj zapleteno igro na srečo. Igralni avtomati imajo tri kolesa s po sedmimi simboli. Ko potegnemo ročico, se kolesa ustavijo neodvisno drug od drugega, simboli na posameznih kolesih pa so enako verjetni. Stavimo vedno 1 žeton. Če se ujemajo vsi trije simboli, nam igralni avtomat vrne stavo in še 4 žetone. Če se ujemata dva simbola, nam avtomat vrne stavo in še en žeton. V vseh drugih primerih stavo izgubimo. Po vsaki igri imamo torej možnost, da smo bogatejši za 4 žetone, 1 žeton ali da smo revnejši za 1 žeton. S kakšno škatlico bi lahko

predstavili to igro? Spomniti se moramo na Galilejevo rešitev iz prvega uvodnega primera. Razmišljati moramo najprej o vseh možnih izidih.

1 1 1	1 1 2	1 1 3	1 1 4	1 1 5	1 1 6	1 1 7
1 2 1	1 2 2	1 2 3	1 2 4	1 2 5	1 2 6	1 2 7
...	...	...	...	...	...	...

Tabela 3.4: Nekaj možnih izidov pri igralnem avtomatu.

Po krajšem razmisleku ugotovimo, da je vseh možnih izidov $7 \cdot 7 \cdot 7 = 343$. Koliko je med temi možnimi izidi takšnih, ki prinesejo dobitek 4 žetonov? Takih dobitkov je očitno 7, ker je 7 različnih simbolov. Koliko pa je izidov, ko žeton izgubimo? Prešteti moramo vse možne izide, pri katerih so vsi trije simboli različni.



Sl. 3.3: Predstavitev igralnega avtomata s škatlico in lističi.

Ko malo premislimo, vidimo, da je takih trojic $7 \cdot 6 \cdot 5 = 210$. Utemeljitev je, da si lahko prvo število prosto izberemo, za kar imamo 7 možnosti. Ko smo si prvo število izbrali, se drugo ne sme ujemati, torej imamo 6 možnosti za drugo število, za tretje pa nam potem ostane 5 možnosti. Izidov, ko se ujemata dva simbola, je potem še toliko, kolikor jih ostane od vseh možnih izidov, torej $343 - 7 - 210 = 126$.

Škatla, ki bi predstavljala to igro, bi potem imela 343 lističev, od tega 7 s številom 4, 126 s številom 1 in 210 s številom -1.

Mnoge naključne pojave, pri katerih je izid število, si lahko predstavimo kot naključno izbiranje lističev z možnimi številkami iz škatle. Pri tem predpostavljamo, da imajo vsi lističi enako verjetnost, da bodo izbrani. Če izbiranja ponavljamo, predpostavljamo, da smo izbran listič pred ponovnim izbiranjem vedno vrnili na svoje mesto.

3.3 Normalna aproksimacija

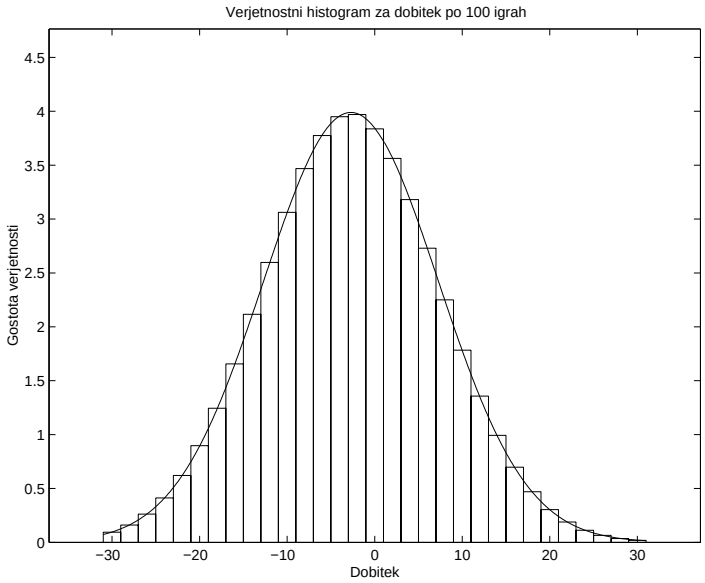
Predstavljajte si spet, da igrate ruleto in vsakič stavite na rdeče. Pravila igre so taka, da vam tedaj, ko se kroglica ustavi na rdečem, vrnejo začetno stavo, povrhu pa dobite še enkrat toliko denarja. Na koncu ste torej za svojo stavo bogatejši. Če se kroglica ustavi na črnem ali zelenem izseku, ste stavo izgubili. V prejšnjem razdelku smo si predstavili eno igro rulete kot naključno izbiranje lističa iz škatlice na sliki 3.2.

Če imate kockarsko žilico, rulete ne boste igrali samo enkrat, ampak boste igro mnogokrat ponovili. To je isto, kot če bi ponavljali izbiranje lističa iz škatlice. Kot rezultat bi dobili zaporedje, sestavljeno iz števil 1 in -1. Čisti dobiček na koncu vašega igranja rulete bi bil predstavljen z vsoto števil na izbranih lističih. Ta čisti dobiček je seveda lahko tudi negativno število in v tem primeru bi mu v vsakdanjem jeziku rekli čista izguba. Vzemimo kot primer, da bi igrali 100-krat. Spodaj je računalniško simulirano zaporedje 100 iger na ruleti. V zgornjem zaporedju je dobiček na koncu 6. Lahko pa bi bil tudi kaj drugega. V načelu so možni vsi dobički od -100 do 100, niso pa seveda vsi enako verjetni. Kolikšne so verjetnosti posameznih končnih dobičkov, si lahko predstavimo z novim grafom, ki mu bomo rekli *verjetnostni histogram*.

Za natančen izračun verjetnosti posameznih možnih končnih dobičkov potrebujemo nekaj matematike, ki je tukaj ne bomo obravnavali. Te verjetnosti pa bomo znali izračunati z uporabo normalne krivulje. Oglejmo si bolj natančno verjetnostni histogram na sliki 3.4.

−1	1	1	1	−1	1	−1	1	1	1
1	1	1	−1	1	1	1	1	1	1
−1	−1	−1	1	−1	1	1	1	−1	1
1	−1	1	−1	−1	1	1	1	−1	1
1	1	−1	−1	1	1	−1	1	1	−1
1	−1	−1	1	1	1	1	−1	1	1
−1	−1	−1	−1	−1	−1	1	−1	−1	−1
−1	1	−1	1	−1	1	−1	−1	1	1
−1	−1	1	1	1	1	−1	1	−1	1
−1	1	−1	−1	−1	−1	−1	−1	−1	−1

Tabela 3.5: Eden od možnih izidov pri 100 stavah na rdeče.



Sl. 3.4: Verjetnostni histogram za 100 iger pri ruleti.

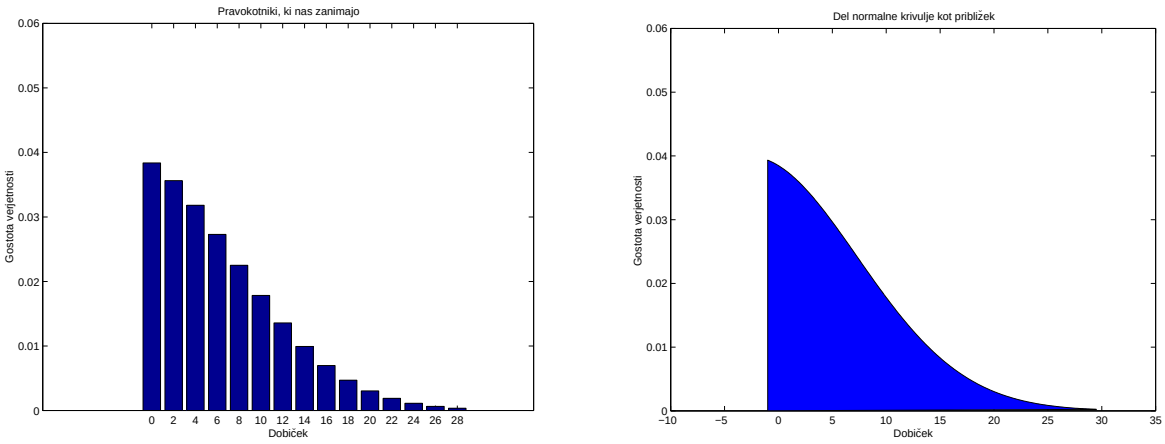
Ploščine v verjetnostnem histogramu predstavljajo verjetnosti. Pravokotnik nad -10, recimo, ima osnovnico 2 in je visok nekaj več kot 0,03, torej je verjetnost, da bomo izgubili 10 enot v 100 igrah, približno 0,06 ali 6%. Histogram moramo interpretirati kot v prvem poglavju. Ploščina tukaj predstavlja verjetnost. Pogosto nas bo zanimala ploščina skupine pravokotnikov v verjetnostnem histogramu. Pri ruleti ploščina pravokotnikov nad 0 ali desno predstavlja verjetnost, da po 100 igrah ne bomo imeli izgube.

V verjetnostnem histogramu ploščine pravokotnikov predstavljajo verjetnost. Če je osnovnica pravokotnika med a in b, potem ploščina predstavlja verjetnost, da bo slučajno število med a in b.

Najzanimivejše je opažanje, da se našemu verjetnostnemu histogramu zelo lepo prilega normalna krivulja. To dejstvo je v podobnih situacijah prvi uporabil francoski matematik Abraham De Moivre (1667–1754) za računanje verjetnosti. Njegov razmislek je bil naslednji: pogosto nas bo zanimala ploščina pravokotnikov v verjetnostnem histogramu. Te ploščine je težko izračunati, lahko pa jih aproksimiramo s ploščino pod primerno izbrano normalno krivuljo.

Verjetnostni histogrami za vsote naključno izbranih števil iz poljubne škatlice se prilegajo normalni krivulji s primerno izbranima parametroma.

Pri ruleti se lahko vprašamo, kolikšna je verjetnost, da po 100 igrah ne bomo v izgubi. Zanima nas torej vsota ploščin vseh pravokotnikov od tistega nad 0 naprej. Na sliki 3.5 je ponazorjena De Moivrova ideja.



Sl. 3.5: Pravokotniki v verjetnostnem histogramu, katerih ploščina nas zanima, in približek z normalno krivuljo.

V načelu znamo izračunati ploščino pod poljubno normalno krivuljo, če le poznamo povprečje in standardni odklon. Želena parametra dobimo po formulah, ki so v spodnjem okvirčku.

Normalna krivulja, ki se najbolje prilega verjetnostnemu histogramu za vsoto naključno izbranih števil iz škatlice, ima parametra

$$EV = n \cdot \mu \qquad \text{in} \qquad SE = \sqrt{n} \cdot \sigma \,,$$

kjer je μ povprečje števil v škatlici, σ standardni odklon za števila v škatlici, n pa število izbranih lističev. Količini EV pravimo pričakovana vrednost, količini SE pa standardna napaka. Izraza sta povzeta iz angleških izrazov *expected value* in *standard error*.

Iz formule je razvidno, da moramo poznati povprečje in standardni odklon števil v škatli. Računanje standardnega odklona v škatli s samo dvema različnima številoma nam olajša naslednje preprosto pravilo:

Če so v škatlici samo lističi s številoma a in b, potem je standardni odklon števil v škatlici enak

$$\sigma = |b-a| \cdot \sqrt{p \cdot (1-p)}\,,$$

kjer je p delež lističev s številom a (ali b, v obeh primerih dobimo isti rezultat).

Izračunajmo zdaj verjetnost, da po 100 igrah rulete s stavami po 1 enoto na rdeče ne bomo v izgubi. Dobimo

$$EV = -100 \cdot \frac{1}{37} = -2.7027 \qquad \text{in}$$

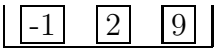
$$SE = \sqrt{100} \cdot \sqrt{\frac{18}{37} \cdot \frac{19}{37}} = 9.9963\,.$$

Ploščino desno od 0 tako dobimo lahko iz tabele za normalno porazdelitev. Dobiček 0 moramo pretvoriti v standardne enote in dobimo

$$z = \frac{0-(-2,7027)}{9,9963} = 0,27\,.$$

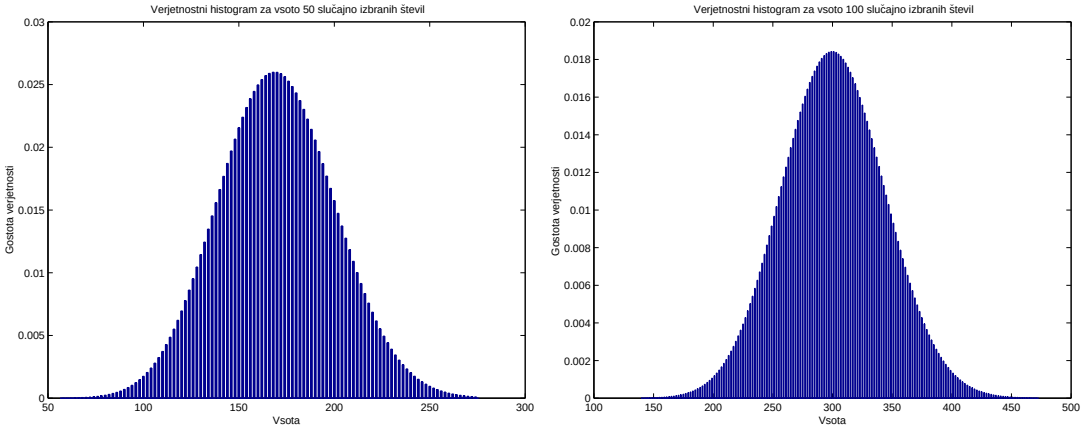
V verjetnosti uporabljamo oznako z za standardne enota namesto oznake s, ki smo jo uporabljali v prvem poglavju. Iskana verjetnost bo potem enaka ploščini desno od 0,27 pod normalno krivuljo, to pa je približno 0,393. Teoretično prava verjetnost je 0,432, tako da nam je zgornji postopek dal kar dobre približke.

De Moivrov razmislek pa ne velja samo za ruleto, temveč za vse slučajne količine, ki nastanejo kot vsote naključno in neodvisno izbranih števil. Da se o tem prepričamo, si oglejmo naslednji primer. Škatla naj bo preprosta, kot je na sliki 3.6.



Sl. 3.6: Asimetrična škatlica.

Oglejmo si verjetnostne histograme za vsoto $n = 50$ in $n = 100$ lističev, izbranih iz škatle na sliki 3.6.



Sl. 3.7: Verjetnostni histogrami za vsoto 50 in 100 izbiranj.

Iz verjetnostnih histogramov na sliki 3.7 lahko razberemo, da se tudi v tem primeru za dovolj velike vsote verjetnostni histogrami lepo prilegajo normalni krivulji.

Verjetnostni histogrami za vsote naključno izbranih števil iz poljubne škatlice se prilegajo normalni krivulji s primerno izbranima parametroma. Ujemanje je tem boljše, čim večje je število izbiranj.

PRIMER: Letalske družbe pogosto prodajo več kart, kot je sedežev na letalu, ker pričakujejo, da si bo nekaj potnikov tik pred zdajci premislilo ali bodo zadržani. Po drugi strani pa družbe ne želijo, da bi prepogosto kdo od potnikov z veljavno karto ostal brez sedeža. Odgovoriti je treba na vprašanje, koliko več kart lahko prodamo, da bo verjetnost tega, da kdo ostane brez sedeža, čim manjša.

Situacijo poskusimo modelirati s primerno škatlico. Predpostavimo, da se potnik s kupljeno karto pojavi z verjetnostjo 0,9, neodvisno od drugih potnikov. Verjetnost 0,9 letalske družbe dobijo iz izkušenj z mnogimi poleti. Z drugimi besedami, predpostavljamo, da dejansko na polet v povprečju pride le 90% potnikov, ki so kupili karto. Če recimo letalska družba proda 550 kart za letalo s 500 sedeži, bi število potnikov, ki bodo prišli, lahko predstavili kot vsoto 550 naključno izbranih števil iz škatlice:

1	1	1	1	1	1	1	1	1	0

Za računanje verjetnosti potrebujemo povprečje števil v škatlici in standardni odklon. Povprečje je 0,9, za standardni odklon pa uporabimo pravilo za škatlice, v katerih so lističi s samo dvema različnima številoma, in dobimo

$$\sigma = \sqrt{\frac{1}{10} \cdot \frac{9}{10}} = 0,3\,.$$

Parametra normalne krivulje, ki se verjetnostnemu histogramu najbolje prilega, sta

$$EV = n \cdot \mu = 550 \cdot 0,9 = 495$$

in

$$SE = \sqrt{n} \cdot \sigma = \sqrt{550} \cdot 0,3 = 7,03\,.$$

Verjetnost, da bo potnikov preveč, izračunamo kot ploščino pod normalno krivuljo desno od 500. V standardnih enotah dobimo

$$z = \frac{500 - 495}{7,03} = 0,71\,.$$

Ploščina pod normalno krivuljo desno od 0,71 je 0,24 ali v odstotkih 24%. Pričakovali bi lahko, da pri 24% poletov pride do težav z "odvečnimi" potniki.

Vprašamo se lahko tudi drugače. Recimo, da ne želimo, da bi v več kot 5% primerov prišlo preveč potnikov. Koliko kart lahko največ prodajamo? Označimo iskano število kart z n. Parametra normalne krivulje, ki jo bomo uporabili, sta

$$EV = n \cdot 0,9$$

in

$$SE = \sqrt{n} \cdot 0,3\,.$$

Tudi zdaj pretvorimo 500 v standardne enote in dobimo

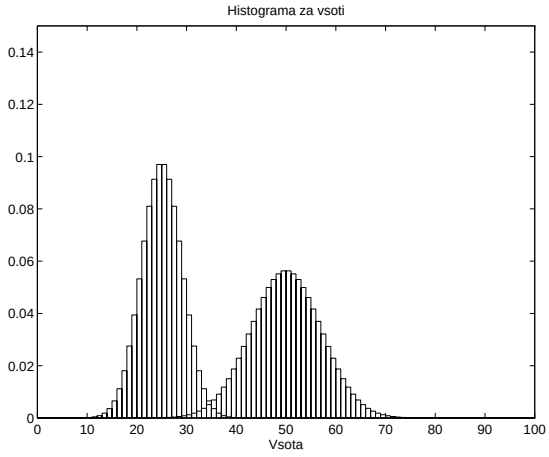
$$z = \frac{500 - 0,9 \cdot n}{\sqrt{n} \cdot 0,3}\,.$$

Želimo, da bi bila ploščina desno od 500 le 5% celotne ploščine, zato bi moral biti z enak $z = 1,65$. Rešiti moramo enačbo za n. Rešitev je $n = 543$, ki jo najdemo tako, da označimo $m = \sqrt{n}$ in rešimo kvadratno enačbo za m.

1. Na spodnji sliki sta histograma za vsoti 25 neodvisnih izbiranj iz ene od naslednjih dveh škatel:

(i) | 0 | 1 | 2 |

(ii) | 0 | 1 | 2 | 3 | 4 |



Histograma za vsote neodvisnih izbiranj iz škatel (i) ali (ii).

a. Kateri od zgornjih dveh histogramov pripada škatli (i) in kateri škatli (ii)? Utemeljite odgovor.

b. Izračunajte približno verjetnost, da bo vsota 25 izbiranj iz škatle (ii) večja ali enaka 40.

Rešitev:

a. Škatla (ii) ima večji standardni odklon kot škatla (i), zato bo imel tudi histogram za vsoto 25 izbiranj večji SE za škatlo (ii) kot za škatlo (i). Škatli (ii) pripada nižji in širši histogram.

b. *Povprečje škatle je* 2*, standardni odklon pa* $\sqrt{2}$*. Sledi* $EV = 50$ *in* $SE = 7,07$*. Pretvorimo* 40 *v standardne enote. Dobimo*

$$z = \frac{40 - 50}{7,07} = -1,41\,.$$

Iz tabele preberemo, da je iskana verjetnost približno enaka $92,1\%$*.*

2. Recimo, da imate na voljo 100,000 USD, ki jih želite vložiti v delnice. Delnice, v katere lahko investirate, za vloženih 1000 USD prinesejo dobiček 200 USD, 100 USD, 0 USD ali -100 USD(izgubo), vsakega z verjetnostjo 0,25. Izbirate lahko med 100 različnimi delnicami. Vrednost ene delnice se spreminja neodvisno od drugih delnic. Recimo, da denar lahko vložite na dva načina:

   1. Vseh 100.000 USD vložite v eno samo delnico.
   2. Vložite po 1.000 USD v vsako od 100 delnic.

   a. Zapolnite manjkajoče besede: V primeru 2 je dobiček enak _____ števil, ki jih _____-krat izvlečemo iz škatlice z _____ lističi. Števila na lističih so _____.

   b. Kolikšna približno je verjetnost, da bo v primeru 2 dobiček 8000 USD ali več?

   c. V primeru 1. je verjetnost, da bo dobiček večji od 8000 USD enaka 1/2, verjetnost izgube pa je 1/4. Kolikšna je verjetnost izgube v primeru 2?

Rešitev:

a. *V primeru 2. je dobiček enak vsoti števil, ki jih* 100*-krat naključno izvlečemo iz škatlice s* 4 *lističi. Števila na lističih so* $200, 100, 0, -100$*.*

b. *Povprečje škatle je enako* 50 *USD, standardni odklon pa* $111,8$ *USD. Sledi* $EV = 5000$ *in* $SE = 1118$. *Pretvorimo* 8000 *USD v standardne enote. Dobimo*

$$z = \frac{8000 - 5000}{1118} = 2,68\,.$$

Iskana verjetnost je ploščina pod normalno krivuljo desno od $2,6$*, kar je približno* $0,3\%$.

c. *V tem primeru moramo v standardne enote pretvoriti število* 0. *Dobimo*

$$z = \frac{0 - 5000}{1118} = -4,4\,.$$

Iskana verjetnost je enaka ploščini pod normalno krivuljo levo od $-4,4$*, kar je praktično enako* 0. *S tem, da smo svoj vložek razpršili med več delnic, smo občutno znižali tveganje v smislu, da je verjetnost izgube zanemarljiva. Opozorilo: Sklep velja samo za ta preprosti model, kjer predpostavljamo, da se delnice obnašajo neodvisno druga od druge. V realnem življenju moramo biti s takimi predpostavkami previdni.*

3. V skupini 25 ljudi je vsak 1000-krat vrgel kovanec neodvisno od drugih in na koncu zabeležil število padlih grbov. Privzemite, da sta verjetnosti, da pade grb ali številka, enaki. V jeziku tega poglavja predpostavljamo, da izbiramo iz škatle, ki vsebuje le listka z 0 in 1.

   a. Izračunajte verjetnost, da bo prva oseba vrgla med 470 in 530 grbi.

   b. Recimo, da so ti ljudje svoja števila padlih grbov pretvorili v standardne enote po formuli

$$z = \frac{\#\text{ grbov} - EV}{SE}\,.$$

   Za celotno skupino so tako dobili 25 številk, ki so v enem od spodnjih zaporedij. V katerem? Utemeljite svojo izbiro!

   (i) $0,758$, $0,948$, $0,126$, $1,454$, $1,264$, $2,719$, $1,391$, $3,352$, $1,011$, $2,846$, $1,011$, $0,758$, $0,063$, $1,201$, $0,126$, $0,758$, $2,909$, -$0,063$, $0,948$, $2,150$, $0,822$, $1,707$, $1,391$, -$1,011$, -$0,758$

(ii) 1, 2540, -1, 5937, -1, 4410, 0, 5711, -0, 3999, 0, 6900, 0, 8156, 0, 7119, 1, 2902, 0, 6686, 1, 1908, -1, 2025, -0, 0198, -0, 1567, -1, 6041, 0, 2573, -1, 0565, 1, 4151, -0, 8051, 0, 5287, 0, 2193, -0, 9219, -2, 1707, -0, 0592, -1, 0106

(iii) -1, 934, -1, 397, 1, 350, 0, 865, -1, 571, 0, 929, 1, 482, 2, 106, 0, 763, -0, 053, 0, 993, -1, 648, -2, 186, 0, 946, -0, 182, 2, 178, 1, 766, 1, 829, 0, 702, 0, 310, 1, 899, -1, 379, 1, 003, 3, 356, -1, 600;

Rešitev:

a. *Izračunamo* $EV = 500$ *in* $SE = 15,8$. *Števili* 470 *in* 530 *pretvorimo v standardne enote in dobimo* $z = -1,90$ *in* $z - 1,90$. *Iskana verjetnost je enaka ploščini pod normalno krivuljo med tema dvema številoma, kar je* 94%.

b. *Za normalno porazdelitev pričakujemo, da bi morala biti približno polovica števil pozitivnih, polovica pa negativnih. Možnost (i) tako odpade. Standardni odklon dobljenih števil bi moral biti okrog* 1. *Pri možnosti (iii) je ta standardni odklon precej večji. Ostane torej možnost (ii).*

4. Iz velike škatle vlečemo listke neodvisno drug od drugega. Vsebina škatle ni natančno znana, imamo pa nekaj informacij.

a. Recimo, da je znano, da je povprečje števil v škatli enako 0. Kolikšna je verjetnost, da bo po velikem številu izvlečenih listkov vsota izvlečenih števil enaka 0 ali večja? Utemeljite svoj odgovor.

b. Recimo, da je znano, da je povprečje števil v škatli enako 0 in da so na listkih samo števila 1 in -1. Po kolikšnem številu izvlečenih listkov bo verjetnost, da se vsota števil na njih od 0 razlikuje za 100 ali več, enaka 99%?

c. Recimo, da je znano, da so na listkih samo števila 1 in -1. Po 10.000 izvlečenih listkih je bila vsota števil na listkih 642. Bi verjeli, da je odstotek listkov s številom 1 enak 50%?

Rešitev:

a. *Histogram za vsoto po velikem številu izbiranj bo približno normalen, ker je povprečje* 0*, pa bo* $EV = 0$*. Verjetnost bo torej* 50 %*.*

b. *Standardni odklon opisane škatle je 1. Poleg tega mora veljati, da je* 100 *enako* $SE \cdot 2,56$*, ali* $\sqrt{n} \cdot 2,56 = 100$*. Sledi* $n \approx 1526$*.*

c. *Če je odstotek listkov z* 1 *enak* 50 %*, je povprečje škatle enako* 0*, standardni odklon pa* 1*. Po* 10.000 *izbiranjih je* $SE = 100$*. Število* 642 *je več kot* 6 *standardnih napak od pričakovane vrednosti. Taka vsota je pri danih predpostavkah nemogoča. No moremo verjeti, da je odstotek listkov s številom* 1 *enak* 50 %*.*

5. Zavarovalnice obravnavajo prihajajoče zahtevke za izplačila kot vlečenje listkov iz škatlice z veliko števili neodvisno drug od drugega. Števila na listkih predstavljajo višine zahtevkov. Iz izkušenj je znano, da je povprečje škatlice 2,4 in standardni odklon 1,6 (v enotah po 1000 USD).

   a. Kolikšna je verjetnost, da bo skupna višina zahtevkov po 10.000 izvlečenih listkih presegala 28.000 (v enotah po 1000 USD)?

   b. Zavarovalnica ima v letu 10.000 strank. Tveganje, da bo skupna višina zahtevkov v enem letu presegla skupno višino premij, ki jih stranke vplačajo, želijo obdržati pod 0,5%. V zavarovalnici pričakujejo, da bo v letu prispelo 2000 zahtevkov. Kolikšno višino premije naj zavarovalnica določi svojim strankam?

Rešitev:

a. *Po* 10.000 *izbiranjih pričakujemo vsoto* $EV = 24.000$ *s standardnim odklonom* $SE = 1600$*. Vsota* 28.000 *je v standardnih enotah enaka*

$$z = \frac{28.000 - 24.000}{1600} = 2,5\,.$$

Iskana verjetnost je enaka ploščini pod normalno krivuljo desno od $2,5$*, kar je* $0,6\%$*.*

b. *Skupna višina zahtevkov bo kot vsota* 2000 *izbiranj iz opisane škatle. Izračunamo* $EV = 4.800$ *in* $SE = 71,6$*. Označimo skupno vsoto premij s* C*. Veljati mora*

$$z = \frac{C - EV}{SE} = 2,56$$

ali

$$C = EV + 2,56 \cdot SE = 4983\,.$$

Ta celotni znesek porazdelimo med vseh 10.000 *zavarovancev. Premija bi tako znašala* 498 *USD.*

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

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:

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

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

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

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

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.

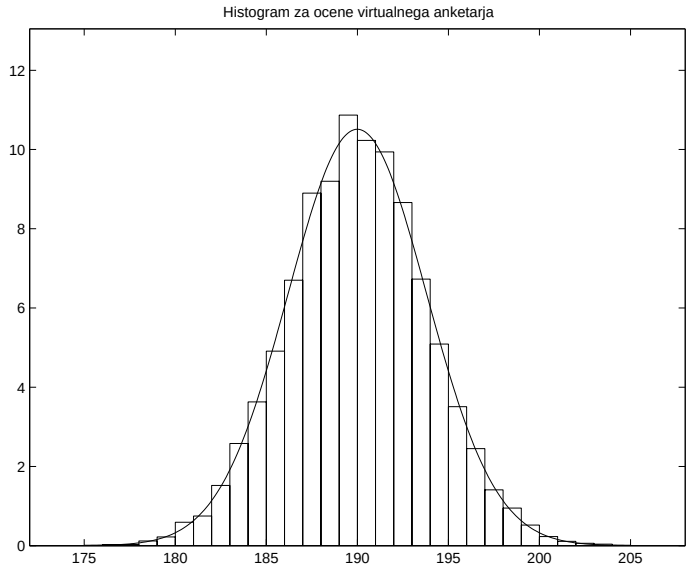
Za računalniško simulacijo si izberimo konkreten primer. Populacija naj bo velikosti $N = 1.000.000$, izbirali pa bomo vzorec velikosti $n = 1000$. V populaciji naj bo

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.

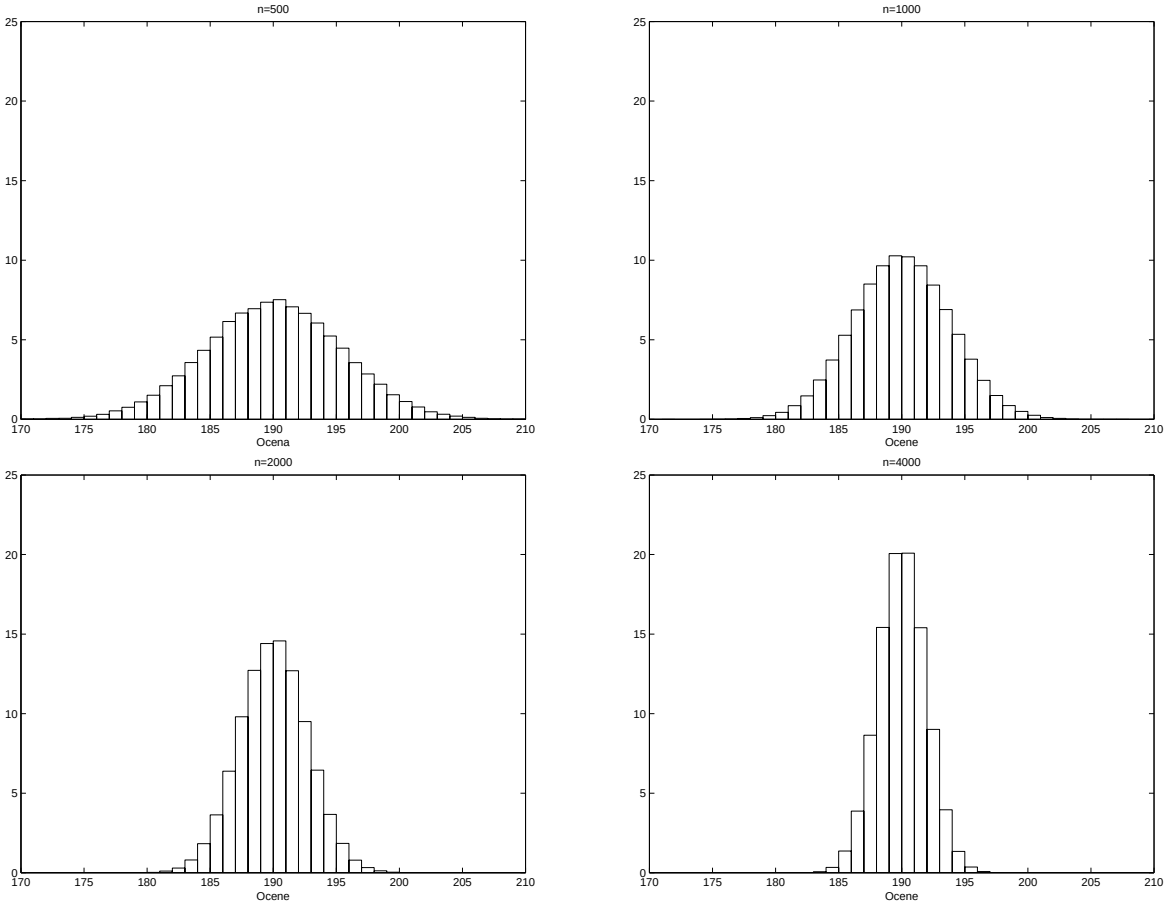
Histogram za vzorčne ocene se za enostavno slučajno vzorčenje dobro prilega normalni krivulji s primernima parametroma. To velja ne glede na to, ali ocenjujemo populacijsko povprečje ali odstotek enot v populaciji z neko lastnostjo. To dejstvo je ključnega pomena za raziskovanje zanesljivosti vzorčnih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

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

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

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

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

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:

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

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%.

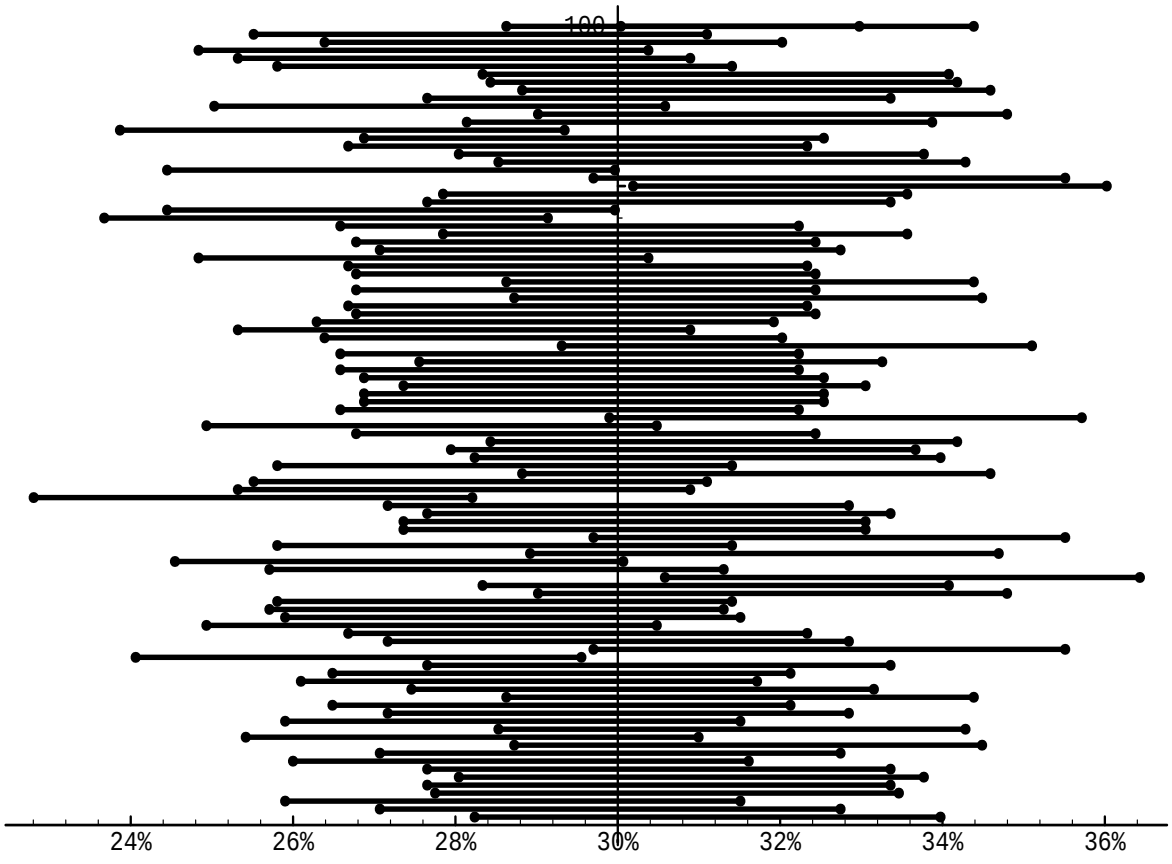
Verjetnost, da interval zaupanja ne pokrije prave vrednosti, si, kot rečeno, lahko izberemo vnaprej. Udomačen izraz za to izbrano verjetnost je *stopnja tveganja*, ki jo pogosto označimo z grško črko α, in govorimo o intervalih zaupanja pri dani stopnji tveganja. Pogosto si namesto verjetnosti, da interval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

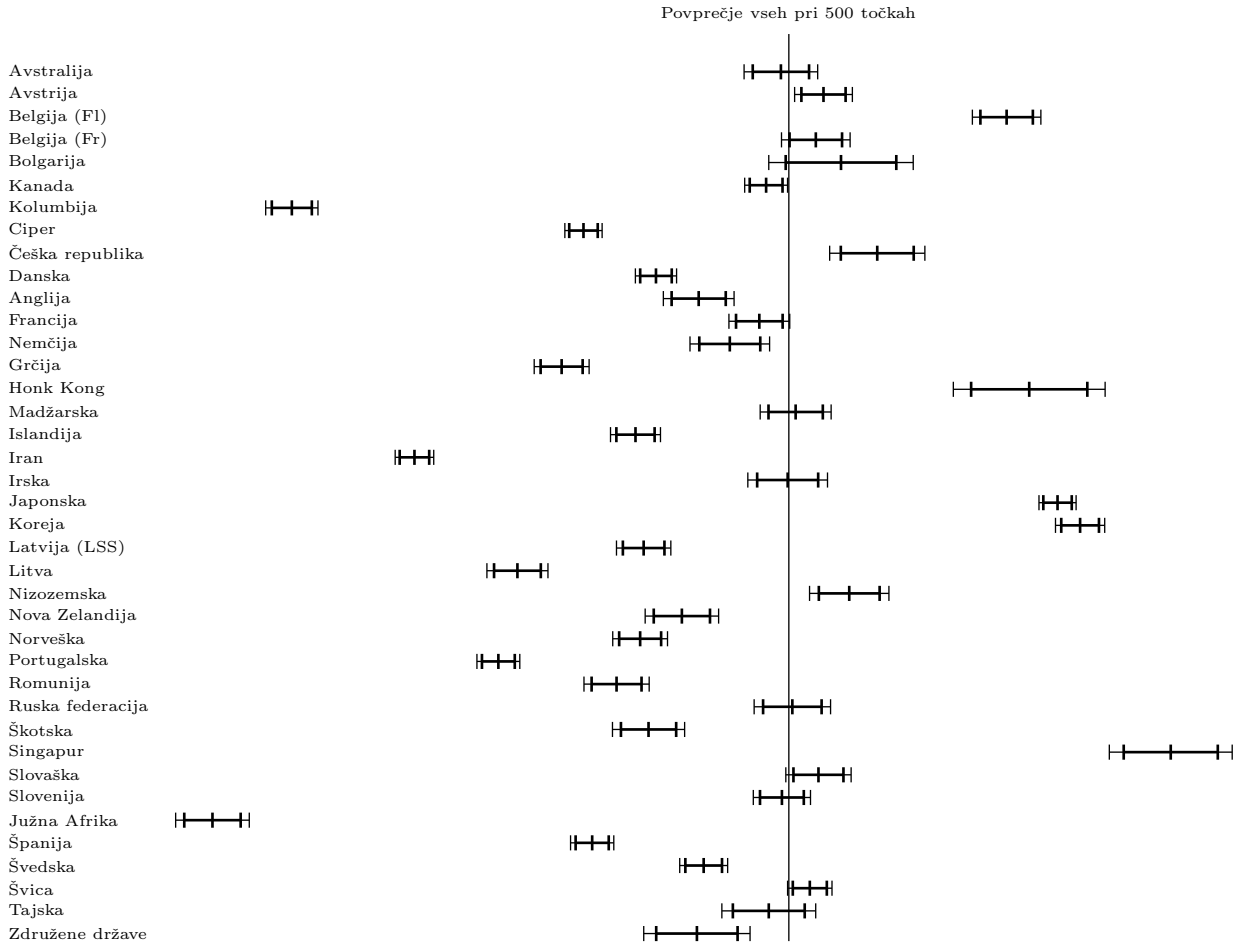


Sl. 4.3: 100 intervalov zaupanja virtualnega anketarja

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.



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* $100\% - 38,2\% = 61,8\%$*.*

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!

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*

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.*

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* $1,96$*.*

Vzorec pisnega izpita

V tem poglavju je primer pisnega izpita za predmet Poslovna statistika. Format dejanskega izpita bo enak in bo vedno vseboval eno nalogo iz opisnih statistik, dve iz regresije, eno iz verjetnosti in dve iz vzorčenja.

POSLOVNA STATISTIKA

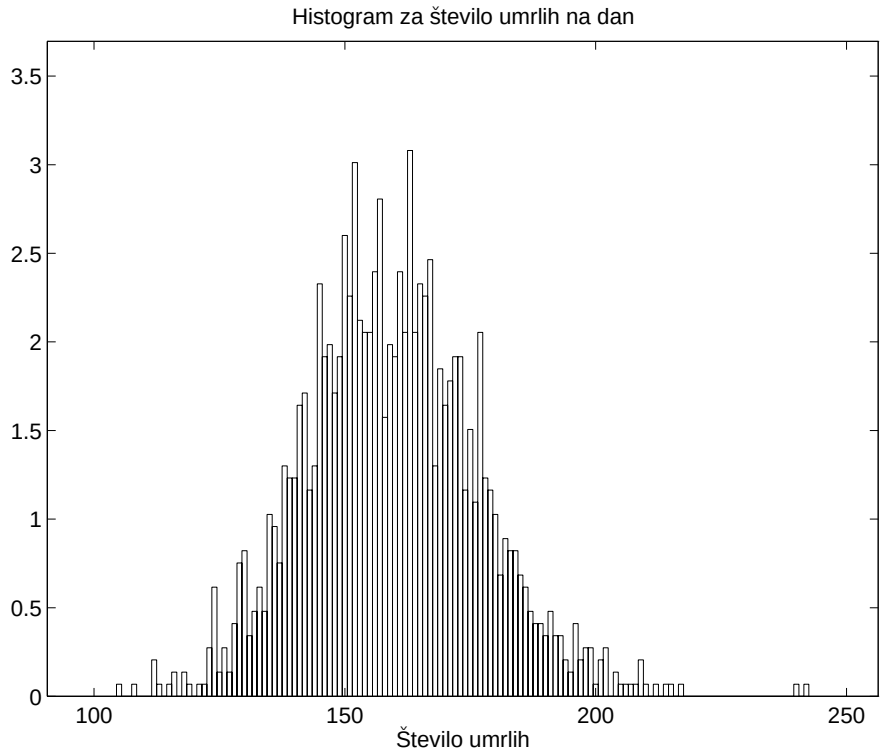
Pisni izpit

10. april 1997

Navodilo

Preden se lotite reševanja, pazljivo preberite besedilo naloge. Veljale bodo samo rešitve na papirju, kjer so naloge. Rešitev naloge mora zajemati vse potrebne izračune in utemeljitve. Na vprašanja odgovorite s celimi stavki. Nalog je 6 in vsaka je vredna 20 točk, torej skupaj 120 točk. Za reševanje imate 90 minut časa.

1. (20) Podatki za to nalogo so števila umrlih za vsak dan med 1. januarjem 1961 in 31. decembrom 1964 v mestu Los Angeles. V tem obdobju je bilo 1461 dni. Na sliki 1 je histogram za te podatke.



Sl. 1: Histogram za dnevno število umrlih za 1461 dni.

a. (5) Kaj je po vašem mnenju v zgornjem primeru populacija, kaj enota in kaj spremenljivka?

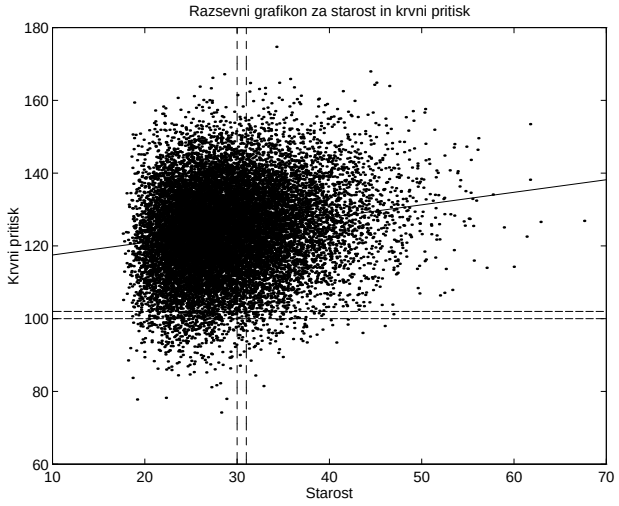
b. (5) Ploščina stolpca nad 200 predstavlja odstotek. Odstotek česa?

c. (5) Katere enote bi morale biti na navpični osi?

d. (5) Povprečje za zgornje podatke je 159,4 in standardni odklon 17,4. Uporabite ti števili za oceno odstotka dni, v katerih je umrlo 150 ali več ljudi.

2. (20) V študijo vpliva kontracepcijskih tablet na krvni pritisk je bilo vključenih 17.500 žensk v starosti od 17 do 58 let, ki so jemale kontracepcijske tablete. Na sliki 2 je razsevni grafikon za starost in krvni pritisk teh 17.500 žensk. Razsevni grafikon je homoshedastičen. Podatki so naslednji:

$\bar{x} = 29,4$ leta $\quad \sigma_x = 12$ let
$\bar{y} = 124$ mm $\quad \sigma_y = 14$ mm
$r = 0,32$



Sl. 2: Razsevni grafikon za starost in krvni pritisk.

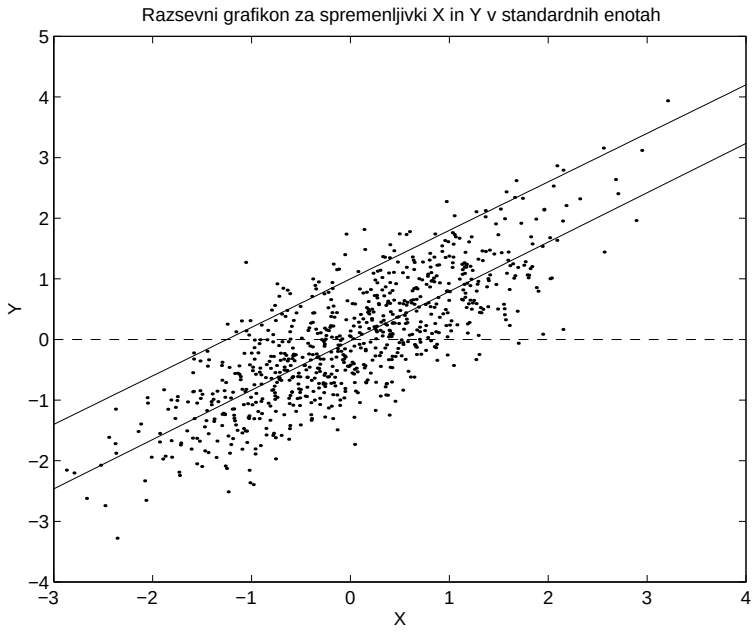
a. (5) Ocenite povprečni krvni pritisk pri skupini žensk v navpični rezini med 30 in 32 let na sliki.

b. (10) Kolikšen odstotek žensk iz a. ima krvni pritisk nad 120 mm?

c. (5) V vodoravni rezini na sliki so ženske, ki imajo krvni pritisk med 100 mm in 102 mm. Ocenite povprečno starost teh žensk.

3. (20) Spremenljivki X in Y najprej pretvorimo v standardne enote, potem pa za pretvorjene podatke narišemo razsevni grafikon. Slika 3 prikazuje ta razsevni grafikon, spodnja tabela pa ustrezajoče podatke. Histograma za spremenljivki X in Y se dobro prilegata normalnim krivuljam.

$\bar{x} = 0 \qquad \sigma_x = 1$
$\bar{y} = 0 \qquad \sigma_y = 1$
$r = 0,8$



Sl. 3: Razsevni grafikon za spremenljivki X in Y, pretvorjeni v standardne enote.

a. (10) Kolikšen odstotek točk bi pričakovali nad vodoravno (črtkano) premico, ki gre skozi točko s koordinatama (0,0)? Utemeljite odgovor!

b. (10) Kolikšen odstotek točk bi pričakovali nad premico, ki je vzporedna regresijski premici in za 1 nad njo v navpični smeri (zgornja premica na sliki)?

4. (20) Zavarovalnice pri avtomobilskem zavarovanju razmišljajo takole: recimo, da zavarujemo 100.000 ljudi. Vemo, da bo 19% od teh, torej 19.000, vložilo zahtevke. Zahtevki bodo različno visoki in jih vnaprej ne moremo napovedati. Lahko pa si predstavljamo, da bo celotna vsota zahtevkov kot vsota 19.000 naključno izbranih števil iz velike škatle, za katero vemo povprečje in standardni odklon.

a. (5) V Veliki Britaniji je povprečje škatle enako 1200 GBP in standardni odklon 900 GBP. Izračunajte EV in SE.

b. (5) Izračunajte verjetnost, da bo vsota 19.000 naključno izbranih števil iz škatle večja od 23.085.329 GBP.

c. (5) Če zavarovalnica postavi premijo na 200 GBP, bo 100.000 zavarovancev skupno plačalo 20.000.000 GBP. Izračunajte verjetnost, da bo vsota 19.000 naključno izbranih števil večja od 20.000.000 GBP, torej verjetnost, da bo zavarovalnica imela izgubo.

d. (5) Kolikšna bi morala biti premija, da bi bila verjetnost za izgubo le 1%?

5. (20) Za potrebe določanja cen življenjskih potrebščin Urad za statistiko RS izbere vzorec velikosti $n = 3270$ gospodinjstev iz populacije vseh gospodinjstev v Sloveniji. Privzemite, da je vzorec enostavni slučajni. Popravne faktorje povsod zanemarite.

a. (10) Na podlagi vzorca ocenimo, da gospodinjstva v povprečju za hrano porabijo 23,4% razpoložljivih sredstev. Vzorčni standardni odklon je $\hat{\sigma} = 12,6\%$. Ali lahko trdimo, da z vzorcem te velikosti z verjetnostjo približno 95% ocenimo povprečni odstotek izdatkov za hrano na 0,44% natančno? Odgovor utemeljite z računom ali v stavku.

b. (10) Na podlagi istega vzorca gospodinjstev kot zgoraj ocenimo odstotek gospodinjstev, v katerih je vsaj en član brezposeln, z 9%. Kolikšna je standardna napaka te ocene?

6. (20) Župan nekega mesta se je odločil, da bo ponovno kandidiral, če so njegove možnosti za zmago dobre. V ta namen si je iz denarja davkoplačevalcev privoščil anketo. Privzemite, da so anketarji izbrali enostavni slučajni vzorec volivcev v mestu velikosti $n = 900$.

a. (5) Recimo, da se je za župana izreklo 52,4% anketiranih. Županovi svetovalci menijo, da naj kandidira, ker anketa kaže, da je odstotek volivcev, ki podpirajo župana, z gotovostjo večji od 50%. Ali imajo prav?

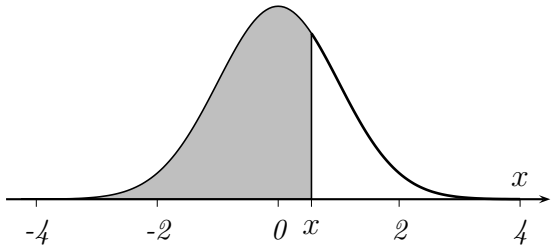
b. (5) Izračunajte zgornjo in spodnjo mejo zaupanja pri stopnji tveganja $\alpha = 0,05$ za odstotek volivcev, ki podpirajo župana.

c. (5) Kolikšna je približno verjetnost, da se ocena iz a. razlikuje od dejanskega odstotka za manj kot 4% ?

d. (5) Recimo, da župana dejansko podpira 49% volivcev. Kolikšna je verjetnost, da bo večina anketirancev v vzorcu velikosti $n = 900$ podprla župana?

Tabela za normalno porazdelitev

Tabelo za normalno porazdelitev potrebujemo pri obravnavanju zanesljivisti vzorčnih ocen in za določanje intervalov zaupanja. Tabela za vsak x prikazuje ploščino levo od x, kot je označno na spodnji sliki.



Navodilo*: Tabela na naslednji strani pove odstotek ploščine levo od dane vrednosti x pod normalno krivuljo. Recimo, da nas zanima odstotek ploščine levo od 1,25. V stolpcu, označenem z x, najdemo 1,25 in v sosednjem predalčku odčitamo 89,44%.*

x	P	x	P	x	P	x	P
-4,00	0,00	-2,00	2,28	0,00	50,00	2,00	97,72
-3,95	0,00	-1,95	2,56	0,05	51,99	2,05	97,98
-3,90	0,00	-1,90	2,87	0,10	53,98	2,10	98,21
-3,85	0,01	-1,85	3,22	0,15	55,96	2,15	98,42
-3,80	0,01	-1,80	3,59	0,20	57,93	2,20	98,61
-3,75	0,01	-1,75	4,01	0,25	59,87	2,25	98,78
-3,70	0,01	-1,70	4,46	0,30	61,79	2,30	98,93
-3,65	0,01	-1,65	4,95	0,35	63,68	2,35	99,06
-3,60	0,02	-1,60	5,48	0,40	65,54	2,40	99,18
-3,55	0,02	-1,55	6,06	0,45	67,36	2,45	99,29
-3,50	0,02	-1,50	6,68	0,50	69,15	2,50	99,38
-3,45	0,03	-1,45	7,35	0,55	70,88	2,55	99,46
-3,40	0,03	-1,40	8,08	0,60	72,57	2,60	99,53
-3,35	0,04	-1,35	8,85	0,65	74,22	2,65	99,60
-3,30	0,05	-1,30	9,68	0,70	75,80	2,70	99,65
-3,25	0,06	-1,25	10,56	0,75	77,34	2,75	99,70
-3,20	0,07	-1,20	11,51	0,80	78,81	2,80	99,74
-3,15	0,08	-1,15	12,51	0,85	80,23	2,85	99,78
-3,10	0,10	-1,10	13,57	0,90	81,59	2,90	99,81
-3,05	0,11	-1,05	14,69	0,95	82,89	2,95	99,84
-3,00	0,13	-1,00	15,87	1,00	84,13	3,00	99,87
-2,95	0,16	-0,95	17,11	1,05	85,31	3,05	99,89
-2,90	0,19	-0,90	18,41	1,10	86,43	3,10	99,90
-2,85	0,22	-0,85	19,77	1,15	87,49	3,15	99,92
-2,80	0,26	-0,80	21,19	1,20	88,49	3,20	99,93
-2,75	0,30	-0,75	22,66	1,25	89,44	3,25	99,94
-2,70	0,35	-0,70	24,20	1,30	90,32	3,30	99,95
-2,65	0,40	-0,65	25,78	1,35	91,15	3,35	99,96
-2,60	0,47	-0,60	27,43	1,40	91,92	3,40	99,97
-2,55	0,54	-0,55	29,12	1,45	92,65	3,45	99,97
-2,50	0,62	-0,50	30,85	1,50	93,32	3,50	99,98
-2,45	0,71	-0,45	32,64	1,55	93,94	3,55	99,98
-2,40	0,82	-0,40	34,46	1,60	94,52	3,60	99,98
-2,35	0,94	-0,35	36,32	1,65	95,05	3,65	99,99
-2,30	1,07	-0,30	38,21	1,70	95,54	3,70	99,99
-2,25	1,22	-0,25	40,13	1,75	95,99	3,75	99,99
-2,20	1,39	-0,20	42,07	1,80	96,41	3,80	99,99
-2,15	1,58	-0,15	44,04	1,85	96,78	3,85	99,99
-2,10	1,79	-0,10	46,02	1,90	97,13	3,90	100,00
-2,05	2,02	-0,05	48,01	1,95	97,44	3,95	100,00

VIRI

[1] M. H. DeGroot, S. E. Fienberg & J. B. Kadane, *Statistics and the Law*, Wiley Interscience, 1994.

[2] D. Freedman, R. Pisani, R. Purves, *STATISTICS*, 3rd edition, Norton, 1998.

[3] Beaton, A. E., Mullis, I. V. S, Martin, M. O., Gonzales, E. J., Kelly, D. L., Smith, T. A., *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College, 1996.

[4] Beaton, A. E., Mullis, I. V. S, Martin, M. O., Gonzales, E. J., Kelly, D. L., *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College, 1996.

[5] R. L. Scheafer, M. Gnanadesikan, A. Watkins in J. Witmer, *Activity Based Statistics*, Instructor Resources, Springer-Verlag, 1996.

[6] R. L. Scheafer, M. Gnanadesikan, A. Watkins in J. Witmer, *Activity Based Statistics*, Student Resources Resources, Springer–Verlag, 1996.

[7] *Statistični letopis 1996*, Zavod Republike Slovenije za statistiko, 1996.

[8] S. M. Stiegler, *The History of Statistics*, The Belknap Press of Harvard Univeristy Press, 1986.

[9] S. M. Stiegler, *The History of Statistics*, Eight Centuries of sampling inspection: the trial of the Pxy, *Journal of the American Statistical Association*, **72**, str. 493-500, 1977.

Stvarno kazalo