Statistika: Znanost o podatkih

David S. Moore, Purdue University

30. oktober 2010

Kazalo

Uvod

Statistika je znanost, ki se ukvarja z zbiranjem, organiziranjem in interpretacijo numeričnih dejstev, ki jih imenujemo **podatki**. V vsakdanjem življenju smo zasuti s podatki. Ko slišimo besedo *statistika*, nas večina pomisli na drobce informacij, ki se pojavljajo v novicah: na športne rezultate, informacije o prodaji uvoženih avtomobilov, rezultate najnovejših anket o bližnjih volitvah, podatke o povprečni temperaturi na današnji dan. V oglasnih sporočilih velikokrat slišimo, kako določeni podatki potrjujejo, da je neki izdelek boljši od tistih, ki jih ponuja konkurenca. V javnih razpravah o ekonomiji, šolstvu in sociali vse strani podpirajo svoje argumente s podatki. V resnici pa uporabnost statistike presega te vsakodnevne primere.

Podatki imajo pomembno vlogo pri delu mnogih, zato je izobraževanje na področju statistike pomemben korak pri usposabljanju za celo vrsto poklicev. Ekonomisti, finančni svetovalci in vodstveni kader v politiki in gospodarstvu proučujejo aktualne podatke o brezposelnosti in inflaciji. Zdravniki morajo razumeti izvor in zanesljivost podatkov, ki se pojavljajo v medicinskih revijah. Poslovne odločitve temeljijo na raziskavah tržišča, ki razkrijejo želje potrošnikov. Inženirji zbirajo podatke o kvaliteti in zanesljivosti proizvedenih izdelkov. Večina akademskih področij uporablja števila in se zato poslužuje tudi statističnih metod.

Pobegniti podatkom je ravno tako nemogoče kot se izogniti uporabi besed. Ampak tako kot so besede na papirju brez pomena za nekoga, ki ne zna brati, in lahko zmedejo slabo izobražene, se tudi podatki ne razložijo sami od sebe, temveč zahtevajo določeno razumevanje. Pisec lahko besede zloži v prepričljive argumente ali pa v nepovezane nesmisle, in tudi podatki so lahko privlačni, zavajajoči ali pa enostavno irelevantni. Statistična pismenost, t.j. sposobnost sledenja in razumevanja argumentov, ki izhajajo iz podatkov, je pomembna za vsakogar izmed nas.

Poglavje 1

Pridobivanje podatkov

Novice so polne številk. Televizijski napovedovalec pove, da se je stopnja nezaposlenosti zmanjšala na 4.7%. Raziskava trdi, da je 45% Američanov zaradi kriminala strah ponoči zapustiti domove. Od kod pridejo te številke? Ne vprašamo vseh ljudi, če so zaposleni ali ne. Raziskovalne agencije vprašajo le nekaj posameznikov, če zaradi strahu pred ropi ostajajo ponoči doma.

Vsak dan se v novicah pojavi nov naslov. Eden od teh trdi: *Aspirin preprečuje srčne infarkte*. Nadaljnje branje razkrije, da je raziskava obravnavala 22 tisoč zdravnikov srednjih let. Polovica zdravnikov je vsak drugi dan vzela aspirin, druga polovica pa je dobila neaktivno tableto. V skupini, ki je jemala aspirin, je 139 zdravnikov doživelo srčni infarkt. V drugi skupini je bilo v enakem časovnem obdobju 239 infarktov. Ali je ta razlika dovolj velika, da lahko trdimo, da aspirin res preprečuje srčne infarkte?

Da bi ubežali neprijetnostim kot sta nezaposlenost in srčni infarkt, prižgimo televizijo. V pogovorni oddaji voditelj povabi gledalce, da sodelujejo v anketi. Tema pogovora je dobrodelnost in voditelja zanima, če gledalci redno prispevajo denar ali oblačila v dobrodelne namene. Med oddajo sprejmejo 50 tisoč klicev in 83% gledalcev trdi, da redno sodelujejo v tovrstnih akcijah. Ali je res, da smo tako zelo humanitarno osveščeni?

Zanesljivost teh številk je v prvi vrsti odvisna od njihovega izvora. Podatkom o nezaposlenosti lahko zaupamo, v tistih 83% iz pogovorne oddaje pa najbrž lahko utemeljeno podvomimo. To poglavje pojasni, zakaj. Naučili se bomo prepoznati dobre in slabe metode pridobivanja podatkov. Razumevanje metod, s katerimi lahko pridobimo zaupanja vredne podatke, je prvi (in najpomembnejši) korak k pridobi-

vanju sposobnosti odločanja o pravilnosti sklepov, ki jih izpeljemo na osnovi danih podatkov. Izpeljava zaupanja vrednih metod za pridobivanje podatkov je področje, kjer vstopimo v svet statistike, znanosti o podatkih.

1.1 Vzorčenje

Statistični urad želi ugotoviti, kakšen je odstotek nezaposlenosti med delavci. Raziskava želi določiti delež ljudi, ki ponoči ostajajo doma, ker se bojijo kriminala. Kontrolor kakovosti mora oceniti, koliko odstotkov proizvedenih ležajev ima pomanjkljivosti. V vseh teh situacijah želimo zbrati podatke o veliki skupini ljudi ali stvari. Ne moremo kontaktirati vsakega delavca ali pregledati vsakega ležaja posebej, ker je to predrago in preveč zamudno. Zberemo torej podatke o manjšem delu skupine in od tod sklepamo na lastnosti celotne skupine.

Celoto posameznikov, o kateri želimo zbrati podatke, imenujemo **populacija**. Populacijo lahko sestavljajo ljudje, živali ali stvari. **Vzorec** je del populacije, ki ga dejansko pregledamo, da bi zbrali podatke.

Velikokrat sklepamo o lastnostih celotne skupine na podlagi vzorcev. Vsakdo je že kdaj poskusil žlico juhe in si na podlagi tega ustvaril mnenje o celotni skledi juhe. Ampak juha v skledi je homogena, zato okus ene žlice res predstavlja celotno skledo. Izbiranje reprezentativnega vzorca velike in raznolike populacije ni tako enostavno. Najprej moramo natančno opredeliti populacijo, ki jo želimo opisati, nato pa moramo povedati, kaj želimo izmeriti. Kot pokaže primer, sta ta začetna koraka lahko zelo zapletena.

Primer. (Stopnja nezaposlenosti) Združene države Amerike dobijo ocene nezaposlenosti iz raziskav, ki vsak mesec zajamejo približno 50 tisoč gospodinjstev. Da bi izmerili nezaposlenost, moramo najprej določiti populacijo, ki jo želimo opisati. Katere starostne skupine bomo vključili? Ali bomo upoštevali tudi nelegalne priseljence in zapornike? Kaj pa redne študente? Odločimo se, da za populacijo vzamemo vse prebivalce (ne glede na državljanstvo), ki so starejši od 16 let, in niso na primer v zaporu.
Drugi korak je težji: Kaj pomeni, da je nekdo nezaposlen? Nekdo, ki ne išče dela, na primer redni študent, ne sme soditi v to skupino samo zato, ker ni plačan za svoje delo. Izpraševalec izbrane posameznike najprej vpraša, če so na voljo in so v

zadnjih štirih tednih iskali delo. Če je odgovor negativen, ni posameznik ne zaposlen ne nezaposlen - ni del delovne sile.
Kadar pa posameznik spada med delovno aktivno prebivalstvo, izpraševalec nadaljuje z vprašanji o zaposlitvi. Vsako delo za plačilo v lastnem podjetju v zadnjem tednu šteje za zaposlitev, kar velja tudi za vsaj 15 ur neplačanega dela v okviru družinskega podjetja. Prav tako spadajo med zaposlene ljudje, ki imajo službo, a trenutno ne delajo zaradi počitnic, stavk ali drugih tovrstnih razlogov. Torej 4.7% stopnja nezaposlenosti pomeni, da je bilo 4.7% posameznikov v izbranem vzorcu nezaposlenih po zgornjih definicijah delovne sile in nezaposlenosti. ♦

1.2 Slabe metode vzorčenja

Kako lahko izberemo vzorec, da bo res reprezentativen (tj. bo dobro predstavljal celotno populacijo)? Najlažji - pa ne najboljši - način, da pridemo do vzorca, je izbira posameznikov, ki so lahko dosegljivi. Če nas na primer zanima, koliko ljudi ima zaposlitev, se lahko odpravimo v nakupovalni center in povprašamo mimoidoče, če so zaposleni. Vzorec, ki ga sestavljajo najlažje dosegljivi posamezniki, imenujemo **priročni vzorec**. Z uporabo priročnih vzorcev običajno pridemo do nereprezentativnih podatkov.

Primer. (Priročni vzorci) Vzorec, sestavljen iz obiskovalcev nakupovalnega centra, omogoča hitro in poceni izvedbo ankete. Ampak ljudje, ki jih srečamo v nakupovalnih središčih, so najbrž bolj uspešni od povprečnega državljana. Precej verjetno je tudi, da so študentje ali pa upokojenci. Še več, ko se odločamo, koga bomo intervjuvali, običajno izbiramo bolje oblečene, urejene ljudi, izogibamo pa se tistim, ki so oblečeni slabo ali pa izgledajo neprijazno ali agresivno. Skratka, na ta način ne bomo prišli do vzorca, ki bi dobro predstavljal celotno populacijo, zato dobljeni rezultati ne bodo odražali dejanske stopnje nezaposlenosti. ♦

Vzorec iz nakupovalnega središča bo gotovo vseboval preveč predstavnikov srednjega razreda in upokojencev in premalo predstavnikov revnejših slojev. To se bo zgodilo vsakič, ko bomo vzorec izbirali na ta način. Ta napaka je torej sistematična in se pojavi zaradi slabo izbrane metode vzorčenja, ne gre zgolj za enkraten ponesrečen poskus. Takšno sistematično razliko med dobljenimi rezultati in dejanskim stanjem imenujemo **pristranskost**.

Načrt raziskave je **pristranski**, če se sistematično nagiba k določenim izidom.

Primer. (Telefonsko glasovanje) Televizijski programi se radi poslužujejo telefonskega glasovanja, da bi prikazali javno mnenje. Gledalcem zastavijo vprašanje in jih prosijo, naj pokličejo na eno od številk, če je njihov odgovor DA, in na drugo, če je odgovor NE. Vsak klic gledalec plača. Ameriška televizijska hiša ABC je nekoč v oddaji *Nightline* vprašala gledalce, če menijo, da bi moral sedež Združenih narodov ostati v ZDA. Odzvalo se je več kot 186 tisoč gledalcev in 67% je bilo mnenja, da ne.
Ljudje, ki žrtvujejo svoj čas in denar, da odgovorijo na tovrstne ankete, ne predstavljajo celotne populacije. Pravzaprav so to običajno tisti ljudje, ki kličejo v radijske oddaje. Bolj je verjetno, da se bodo odzvali ljudje, ki imajo o obravnavani tematiki trdno stališče, še zlasti tisti z izrazito negativnim mnenjem. Ni torej presenetljivo, da je podobna raziskava s pravilneje izbranim vzorcem pokazala, da 72% odraslih želi sedež Združenih narodov obdržati v ZDA. ♦

Telefonska glasovanja so primer *prostovoljnega vzorca*. Tak vzorec lahko brez težav privede do 67% negativnih odgovorov, medtem ko je dejansko stanje bližje 72% pozitivnih odgovorov.

Prostovoljni vzorec je sestavljen iz posameznikov, ki se odzovejo na splošen poziv k sodelovanju. Ti vzorci so pristranski, ker je večja verjetnost, da se odzovejo posamezniki z odločnimi (predvsem negativnimi) prepričanji.

1.3 Enostavni slučajni vzorci

V primeru prostovoljnega vzorca se ljudje sami odločijo za sodelovanje, v priročnem vzorcu je izbira na strani izpraševalca, v obeh primerih pa osebna izbira povzroči pristranskost. Statistik se temu izogne tako, da uporabi neosebno slučajno izbrani vzorec. Slučajno izbrani vzorec izključi možnost favoriziranja tako s strani izpraševalca kot s strani anketirancev. Pristranskosti se izognemo, ker imajo vsi posamezniki v populaciji enako možnost, da so izbrani.

Najpreprostejši način za slučajno izbiro vzorca je princip loterije: imena (populacijo)

napišemo na listke, jih damo v klobuk, dobro premešamo in nato izvlečemo pest listkov (vzorec). Tako dobimo *enostavni slučajni vzorec*.

Enostavni slučajni vzorec velikosti n sestavlja n posameznikov, ki jih iz populacije izberemo tako, da ima vsaka skupina n posameznikov enako možnost, da je izbrana.

Primer z listki v klobuku nam pomaga, da si predstavljamo enostavni slučajni vzorec. Ista predstava nam pomaga razumeti, da je ta metoda boljša kot priročno ali prostovoljno vzorčenje, ker nobenemu delu populacije ni naklonjena bolj kot kateremu drugemu. Ampak pisanje imen na listke in žrebanje je počasno in nepraktično. Še posebej to pride do izraza, kadar delamo z velikimi vzorci, na primer z vzorcem 50 tisoč posameznikov pri raziskavi nezaposlenosti. Postopek lahko pospešimo z uporabo *tabele naključnih števil*. V praksi statistiki uporabljajo računalnike, ki namesto njih opravijo delo, za majhne vzorce pa lahko to storimo tudi na roke.

Tabela naključnih števil je daljše zaporedje števk $0, 1, \ldots, 9$, ki ima naslednji dve lastnosti:

(1) Vsak element tabele je z enako verjetnostjo katerakoli od teh desetih števk.

(2) Elementi tabele so med seboj neodvisni. To pomeni, da poznavanje kateregakoli elementa ali dela tabele ne pove ničesar o nobenem drugem delu tabele.

Tabela 1.1 je primer tabele naključnih števil. Števke so v tabeli zbrane v skupinah po pet zaradi lažje berljivosti, vrstice pa so oštevilčene, da se lahko nanje sklicujemo. Tako grupiranje in označevanje je dodano zgolj zaradi večje preglednosti. V resnici je celotna tabela le en dolg niz naključno izbranih števk. S pomočjo take tabele lahko izberemo preprosti slučajni vzorec v dveh korakih:

(1) Označevanje. Vsakemu pripadniku populacije dodelimo številsko oznako iste dolžine. Z dvomestnimi oznakami lahko torej popišemo do 100 pripadnikov, s trimestnimi do 1000 in tako naprej.

(2) Uporaba tabele. Iz tabele preberemo zaporedne števke v skupinah, katerih dolžine so enake prej izbrani dolžini oznak. Vzorec sestavimo iz posameznikov

z oznakami, ki jih dobimo iz tabele. Na ta način imajo vsi posamezniki enake možnosti, da so izbrani, ker imajo vse oznake iste dolžine enako možnost, da se pojavijo v tabeli. Tako je na primer vsak par števk iz tabele z enako verjetnostjo enak eni od 100 možnih dvomestnih oznak $00, 01, \ldots, 99$. Če v tabeli naletimo na skupino števk, ki ni bila uporabljena za označitev posameznikov, jo izpustimo. Prav tako preskočimo morebitne ponovitve.

Primer. (Pregled avtomobilov) Proizvajalec avtomobilov želi izmed zadnjih 50 izdelanih avtomobilov izbrati 5 primerkov za natančen tehnični pregled. Zakaj bi najverjetneje prišlo do pristranskosti, če bi lahko avtomobile izbrali delavci? Da se pristranskosti izognemo, uporabimo preprosti slučajni vzorec.

(1) Vsakemu od avtomobilov dodelimo številsko oznako. Vse oznake bodo dvomestne, začnemo pa z 00. Oznake so torej od 00 do 49. Prav tako bi lahko izbrali oznake od 01 do 50.

(2) V tabeli si izberemo vrstico 140 (lahko pa bi izbrali tudi katerokoli drugo):

73063 63623 29388 89507 78553 62792 89343 27401.

Ker so naše oznake dvomestne, beremo iz tabele po dve števili naenkrat. Pare, ki se med oznakami ne pojavijo, ignoriramo. Na začetku torej izpustimo 73. Ignoriramo tudi ponovljene oznake, torej drugo pojavitev 36 v tej vrstici, ker seveda ne moremo izbrati istega avtomobila dvakrat. Naš vzorec tako vsebuje avtomobile z oznakami 06, 36, 23, 29 in 38. ♦

Primer. (Vzorčenje stanovanjskih enot) Večina nacionalnih raziskav izbira vzorce v več stopnjah. Ameriški statistični urad na primer vsak mesec izvede raziskavo prebivalstva z naslednjim vzorcem:

(1) Državo razdelijo na 2007 geografskih enot, imenovanih *osnovne enote vzorčenja*. Od tod izberejo vzorec 754 enot. Ta vzorec vsebuje 428 enot z največ prebivalstva in slučajni vzorec ostalih.

(2) Vsaka od enot je razdeljena na manjša področja, imenovana *bloki*. Ti bloki so naprej klasificirani v *plasti* glede na etnične in druge lastnosti. V okviru vsake od plasti nato izberejo slučajni vzorec blokov.

(3) Stanovanjske enote znotraj vsakega bloka so razdeljene v gruče, od katerih je vsaka sestavljena iz štirih sosednjih enot. Od tu izberejo slučajni vzorec gospodinjstev, ki so nato anketirana.

101	03918	86495	47372	21870	28522	99445	38783	83307
102	10041	35095	66357	64569	08993	20429	28569	63809
103	43537	58268	80237	17407	89680	04655	24678	61932
104	64301	47201	31905	60410	80101	33382	95255	10353
105	43857	42186	77011	93839	28380	49296	63311	49713
106	91823	39794	47046	78563	89328	39478	04123	19287
107	34017	87878	35674	39212	98246	29735	09924	27893
108	49105	00755	39242	50472	39581	44036	54518	46865
109	72479	02741	75732	99808	02382	77201	44932	88978
110	84281	45650	28016	77753	39495	41847	19634	82681
111	61589	35486	59500	20060	89769	54870	75586	07853
112	25318	01995	87789	41212	74907	90734	31946	24921
113	40113	37395	51406	98099	43023	70195	07013	72306
114	58420	43526	15539	24845	15582	16780	95286	69021
115	18075	45894	09875	42869	20618	07699	80671	54287
116	52754	73124	93276	71521	59618	44966	37502	15570
117	05255	53579	08239	99174	75548	95776	42314	13093
118	76032	35569	28738	38092	74669	00749	17832	64855
119	97050	31553	32350	51491	53659	89336	36912	05292
120	29030	43074	84602	95131	22769	44680	68492	33987
121	28124	29686	63745	12313	15745	11570	20953	17149
122	97469	41277	90524	36459	22178	63785	20466	67130
123	91754	40784	38916	12949	76104	20556	34001	59133
124	84599	29798	57707	57392	91757	76994	43827	69089
125	06490	42228	94940	10668	62072	58983	10263	08832
126	30666	02218	89355	76117	75167	69005	42479	79865
127	87228	15736	08506	29759	74257	85594	75154	48664
128	45133	49229	32502	99698	68202	44704	39191	73740
129	55713	98670	57794	64795	27102	83420	26630	95009
130	20390	38266	30138	61250	07527	02014	43972	49370
131	13400	68249	32459	41627	56194	93075	50520	96784
132	08900	87788	73717	19287	69954	45917	80026	55598
133	86757	47905	16890	99047	78249	73739	97076	00525
134	19862	54700	18777	22218	25414	13151	54954	80615
135	96282	11576	59837	27429	60015	40338	39435	94021
136	17463	26715	71680	04853	55725	87792	99907	67156
137	44880	55285	95472	57551	24602	98311	63293	58110
138	61911	78152	96341	31473	58398	61602	38143	93833
139	07769	22819	58373	88466	71341	32772	93643	92855
140	73063	63623	29388	89507	78553	62792	89343	27401
141	24187	60720	74055	36902	22047	09091	79368	35408
142	06875	53335	91274	87824	04137	77579	54266	38762
143	23393	37710	46457	03553	58275	11138	18521	59667
144	00980	73632	88008	10060	48563	31874	90785	78923
145	46611	39359	98036	25351	88031	72020	13837	03121
146	56644	79453	49072	30594	73185	81691	29225	70495
147	98350	36891	04873	71321	29929	37145	95906	41005
148	17444	61728	86112	76261	92519	61569	65672	95772
149	45785	21301	89563	23018	60423	50801	70564	45398
150	54369	08513	36838	19805	67827	74938	66946	01206

Tabela 1.1: Tabela naključnih števil.

V zadnjem koraku moramo torej izbrati enostavni slučajni vzorec treh od 189 gruč iz vsakega bloka. Za gruče potrebujemo trimestne oznake, priredimo jim števila med 001 in 189 (lahko pa bi se odločili tudi za števila med 000 in 188). Nato preberemo števila iz tabele naključnih števil v skupinah po tri. Če se na primer odločimo, da bomo začeli v vrstici 135, bodo izbrane gruče 157, 001 in 117. Večino trimestnih skupin iz tabele moramo pri tem preskočiti, ker ne označujejo nobene gruče. ♦

Takemu vzorcu pravimo *večstopenjski slučajni vzorec* in ima več praktičnih prednosti pred enostavnim slučajnim vzorcem. Ne potrebujemo seznama vseh gospodinjstev v državi, samo seznam tistih, ki so v blokih, izbranih v drugem koraku. Če je potrebno, lahko ta seznam sestavimo z obiskom teh lokacij. Še več, gospodinjstva, ki jih anketiramo, so grupirana na relativno majhnem številu lokacij, kar zmanjša potne stroške anketarjev. Cena, ki jo plačamo za to praktičnost, pa je večja kompleksnost pri izboru vzorcev in pri interpretaciji pridobljenih podatkov. Ker je enostavno slučajno vzorčenje osnovno načelo, ki stoji za vsakim slučajnim vzorčenjem, in ker je tudi osnova za bolj kompleksne vzorce, bo osrednja tema našega študija.

1.4 Statistično ocenjevanje

Vzorec izberemo z namenom, da bi dobili podatke o populaciji. Če je vzorec izbran slučajno, pričakujemo, da bo imel podobne lastnosti kot celotna populacija. Rezultat, ki ga je dal dani vzorec, torej uporabimo za *oceno* lastnosti celotne populacije.

Primer. (Statistično ocenjevanje) Agencija Gallup je vzorcu 1493 ljudi zastavila naslednje vprašanje: *"Vas je ponoči strah zapustiti stanovanje zaradi kriminala?"* Pri tem je 672 posameznikov odgovorilo pritrdilno. Pozitivnih odgovorov je bilo torej

$$\frac{672}{1493} = 0{,}45 = 45\%.$$

Populacija, vključena v raziskavo, je bila sestavljena iz polnoletnih prebivalcev. Ne vemo, kolikšen odstotek populacije bi odgovoril pritrdilno, če bi jim zastavili to vprašanje. Ker pa so imeli vsi posamezniki enako možnost biti vključeni v raziskavo, pričakujemo, da izbrani vzorec predstavlja populacijo. Ocenjujemo torej, da se približno 45% populacije ponoči ne upa zdoma zaradi kriminala. ♦

Ni prav verjetno, da je delež tistih, ki si ponoči ne upajo zapustiti domov zaradi kriminala, točno 45%. Trdimo lahko le, da je rezultat, ki ga je dal vzorec, verjetno precej blizu resnični vrednosti. Če bi agencija Gallup izbrala nek drug vzorec, bi

ga sestavljali drugi posamezniki, ki bi verjetno imeli nekoliko drugačen pogled na kriminal. Če bi 641 vprašanih odgovorilo pritrdilno, bi ocenili, da je ustrezni delež enak

$$\frac{641}{1493} = 0{,}43 = 43\%.$$

Temu pravimo *vzorčna spremenljivost*: kadar izbiramo različne vzorce iste populacije, se dobljeni rezultati razlikujejo. Slučajno vzorčenje se sicer izogne pristranskosti, vzorčni spremenljivosti pa ne.

Pri nekem vzorcu dobimo 45%, pri nekem drugem pa 43%. Ali je možno, da pri drugih vzorcih dobimo 13% ali pa 89%? Ali lahko zaupamo rezultatom, ki jih da vzorec, čeprav vemo, da bi pri drugačnem vzorcu dobili drugačen rezultat? Lahko. Da bi razumeli zakaj, si moramo vzorčno spremenljivost ogledati podrobneje.

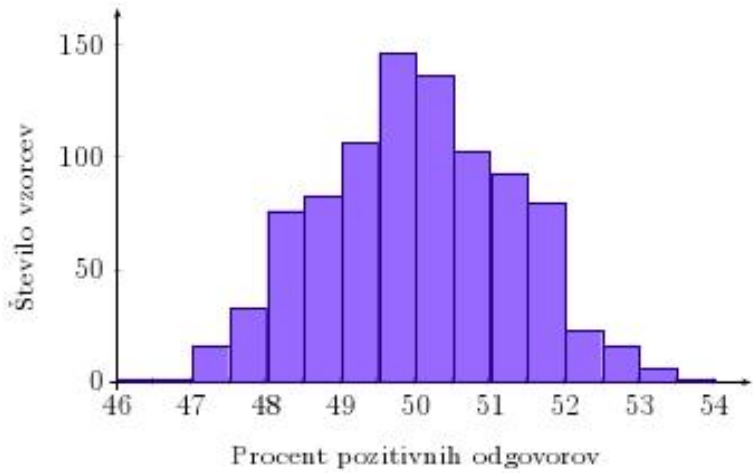
Obstajajo različne vrste spremenljivosti. Razlike v rezultatih, ki jih dobimo, če anketarja pošljemo v nakupovalni center, so nepredvidljive. Pri slučajnem vzorčenju pa te razlike pokažejo določeno regularnost. Razlog se skriva v načinu izbire slučajnega vzorca. Dolgoročni rezultati v tem primeru niso slučajni. Te vrste pravilnosti lahko opazimo v igrah na srečo, na primer pri metu kovanca. Pravzaprav je 1493 metov kovanca enakovredno izbiri 1493 posameznikov iz dane populacije, če je mnenje v tej populaciji deljeno enakomerno in glava pomeni pritrdilen, cifra pa nikalen odgovor. Rezultati se v obeh primerih spreminjajo, vendar lahko napovemo, kako močno se bodo spreminjali, ker se na dolgi rok pojavijo pravilni vzorci. Oglejmo si spremenljivost rezultatov pri veliko slučajnih vzorcih.

Primer. (Poskus z vzorčenjem) Vrnimo se h Gallupovi raziskavi. Recimo, da bi točno polovica populacije odgovorila pritrdilno (čeprav Gallup seveda tega ne ve). Ali smo lahko prepričani, da bo vzorec 1493 posameznikov dal rezultat, ki bo blizu te vrednosti?

Da bi odgovorili na to vprašanje, izberimo 1000 enostavnih slučajnih vzorcev iz te populacije, in si oglejmo rezultate, ki jih dobimo s temi vzorci. S prvim vzorcem smo dobili 50,2% pozitivnih odgovorov, z drugim 49,2%, s tretjim 50,4%, in tako dalje. Rezultati se spreminjajo.

Vse dobljene rezultate prikažemo v **histogramu**. Višina vsakega stolpca prikazuje, kako pogosto so se pojavili odgovori, ki jih pokriva baza tega stolpca. Na primer, višina stolpca, ki pokriva vrednosti med 48% in 48,5% je 80, ker se je v 80 od 1000 primerov zgodilo, da je bil delež pozitivnih odgovorov med 48% in 48,5%. (Podrobneje si bomo histograme ogledali v naslednjem poglavju.) Če proučimo ta histogram, bomo razumeli, zakaj lahko zaupamo ocenam, ki jih dobimo iz slučajnih

vzorcev. ♦



Slika 1.1: Rezultat 1000 vzorčenj z vzorci velikosti 1493 iz populacije, v kateri bi 50% vprašanih odgovorilo pritrdilno.

Iz histograma je razvidno, da je vseh 1000 vzorcev dalo rezultate med 46% in 54%. Drugače povedano, vsi rezultati so se od točne vrednosti razlikovali za manj kot 4 odstotne točke. Poleg tega so v histogramu vidne jasne pravilnosti v rezultatih. Sredina je pri 50%. Tam so stolpci najvišji in se znižujejo, ko se od sredine oddaljujemo. To pomeni, da so rezultati, ki so bližje dejanski vrednosti, pogostejši, rezultati, ki se od dejanske vrednosti bolj razlikujejo, pa so redkejši. Srednjih 95% rezultatov leži med 47,4% in 52,6%. Kaže, da bo torej vzorec velikosti 1493 skoraj vedno dal rezultat, ki se bo razlikoval od dejanske vrednosti kvečjemu za ±4%, v večini primerov (v 95% primerov) pa se bo rezultat razlikoval kvečjemu za ±2,6%. S precejšnjo gotovostjo lahko torej trdimo, da bomo z enim takim vzorcem prišli do rezultata, ki bo blizu resničnemu stanju v populaciji. To je dobra novica. V ZDA je skoraj 200 milijonov odraslih, vendar zadošča, da naključno izberemo le 1493 posameznikov, pa lahko precej natančno opišemo mnenje celotne populacije.

Pravilna oblika histograma na sliki ni naključna. Kadar vzorce izbiramo naključno, bo histogram, ki ga dobimo iz velikega števila vzorcev, vedno te oblike. Ni potrebno obravnavati na tisoče vzorcev, da bi ugotovili, za katero obliko gre. Matematična teorija verjetnosti nam omogoča, da jo že vnaprej določimo. Več o tem se bomo naučili v kasnejših poglavjih. Tukaj sta osnovni dejstvi, ki pojasnita, zakaj lahko zaupamo ocenam, ki jih dobimo z vzorci:

- Pri velikem številu slučajnih vzorcev *se rezultati kopičijo okoli dejanske vre-*

dnosti. V prejšnjem primeru je bila ta vrednost 50%, ker je bil dejanski delež pozitivnih odgovorov v celotni populaciji 50%. Če bi vzorce jemali iz populacije, v kateri bi bil ta delež enak 40%, bi se rezultati nakopičili okoli vrednosti 40%. Središče histograma tako dokazuje, da pri slučajnem vzorčenju ne prihaja do pristranskosti. Vsak posamezni vzorec lahko da rezultat, ki je manjši ali večji od dejanske vrednosti, ni pa nobene sistematične težnje po previsokih ali prenizkih rezultatih.

- *Širina histograma je določena z velikostjo vzorca*. Pri večjih vzorcih se dobljeni rezultati zbirajo bližje dejanski vrednosti kot v primeru manjših vzorcev. Večji kot je torej vzorec, bolj smo lahko prepričani, da je dobljeni rezultat dober približek za dejansko vrednost. Javnomnenjske raziskave običajno vključijo med tisoč in dva tisoč ljudi. Raziskava o prebivalstvu, ki jo vsak mesec izvaja ameriški statistični urad, zajame vzorec 50 tisoč ljudi, ker želi ameriška vlada zelo natančno oceno stopnje nezaposlenosti.

Natančnost rezultata običajno opišemo z *mejo napake*. Ne moremo z gotovostjo trditi, da so rezultati, dobljeni z izbranimi vzorci, tako blizu dejanski vrednosti, kot pravi meja napake. Navsezadnje vzorce izbiramo naključno in lahko bi se zgodilo, da bi izbrali zelo ponesrečeno. V primeru raziskave o kriminalu bi lahko izbrali ravno tistih 1493 posameznikov, ki živijo na območju visoke stopnje kriminala, ne pa vzorca, ki bi dobro predstavljal populacijo. Pri takem vzorcu bi blizu 100% vprašanih odgovorilo, da se boji kriminala. Ustrezen histogram pokaže, da se to ne zgodi skoraj nikoli, če je dejanska vrednost 50%. Običajno dobimo mejo napake iz sredinskih 95% rezultatov.

Meja napake nam običajno pove, kako blizu dejanskemu rezultatu bo rezultat v 95% vzorcev, izbranih po metodi, ki je bila uporabljena za vzorec, s katerim smo prišli do ocene.

V časopisu preberemo: "Najnovejše analize kažejo, da se samo 34% Američanov strinja s predsednikovim načinom vodenja države." Meja napake za to raziskavo je $\pm 3\%$. To pomeni: "Ta rezultat smo dobili po metodi, katere rezultati se od resnične vrednosti v 95% primerov razlikujejo za manj kot 3%." Ta konkretni rezultat je lahko eden izmed tistih 5% primerov, ki se od dejanske vrednosti razlikujejo za več, vendar pa nam da poznavanje meje napake približno idejo o natančnosti raziskave.

Primer. (Meja napake v Gallupovih raziskavah) V našem primeru vzorčenja smo izbrali veliko enostavnih slučajnih vzorcev. Agencija Gallup in ameriški

statistični urad uporabljata kompleksneje načrtovane vzorce. Ker pa ti prav tako temeljijo na naključni izbiri, so oblike histogramov tudi v teh primerih podobne tistemu, ki smo si ga ogledali. Za Gallupove raziskave velja, da je meja napake

- približno ±5% pri vzorcih velikosti okoli 600,
- približno ±4% pri vzorcih velikosti okoli 1000 in
- približno ±3% pri vzorcih velikosti okoli 1500.

Agencija Gallup je anketirala 1514 odraslih in ugotovila, da 53% vprašanih nasprotuje uvedbi daljšega šolskega leta. Meja napake je ±3%. Lahko smo torej precej prepričani, da med 50% (53% − 3%) in 56% (53%+3%) vseh odraslih nasprotuje podaljšanju šolskega leta. ♦

1.5 Eksperimenti

Raziskave vzorcev zberejo informacije o delu populacije z namenom, da bi lahko sklepale o celotni populaciji. Kadar je naš cilj opisati populacijo, je statistično vzorčenje prava izbira.

Zdaj pa predpostavimo, da želimo proučevati reakcijo na dražljaj, da bi ugotovili, kako ena spremenljivka vpliva na drugo, ko spreminjamo obstoječe pogoje. Ali bo nov učni načrt za matematiko prispeval k izboljšanju rezultatov v šestem razredu na nacionalnih preverjanjih znanja? Ali bo redno jemanje majhnih količin aspirina zmanjšalo nevarnost srčnega infarkta? Ali kajenje med nosečnostjo znižuje otrokov IQ? Študije, ki zgolj opazujejo in opisujejo, niso učinkovite pri iskanju odgovorov na ta vprašanja. Boljše odgovore nam dajo *eksperimenti*.

Opazovalna študija, na primer vzorčna analiza, opazuje posameznike in meri spremenljivke, ki jo zanimajo, vendar ne poskuša vplivati na odzive. **Eksperiment** načrtno podvrže posameznike določeni *terapiji* z namenom opazovati njihove odzive.

Eksperimentov se poslužujemo, kadar želimo izmeriti vpliv ene spremenljivke na drugo. Z izbiro določene terapije in omejevanjem ostalih vplivov lahko natančno določimo vzroke in posledice. Vzorčna analiza lahko pokaže, da sta dve spremenljivki povezani, ne more pa dokazati, da je ena od njiju vzrok druge. Tako kot

nam statistika predlaga metode vzorčenja, nam pove tudi nekaj o tem, kako naj bi načrtovali eksperimente.

Primer. (Nenadzorovani eksperiment) V Osnovni šoli kralja Matjaža so zaskrbljeni nad slabo matematično pripravljenostjo učencev, zato sestavijo nov ambiciozen učni načrt. Po treh letih izvajanja novega programa opazijo pri učencih, ki so končali šesti razred, za 10% boljši uspeh kot pri starem načrtu in proglasijo nov učni načrt za uspešnega.

Načrt tega eksperimenta je preprost. Skupina osebkov (učencev) je bila podvržena terapiji (novemu učnemu načrtu), nato pa so opazovali rezultate (dosežek na preverjanju znanja). Torej:

$$\text{Nov učni načrt} \rightarrow \text{spremljaj rezultate}$$

ali splošneje

$$\text{Terapija} \rightarrow \text{spremljaj odziv.}$$

♦

Večina laboratorijskih eksperimentov je zasnovana podobno: uporabi določeno terapijo in izmeri odziv. V kontroliranem laboratorijskem okolju se najbolje obnesejo preprosto zasnovani poskusi. Terenske raziskave in raziskave na ljudeh pa so podvržene bolj raznolikim vplivom in večji raznolikosti osebkov. Ni vedno mogoče nadzorovati zunanjih dejavnikov, ki lahko vplivajo na rezultat. Z večjo raznolikostjo se veča tudi potreba po statističnem načrtu. V Osnovni šoli kralja Matjaža je zaskrbljenost glede izobrazbe pripeljala do številnih hkratnih sprememb, ki bi lahko vplivale na dosežke učencev. Učitelji so bili deležni dodatnega izobraževanja, svet staršev je organiziral inštrukcije za individualno pomoč, javna zaskrbljenost je povzročila, da so starši postali bolj pozorni na otrokov napredek in učitelji so začeli predpisovati več domačih nalog.

V teh okoliščinah bi se matematična pripravljenost povečala tudi brez novega učnega načrta. Pravzaprav je lahko novi učni načrt celo slabši od starega. Eksperiment, ki ga je izvedla šola, ne more razločiti med vplivi sprememb s strani staršev in učiteljev in vplivi novega učnega načrta. Novi učni načrt je torej *pomešan* z ostalimi spremembami, ki so se pojavile hkrati z njim.

Spremenljivke, ki so ali pa niso del študije, so **pomešane**, kadar ne moremo razločiti njihovih vplivov na izid.

1.6 Slučajeni primerjalni eksperimenti

Zapletom s pomešanimi spremenljivkami se lahko izognemo, če naredimo primerjalni eksperiment, pri katerem ena skupina učencev sledi novemu učnemu načrtu, druga pa staremu. Drugo skupino učencev imenujemo **kontrolna skupina**. Spremembe v odnosu in sodelovanju staršev, izobraževanje učiteljev in drugi vplivi so zdaj prisotni v enaki meri v obeh skupinah učencev, zato lahko vidimo vplive novega učnega načrta pri primerjavi s kontrolno skupino. Večina dobro načrtovanih eksperimentov primerja eno ali več terapij.

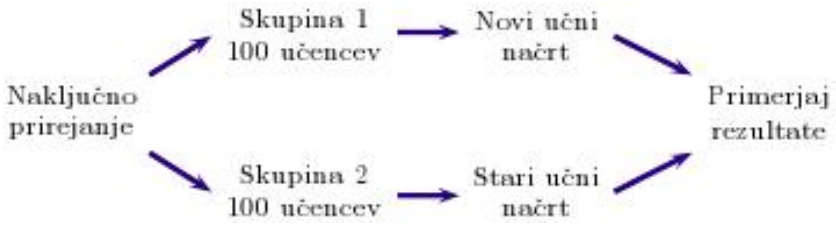
Vendar pa samo primerjava ni dovolj, da bi prišli do rezultatov, ki jim lahko zaupamo. Če podvržemo terapijama skupini, ki se bistveno razlikujeta ob začetku eksperimenta, bodo rezultati pristranski. Če na primer dovolimo, da se učenci sami odločijo, kateremu načrtu bodo sledili, se bodo za novi načrt odločili predvsem tisti, ki jih matematika zanima, in ti bodo najverjetneje dosegali dobre rezultate. Osebna izbira bo privedla do pristranskih rezultatov, podobno kot v primeru prostovoljnega telefonskega glasovanja. Tudi rešitvi teh dveh problemov sta podobni: skupini oziroma vzorec izberemo naključno.

Primer. (Slučajeni primerjalni eksperiment)[1] V Osnovni šoli kralja Matjaža se odločijo, da bodo primerjali napredek 100 učencev, ki sledijo novemu učnemu načrtu, z napredkom 100 učencev, ki ostanejo pri starem. Učence, ki bodo sledili novemu načrtu izberemo kot enostavni slučajni vzorec moči 100 izmed populacije 200 posameznikov. Preostalih 100 učencev bo sledilo staremu načrtu.

Na ta način dobimo **slučajeni primerjalni eksperiment** z dvema skupinama. Na sliki 1.2 je grafično predstavljen načrt eksperimenta.

Postopek izbiranja je povsem enak kot pri vzorčenju: označimo posameznike in uporabimo tabelo. Najprej učencem priredimo oznake med 000 in 199. Potem iz tabele naključnih števil prebiramo po tri zaporedne števke. Prvih 100 oznak, ki se pojavijo v tabeli, predstavlja tistih 100 učencev, ki bodo sledili novemu programu. Kot običajno ignoriramo ponavljanja in skupine, ki ne predstavljajo nobene oznake. Če na primer začnemo v vrstici 125, ima prvih nekaj izbranih učencev oznake 064, 106, 102, 022 in 188. ♦

[1] Gre za prevod angleškega termina *randomized*. Več o slučajenih primerjalnih eksperimentih in verjetnosti nasploh lahko radovedni bralec najde v knjigi J.S. Rosenthala, *Ko strela udari*, ki je v prevodu izšla pred kratkim v zbirki Sigma. (Op. prev.)



Slika 1.2: Načrt slučajenega primerjalnega eksperimenta, ki ugotavlja uspešnost novega matematičnega učnega načrta.

Pod žarometom

Sir Ronald A. Fisher

1890 - 1962

Ideje in metode, ki jih danes poznamo pod imenom *statistika*, so v devetnajstem in dvajsetem stoletju izumili ljudje, ki so se ukvarjali s problemi, pri katerih je bilo potrebno analizirati velike količine podatkov. Astronomija, biologija, družboslovne vede in celo geodezija lahko trdijo, da so igrale pomembno vlogo pri rojstvu statistike. Če pa si kdo zasluži naziv "oče statistike", je to Sir Ronald A. Fisher.

Fisherjevi zapiski so opredelili statistiko kot posebno področje proučevanja, katerega metode lahko uporabimo v številnih panogah. Sistematiziral je matematično teorijo statistike in izumil številne nove metode. Slučajeni primerjalni eksperiment je najbrž njegov največji dosežek.

Kot druge statistične pionirje so tudi Fisherja gnale zahteve praktičnih problemov. Od leta 1919 naprej je delal na področju kmetijstva na terenu v Rothamstedu v Angliji. Kako naj razporedimo sajenje različnih vrst pridelka ali pa uporabo različnih gnojil, da jih lahko primerjamo? Ker se rodovitnost in druge lastnosti spreminjajo, ko se premikamo po polju, so eksperimentatorji uporabljali zamotane vzorce, ki so posnemali šahovnico. Fisher je imel boljšo idejo: namenoma uredimo parcele naključno.

Slučajeni primerjalni eksperimenti so relativno nova ideja. V dvajsetih letih dvajsetega stoletja jih je vpeljal Sir R. A. Fisher. Eksperiment Osnovne šole kralja Matjaža je *primerjalni*, ker primerja dve različni terapiji (dva matematična učna načrta). Je *slučajeni*, ker izberemo osebke, ki bodo sledili posamezni terapiji, naključno. Naključna izbira ustvari skupini, ki sta pred začetkom eksperimenta podobni. Ker skupini primerjamo, spremenljivke, ki bi jih lahko pomešali, delujejo na obe skupini

hkrati. Edina razlika med skupinama je v učnem načrtu. Če torej opazimo razlike v napredku, mora tičati vzrok teh razlik v učnem načrtu. To je osnovna ideja na kateri temelji slučajeni primerjalni eksperiment. Kasneje bomo videli, da je potrebno poskrbeti še za nekaj malenkosti, vendar pa nam ta osnovna ideja pokaže, zakaj lahko s takimi eksperimenti res dokažemo, da so različni izidi posledica razlike v terapijah. Slučajene primerjalne eksperimente uporabljamo, kadar obstaja nevarnost, da bi se okoljske spremenljivke (kot je v primeru šole kralja Matjaža obnašanje staršev in učiteljev) pomešale s terapijami. Oglejmo si še en primer, v katerem primerjamo tri terapije.

Primer. (Varčevanje z energijo) Elektro Slovenije je predstavilo programe, ki ljudi spodbujajo k varčevanju z energijo. V podjetju razmišljajo, da bi v vsako gospodinjstvo namestili elektronski števec, ki bi prikazoval višino mesečnega računa ob predpostavki, da se nadaljuje s trenutno porabo energije. Ali bi taki števci prispevali k zmanjšanju porabe? Bi lahko isto dosegli s cenejšimi metodami? Odločijo se za eksperiment, ki bo odgovoril na ta vprašanja.

Cenejša je odločitev za brošure s tabelo in navodili za spremljanje porabe. Eksperiment primerja ta dva pristopa, vključuje pa tudi kontrolno skupino. Kontrolna skupina dobi informacije o varčevanju z energijo, ne dobi pa nobene pomoči pri spremljanju porabe. Merimo celotno letno porabo energije. Podjetje je našlo 60 enodružinskih gospodinjstev v istem mestu, ki so pripravljena sodelovati, zato v vsako od treh skupin naključno izberemo po 20 gospodinjstev. Struktura eksperimenta je prikazana na sliki 1.3.

Naključno razdelitev izvedemo tako, da priredimo gospodinjstvom oznake med 01 in 60, s pomočjo tabele izberemo enostavni slučajni vzorec 20 gospodinjstev, ki bodo prejeli števce, nato pa še 20, ki bodo dobili brošure. Preostalih 20 gospodinjstev sestavlja kontrolno skupino. ♦

Slučajeni primerjalni eksperimenti so pogosti v industrijskem in akademskem raziskovanju. Prav tako so pogosto uporabljeni pri medicinskih raziskavah. Po predpisih mora biti na primer varnost in učinkovitost novega zdravila dokazana s slučajenim primerjalnim eksperimentom. Oglejmo si en tak primer.

Primer. (Študija na zdravnikih) Sumimo, da redno jemanje aspirina zmanjša nevarnost srčnega infarkta. Možno je tudi, da redni odmerki beta karotena (ki ga telo spremeni v vitamin A), pomagajo preprečevati nekatere vrste raka. Ti trditvi so preverjali z eksperimentom, ki so ga izvedli na 22 tisoč zdravnikih, starejših od 40 let. Vsak posameznik je v obdobju nekaj let vsak dan vzel eno tableto. Terapije

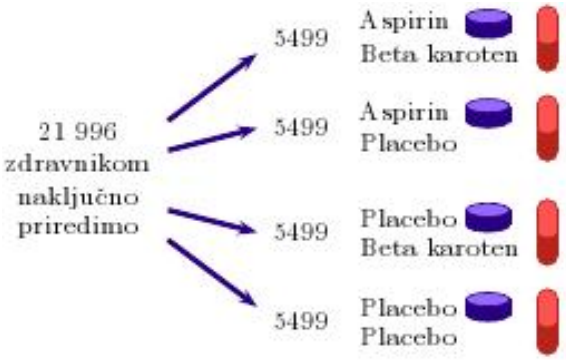
Slika 1.3: Načrt slučajenega primerjalnega eksperimenta, ki primerja tri načine spodbujanja k varčevanju z elektriko.

so bile naslednje: samo aspirin, samo beta karoten, oba ali nobeden. Posamezniki so bili naključno dodeljeni eni od terapij. ♦

Na tem primeru spoznamo nekaj novih idej, ki so pomembne pri načrtovanju eksperimentov. Prva je pomembnost **placebo efekta**. *Placebo* je lažna terapija, tableta, ki ne vsebuje aktivnih sestavin, vendar je po videzu in okusu enaka pravemu zdravilu. *Placebo efekt* je nagnjenje osebkov, da se pozitivno odzovejo na vsako terapijo, tudi na placebo. Če na primer primerjamo osebke, ki so dobili aspirin, s tistimi, ki ne dobijo nobene tablete, vplivata na prvo skupino tako aspirin kot placebo efekt, zato so pozitivni učinki aspirina pomešani z učinkom placeba. Da to preprečimo, moramo tudi ostale osebke podvreči neki terapiji. V zgornjem primeru so vsi zdravniki jemali tablete, ki so izgledale enako. Nekatere so vsebovale aspirin ali beta karoten, druge so služile kot placebo. Na sliki 1.4 je prikazan načrt eksperimenta.

Študija na zdravnikih je bila primer **dvojno slepega eksperimenta**: niti osebki sami niti eksperimentatorji, ki delajo z njimi, ne vedo, katere terapije so bili deležni osebki. Posamezniki bi lahko drugače odreagirali, če bi vedeli, da dobivajo "samo placebo". Prav tako bi to lahko vplivalo na eksperimentatorje, ki so anketirali in pregledovali posameznike. Torej so tako osebki kot pregledovalci ostajali "slepi", le statistik je imel podatke o tem, kdo je prejemal katero zdravilo.

Nazadnje pa je študija na zdravnikih bolj dovršen eksperiment kot naši prejšnji primeri. V primeru Osnovne šole kralja Matjaža in porabe energije smo primerjali vrednosti ene same spremenljivke (Kateri učni načrt? Kateri varčevalni program?). V študiji na zdravnikih pa si ogledamo dve različni eksperimentalni spremenljivki: aspirin ali ne, beta karoten ali ne? Eksperiment z dvema spremenljivkama, običajno



Slika 1.4: Načrt študije na zdravnikih, eksperimenta z dvema dejavnikoma.

imenovan *eksperiment z dvema dejavnikoma*, nam omogoča, da proučujemo *interakcijo* (ali skupni učinek) dveh zdravil, hkrati pa tudi učinek vsakega posebej. Beta karoten bi na primer lahko okrepil učinek aspirina na srčne infarkte. S primerjavo teh štirih skupin lahko proučimo vse te možne interakcije. Kljub vsemu pa je načrt eksperimenta na sliki podoben našim prejšnjim eksperimentom, ker obdržimo osnovni ideji uporabe naključij in primerjave večih terapij.

1.7 Statistični dokazi

Primerno načrtovan eksperiment je po mnenju statistika tisti, ki uporablja načela *primerjave* in *slučajnosti*: primerjave večih terapij in slučajnega prirejanja teh terapij osebkom.

Prihodnje zdravstveno stanje zdravnikov iz zgornje študije je lahko odvisno od starosti, preteklega zdravstvenega stanja, čustvenega stanja, kadilskih navad in še od veliko drugih znanih in neznanih spremenljivk. Slučajna izbira bo v povprečju uravnotežila skupine glede na vse te spremenljivke. Ker so skupine izpostavljene povsem enakemu okolju z izjemo dejanske vsebine tablet, lahko trdimo, da so razlike v številu infarktov ali rakavih obolenj posledica zdravil. To je osnovna ideja slučajenega primerjalnega eksperimenta.

Bodimo nekoliko bolj previdni: vsaka razlika med skupinami je bodisi rezultat terapije bodisi je rezultat ponesrečene naključne izbire. Lahko bi se na primer zgodilo, da bi bilo čisto po naključju v eni od skupin več tistih zdravnikov, ki bodo doživeli srčni infarkt. Problem je povsem enak tistemu, na katerega smo naleteli pri slučajnem vzorčenju, kjer bi lahko v enostavni slučajni vzorec čisto naključno izbrali same re-

publikance. Tako kot pri vzorčenju nas tudi tukaj rešuje regularnost, ki se skriva za naključji.

Če velikokrat ponovimo slučajno izbiranje skupin in ne uporabimo različnih terapij, bodo razlike med skupinami približno sledile pravilni šabloni. Ta nam pove, kako velike bodo najverjetneje razlike v rezultatih med skupinami, če nanje ne vplivajo nobeni drugi dejavniki. Če so razlike, ki jih opazimo, tako velike, da se skoraj nikoli ne bi pojavile ob odsotnosti drugih dejavnikov, smo prepičani, da je razlog v uspešnosti terapij. Delovanje terapij torej ne dokažejo vse razlike v rezultatih, zgolj tiste, ki so dovolj velike, da jih ne moremo pripisati naključju. Razlike med skupinami, ki so dovolj velike, da redko nastopijo naključno, imenujemo *statistično pomembne*.

Vsak opaženi učinek, ki je prevelik, da bi ga lahko prepričljivo pripisali naključju, se imenuje **statistično pomemben**.

Tako kot pri vzorčenju tudi tu večje število osebkov poveča naše zaupanje v rezultate. Študija na zdravnikih je vključevala 22 tisoč osebkov, da bi lahko s precejšnjo gotovostjo trdili, da so vse medicinsko pomembne razlike med skupinami opažene in posledica aspirina oziroma beta karotena. Izkazalo se je, da je bilo v skupini, ki je jemala aspirin, bistveno manjše število srčnih infarktov v primerjavi s skupino, ki je jemala placebo. Posledično zdravniki moškim po petdesetem letu velikokrat priporočajo redno jemanje manjših količin aspirina. Po drugi strani pa jemanje beta karotena ni bistveno zmanjšalo nastanka rakavih obolenj.

Osnovna logika eksperimentov, statistično načrtovanje eksperimentov in zakonitosti, ki vladajo naključjem, nam skupaj dajo prepričljive dokaze o vzrokih in efektih. Le eksperimenti lahko dajo povsem prepričljive dokaze o vzročnosti.

Primer. (Kajenje in zdravje) Oglejmo si še statistične dokaze, ki povezujejo kajenje z nastankom pljučnega raka. Ne moremo neki skupini ljudi ukazati, naj kadi ali pa ne kadi. Neposredni eksperiment torej ni mogoč. Najbolj natančne študije so izbrale vzorce kadilcev in nekadilcev, jih opazovale več let in nazadnje zabeležile vzrok smrti. To so *dolgoročne študije*, ker osebke spremljajo skozi čas. Dolgoročne študije so primerjalne, niso pa eksperimenti, ker se osebki sami odločijo, če bodo kadili ali ne. Velika dolgoročna študija britanskih zdravnikov je pokazala, da je umrljivost za rakom pljuč med kadilci dvajsetkrat večja kot pri nekadilcih. Druga študija na Američanih, starih med 40 in 79 let, je ugotovila enajstkrat večjo smrtnost

pri kadilcih. Te in številne druge opazovalne študije kažejo na tesno povezavo med kajenjem in pljučnim rakom. ♦

Povezava med kajenjem in pljučnim rakom je statistično pomembna. To pomeni, da je veliko večja kot bi bilo pričakovati, če bi bila naključna. Lahko smo torej prepričani, da povezava med kajenjem in rakom na pljučih ni zgolj slučajna. Vendar pa nam opazovanje vzorcev ne more povedati, kateri dejavniki še povezujejo ta dva pojava. Morda obstaja genetska predispozicija, ki je odgovorna tako za zasvojenost z nikotinom kot za razvoj raka. V tem primeru bi opazili močno povezavo med obema pojavoma, pa četudi kajenje samo ne bi imelo nobenega učinka na pljuča.

Statistični dokaz, ki kaže na kajenje kot vzrok za pljučnega raka je tako močen kot je neeksperimentalni dokaz lahko. Za začetek so povezavo potrdile številne raziskave v številnih državah. S tem so odpisani dejavniki, ki so značilni za posamično skupino ljudi ali pa za posamično študijo. Poleg tega obstaja *odmerek-odziv* povezava: ljudje, ki pokadijo več cigaret, imajo večjo verjetnost, da zbolijo za pljučnim rakom, kot tisti, ki kadijo manj, in opustitev kajenja zmanjša tveganje. Tretjič, znano je, na kakšne načine bi lahko kajenje povzročalo raka: cigaretni dim vsebuje katran, za katerega so z eksperimentom pokazali, da pri živalih povzroča tumorje. In končno, na voljo ni nobene verjetne alternativne razlage. Hipoteza o genetski povezavi na primer ne more razložiti povišanega števila primerov pljučnega raka pri ženskah, ki se je pojavilo, ko je vse več in več žensk začelo kaditi. Pljučni rak, ki je bil dolgo časa vodilni vzrok vseh z rakom povezanih smrti pri moških, je zdaj prehitel raka dojk kot najbolj smrtnonosna oblika raka pri ženskah.

Ti dokazi so prepričljivi, niso pa tako močni kot odločilni statistični dokazi, ki jih dobimo s slučajenim primerjalnim eksperimentom.

1.8 Statistika v praksi

Za pametno uporabo statistike se skriva več kot le poznavanje statističnih tehnik, kakršni sta enostavni slučajni vzorec in slučajeni primerjalni eksperiment. Ti načini pridobivanja podatkov se izognejo pastem prostovoljnih vzorcev ali nekontroliranih eksperimentov. Vendar pa obstajajo še druge pasti, ki lahko zmanjšajo uporabnost podatkov tudi kadar uporabimo zanesljive statistične metode.

Primer. (Neodziv pri vzorčenju) Izbira slučajnega vzorca je prvi korak pri izvajanju raziskave, ki zadeva veliko število ljudi. Posameznike iz vzorca je potem

potrebno kontaktirati in jih pripraviti k sodelovanju. To ni lahko. Nekateri ljudje so redko doma. Drugi ne želijo govoriti z anketarjem. *Neodziv* se pojavi, kadar je posameznik, ki je bil izbran v vzorec, nedosegljiv ali pa noče sodelovati. Preden preveč zaupamo rezultatom raziskave bi se morali vprašati, kako velik je bil neodziv. Pogosto je neodziv tudi 30% in več. Višji je v mestih kot na podeželju, zato lahko neupoštevanje posameznikov, ki niso odgovorili, povzroči pristranskost. Raziskave mnenja običajno zamenjajo takega posameznika z drugim iz istega območja, da bi zmanjšale pristranskost.

Celo pri štetju prebivalstva ZDA iz leta 1990, kjer so imeli vso podporo vlade, so imeli probleme z neodzivom. Štetje prebivalstva ne uporablja vzorca, temveč želi prešteti čisto vse ljudi v državi. Zaradi neodziva do nekaj pristranskosti pride v primeru mest in manjšin, čeprav so popisovalci do šestkrat poskušali kontaktirati vsakega posameznika. Ocenjujejo, da je to štetje izpustilo 1.8% celotnega prebivalstva, vendar pa je pri tem izpustilo 4.4% Afroameričanov in kar 5% prebivalcev latinskoameriškega porekla.

Za izvedbo štetja leta 2000 so predlagali, da namesto ponovnega obiska posameznikov, ki se niso odzvali, izberejo slučajni vzorec teh gospodinjstev in še vzorec 750 tisoč izmed vseh gospodinjstev. Končni rezultat bi potem določili hkrati iz začetnega štetja in izbranega vzorca. Vrhovno sodišče je odločilo, da vzorčenja ne morejo uporabiti za štetje, ki določa, koliko mest v kongresu dobi vsaka od zveznih držav, lahko pa ga uporabljajo za druge namene. ⧫

Primer. (Ali je eksperiment realen?) Študija na zdravnikih je predpisala tablete osebkom, ki so sicer živeli kot običajno. Veliko eksperimentov pa je izvedenih v umetnih okoljih. Psiholog, ki proučuje vpliv stresa na skupinsko delo, opazuje skupino študentov, ki v laboratoriju pod različnimi pogoji izvajajo določeno nalogo. Študentje vedo, da je naloga "samo eksperiment"in da bo stres trajal samo eno uro. Ali se torej lahko zaključki, do katerih pridemo s takim eksperimentom, nanašajo tudi na stres v resničnem življenju? Inženir izvede pomanjšan poskus v laboratoriju, da bi določil pritisk in temperaturo, ki dasta največji izkopiček pri kompleksni kemijski reakciji. Ali bo rezultat enak na večji ravni v tovarni?

To niso statistična vprašanja. Psiholog in inženir morata uporabiti svoje znanje psihologije in inženiringa, da ocenita, do kakšne mere so njuni rezultati uporabni. Statistični načrt nam omogoča, da tem eksperimentom zaupamo, ne moremo pa jih posplošiti na druge situacije. ⧫

Pod žarometom

Eksperimenti in etika

Dr. Charles Hennekens, direktor študije na zdravnikih, *Physicians' Health Study*, je moral razmisliti o ciljih, zasnovi in izvedbi te obsežne študije. Pri izvajanju takega eksperimenta pa se pojavijo še druga vprašanja. Dr. Hennekensa smo povprašali o etiki eksperimentiranja s človekovim zdravjem:

> O etičnih pomislekih glede slučajenih poskusov se je že veliko govorilo. Obstajajo primeri, kjer bi bili taki slučajeni poskusi neetični. Ko so predstavili penicilin kot zdravilo za pljučnico, ki je bila usodna v skoraj 100% primerov, se je smrtnost bistveno zmanjšala. Zagotovo bi bilo neetično izvesti slučajeni poskus in s tem začasno onemogočiti zdravljenje določenemu delu ljudi.
>
> Med odločitvama za in proti slučajenemu poskusu je tanka črta. Po eni strani mora obstajati dovolj veliko prepričanje v potencialno delovanje agensa, da lahko upravičimo njegovo uporabo na polovici osebkov. Po drugi strani mora biti dvom v njegovo delovanje dovolj velik, da upravičimo dejstvo, da ga drugi polovici osebkov ne damo, temveč namesto tega dobijo placebo. Prav te okoliščine so se po našem mnenju pojavile v primeru hipotez o aspirinu in beta karotenu.

Ko načrtujemo statistično študijo, se soočamo tudi z *etičnimi vprašanji*. Ali znanje, ki ga lahko dobimo z eksperimentom, upraviči možna tveganja, ki so jim podvrženi osebki? Pri študiji na zdravnikih so le-ti privolili v jemanje tablete, pri čemer so se zavedali možnih tveganj. Ko je postalo jasno, da so bili posamezniki, ki so jemali aspirin, manj pogosto podvrženi srčnim infarktom, so z eksperimentom prenehali, da so lahko vsi osebki izkoristili to novo znanje. V okvirju direktor te raziskave razloži, zakaj so slučajeni primerjalni eksperimenti temelj medicinskih raziskav in kdaj so takšni eksperimenti upravičeni. Praktični in etični problemi niso nikoli daleč, kadar statistiko uporabimo pri realnih problemih.

1.9 Slovarček

dvojno slepi eksperiment (ang. double-blind experiment) Eksperiment, pri katerem niti testirani osebki niti izvajalci eksperimenta med samim testiranjem

ne vedo, kateri od terapij je bil testirani osebek podvržen.

eksperiment (ang. experiment) Študija, v kateri so ljudje, živali ali objekti podvrženi terapijam, katerih vpliv želimo proučiti.

enostavni slučajni vzorec (ang. simple random sample) Slučajno izbrani vzorec: vsi možni vzorci iste velikosti imajo enako možnost, da so izbrani.

histogram (ang. histogram) Grafični prikaz pogostosti različnih izidov s pomočjo stolpcev; višina vsakega stolpca je število pojavitev posameznega izida.

kontrolna skupina (ang. control group) Skupina testnih osebkov, na kateri izvajamo standardno terapijo (ali pa ne izvajamo nobene terapije: glej placebo efekt).

meja napake (ang. margin of error) Pove, kako blizu resničnemu stanju bi bil vzorec v 95% vseh vzorcev, ki bi bili pridobljeni po enaki metodi kot uporabljeni vzorec.

mešanje (ang. confounding) Dve spremenljivki sta pomešani, kadar ne moremo razločiti njunih vplivov na izid.

opazovalna študija (ang. observational study) Študija, ki opazuje posameznike in meri spremenljivke, ki nas zanimajo, vendar ne poskuša vplivati na odzive.

placebo efekt (ang. placebo effect) Efekt navidezne terapije, npr. zdravilo brez zdravilnih učinkovin, ki je po videzu in okusu enako kot pravo zdravilo.

populacija (ang. population) Celotna skupina ljudi ali stvari, o kateri želimo zbrati informacije.

priročni vzorec (ang. convenience sample) Vzorec, sestavljen iz osebkov, ki jih je najlažje doseči, npr. mimoidoči na ulici; tak vzorec običajno ni nepristranski.

pristranskost (ang. bias) Sistematična napaka, zaradi katere navadno opazovanja odstopajo od resnice v isti smeri pri vsaki ponovitvi eksperimenta.

prostovoljni vzorec (ang. voluntary response sample) Vzorec, ki nastane kot odgovor na povabilo k sodelovanju; ti vzorci so ponavadi zelo pristranski.

slučajeni primerjalni eksperiment (ang. randomized comparative experiment) Eksperiment, ki služi primerjavi dveh ali več terapij, v katerem so ljudje, živali ali stvari podvrženi slučajno izbrani terapiji.

statistična pomembnost (ang. statistical significance) Opazovani pojav je statistično pomemben, če je tako pogost, da ni verjetno, da bi se v dani populaciji pojavil zgolj po naključju brez dejanskega vzroka.

tabela naključnih števil (ang. table of random digits) Tabela, katere elementi so števke 0,1,...,9, izbrane popolnoma naključno; vsak element tabele je z enako verjetnostjo katerokoli od teh števil in noben element nam ne da informacije o ostalih.

vzorec (ang. sample) Del populacije, ki ga dejansko opazujemo in ga uporabimo za izpeljavo zaključkov o celotni populaciji.

1.10 Dodatna literatura

- Cobb, G. W. *Design and Analysis of Experiments*, Springer, New York, 1998. Prvo poglavje tega zahtevnejšega besedila je zanimiv sestavek o načrtovanju eksperimentov.
- Kalton, G. *Introduction to Survey Sampling*, Sage Publications, Newbury Park, California, 1983. Podroben, a relativno nezahteven uvod v statistiko vzorčnih analiz.
- Moore, D. S. *Statistics: Concepts and Controversies*, W.H. Freeman, New York, 1997. Prvo in drugo poglavje vsebujeta obširno razpravo na podobnem nivoju kot ta knjiga.
- Moore, D. S. *The Basic Practice of Statistics*, W.H. Freeman, New York, 1999. Tretje poglavje jasno obravnava pridobivanje podatkov v besedilu o praktični statistiki na podobnem nivoju kot ta knjiga.
- Tanur, J. M.: Samples and Surveys, *Perspectives on Contemporary Statistics*, Mathematical Association of America, Washington, D.C. 1992, str. 55–70. Ta sestavek opisuje uporabo vzorčnih analiz na dokaj nezahtevnem nivoju.

Veliko pomembnih vzorčnih analiz v Združenih državah Amerike izvajata

- Urad za statistiko dela,
`http://stats.bls.org`,

- Urad za popis prebivalstva
  www.census.gov.

Stran Agencije Gallup, www.gallup.gov, ima veliko informacij o izvajanju raziskav.

Pomembne medicinske raziskave, ki večinoma temeljijo na slučajenem primerjalem eksperimentu, se pojavljajo v

- *Journal of the American Medical Association*,
  www.ama-assn.org/public/journals/jama/,
- *New England Journal of Medicine*,
  www.nejm.org.

1.11 Preverjanje znanja

(1) Trgovsko podjetje anketira 50 nakupovalcev, ki jih naključno izbere med 3500 obiskovalci enega od 23 nakupovalnih centrov. Vzorec v tem primeru sestavlja

(a) 50 izbranih nakupovalcev.

(b) 3500 obiskovalcev.

(c) 23 nakupovalnih centrov.

(2) V volitvah za župana sodeluje 5 kandidatov in 45 tisoč volilcev. Časopis anketira 750 volilcev, ko zapuščajo volišča. Populacija je v tem primeru sestavljena iz

(a) 5 kandidatov.

(b) 45 tisoč volilcev.

(c) 750 izbranih volilcev.

(3) Pred trgovino z zdravo prehrano izvajamo raziskavo o pozitivnih učinkih vitaminov. To je primer

(a) priročnega vzorca.

(b) mešanja.

(c) placebo efekta.

(4) S pomočjo naslednjega seznama naključnih števil izberi 3 ljudi z abecednega seznama 20 ljudi: 17463 26715 71680 64853. Izbrani ljudje imajo oznake

(a) 17, 46, 32.

(b) 17, 15, 16.

(c) 17, 15, 06.

(5) Če ima Anin vzorec mejo napake ±3%, Betin pa pri isti populaciji ±6%, potem

(a) je Anin vzorec dal manjši približek dejanske vrednosti kot Betin.

(b) je Betin vzorec pristranski.

(c) je Anin vzorec večji.

(6) Vsak od petih študentov je izbral vzorec 100 študentov. Vprašali so jih po najljubši vrsti brezalkoholne pijače. Vseh pet rezultatov je bilo različnih. Razlog za to je

(a) pristranskost.

(b) vzorčna spremenljivost.

(c) uporaba kontrolne skupine.

(7) Pri preiskušanju zdravil smo naključno izbrali vzorce za vsako od treh terapij. Niti osebki niti eksperimentator ne vedo, katero terapijo je prejel vsak posameznik. To je primer (I) slučajenega primerjalnega eksperimenta in/ali (II) dvojno slepega eksperimenta.

(a) Samo (I).

(b) Samo (II).

(c) (I) in (II).

1.12 Naloge

Vzorčenje

(1) Sociolog želi dobiti mnenja zaposlenih odraslih žensk o sofinanciranju vrtcev. Pridobi si seznam 520 članov lokalnega združenja poslovnih žensk, naključno izbere 100 izmed njih in jim pošlje vprašalnik. V odgovor dobi samo 68% izpolnjenih vprašalnikov. Kaj je v tem primeru populacija? Kaj je vzorec?

(2) Gospodinje včasih vlagajo zelenjavo v prazne kozarce majoneze, da prihranijo pri nakupu novih kozarcev. Revijo *Organsko vrtnarjenje* je zanimalo, kolikšen odstotek teh kozarcev pri tem poči. Zbrali so 100 kozarcev in vanje vložili paradižnik, pri čemer so počili trije kozarci. Določi populacijo in vzorec.

(3) Politika zanima, če bo zakonodajna skupščina podprla predlagani zakon o omejevanju orožja. Njegovo osebje mu je sporočilo, da so dobili pisma 361 predstavnikov, od katerih jih 323 nasprotuje zakonu. Kaj je populacija? Kaj je vzorec? Ali ta vzorec dobro predstavlja populacijo? Odgovor utemelji.

Slabe metode vzorčenja

(4) Revija za zdravo prehrano in naravno medicino želi dokazati, da uživanje velikih količin vitaminov pozitivno vpliva na zdravje. Uredniki prosijo bralce, ki so redno jemali vitamine v velikih količinah, da sporočijo svoje izkušnje. Od 2754 prejetih pisem jih 93% poroča o pozitivnih učinkih. Ali je bolj verjetno, da je delež 93% manjši, večji ali enako velik kot delež odraslih, ki bi opazili pozitivne očinke jemanja vitaminov? Zakaj?

(5) Leta 1995 je revija *USA Weekend* objavila oglas, podoben spodnjemu. V drobnem tisku je bilo navedeno, da vsak klic stane 50 centov. Ali lahko zaupamo rezultatom te ankete? Odgovor utemelji.

> Ali bi želeli v primeru neozdravljive bolezni
> imeti pravico do evtanazije?
> DA: 1-900-255-2257 NE: 1-900-255-2258

(6) V televizijski oddaji so izvedli anketo o dobrodelnosti. Prejeli so 50 tisoč klicev in 83% gledalcev je trdilo, da redno darujejo denar in oblačila dobrodelnim organizacijam. Razloži, zakaj je ta vzorec skoraj gotovo pristranski. V katero smer je pristranski? Se pravi, ali je odstotek ljudi, ki darujejo v humanitarne namene večji ali manjši od 83%, ki jih je dal ta vzorec.

(7) Mestna policija v Miamiju želi dobiti mnenje črnske skupnosti o delu policije. Sociolog pripravi vprašalnik in izbere vzorec 300 naslovov v pretežno črnskih četrtih. Uniformirani temnopolti policaj obišče vsakega od naslovov in želi govoriti z odraslim članom gospodinjstva. Razloži, zakaj pričakuješ, da bodo rezultati te raziskave pristranski. Na kakšen način se bodo razlikovali od resničnega mnenja populacije.

Enostavni slučajni vzorci

(8) Podjetje želi razumeti odnos pripadnikov manjšin med vodilnimi uslužbenci do sistema ocenjevanja vodstvenih delavcev. Spodaj je seznam vseh vodilnih uslužbencev, ki pripadajo kakšni od manjšin. Uporabi tabelo naključnih števil in s seznama izberi 6 ljudi, ki bodo podrobneje izprašani.

Agarwal	Dewald	Huang	Puri
Anderson	Fernandez	Kim	Richards
Baxter	Flemming	Liao	Rodrigez
Brown	Gates	Mouring	Santiago
Bowman	Goel	Naber	Shen
Castillo	Gomez	Peters	Vega
Cross	Hernandez	Pliego	Wang

(9) Kakšne vrste programov ponujajo fakultete najboljšim maturantom? Odločiš se zbrati informacije s petih naključno izbranih oddelkov. Uporabi tabelo naključnih števil in začni v vrstici 132, da izbereš enostavni slučajni vzorec petih oddelkov s spodnjega seznama.

Računalništvo	Kemija	Komunikacije
Ekonomija	Geografija	Slovenščina
Tuji jeziki	Medicina	Matematika
Zgodovina	Politologija	Filozofija
Fizika	Statistika	Psihologija
Sociologija	Biologija	

(10) Študentka želi ugotoviti, kaj menijo predavatelji o ustanovitvi nacionalnega odbora za visoko šolstvo, ki bi nadzoroval vse fakultete v državi. Na fakulteti je zaposlenih 380 predavateljev.

(a) Kaj je v tem primeru populacija?

(b) Natančno razloži, kako bi izbral(a) enostavni slučajni vzorec 50 predavateljev.

(c) Uporabi vrstico 135 tabele naključnih števil, da izbereš prvih pet oseb tega vzorca.

(11) Število študentov politologije na Ljubljanski univerzi se je bistveno povečalo, število predavateljev pa je ostalo nespremenjeno. Študentski časopis namerava anketirati 25 od 450 študentov politologije, da bi ugotovili, kaj menijo o velikosti razredov in drugih s tem povezanih problemih. Predlagaš jim enostavni slučajni vzorec. Natančno razloži, kako bi izbral(a) ta vzorec. Nato uporabi vrstico 120 iz tabele naključnih števil, da izbereš prvih pet oseb tega vzorca.

(12) Katere od naslednjih trditev o tabeli naključnih števil so pravilne? Odgovore utemelji.

(a) V vsaki vrstici, ki vsebuje 40 števil, so natanko štiri ničle.

(b) Vsak par števk ima natanko $\frac{1}{100}$ možnosti, da je enak 00.

(c) Skupina 0000 se nikoli ne pojavi v tabeli, ker tako zaporedje ni naključno.

Statistično ocenjevanje

(13) Pet vstopnic za rock koncert moramo razdeliti med 25 razgrajaških članov našega kluba. Ta naloga bo služila kot primer vzorčne spremenljivosti.

(a) Naključno izberi pet srečnežev s spodnjega seznama, ki bodo dobili vstopnice. Pomagaj si z vrstico 135 iz tabele naključnih števil.

(b) Izkaže se, da je med 25 člani kluba 10 žensk. Njihova imena so na seznamu označena z zvezdicami. Dvajsetkrat naključno izberi po pet imen, vsakič s pomočjo neke druge vrstice iz tabele naključnih števil (pri tem vključi še vzorec iz prejšnje točke). Zabeleži število žensk v vsakem od vzorcev. Nariši histogram, ki predstavi dobljene rezultate. Izračunaj povprečno število žensk v teh 20 vzorcih.

(c) Ali meniš, da bi morali člani kluba posumiti, da gre za diskriminacijo, če nobene od podeljenih vstopnic ne bi dobila ženska?

Agassiz	Darwin	Herrnstein	Myrdal	Vogt*
Binet	Epstein	Jimenez*	Perez*	Went
Blumenbach	Ferri	Lombrosco	Spencer*	Wilson
Chase*	Gupta*	Moll*	Thomson	Yerkes
Chen	Gutierrez	McKim*	Toulmin	Zimmer

(14) Oglaševalska agencija izvaja raziskavo, da bi ugotovila, kako ženske reagirajo na različne pridevnike, ki jih lahko uporabimo za opis avtomobilov. Izberejo 600 žensk iz cele države. Vsaki predvajajo seznam pridevnikov kot na primer *eleganten* ali *prestižen*. Pri vsakem morajo povedati, kako zaželjen se bi jim zdel avto, ki bi ga opisali s tem pridevnikom. Možni odgovori so (1) zelo zaželjen, (2) zaželjen, (3) neopredeljena ali (4) ne zaželjen. Med vprašanimi je 76% odgovorilo, da je avtomobil, ki je opisan kot *eleganten*, zelo zaželjen.

(a) Kaj je v tej raziskavi populacija?

(b) Koliko žensk iz vzorca je odgovorilo, da bi bil eleganten avtomobil zelo zaželjen?

(c) Uporabili so slučajni vzorec po vzoru Gallupovih raziskav. Na katerem intervalu se zelo verjetno nahaja resnični odstotek žensk, ki bi menile, da je takšen avtomobil zelo zaželjen?

(15) V raziskavi mnenja vprašamo 1324 odraslih oseb, če verjamejo, da obstaja življenje na drugih planetih. Med njimi jih 609 odgovori pritrdilno. Kakšen odstotek vzorca verjame v zunajzemeljsko življenje? Agencija oznani, da je meja napake pri tej raziskavi ±3%. Kaj lahko zaključiš o deležu vseh odraslih, ki verjamejo v življenje na drugih planetih.

(16) Nacionalne raziskave mnenja kot je na primer Gallupova, običajno vsak teden izberejo vzorec 1500 ljudi.

(a) Pri vzorcu te velikosti je običajno meja napake približno ±3 odstotne točke. Razloži to trditev nekomu, ki ne ve ničesar o statistiki.

(b) Tik pred predsedniškimi volitvami pa agencije navadno povečajo velikost vzorcev na približno 4000 ljudi. Ali je meja napake zdaj več kot ±3%, manj kot ±3% ali nespremenjena? Zakaj?

(17) Članek v časopisu poroča, da je v nedavni Gallupovi raziskavi 78% vzorca 1108 odraslih oseb reklo, da verjamejo v obstoj nebes. Samo 60% je reklo, da verjamejo v pekel. Članek se zaključi z besedami: *Meja napake vzorca je bila 4 odstotne točke.* Ali lahko z gotovostjo trdimo, da med 56% in 64% odraslih verjame v pekel? Odgovor utemelji.

(18) Naključna števila lahko uporabimo za *simulacijo* rezultatov slučajnega vzorčenja. Recimo, da izberemo enostavni slučajni vzorec velikosti 25 iz velikega

števila srednješolcev in da si 20% ne poišče dela med poletnimi počitnicami. Da simuliramo ta enostavni slučajni vzorec, naj 25 zaporednih števk iz tabele naključnih števil označuje dijake iz našega vzorca. Števili 0 in 1 naj pomenita nezaposlene dijake, ostale števke pa tiste z zaposlitvijo. To je točna imitacija enostavnega slučajnega vzorca, ker 0 in 1 predstavljata ravno 20% izmed 10 enako verjetnih števk.
Simuliraj rezultate petdesetih vzorcev tako, da prešteješ ničle in enice med prvimi 25 števkami v vsaki od 50 vrstic v tabeli naključnih števil. Nariši histogram, ki prikaže dobljene rezultate. Ali je resnično stanje populacije (20% nezaposlenih, tj. 5 v vzorcu 25) blizu sredine tega histograma? Kaj sta največje in najmanjše dobljeno število nezaposlenih dijakov med temi 50 vzorci? Kolikšen odstotek vzorcev je imel med 4 in 6 nezaposlenih?

Eksperimenti

Študije iz nalog 19, 20 in 21 lahko dajo napačne rezultate zaradi mešanja zunanjih vplivov z obravnavanimi terapijami. V vsakem od primerov razloži, kako bi mešanje lahko vplivalo na rezultate.

(19) Študentka meni, da bo pitje zeliščnega čaja izboljšalo zdravje pacientov v domu za ostarele. Z nekaj prijatelji redno obiskuje večji dom in streže zeliščni čaj delu varovancev. Drugega dela varovancev ne obiskujejo. Po šestih mesecih ugotovijo, da je bilo skupno število bolniških dni v prvi skupini manjše kot v drugi.

(20) Učitelj je prepričan, da učenje tujih jezikov izboljša znanje slovenščine. Preuči rezultate na neki srednji šoli in ugotovi, da so učenci, ki so si izbrali še kak tuj jezik, dosegali boljše ocene tudi pri slovenščini.

(21) Članek v neki reviji je poročal, da so ženske, ki dojijo svoje otroke, bolj povezane z njimi, kot tiste, ki jih hranijo po steklenički. Avtorica je od tod sklepala, da ima dojenje pozitiven učinek na materin odnos do otroka.

(22) V ZDA so ugotavljali, da ženske volijo demokrate raje kot moški. Politolog je vprašal skupino moških in skupino žensk, za koga so glasovali na zadnjih volitvah. Natančno razloži, zakaj to ni eksperiment.

(23) Nekateri menijo, da telesna vadba dvigne bazični metabolizem za 12 do 24 ur in tako omogoča topljenje maščob tudi po končani vadbi. Raziskovalec prosi

udeležence, da več ur hitro hodijo po tekalni stezi. Pri tem meri njihov bazični metabolizem pred, takoj po in 12 ur po vadbi. Ali je to eksperiment? Zakaj ali zakaj ne?

(24) Študija povezave med telesno pripravljenostjo in vodstvenimi sposobnostmi uporabi za vzorec vodstvene delavce srednjih let, ki so se prostovoljno prijavili za program vadbe. Razdelijo jih v dve skupini glede na njihovo telesno pripravljenost, ki jo ocenijo z zdravniškim pregledom. Vsi dobijo psihološki test, ki izmeri njihove vodstvene sposobnosti, nato pa primerjajo rezultate. Je to opazovalna študija ali eksperiment? Odgovor utemelji.

Slučajeni primerjalni eksperimenti

(25) Anemija srpastih eritrocitov je dedna bolezen rdečih krvnih celic, ki v ZDA prizadene predvsem Afroameričane. Povzroči lahko močne bolečine in številne zaplete. Leta 1992 je nacionalni zdravstveni inštitut začel raziskavo s hidroksikarbamidom. V raziskavo je bilo vključenih 300 odraslih oseb, ki so imele v preteklem letu vsaj tri napade bolečine, ki jih je povzročila anemija srpastih eritrocitov.

(a) Zakaj ne bi dobili dobre informacije o učinkovitosti zdravila, če bi dali hidroksikarbamid vsem 300 osebkom?

(b) Raziskava je bila zasnovana kot slučajeni primerjalni eksperiment, ki je primerjal hidroksikarbamid in placebo. Povzemi ustrezen načrt. Pri tem si pomagaj s sliko 1.2. Ne pozabi vključiti velikosti skupin in opis rezultatov, ki jih boš pregledal(a).

(c) Leta 1995 so eksperiment predčasno končali, ker je bilo v skupini, ki je prejemala hidroksikarbamid, pol manj bolečinskih napadov kot v kontrolni skupini. Razloži, zakaj je dal ta eksperiment močan dokaz, da je hidroksikarbamid uspešno zdravilo za anemijo srpastih celic.

(26) Nekateri finančni svetovalci verjamejo, da grafi preteklih trendov cen napovedujejo, kakšne bodo cene v prihodnosti. Večina ekonomistov se ne strinja. V eksperimentu, ki razišče, kako uporaba grafov vpliva na rezultate, študentje trgujejo z virtualnim denarjem v računalniški simulaciji. Študentov je dvajset, poimenujemo jih $A, B, \ldots, T$ (s črkami angleške abecede). Njihov cilj je zaslužiti čimveč denarja in najboljše nagradimo z manjšimi nagradami.

Študentje imajo na razpolago zgodovino predhodnega trgovanja s to valuto, nekateri imajo tudi programsko opremo, ki prikaže trende. Opiši načrt tega eksperimenta in uporabi tabelo naključnih števil, kjer to eksperiment zahteva.

(27) Študentje imajo pri predavanju iz matematike izbiro med običajnim predavanjem in učenjem po lastni presoji. Fakulteta želi primerjati rezultate pri obeh možnostih. Nekdo predlaga, da bi obema skupinama dali enaka končna izpita in primerjali povprečna rezultata v obeh skupinah.

(a) Razloži, zakaj so zaradi mešanja rezultati takšne raziskave brez pomena.

(b) Med študenti jih je 30 pripravljeno slediti kateremukoli od teh dveh programov. Načrtuj eksperiment, ki bo primerjal ti dve metodi. Pomagaj si z vrstico 108 iz tabele naključnih števil.

(28) Članek v *New England Journal of Medicine*, ki predstavlja končne rezultate študije na zdravnikih, se začne z naslednjimi besedami: *Študija na zdravnikih je slučajeni, dvojno slepi eksperiment s placebom, s katerim smo želeli ugotoviti, če majhne količine aspirina (325mg vsak drugi dan) zmanjšajo možnost srčnih infarktov in če beta karoten zmanjšuje možnost nastanka rakavih obolenj.* Od zdravnikov se pričakuje, da bodo to razumeli. Razloži zdravniku, ki ne zna statistike, kaj pomenijo *slučajeni*, *dvojno slepi* in *placebo*.

(29) V časopisu prebereš, da *so se pri kontroliranih znanstvenih raziskavah nezdravstvene terapije kot sta meditacija in molitev izkazale za učinkovite pri zdravljenju visokega krvnega pritiska, nespečnosti, čirov in astme.* Razloži, kaj je mišljeno s *kontroliranimi znanstvenimi raziskavami* in zakaj lahko take raziskave pokažejo pozitivne učinke meditacije in molitev.

(30) Spodaj je seznam 20 pacientov, ki so privolili k sodelovanju v preiskusu kirurškega zdravljenja angine. Načrtuj eksperiment, ki bo kirurško zdravljenje primerjal s placebom (lažno operacijo) in uporabi tabelo naključnih števil, da izvedeš ustrezno naključno delitev.

Ashley	Cravens*	Lippmann	Strong*
Bean*	Dorfman	Mark*	Tobias
Block	Garcia	Morton*	Valenzuela*
Chen	Huang*	Popkin	Washington
Chavez*	Kidder	Sosa	Williams

(31) Raziskovalci iz prejšnje naloge niso vedeli, da bodo osebki, označeni z zvezdico, med izvajanjem eksperimenta doživeli usodni srčni infarkt. Opazujemo lahko, kako se vzorčna spremenljivost obnaša pri slučajenem eksperimentu, če pogledamo, koliko od teh osem osebkov je bilo v skupini, ki je bila podvržena novi kirurški terapiji. Dvajsetkrat naključno izberi 10 osebkov v to skupino in si zapisuj, koliko jih je označenih z zvezdico v vsakem od izborov. Nato nariši histogram, ki prikazuje število izbranih osebkov, ki so doživeli srčni napad. Kolikšno je povprečje po 20 poskusih?

(32) Razloži, zakaj je bolje uporabiti več tisoč osebkov namesto samo zgornjih dvajset.

(33) Razloži, kako bi naključno razdelil 20 osebkov iz zgornjih primerov med štiri terapije iz študije na zdravnikih. Slika 1.4 opisuje te terapije. Izberi po pet osebkov v vsako od skupin. Uporabi vrstico 120 iz tabele naključnih števil.

Statistični dokazi

(34) S slučajenim primerjalnim eksperimentom so preverjali, če dodatek kalcija v prehrani znižuje krvni pritisk pri zdravih moških. Osebki so 12 tednov prejemali kalcij ali placebo. Raziskovalci so ugotovili, da *je bil krvni pritisk v skupini, ki je jemala kalcij, bistveno nižji v primerjavi s kontrolno skupino.* Pri tem *bistveno* pomeni *statistično pomembno.* Razloži, kaj pomeni statistična pomembnost v primeru tega eksperimenta, zdravniku, ki ne zna statistike.

(35) Finančni oddelek univerze naredi med študenti anketo o njihovih zaposlitvah in zaslužkih. Poročilo pravi, da so *v akademskem letu opazili statistično pomembno razliko v zaslužku med spoloma, pri čemer so moški v povprečju zaslužili več. Nobene razlike ni bilo opaziti med zaslužki črncev in belcev.* Razloži oba zaključka nekomu, ki ne zna statistike.

Statistika v praksi

(36) Pogosta oblika neodziva pri telefonskih raziskavah je "*zvoni, se ne javi*". To pomeni, da smo sicer klicali aktivno telefonsko številko, vendar se ni nihče oglasil. Italijanski nacionalni inštitut za statistiko je opazoval neodziv pri raziskavah v italijanskih gospodinjstvih od januarja do velikonočnih praznikov

ter preko julija in avgusta. Vsi klici so bili opravljeni med sedmo in deseto zvečer, vendar je bilo v enem od obdobij $21,4\%$ "zvoni, se ne javi" neodziva, v drugem pa $41,5\%$. Katero od obdobij je po tvojem mnenju imelo večji delež neodziva? Zakaj? Razloži, zakaj so zaradi visokega deleža neodziva rezultati raziskave manj zanesljivi.

(37) Način zastavljanja vprašanj lahko močno vpliva na izid raziskave. Spodaj sta primera dveh različih formulacij istega vprašanja.

(a) Bi morali sprejeti zakon, ki bi preprečil vse možnosti, da bi interesne skupine dajale kandidatom velike količine denarja?

(b) Bi morali sprejeti zakon, ki bi interesnim skupinam preprečeval prispevanje h kampanjam, ali imajo skupine pravico prispevati sredstva kandidatom, ki jih podpirajo?

Pri enem od teh vprašanj je bilo 40% vprašanih za prepoved prispevkov, pri drugem jih je bilo za prepoved 80%. Kateremu vprašanju pripada vsak od teh deležev? Razloži, zakaj sta bila rezultata tako različna.

(38) Ali sredinske zavorne luči, ki so obvezne za vse avtomobile, ki so jih v ZDA prodali po letu 1986, res zmanjšajo nevarnost naleta? Slučajeni primerjalni eksperimenti na izposojenih in poslovnih avtomobilih, ki so jih izvedli pred uvedbo obveznih luči, so pokazali, da so se tovrstne nesreče pri uporabi luči zmanjšale za 50%. Na žalost pa je uvedba obveznih luči pri vseh avtomobilih pripeljala le do 5% zmanjšanja. Razloži, zakaj eksperiment ni realistično posnemal stanja po uvedbi luči.

Dodatne naloge

(39) Gospa Caucus je kandidatka svoje stranke v drugem kongresnem okrožju v Indiani. Stranka želi vedeti, kolikšen delež registriranih volilcev bi glasoval za gospo Caucus, če bi bile volitve jutri. Agencija kontaktira 800 volilcev, od katerih jih 456 izjavi, da bi glasovali za gospo Caucus. Kaj je v tem primeru populacija? Kaj je vzorec? Razloži nekomu, ki ne zna statistike, zakaj je vzorec z 800 volilci boljši kot tak z 200 volilci.

(40) Izbiranje vzorca s seznama, ki vsebuje le del populacije, je pogost vzrok pristranskosti pri vzorčenju. V vsakem od naslednjih primerov razloži, zakaj bi lahko bil tak vzorec pristranski.

(a) Da bi dobili javno mnenje o predlogu, ki bi znižal socialno podporo, v agenciji izberejo vzorec tako, da pokličejo s pomočjo računalnika naključno izbrane telefonske številke.

(b) Da bi ugotovila, kako bodo na ta isti predlog odreagirali njeni volilci, članica kongresa pošlje vprašalnik vsem registriranim volilcem v svojem okrožju.

(41) V neki reviji večkrat povabijo bralce, da sodelujejo pri anketah. Ob neki priložnosti so jim postavili naslednje vprašanje: "Če bi se imeli možnost ponovno odločati, bi imeli otroke?" Od skoraj 10 tisoč staršev, ki so se odzvali na anketo, jih je 70% odgovorilo, da ne. Kmalu po tem je nacionalna raziskava postavila enako vprašanje vzorcu 1400 staršev. Pri tem so prejeli 90% pozitivnih odgovorov. Kateremu od teh vzorcev gre bolj zaupati? Zakaj?

(42) Janino malo računovodsko podjetje ima 30 strank. Jana želi anketirati vzorec petih strank, da bi ugotovila, kako bi lahko povečali zadovoljstvo z njihovimi storitvami. Da bi se izognila pristranskosti, izbere enostavni slučajni vzorec velikosti 5. Začni z vrstico 123 tabele naključnih števil in izberi enostavni slučajni vzorec z Janinega seznama.

A-1 Plumbing	Accent Printing	Baloons, Inc.
Anderson Construction	Bailey Trucking	Blue Prints Specialties
Bennett Hardware	Best's Camera Shop	Computer Answers
Central Tree Service	Classic Flowers	Hernandez Electronics
Darlene's Dolls	Fleish Reality	Action Sport Shop
Johnson Commodities	JL Records	Keiser Construction
Liu's Chinese Restaurant	Magic Tan	Peerless Machine
Photo Arts	River City Books	Riverside Tavern
Rustic Boutique	Satellite Services	Scotch Wash
Sewer's Center	Tire Specialties	Von's Video Store

(43) Davčni urad v ZDA namerava pregledati enostavni slučajni vzorec davčnih napovedi za vsako od držav. Med drugim jih zanima delež olajšav. Celotno število napovedi se spreminja od države do države: od preko 13 milijonov v Kaliforniji do manj kot 220 tisoč v Wyomingu.

(a) Ali se bo meja napake pri ocenjevanju deleža olajšav spreminjala od države do države, če uporabimo enostavni slučajni vzorec 2000 napovedi?

(b) Ali se bo meja napake spreminjala od države do države, če uporabimo enostavni slučajni vzorec, ki zajema 1% vseh napovedi v vsaki državi? Odgovor utemelji.

(44) Na zadnjem koraku pri štetju prebivalstva izberemo naslove znotraj majhnih območij, imenovanih bloki. Uporabimo *sistematično slučajno vzorčenje*, ki ga bomo predstavili na tem primeru. Recimo, da moramo izbrati 4 naslove izmed 100. Ker je $\frac{100}{4} = 25$, si lahko mislimo, da je naš seznam sestavljen iz štirih seznamov s po 25 naslovi. Izberemo prvega od petindvajsetih naključno s pomočjo tabele naključnih števil. Vzorec vsebuje ta naslov in naslove, ki so od njega oddaljeni za 25, 50 in 75 mest. Če na primer izberemo 13, potem vzamemo v vzorec naslove z oznakami 13, 38, 63 in 88.

(a) Uporabi tabelo naključnih števil, da izbereš sistematični slučajni vzorec petih naslovov s seznama 200 naslovov. Uporabi vrstico 120.

(b) Kot pri enostavnem slučajnem vzorcu imajo tudi pri sistematičnem slučajnem vzorcu vsi posamezniki enako možnost, da so izbrani. Razloži, zakaj je to res, potem pa še, zakaj kljub temu sistematični slučajni vzorec ni enostavni slučajni vzorec.

(45) Eksperiment, s katerim naj bi potrdili, da meditacija znižuje stres, je potekal takole: Eksperimentator je anketiral osebke in si zabeležil njihov nivo stresa. Nato so osebke naključno razporedili v dve skupini. Eksperimentator je eno od skupin naučil meditirati, kar so nato počeli en mesec. Drugi skupini so samo rekli, naj se bolj sprostijo. Po enem mesecu je eksperimentator spet izprašal vse osebke in ocenil nivo stresa. Ta je bil nižji v skupini, ki je meditirala. Psihologi trdijo, da so rezultati sumljivi, ker ocenjevanje nivoja stresa ni bilo slepo. Razloži, kaj to pomeni, in kako bi to lahko povzročilo pristranskost rezultatov.

(46) Ignoriraj praktične težave in moralne pomisleke in osnuj eksperiment, ki bi odgovoril na vprašanje ali kajenje povzroča pljučnega raka.

(47) Pri testiranju učinka obstojnih pesticidov bodo raziskovalci 60 dni hranili podgane s hrano, ki bo vsebovala DDT. Nato bodo izmerili njihovo odzivnost živčevja, da bi ugotovili, kakšen je vpliv DDT.

(a) Razloži, zakaj bi morali raziskovalci opazovati tudi kontrolno skupino, ki bi dobivala sicer enako, a nekontaminirano hrano.

(b) Recimo, da je na voljo 20 novorojenih podgan. Načrtuj eksperiment in s pomočjo tabele naključnih števil (vrstica 123) izberi vzorec.

(48) Zanima nas, če bo zaradi organiziranega dnevnega varstva podjetje pritegnilo več žensk, tudi neporočenih. Načrtuješ eksperiment, ki bo odgovoril na to vprašanje. Pripraviš reklamni material za dve namišljeni podjetji, ki se ukvarjata s podobnimi posli in se nahajata v istem kraju. Brošura podjetja A ne omenja varstva. Podjetje B ima dve verziji brošure, ki sta skoraj enaki, le da ena opisuje, kako je organizirano varstvo za otroke. V raziskavi je udeleženih 40 neporočenih enako izobraženih žensk, ki iščejo zaposlitev. Vsaka bo prebrala reklamni material za obe podjetji in izbrala tisto, za katero bi raje delala. Vsako od obeh verzij brošure podjetja B bo dobila polovica osebkov. Pričakuješ, da bo večji delež tistih, ki bodo dobile brošuro z opisom varstva, izbral podjetje B.

(a) Opiši, kako bo potekal eksperiment.

(b) V spodnji tabeli so imena sodelujočih. Začni z vrstico 131 v tabeli naključnih števil in izberi vzorec. Naredi seznam udeleženk, ki bodo dobile tisto brošuro podjetja B, ki vključuje opis organiziranega varstva.

Abrams	Danielson	Gutierrez	Lippman	Rosen
Adamson	Durr	Howard	Martinez	Sugiwara
Afifi	Edwards	Hwang	McNeill	Thompson
Brown	Fluharty	Iselin	Morse	Travers
Cansico	Garcia	Janle	Ng	Turing
Chen	Gerson	Kaplan	Quinones	Ullmann
Cortez	Green	Kim	Rivera	Williams
Curzakis	Gupta	Lattimore	Roberts	Wong

(49) Fizz Laboratories, farmacevtsko podjetje, je razvilo novo zdravilo za lajšanje bolečin. Na voljo je šestdeset pacientov z artritisom. Vsak bo podvržen terapiji in eno uro kasneje jih bomo vprašali, za približno koliko odstotkov so se zmanjšale bolečine.

(a) Zakaj ne bi dali zdravila vsem pacientom in si zabeležili odgovore?

(b) Načrtuj eksperiment, ki bo primerjal učinkovitost novega zdravila z aspirinom in placebom.

(c) Ali boš pri tem pacientom povedal(a), katero zdravilo so dobili? Kako bi to vplivalo na njihove reakcije?

(d) Če pacientom ne povemo, katero zdravilo so dobili, imamo slepi eksperiment. Ali bi moral biti ta eksperiment tudi dvojno slepi? Razloži.

(50) Ali je število dni, ki ga pismo potrebuje, da pride do cilja, odvisno od tega, kateri dan v tednu je bilo poslano in ali je naslov vseboval tudi poštno številko? Na kratko opiši, kako bi načrtoval(a) eksperiment z dvema dejavnikoma, ki bi odgovoril na to vprašanje. Natančno opiši vse terapije in povej, kako boš obravnaval(a) zunanje dejavnike kot je na primer ura, ko je bilo pismo odposlano.

(51) Prejšnja naloga prikazuje uporabo statistično osnovanega eksperimenta pri iskanju odgovorov na vsakdanja vprašanja. Izberi si vprašanje, ki te zanima in na katero bi se dalo odgovoriti s takim eksperimentom. Natančno opiši, kako bi načrtoval(a) tak eksperiment.

(52) Koruza je pomembna kot sestavni del krme za številne domače živali. Običajna koruza vsebuje malo aminokisline, imenovane lizin. Živali bodo morda rasle hitreje, če bodo jedle nove vrste koruze, ki vsebujejo povečane količine lizina. Raziskovalci izvedejo eksperiment, da bi primerjali novo vrsto koruze, imenovano *floury-2*, z običajno koruzo. Pripravijo krmo iz mešanic koruze in soje, pri čemer ima lahko koruza tri različne vsebnosti beljakovin: 12%, 16% in 20%. Skupno imamo torej šest možnih diet. Vsako dieto predpišemo desetim en dan starim piščančkom in zabeležimo, koliko teže so pridobili v 21 dneh.

(a) Ta eksperiment ima dva dejavnika. Katera?

(b) Opiši načrt eksperimenta. Uporabi slučajenje (ni pa ga potrebno eksplicitno izvesti).

(53) Številne povezave na spodnjem delu elektronskih vezij so lotane tako, da plošče potujejo skozi val staljenega lota. Inženir želi ugotoviti, kako hitrost tekočega traka vpliva na kakovost lotanja. Primerjati želi hitrosti 20, 25 in 30 čevljev na minuto. Spremenljivka, ki jo opazuje, je število nepravilno zalotanih povezav izmed 2000 povezav na plošči.

(a) Inženir namerava izdelati 10 plošč pri vsaki od hitrosti. Zakaj bi moral izbrati hitrosti naključno za 30 plošč, namesto da bi prvih 10 izdelal pri 20 čevljih na minuto in tako naprej?

(b) Opiši, kako bo osnovan slučajeni primerjalni eksperiment. Začni tako, da označiš plošče s števili od 1 do 30 glede na vrstni red, v katerem bodo izdelane.

(c) Začni v vrstici 130 tabele naključnih števil, da izpelješ zahtevano slučajenje. Napiši seznam ustreznih hitrosti.

(54) Psiholog poroča, da je bil "v našem vzorcu etnocentrizem značilno večji med ljudmi, ki so hodili v cerkev, kot med tistimi, ki niso". Pojasni, kaj to pomeni, nekomu, ki ne zna statistike. Pri tem ne uporabi besede "značilno".

(55) Cigaretna industrija se je odločila, da morajo izgledati modeli, ki oglašujejo njene izdelke, stari vsaj 25 let. Vendar pa so raziskave pokazale, da potrošniki menijo, da je med temi modeli veliko mlajših. Spodaj je citat iz raziskave, v kateri so ljudi spraševali, če menijo, da razne znamke cigaret uporabljajo pri oglaševanju modele različnih starosti:

> Statistična analiza je pokazala, da je vrsta cigaret zelo značilna, kar kaže na to, da povprečna ocenjena starost modelov ni enaka pri obravnavanih 12 znamkah. Kot smo lahko videli, so nekatere znamke, na primer *Lucky Strike Lights*, *Kool Milds* in *Virginia Slims*, uporabljale domnevno mlajše modele.
>
> [Vir: M.B.Maziz et al., *Perceived age and attractiveness of models in cigarette advertisements*, Journal of Marketing, 56(January 1992): 22–37.]

Razloži nekomu, ki ne zna statistike, kaj pomeni "zelo značilna" in zakaj je to dober dokaz razlik med oglaševalci teh znamk, čeprav so vprašani videli le del oglasov.

1.13 Tehnološki kotiček

Generiranje naključnih števil

Obstaja veliko načinov, kako naključno generirati števila. Uporabimo lahko tabelo naključnih števil. Lahko vlečemo karte iz kupčka, pri čemer zanemarjamo figure. Lahko koga prosimo, da si izmišlja števila. Lahko jih tudi ustvarimo s pomočjo preglednic. Spodnja preglednica vsebuje naključna števila med 0 in 99 iz vsakega od teh virov.

Prvi stolpec smo dobili z uporabo vrstice 142 v tabeli naključnih števil. Drugega smo dobili tako, da smo nekoga prosili, naj si izmisli nekaj števil med 0 in 99. Pri

	A	B	C	D
1	**Tabela**	**Človek**	**Karte**	**Preglednica**
2				
3	6	34	17	0
4	87	54	3	36
5	55	87	94	21
6	33	31		2
7	35	9		69
8	91			78
9	27			87
10	48			74
11	78			1
12	24			55
13	4			47
14	13			14
15	77			6
16	75			3
17	79			46
18	54			67
19	26			45
20	63			17
21	87			30
22	62			97

tretjem smo si pomagali z igralnimi kartami: figure (fanta, kraljico in kralja) smo izpustili, namesto 10 smo prebrali 0, pri vseh ostalih pa smo upoštevali vrednosti, napisane na karti. Rezultate v tabeli dobimo tako, da karte premešamo, izberemo eno, jo vrnemo nazaj v kup in to ponavljamo. V tem konkretnem primeru smo izvlekli asa, sedmico, desetko, trojko, fanta, devetko in štirico.

Četrti stolpec uporabi generator, ki ga najdemo v programih za delo s preglednicami, `=RandBetween(0,99)`. Vsakič ko program izvede ta ukaz, dobimo neko število med 0 in 99. Čeprav vsak element v četrtem stolpcu dobimo z istim ukazom, so vrednosti različne. (Pravzaprav ti programi včasih neprenehoma evalvirajo formule, tako da bodo morda vsa ta okenca izračunana ponovno vsakič, ko bomo kjerkoli v preglednici dodali novo polje. To lastnost je mogoče izklopiti.)

Naloga 1 Naredi in dokončaj preglednico, podobno zgornji.

Uporaba histogramov za prikaz podatkov

Histogrami so en način za prikaz in primerjavo podatkov. Da bi narisali histogram s pomočjo te preglednice, najprej določimo kategorije, na katere bomo razdelili podatke. V preglednico dodamo stolpec, kakršen je na sliki 1.5. Nato izberemo

`Data Analysis Tool`→ `Histogram`.

Kategorije
19
39
59
79
99

Slika 1.5: V preglednico dodamo stolpec, v katerem navedemo željene kategorije.

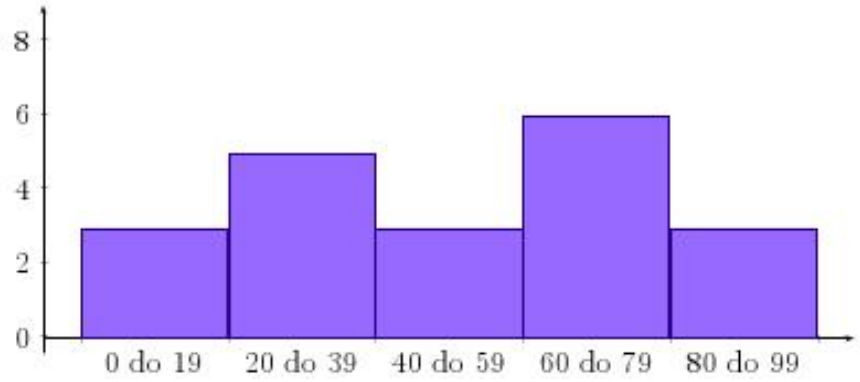
Izberemo stolpec podatkov, ki jih želimo organizirati, nato pa stolpec s kategorijami. Program izdela frekvenčno tabelo, v kateri drugi stolpec pove, koliko podatkov je v vsaki od kategorij. Da bi narisali graf, najprej preimenujemo prvi stolpec v opise kategorij. Za primera podatkov, dobljenih iz tabele naključnih števil oziroma s pomočjo programa, dobimo tabeli 1.2 (a) in (b).

Razred	Frekvenca		Razred	Frekvenca
0 do 19	3		0 do 19	7
20 do 39	5		20 do 39	3
40 do 59	3		40 do 59	4
60 do 79	6		60 do 79	4
80 do 99	3		80 do 99	2
več	0		več	0

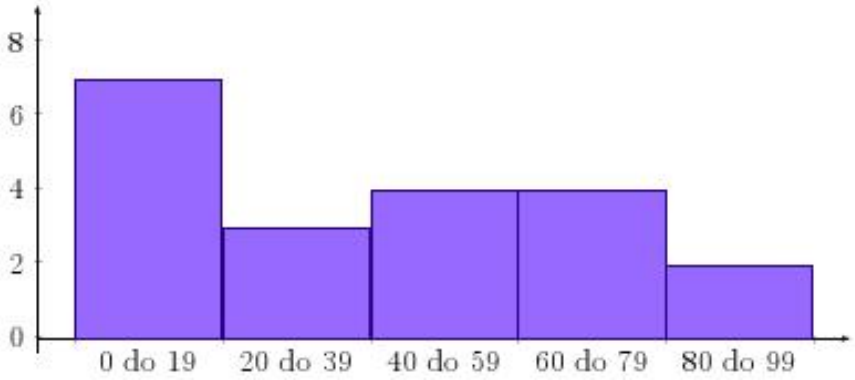
Tabela 1.2: Frekvenčni tabeli za podatke, dobljene z (a) uporabo tabele naključnih števil in (b) s programom.

Nato izberemo oba stolpca, kliknemo na ikono `Graph` in sledimo navodilom za izdelavo grafa. (Stolpci histograma morajo biti sosednji. Če so med njimi presledki, jih izberi, klikni `Format` v meniju, izberi `Options` za `Selected Data Series` in nastavi `Gap Width` na 0.)

Čeprav sta tabela naključnih števil in generator v programu za delo s preglednicami oba nepristranska, se pojavi naravna spremenljivost. Opazimo, da ustrezna histograma na slikah 1.6 in 1.7 izgledata različno in (morda) nekoliko neuravnoteženo.



Slika 1.6: Histogram za podatke, dobljene s tabelo naključnih števil.



Slika 1.7: Histogram za podatke, dobljene z generatorjem iz programa za delo s preglednicami.

Naloga 2 Izdelaj histograma za preglednico, ki si jo naredil(a) v prejšnji nalogi. Primerjaj dobljena histograma. Ali gre razlike v celoti pripisati spremenljivosti?

Naloga 3 Naredi preglednico, v kateri je prvi stolpec seznam študentov v tvojem razredu. Z uporabo generatorja naključnih števil, ki je vključen v program za delo s preglednicami, vsaki osebi priredi enega od treh načinov vadbe: tek, dvigovanje uteži ali nič vadbe.

Naloga 4 Recimo, da je tvoj razred okužen z neznanim smrtonosnim virusom. Obstajajo tri vrste poskusnih zdravil, ki morda delujejo posamično ali v kombinaciji. Napravi preglednico, v kateri prvi stolpec sestavljajo imena študentov, naslednje tri pa tri poskusna zdravila. Naključno določi, katero kombinacijo zdravil bo prejel vsak od študentov.

Raziskovanje

Ko se število podatkov povečuje, naravna spremenljivost postaja vse manj izrazita. Uporabi generator naključnih števil, ki je vgrajen v program, da izdelaš skupine 50, 100 in 500 naključnih števil med 0 in 99. Za vsako od skupin nariši histogram in jih primerjaj.

1.14 Pisni projekti

(1) Pojdi na spletno stran Agencije Gallup (`www.gallup.com`). Tam najdeš arhive nedavnih objav Gallupovih anket. Izberi si temo, ki te zanima, in povzemi rezultate ankete. Zdaj si podrobno oglej ustrezno objavo. Ali natančno navedejo vprašanje, ki so ga zastavili? Mejo napake? Opozorila o neodzivu in drugih virih dodatnih napak? Nazadnje se vrni na Gallupovo domačo stran in preberi sestavek o tem, kako izvajajo ankete. Napiši kratek povzetek uporabljene metode vzorčenja.

(2) Članki v tisku velikokrat opisujejo medicinska spoznanja, ki temeljijo na eksperimentih. Zaključki študije na zdravnikih, da redno jemanje aspirina preprečuje srčne infarkte, so primer take objave. Veliko takih poročil se pojavi v *Journal of the American Medical Association* ali v *New England Journal of Medicine*. V časopisju objavijo članke na dan, ko izideta reviji.

Poišči članek v reviji, ki objavlja rezultate nedavnih medicinskih raziskav. Opiši namen in potek raziskave. Ali je šlo za eksperiment? Do kakšnih zaključkov so prišli in do kakšne mere jim gre verjeti?

Dodaj še kratko kritiko načina predstavitve rezultatov v članku. Ali članek omenja kontrolno skupino? Ali omenja slučajno izbiro osebkov? Če članek opisuje opazovalno študijo, ali vključuje opozorilo pred prevelikim posploševanjem?

Neobvezno: Poišči poročilo v medicinski reviji v knjižnici ali na spletu. Uporabi dejansko poročilo, da napišeš kritiko.

(3) Izberi aktualno temo, ki zanima študente na tvoji šoli. Pripravi kratko anketo (ne več kot pet vprašanj), da zbereš mnenja o tej temi. Izberi vzorec približno 25 študenov, jim daj vprašalnik in na kratko povzemi ugotovitve. Napiši še kratek esej o svojih izkušnjah glede načrtovanja in izvedbe raziskave, ki si jih s tem pridobil(a).

(Čeprav je 25 študentov premalo, da bi lahko bil(a) statistično gotov(a) glede rezultatov, je namen projekta predvsem praktično delo. Najprej moraš določiti populacijo: če ne moreš doseči širše skupine študentov, uporabi študente, ki obiskujejo ta predavanja. So vprašani menili, da so vprašanja jasna? Si postavil(a) vprašanja tako, da je bilo enostavno tabelirati odgovore? Si morda na koncu želel(a), da bi izbral(a) drugačna vprašanja?)

(4) Ankete imajo pomembno vlogo v ameriški politiki. Osredotočimo se na ankete, ki poskušajo napovedati izide volitev. Obstajata dve vrsti takih anket. Ankete, ki jih opravimo pred volitvami, sprašujejo ljudi, za koga bi glasovali, če bi bile volitve tisti dan. *Vzporedne volitve* anketirajo volilce v trenutku, ko zapuščajo volišča. Televizijske hiše uporabljajo vzporedne volitve, da na večer volitev napovedujejo rezultate preden so prešteti vsi glasovi. Nekatere države imajo zakone, ki omejujejo napovedovanja izidov. V Franciji ni dovoljeno objaviti nobenih rezultatov anket v tednu pred predsedniškimi volitvami. Belgija, Italija in Portugalska imajo podobne zakone. Meniš, da je to pametno? Zakaj? (Podrobne komentarje glede volilnih anket lahko najdeš v knjigi *Statistics: Concepts and Controversies*, ki je navedena pod dodatno literaturo.)

(5) Čeprav se ob eksperimentih na ljudeh postavljajo določena etična vprašanja, obstajajo tudi etični pomisleki glede študij, ki zbirajo podatke o ljudeh. Spodaj sta navedena dva pomisleka. O enem izmed njiju napiši krajši esej.

- Vsaka institucija v ZDA, ki prejema državna denarna sredstva, mora imeti posebni revizijski odbor (*Institutional Review Board, IRB*), ki vnaprej pregleda vse študije, ki se izvajajo na ljudeh. IRB je zadolžen za zaščito osebkov. Kako se imenuje ta odbor na tvoji fakulteti? Kdo so člani? Ali so v njem zastopani zunanji člani? Če so, kako so izbrani? Kakšnim smernicam sledijo? Ali imaš predloge, kako bi okrepil(a) zaščito, ki jo nudi odbor?

- Recimo, da na vzorcu izvajaš raziskavo, v kateri ugotavljaš mnenje udeležencev. Etični standardi zahtevajo, da daš potencialnim osebkom nekaj informacij o anketi in pridobiš njihovo informirano soglasje za sodelovanje. Kakšne informacije naj bi dobili? (Morda pričakujejo, da bodo za to porabili deset minut, anketa pa v resnici traja eno uro. Morda vključuje vprašanja o spolnosti ali drogah brez predhodnega opozorila.) Ali bi morali osebkom vedno povedati, kdo sponzorira anketo? (Če da, bo informacija, da anketo sponzorirajo republikanci, vplivala na rezultate?) Ali bi morali anketirancem vedno ponuditi možnost, da jim pošljemo kopijo končnega poročila, da vidijo, kako smo uporabili pridobljene informacije? (To precej stane.)

Poglavje 2

Analiza podatkov

Za sodobno družbo je značilna poplava podatkov. Podatki, ali numerična dejstva, so bistveni pri odločanju na skoraj vseh področjih življenja in dela. Kot druge velike poplave nam poplava podatkov grozi, da nas bo pokopala pod sabo. Moramo jo kontrolirati s premišljeno organizacijo in interpretacijo podatkov. Baza podatkov kakšnega podjetja na primer vsebuje velikansko število podatkov: o zaposlenih, prodaji, inventarju, računih strank, opremi, davkih in drugem. Ti podatki so koristni le v primeru, ko jih lahko organiziramo in predstavimo tako, da je njihov pomen jasen. Posledice neupoštevanja podatkov so lahko hude. Veliko bank je izgubilo na milijarde dolarjev pri nedovoljenih špekulacijah njihovih zaposlenih, ki so ostale skrite med goro podatkov, ki jih odgovorni niso dovolj pozorno pregledali.

Vsaka množica podatkov vsebuje informacije o neki skupini *posameznikov*. Informacije so urejene v *spremenljivke*.

> **Posamezniki** so objekti, ki jih opisuje množica podatkov. To so lahko ljudje, lahko pa so tudi živali ali stvari. **Spremenljivka** je neka lastnost posameznika. Spremenljivka lahko pri različnih posameznikih zavzame različne vrednosti.

Primer. (Baza podatkov v podjetju) Slika 2.1 prikazuje majhen del baze, v kateri korporacija CyberStat hrani podatke o svojih zaposlenih. *Posamezniki* so torej zaposleni. Vsaka vrstica vsebuje podatke o enem posamezniku. Vsak stolpec vsebuje vrednosti ene *spremenljivke* za vse posameznike. Poleg imen je v bazi še pet drugih spremenljivk. Spol, rasa in vrsta dela so spremenljivke, ki razvrščajo delavce in ne zavzamejo številskih vrednosti. Starost in višina plače zavzameta številske

vrednosti. Vidimo, da je starost merjena v letih in višina plače v dolarjih. Večina tabel podatkov ima to obliko, vsaka vrstica je posameznik in vsak stolpec je spremenljivka. Ta množica podatkov je shranjena v programu za delo s preglednicami, ki ima vrstice in stolpce pripravljene za uporabo. Preglednice pogosto uporabljamo za vnos in prenos podatkov in ustrezni programi večinoma vsebujejo razne funkcije za osnovno statistiko. ♦

	A	B	C	D	E	F
1	**Ime**	**Starost**	**Spol**	**Rasa**	**Plača**	**Vrsta dela**
2	Fleetwood, Delores	39	ženski	bela	62,100	menedžment
3	Perez, Juan	27	moški	bela	47,360	strokovno
4	Wang, Ling	22	ženski	azijska	18,250	pisarniško
5	Johnsohn, LaVerne	48	moški	črna	77,600	menedžment

Slika 2.1: Del zbirke podatkov iz programa za delo s preglednicami.

Statistična orodja in ideje nam pomagajo pregledovati podatke, da bi lahko opisali njihove glavne značilnosti. Tako pregledovanje imenujemo *analiza podatkov*. Kot raziskovalec, ki prečka neznano deželo, najprej želimo preprosto opisati, kaj vidimo. V tem poglavju bomo uporabljali števila in slike za raziskovanje podatkov. Tole sta dve načeli, ki nas opremita s taktiko za analizo podatkov:

(1) Najprej pregledamo vsako spremenljivko posebej, nato proučimo povezave med več spremenljivkami.

(2) Začnemo z grafom ali več grafi. Dodajamo numerične povzetke določenih aspektov podatkov.

Ta principa smo uporabili tudi za organizacijo snovi v tem poglavju. Začnemo z obravnavo ene spremenljivke, nato si ogledamo povezave med več spremenljivkami. Na vsakem koraku najprej podatke predstavimo z grafom, nato pa dodamo numerične povzetke.

2.1 Prikaz porazdelitev: Histogrami

Porazdelitev spremenljivke nam pove, katere vrednosti zavzame spremenljivka in kako pogosto zavzame vsako od vrednosti. Analiza podatkov se začne z grafično predstavitvijo porazdelitve vsake od spremenljivk.

Številske spremenljivke pogosto zavzamejo veliko vrednosti. Graf porazdelitve je bolj pregleden, če so vrednosti, ki so si blizu, predstavljene v skupini. Najpogostejši graf porazdelitve številske spremenljivke je **histogram**.

Primer. (Izdelava histograma) Tabela 2.1 prikazuje delež prebivalcev, ki so stari 65 let in več, v vsaki od 50 držav ZDA. Histogram te porazdelitve naredimo takole:

(1) Razdelimo interval, na katerem se nahajajo vrednosti spremenljivke, na enako velike razrede. Podatki v tabeli 2.1 so med 5,2 in 18,5, zato izberemo razrede

5,0 do vključno 6,0,
6,0 do vključno 7,0,
...
18,0 do vključno 19,0.

Pri tem pazimo, da razrede izberemo tako, da je vsak posameznik vsebovan v natanko enem razredu. Država, v kateri je 6,0% prebivalcev starih nad 65 let, bi spadala v prvi razred, država, v kateri bi bil ta delež enak 6,1% pa v drugega.

(2) Za vsak razred preštejemo, koliko posameznikov vsebuje. Podatki so zbrani v spodnji tabeli.

Razred	Število	Razred	Število	Razred	Število
5,1 do 6,0	1	10,1 do 11,0	4	15,1 do 16,0	4
6,1 do 7,0	0	11,1 do 12,0	8	16,1 do 17,0	0
7,1 do 8,0	0	12,1 do 13,0	13	17,1 do 18,0	0
8,1 do 9,0	1	13,1 do 14,0	12	18,1 do 19,0	1
9,1 do 10,0	1	14,1 do 15,0	5		

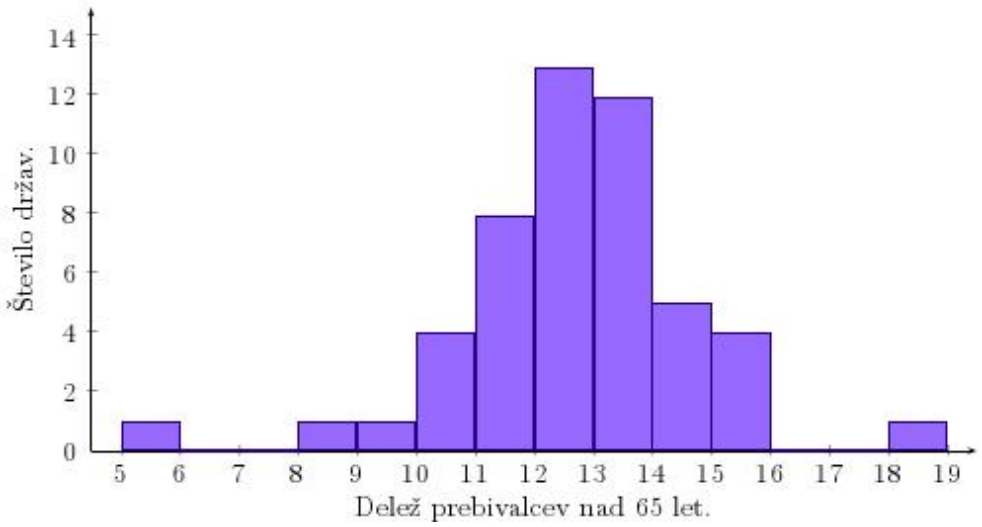
(3) Narišemo histogram. Najprej na vodoravni osi označimo merilo za spremenljivko, katere porazdelitev rišemo. V našem primeru je to "delež prebivalcev nad 65 let". Označimo vrednosti med 5 in 19, ker smo na tem intervalu izbrali naše razrede. Na navpični osi je merilo za števila, ki smo jih dobili s preštevanjem v prejšnji točki. Vsak stolpec predstavlja nek razred, višina stolpca pa je enaka številu posameznikov, ki so v tem razredu. Med stolpci ni vrzeli, razen ko je kakšen od razredov prazen in je zato višina ustreznega stolpca enaka nič. Dobimo histogram na sliki 2.2. ♦

Država	Delež	Država	Delež	Država	Delež
Alabama	13,0	Louisiana	11,4	Ohio	13,4
Aljaska	5,2	Maine	13,9	Oklahoma	13,5
Arizona	13,2	Maryland	11,4	Oregon	13,4
Arkanzas	14,4	Massachusetts	14,1	Pensilvanija	15,9
Kalifornija	10,5	Michigan	12,4	Rhode Island	15,8
Kolorado	11,0	Minnesota	12,4	Južna Karolina	12,1
Conneticut	14,3	Mississippi	12,3	Južna Dakota	14,4
Delaware	12,8	Missouri	13,8	Tennesee	12,5
Florida	18,5	Montana	13,2	Teksas	10,2
Georgia	9,9	Nebraska	13,8	Utah	8,8
Havaji	12,9	Nevada	11,4	Vermont	12,1
Idaho	11,4	New Hampshire	12,0	Virginija	11,2
Illinois	12,5	New Jersey	13,8	Washington	11,6
Indiana	12,6	Nova Mehika	11,0	Zahodna Virginija	15,2
Iowa	15,2	New York	13,4	Wisconsin	13,3
Kansas	13,7	Severna Karolina	12,5	Wyoming	11,2
Kentucky	12,6	Severna Dakota	14,5		

Tabela 2.1: Delež prebivalcev nad 65 let po zveznih državah v ZDA.

Stolpci v histogramu naj bi pokrili celoten obseg vrednosti spremenljivke. Kadar obstajajo med možnimi vrednostmi spremenljivke vrzeli, stolpce razširimo tako, da se srečajo na sredini med dvema možnima vrednostma. Tako bi se na primer v histogramu, ki bi prikazoval starosti predavateljev na fakulteti, stolpca, ki predstavljata starosti med 25 in 29 leti in med 30 in 34 leti, stikala pri 29,5.

V oči nam pade *ploščina* pravokotnih stolpcev. Ker so vsi razredi iste širine, je ploščina določena z višino. Uporabiti moramo lastno presojo, da določimo razrede, od tega pa je odvisna oblika histograma. Če izberemo premalo razredov, ima histogram obliko "nebotičnika", pri katerem so vse vrednosti zbrane v nekaj razredih, predstavljenih z visokimi stolpci. Če izberemo preveč razredov, ima histogram obliko "palačinke", večina razredov vsebuje en element ali pa nobenega. Nobena od izbir ne bo dala dobre predstave porazdelitve. Programska oprema za statistiko bo opravila izbiro namesto nas. Ta izbira je običajno dobra, če želimo, pa jo je moč spremeniti.



Slika 2.2: Histogram deležev državljanov nad 65 let.

2.2 Interpretacija histogramov

Izdelava statističnega grafa ni sama sebi namen. Graf naj bi nam pomagal razumeti podatke. Ko narišemo graf, se vprašajmo, "Kaj vidim?". Ko smo porazdelitev predstavili z grafom, lahko iz njega izluščimo pomembne informacije:

> V vsakem histogramu si najprej ogledamo **celostno sliko** in opazimo morebitna izrazita **odstopanja**. Celostno sliko lahko opišemo z **obliko**, **središčem** in **razponom**. Kmalu se bomo naučili, kako središče in razpon opišemo numerično. Pomembna vrsta odstopanja so **ubežniki**, posamezne vrednosti, ki se ne skladajo s celostno sliko.

Primer. (Opis porazdelitve) Oglejmo si ponovno histogram na sliki 2.2. *Oblika*: Porazdelitev ima en sam *vrh*. Je približno simetrična, tj. oblika je podobna na obeh straneh vrha. *Središče*: Sredina porazdelitve je blizu edinega vrha pri približno 13%. *Razpon*: Razpon je med približno 10% in 16%, če ne upoštevamo štirih najbolj ekstremnih vrednosti.

Ubežniki: Dve vrednosti izstopata v histogramu na sliki 2.2. Potem ko nas histogram opozori nanje, jih brez težav najdemo v tabeli. Na Floridi ima 18,5% prebivalstva več kot 65 let, na Aljaski pa le 5,2%. Ko opazimo ubežnike, začnemo iskati razlago zanje. V nekaterih primerih je vzrok napaka. Lahko bi na primer pri tipkanju

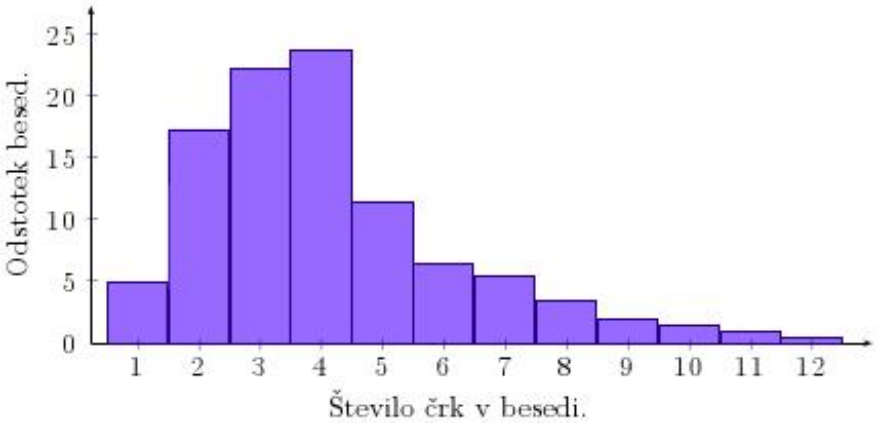
namesto 50 vnesli 5,0. Drugi ubežniki pa opozarjajo na posebne lastnosti nekaterih opažanj. Na Floridi, kjer živi veliko upokojencev, je velik del populacije starejši od 65 let, in na Aljaski, ki je čisto na severu, jih je malo. ♦

Kadar opisujemo porazdelitev, se moramo osredotočiti na glavne značilnosti. Iskati moramo visoke vrhove, ne manjših nihanj v višini stolpcev. Iščemo izrazite ubežnike, ne zgolj največje in najmanjše vrednosti. Iščemo grobo *simetrijo* ali pa očitno *asimetrijo*.

Porazdelitev je **simetrična**, če sta leva in desna stran histograma približno zrcalni sliki druga druge. Porazdelitev je **desno asimetrična**, če je desna stran histograma (tista, kjer so večje vrednosti) precej bolj razpotegnjena kot leva. Je **levo asimetrična**, če je leva stran precej bolj razpotegnjena kot desna.

Porazdelitve realnih podatkov so ponavadi le v grobem simetrične. Sliko 2.2 (brez ubežnikov) tako imamo za približno simetrično. Oglejmo si še primer asimetrične porazdelitve.

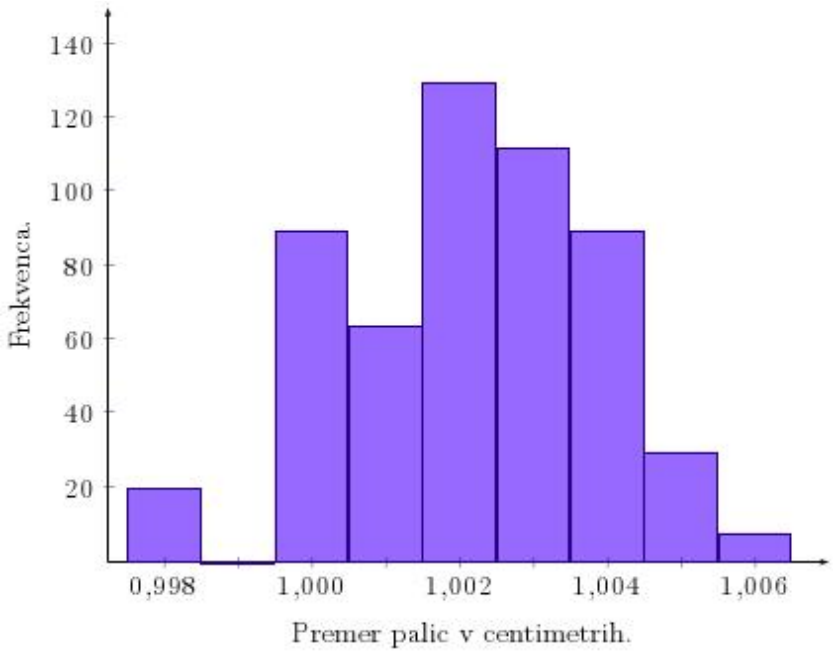
Primer. (Shakespearove besede) Slika 2.3 prikazuje porazdelitev dolžin besed, ki se pojavijo v Shakespearovih dramah. Tudi ta porazdelitev ima en sam vrh, vendar pa je desno asimetrična. Veliko je torej kratkih besed, takih s tremi ali štirimi črkami, daljših, takih z 10, 11 ali 12 črkami, pa je zelo malo. Desni konec diagrama se zato širi precej dlje kot levi. ♦



Slika 2.3: Histogram dolžin besed, uporabljenih v Shakespearovih dramah.

Ubežniki so eno od odstopanj, ki bi jih morali opaziti, ko proučujemo diagram. Naslednji primer prikazuje še eno vrsto odstopanja od splošnih vzorcev.

Primer. (Kontrola kakovosti) Slika 2.4 prikazuje podatke iz študije W. E. Deminga, strokovnjaka za kvaliteto. Gre za podatke o velikosti jeklenih palic, ki se uporabljajo v nekem proizvodnem procesu. Histogram prikazuje premere 500 jeklenih palic, ki so jih izmerili proizvajalčevi inšpektorji. Premere so izmerili na tisočinko milimetra natančno, zato vsak stolpec histograma prikazuje, kako pogosto se je pojavila ustrezna meritev. Opazimo splošni vzorec v velikostih: porazdelitev je približno simetrična s središčem pri 1,002 cm in ostro pada nad in pod to vrednostjo. Opazimo pa tudi odstopanje od tega vzorca: *vrzel* pri 0,999 cm. Palice s premerom, manjšim od 1,000 cm, se ne prilegajo dobro ležajem. Inšpektorji bi jih morali zavreči. Prazni razred pri vrednosti 0,999 cm v histogramu in nepričakovano visok razred pri 1,000 cm, kažeta na to, da inšpektorji palice, ki merijo 0,999 cm, podtikajo k tistim, ki merijo 1,000 cm. Inšpektorji se ne zavedajo, da je lahko tudi samo tisočinka centimetra ključnega pomena. Če bi bili inšpektorji bolje usposobljeni, bi nekaj meritev padlo tudi v razred pri 0,999 cm in ustrezni histogram bi imel pravilnejšo obliko. ♦



Slika 2.4: Demingova ilustracija posledic nepravilnega pregledovanja: histogram z vrzeljo.

Pod žarometom

W. Edwards Deming

Nekateri so mnenja, da je bistvo statistike v razumevanju variacij. Odprava variacij v produktih in procesih je osrednja tema statistične kontrole kakovosti. Ni torej presenetljivo, da je statistik postal vodilni guru na področju kakovosti gospodarstva. V zadnjih desetletjih svojega dolgega življenja je bil W. Edwards Deming (1900-1993) eden od vodilnih svetovnih svetovalcev.

Deming je odraščal v Wyomingu, ZDA, in doktoriral iz fizike na univerzi Yale. Ko je v 30. letih 20. stoletja delal za Oddelek za kmetijstvo, se je seznanil s takrat novim področjem statistike, še posebej s statistično kontrolo procesov, ki jo je izumil Walter Shewhart, AT&T.[a] Leta 1939 se je preselil na Urad za popis prebivalstva kot strokovnjak za vzorčenje.

Delo, s katerim je zaslovel, se je začelo po letu 1946, ko je zapustil državne ustanove. Obiskal je Japonsko, da bi svetoval pri popisu prebivalstva, nato pa se je vrnil, da bi predaval o kontroli kakovosti. Na Japonskem si je pridobil številne privržence, ki so po njem poimenovali najprestižnejšo nagrado za kakovost v industriji. Ko je ugled Japonske proizvodnje rasel, je z njim rasla tudi Demingova slava. Odkrito in celo zajedljivo je dal vodstvom podjetij vedeti, da je večina problemov glede kvalitete sistemskih in da je zanje odgovorna uprava. Spodbujal je vključenost delavcev in neprestano iskanje razlogov za variacije.

[a] *American Telephone & Telegraph Company*, od leta 1885 ameriško telefonsko in telegrafsko podjetje. (Op. prev.)

2.3 Prikaz porazdelitev: Stebelni diagrami

Histogrami niso le grafične predstavitve porazdelitev. Za majhne množice podatkov je izdelava *stebelnega diagrama* hitrejša, poleg tega pa predstavi bolj podrobne informacije.

Za izdelavo **stebelnega diagrama**:

(1) Ločimo vsakega od podatkov na **steblo**, ki je sestavljeno iz vseh števk razen zadnje, in **list**, zadnjo števko. Stebla lahko imajo poljubno mnogo mest, vsak list pa vsebuje le eno števko.

(2) Stebla zapišemo v stolpec, in sicer padajoče z najmanjšim na vrhu. Na desni strani stolpca narišemo navpično črto.

(3) Dodamo vsak list v vrstico na desni strani ustreznega stebla, in sicer v naraščajočem vrstnem redu od stebla navzven.

Primer. (Izdelava stebelnega diagrama) V tabeli 2.1 je celi del vsakega podatka steblo, zadnja števka (desetine) pa je list. Na primer, podatek za Alabamo ima steblo 13 in list 0. Stebla lahko imajo toliko mest, kot je potrebno, vsak list pa mora biti le iz ene števke. Na sliki 2.5 je stebelni diagram za tabelo 2.1. ♦

5	2
6	
7	
8	8
9	9
10	2 5
11	0 0 2 2 4 4 4 4 6
12	0 1 1 3 4 4 5 5 5 6 6 8 9
13	0 2 2 3 4 4 4 578889
14	1 3 4 4 5
15	2 2 8 9
16	
17	
18	5

Slika 2.5: Stebelni diagram deleža prebivalstva nad 65 let.

Stebelni diagram je podoben prevrnjenemu histogramu. Diagram na sliki 2.5 spominja na histogram s slike 2.2. Stebelni diagram pa v nasprotju s histogramom shranjuje tudi vrednosti vsakega podatka. Stebelne diagrame beremo podobno kot histograme: ogledamo si celostno sliko in iščemo morebitne ubežnike.

V histogramu lahko izberemo razrede. Razredi (stebla) v stebelnem diagramu so določeni. Večjo fleksibilnost dosežemo, če podatke zaokrožimo, tako da je zadnja števka po zaokroženju uporabna kot list. To naredimo, kadar imajo podatki preveč mest. Na primer, pri podatkih

$$3{,}468 \quad 2{,}567 \quad 2{,}981 \quad 1{,}095 \quad \ldots$$

bi imeli preveč stebel, če bi za te izbrali prve tri števke. Raje jih zaokrožimo na

$$3{,}5 \quad 2{,}6 \quad 3{,}0 \quad 1{,}1 \quad \ldots$$

preden naredimo stebelni diagram.

2.4 Opis sredine: Povprečje in mediana

Opis porazdelitve skoraj vedno vsebuje podatek o središču ali srednji vrednosti. Najbolj pogosto merilo za središče je običajna aritmetična sredina, *povprečje*.

Povprečje množice podatkov poiščemo tako, da vrednosti seštejemo in vsoto delimo s številom podatkov. Če je teh n podatkov enakih $x_1, x_2, \ldots, x_n$, potem je njihovo povprečje

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Črta nad x označuje, da gre za povprečje vseh vrednosti spremenljivke x. Oznako $\overline{x}$ preberemo kot "x prečna". To je standardna oznaka. Kadar ljudje govorijo o podatkih in zapišejo na primer $\overline{x}$ ali $\overline{y}$, vedno govorijo o povprečju.

Primer. (Računanje povprečja) Švicarska študija je proučila število histerektomij (odstranitev maternice), ki so jih opravili zdravniki v enem letu. Tole so podatki za vzorec 15 zdravnikov:

$$27 \quad 50 \quad 33 \quad 25 \quad 86 \quad 25 \quad 85 \quad 31 \quad 37 \quad 44 \quad 20 \quad 36 \quad 59 \quad 34 \quad 28$$

Stebelni diagram pokaže, da je porazdelitev desno asimetrična in da sta prisotna dva ubežnika z visokima vrednostma:

2	0 5 5 7 8
3	1 3 4 6 7
4	4
5	0 9
6	
7	
8	5 6

Povprečno število histerektomij, ki so jih opravili ti zdravniki, je

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n} = \\ &= \frac{27 + 50 + 33 + \ldots + 28}{15} = \\ &= \frac{620}{15} = 41,3. \end{aligned}$$

V praksi lahko vnesemo podatke v kalkulator in pritisnemo tipko $\overline{x}$. Ni nam potrebno seštevati in deliti. Vseeno pa bi morali vedeti, da je to tisto, kar kalkulator pri tem naredi. ♦

Povprečje je srednja (povprečna) vrednost. Druga možnost za določitev središča podatkov je, da podamo podatek, ki je na sredini, vrednost, od katere je natanko polovica vrednosti manjših in natanko polovica večjih. To je ideja za pojmom *mediane*. Poiščemo jo po temle pravilu:

Mediano M neke porazdelitve poiščemo takole:

(1) Uredimo vse podatke po velikosti od najmanjšega do največjega.

(2) Če je število podatkov n liho, dobimo mediano M tako, da preštejemo $\frac{n+1}{2}$ vrednosti od konca k začetku.

(3) Če je n sodo, je mediana M povprečje sredinskih dveh vrednosti iz urejenega seznama. Tudi v tem primeru se nahaja $\frac{n+1}{2}$ mest od konca seznama.

Pazimo, da pri tem upoštevamo vse podatke, tudi če se kakšna vrednost ponovi večkrat. Prav tako ne smemo pozabiti seznama urediti po velikosti. Sredinsko število v neurejenem seznamu nima nobenega pomena. Formula $\frac{n+1}{2}$ nam pove, na katerem mestu se mediana nahaja, ne pa, koliko je.

Primer. (Računanje mediane) Da bi našli mediano za naš vzorec 15 zdravnikov, najprej vrednosti uredimo:

$$20 \quad 25 \quad 25 \quad 27 \quad 28 \quad 31 \quad 33 \quad 34 \quad 36 \quad 37 \quad 44 \quad 50 \quad 59 \quad 85 \quad 86$$

Skupno imamo $n = 15$ podatkov, torej se mediana nahaja na mestu

$$\frac{n+1}{2} = \frac{16}{2} = 8.$$

Mediana je torej osmi podatek na urejenem seznamu, $M = 34$.

Pri tej študiji so si ogledali tudi vzorec 10 zdravnic. Števila histerektomij, ki so jih opravile te zdravnice, so

$$5 \quad 7 \quad 10 \quad 14 \quad 18 \quad 19 \quad 25 \quad 29 \quad 31 \quad 33$$

Mediana se nahaja na mestu

$$\frac{n+1}{2} = \frac{11}{2} = 5,5.$$

Mesto $5,5$ pomeni "med petim in šestim mestom v urejenem seznamu". Mediana je torej povprečje ustreznih dveh vrednosti:

$$M = \frac{18+19}{2} = 18,5.$$

Tipična zdravnica je torej opravila precej manj histerektomij kot tipični zdravnik. To je bil eden od pomembnih zaključkov te študije. Opazimo še, da je pri lihih n mediana kar nek element s seznama. Kadar je n sod, leži mediana med dvema od vrednosti. ♦

Ta primer prikazuje pomembno razliko med povprečjem in mediano. *Na povprečje močno vpliva nekaj ekstremnih vrednosti.* Posebej je povprečje desno asimetrične porazdelitve večje od mediane. Mediana števila histerektomij, ki so jih opravili zdravniki, je bila 34, vendar pa sta dve zelo veliki vrednosti (85 in 86) v desnem delu porazdelitve dvignili povprečje na $41,3$. V praksi se moramo vedno vprašati, kateri opis središča je boljši, "središčna točka" (mediana) ali "srednja vrednost" (povprečje).

2.5 Opis razpona: Kvartili

Povprečje in mediana sta dve različni merili za središče porazdelitve. Vendar pa je lahko zgolj ta podatek zavajajoč. Urad za štetje prebivalstva je poročal, da je bil leta 1997 povprečni dohodek ameriškega gospodinjstva 37 005 $. Polovica gospodinjstev je imela dohodke pod to vrednostjo in polovica je imela višje dohodke. Vendar pa to število ne pove celotne zgodbe. Dve državi z enako mediano dohodkov sta lahko zelo različni, če se v eni pojavlja ekstremno blagostanje ali revščina, v drugi pa so razlike med gospodinjstvi majhne. Zdravilo, ki ima pravilno povprečno porazdelitev učinkovine, je lahko nevarno, če imajo nekatere serije preveč in druge premalo učinkovine. Poleg središč nas zanimata *razpon* ali *variabilnost* dohodkov ali

moči zdravila. Najenostavnejši uporaben numerični opis porazdelitve je sestavljen iz opisa središča in razpona.

En način za merjenje razpona je, da podamo najmanjšo in največjo vrednost. Na primer, delež prebivalstva nad 65 v ZDA se razprostira med 5,2% na Aljaski in 18,5% na Floridi. Ti posamezni vrednosti nam pokažeta poln razpon teh podatkov, vendar pa lahko gre za ubežnika. Opis razpona lahko izboljšamo, če pogledamo še, kakšen je razpon srednje polovice podatkov. *Kvartili* označujejo srednjo polovico. Preštejemo podatke v urejenem seznamu, začnemo pri najmanjšem. *Prvi kvartil* se nahaja na četrt poti po seznamu. *Tretji kvartil* leži na treh četrtinah seznama. Povedano drugače, prvi kvartil je večji od 25% podatkov, tretji kvartil pa je večji od 75% podatkov. Drugi kvartil je mediana, ki je večja od 50% podatkov. To je ideja kvartilov. Potrebujemo pravilo, s katerim jih natančno opredelimo. Za to pravilo uporabimo pravilo za mediano.

Kvartile izračunamo takole:

(1) Podatke razvrstimo v naraščajoč seznam in poiščemo mediano M.

(2) **Prvi kvartil** Q_1 je mediana tistih podatkov, ki ležijo v urejenem seznamu levo od M.

(3) **Tretji kvartil** Q_3 je mediana tistih podatkov, ki ležijo v urejenem seznamu desno od M.

Primer. (Računanje kvartilov) Števila histerektomij, ki jih je opravil naš vzorec 15 zdravnikov, so (urejeno):

$$20 \quad 25 \quad 25 \quad 27 \quad 28 \quad 31 \quad 33 \quad \mathbf{34} \quad 36 \quad 37 \quad 44 \quad 50 \quad 59 \quad 85 \quad 86$$

Skupno število podatkov je liho, zato je mediana enaka tistemu na sredini, odebeljeno natisnjenemu številu 34. Prvi kvartil je mediana sedmih podatkov, ki ležijo levo od mediane. To je četrti od sedmih podatkov, torej je $Q_1 = 27$. Lahko pa uporabimo tudi pravilo za izračun položaja mediane pri $n = 7$:

$$\frac{n+1}{2} = \frac{7+1}{2} = 4.$$

Tretji kvartil je mediana sedmih podatkov, ki ležijo desno od mediane, $Q_3 = 50$. Kadar je število podatkov liho, mediano celotnega seznama pri računanju kvartilov izpustimo. V vzorcu 10 zdravnic prav tako uredimo podatke:

5 7 10 14 18 | 19 25 29 31 33

Število podatkov je sodo, zato mediana leži med podatkoma iz srednjega para, med peto in šesto vrednostjo, na mestu, označenem z | v seznamu. Prvi kvartil je mediana prvih petih vrednosti, ker ti podatki ležijo levo od mediane. Preveri, da je $Q_1 = 10$ in $Q_3 = 29$. Kadar je število podatkov sodo, pri računanju kvartilov upoštevamo vse vrednosti. ♦

Nekateri programi uporabljajo nekoliko drugačno pravilo za iskanje kvartilov, zato se lahko rezultati, ki jih dobimo z računalnikom, razlikujejo od tistih, ki jih izračunamo sami. To nas ne bo skrbelo. Razlike bodo vedno zanemarljive.

2.6 Povzetek s petimi števili in škatle z brki

Najmanjša in največja vrednost nam povesta le malo o celotni porazdelitvi, dasta pa nam neko informacijo o obeh koncih porazdelitve, ki manjka, če poznamo le Q_1, M in Q_3. Za strnjeno informacijo o središču in razponu hkrati združimo vseh pet števil.

Povzetek s petimi števili neke porazdelitve sestavljajo najmanjša vrednost, prvi kvartil, mediana, tretji kvartil in največja vrednost, zapisani od najmanjšega k največjemu. S simboli:

minimum Q_1 M Q_3 maksimum

Teh pet števil poda dovolj popoln opis središča in razpona. V primeru histerektomij je povzetek s petimi števili za zdravnike enak

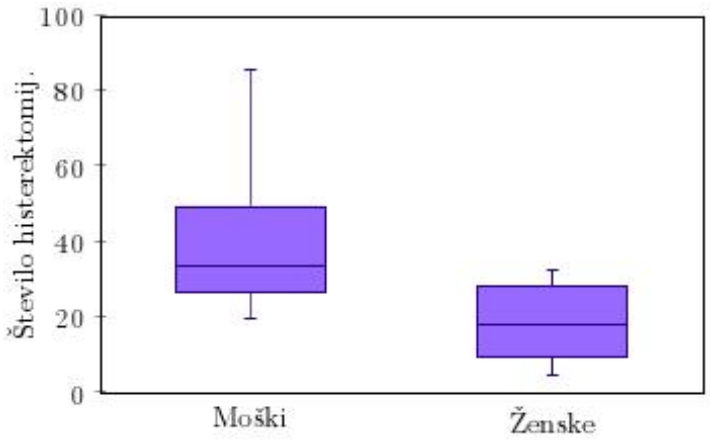
20 27 34 50 86

za zdravnice pa

5 10 18,5 29 33

Povzetek s petimi števili nas privede do nove vrste grafične predstavitve porazdelitev, *škatle z brki*. Na sliki 2.6 je prikazana škatla z brki za primer švicarskih zdravnikov.

Škatla z brki je graf povzetka s petimi števili. Pravokotna škatla je razpeta med obema kvartiloma, črta pa označuje mediano. Dve črti segata iz škatle do največje in najmanjše vrednosti.



Slika 2.6: Vzporedno prikazani škatli z brki za primerjavo števila histerektomij, ki so jih opravili švicarski zdravniki in zdravnice.

Škatle z brki lahko rišemo vodoravno ali pa navpično. V graf moramo vključiti številsko merilo. Ko želimo "prebrati" škatlo z brki, najprej poiščemo mediano, ki označuje središče porazdelitve. Nato si ogledamo razpon. Kvartili nam povedo, kako je porazdeljena srednja polovica podatkov, ekstremi (najmanjša in največja vrednost) pa pokažeta razpon celotne množice podatkov.

Ker so škatle z brki manj podrobne kot histogrami ali stebelni diagrami, so najbolj uporabne pri vzporedni primerjavi večih porazdelitev, podobno kot na sliki 2.6. Takoj opazimo, da so zdravnice v splošnem opravile manj histerektomij kot zdravniki. Pravzaprav je maksimum pri zdravnicah manjši od mediane pri zdravnikih. Vidimo tudi, da je razpon pri zdravnicah manjši. Posebej manjkajo zelo velike vrednosti, ki raztezajo porazdelitev pri zdravnikih.

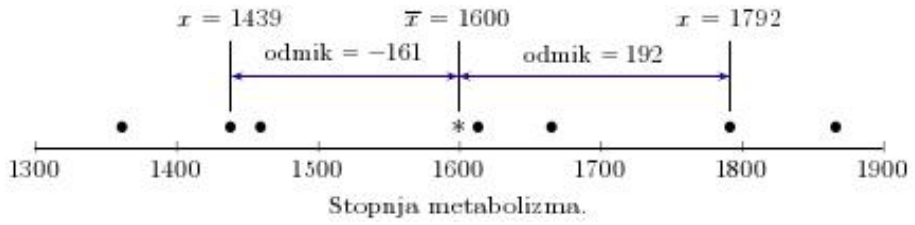
2.7 Opis razpona: Standardni odmik

Čeprav je povzetek s petimi števili najbolj splošno uporaben numerični opis porazdelitve, ni najbolj pogost. To mesto pripada kombinaciji povprečja in *standardnega odmika*. Povprečje je (tako kot mediana) merilo za središče. Standardni odklon pa je (tako kot kvartili in ekstremi iz povzetka s petimi števili) merilo razpona. Standardni odklon in njegova bližnja sorodnica, *varianca*, merita razpon tako, da pogledata, kako daleč od povprečja so vrednosti.

Primer. (Razumevanje standardnega odklona) Stopnja metabolizma je hitrost, s katero telo porablja energijo. Pomembna je pri proučevanju pridobivanja teže, hujšanja in vadbe. Spodaj so podatki o stopnjah metabolizma za sedem moških, ki so sodelovali v raziskavi o hujšanju. (Enote so kalorije na 24 ur. Gre za iste kalorije, ki jih uporabljamo za opis energijske vsebnosti hrane.)

1792 1666 1362 1614 1460 1867 1439

Na sliki 2.7 so prikazani ti podatki kot točke nad številsko premico, povprečje pa je označeno z zvezdico (*). Črti s puščicami označujeta dva od odmikov od povprečja.



Slika 2.7: Varianca in standardni odmik merita razpršenost tako, da pogledata, kako se opažanja razlikujejo od njihovega povprečja.

Ti odmiki pokažejo, kako zelo so podatki razprostrti okoli povprečja. Nekateri odmiki so pozitivni, nekateri negativni. Kvadriramo jih, da postanejo vsi pozitivni. Kvadrati odmikov vrednosti, ki so daleč od povprečja v katerikoli smeri, bodo veliki. Razumna mera za razpon je torej povprečje kvadratov odmikov. To povprečje imenujemo *varianca*. Varianca je velika, če se vrednosti široko razprostrirajo okoli povprečja; majhna je, če so te vrednosti blizu povprečja.

Vendar pa ima varianca napačne enote: če merimo stopnjo metabolizma v kalorijah, bodo enote za varianco stopnje metabolizma kvadratne kalorije. Da dobimo kalorije, varianco korenimo. Kvadratni koren iz variance je *standardni odklon*. ♦

Varianca s^2 neke množice podatkov je povprečje kvadratov odklonov vrednosti od povprečja. Označimo podatke z $x_1, x_2, \ldots, x_n$. Potem je varianca

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}.$$

Standardni odklon s je kvadratni koren iz variance s^2.

V praksi uporabimo funkcije, ki so vgrajene v kalkulatorje, da dobimo standardni odklon za vnešene podatke. Kljub temu si bomo ogledali podroben primer, ki nam bo pomagal razumeti, kako delujeta varianca in standardni odklon.

Primer. (Računanje standardnega odklona) Da bi poiskali standardni odklon za danih sedem stopenj metabolizma, najprej izračunamo povprečje:

$$\overline{x} = \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} = \frac{11200}{7} = 1600.$$

Varianco in standardni odklon začnemo računati pri odklonih, kakršna sta prikazana na sliki 2.7.

Vrednosti	**Odkloni**	**Kvadrati odklonov**
x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1792	$1792 - 1600 = 192$	$192^2 = 36864$
1666	$1666 - 1600 = 66$	$66^2 = 4356$
1362	$1362 - 1600 = -238$	$(-238)^2 = 56644$
1614	$1614 - 1600 = 14$	$14^2 = 196$
1460	$1460 - 1600 = -140$	$(-140)^2 = 19600$
1867	$1867 - 1600 = 267$	$267^2 = 71289$
1439	$1439 - 1600 = -161$	$(-161)^2 = 25921$
	vsota $= 0$	vsota $= 214870$

Varianca je enaka vsoti kvadratov odklonov, deljeni z ena manj kot je število podatkov:

$$s^2 = \frac{214870}{6} = 35\,811,67.$$

Standardni odklon je kvadratni koren iz variance:

$$s = \sqrt{35\,811,67} = 189,24 \text{ cal}.$$

♦

Pomembnejše od samega izračuna so lastnosti, zaradi katerih je standardni odklon koristen:

- s meri razpon okoli povprečja, zato ga smemo uporabiti le takrat, kadar središče podajamo s povprečjem.
- $s = 0$ le takrat, ko *ni odklona*. To se zgodi le v primeru, ko so vse vrednosti enake. Sicer je $s > 0$. Ko postajajo vrednosti bolj razpršene okoli povprečja, se s povečuje.

- s ima iste enote kot ustrezni podatki. Na primer, če merimo stopnjo metabolizma v kalorijah, so tudi enota za s kalorije. To je eden od razlogov, zakaj imamo raje s kot varianco s^2, ki se izraža v kvadratnih kalorijah.

- Kot na povprečje $\overline{x}$ tudi na s močno vplivajo posamične ekstremne vrednosti. Standardni odklon na primeru podatkov o histerektomijah, ki so jih opravili zdravniki, je na primer enak 20,61. (Preveri s kalkulatorjem.) Če izpustimo ekstremni vrednosti 85 in 86, se standardni odklon zmanjša na 10,97.

Zdaj lahko izbiramo med dvema opisoma središča in razpona: med povzetkom s petimi števili in parom $\overline{x}$, s. Ker sta $\overline{x}$ in s občutljiva na ekstremne vrednosti, nas lahko zavedeta, kadar je porazdelitev zelo asimetrična ali kadar ima ubežnike. Še več, ker ima vsaka od strani asimetrične distribucije drugačen razpon, ga ne moremo opisati z enim samim številom kot je s. To nalogo bolje opravi povzetek s petimi števili.

> Povzetek s petimi števili je boljša izbira kot povprečje in standardni odklon za opis asimetričnih porazdelitev ali porazdelitev z ubežniki. Uporabi $\overline{x}$ in s le za primerno simetrične porazdelitve brez ubežnikov.

Čeprav je uporaba standardnega odklona zelo razširjena, ni naravna ali priročna izbira za merjenje razpona porazdelitve. Resnični razlog za popularnost standardnega odklona je v tem, da je naravno merilo za razpon *normalnih porazdelitev*, pomembnega razreda porazdelitev, ki jih bomo spoznali v naslednjem poglavju.

2.8 Prikaz zveze med dvema spremenljivkama

Primeri, ki smo si jih ogledali do sedaj, so obravnavali le eno spremenljivko, na primer število histerektomij, ki so jih opravili zdravniki. Zdaj bomo pregledali podatke za dve spremenljivki, pri čemer bo poudarek na vrsti in moči zveze med spremenljivkama. V ta namen izmerimo obe spremenljivki za isto skupino posameznikov. Velikokrat smo mnenja, da ena od spremenljivk pojasnjuje drugo ali nanjo vpliva.

> **Odzivna spremenljivka** meri izide študije. **Obrazložitvena spremenljivka** razlaga ali vpliva na spremembe odzivne spremenljivke.

Primer. (Poraba zemeljskega plina) Samo bo vgradil sončne kolektorje, da bi zmanjšal stroške ogrevanja hiše. Seveda želi vedeti, v kakšni meri bodo kolektorji pripomogli k manjši porabi plina, zato spremlja porabo pred inštalacijo. Poraba plina je višja ob hladnem vremenu, zato je pomembna zveza med zunanjo temperaturo in porabo.

V tabeli 2.2 so podatki za devet mesecev. Odzivna spremenljivka[1] je povprečna dnevna poraba zemeljskega plina za vse dni v mesecu v kubičnih metrih. Obrazložitvena spremenljivka je povprečno število stopinjskih dni za vse dni v mesecu. (Stopinjski dnevi so mera za potrebo po ogrevanju. Število stopinjskih dni v nekem dnevu dobimo iz povprečne dnevne temperature tako, da za vsako stopinjo pod 65°F dodamo en stopinjski dan. Tako na primer povprečna temperatura 20°F ustreza 45 stopinjskim dnem.[2]) Pogled na števila v tabeli nam pove, da več stopinjskih dni

	Okt	Nov	Dec	Jan	Feb	Mar	Apr	Maj	Jun
Stopinjskih dni	15,6	26,8	37,8	36,4	35,5	18,6	15,3	7,9	0,0
Poraba plina (m^3)	14,72	17,27	22,65	24,07	24,92	13,88	12,74	7,08	3,11

Tabela 2.2: Poraba zemeljskega plina v gospodinjstvu.

(nižje temperature) sovpada z večjo porabo plina. Ampak oblika in moč te zveze nista popolnoma jasni. Da bi prikazali in interpretirali te podatke, potrebujemo primeren diagram. Na sliki 2.8(a) je *razsevni diagram* Samovih podatkov. ♦

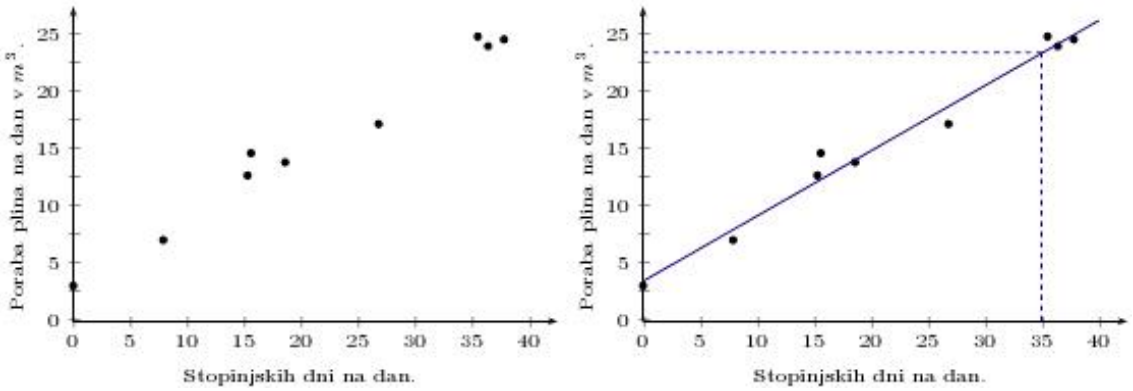
Razsevni diagram pokaže povezavo med dvema numeričnima spremenljivkama, ki ju izmerimo na istih posameznikih. Vrednosti ene od spremenljivk se pojavijo na vodoravni osi, vrednosti druge pa na navpični. Vsak posameznik iz množice podatkov je predstavljen s točko na diagramu, ki jo določata koordinati, dani z ustreznima vrednostma obeh spremenljivk.

Obrazložitveno spremenljivko, če obstaja, vedno narišemo na vodoravno os (x-os) razsevnega diagrama. V navadi je tudi, da imenujemo obrazložitveno spremenljivko

[1] Matematikom sta najbrž bolj domača pojma odvisne in neodvisne spremenljivke. V statistiki včasih neodvisni spremenljivki pravimo regresor, pa tudi obrazložitvena, napovedna, kontrolirana spremenljivka in podobno. Neodvisno spremenljivko včasih imenujemo regresand ali pa merjena, pojasnjena, odzivna spremenljivka. (Op. prev.)

[2] V Sloveniji energetiki uporabljajo izraz *temperaturni primakljaj*. Ta pove, koliko stopinj Celzija je bila povprečna temperatura zunanjega zraka v izbranem dnevu nižja od 20°C. Dogovorjeno je, da se temperaturni primankljaj računa le v dneh, ko je povprečna temperatura nižja od 12°C. Le tedaj je namreč potrebno ogrevanje stavb. (Op. prev.)

x in odzivno y. Kadar nimamo ene obrazložitvene in ene odzivne spremenljivke, lahko poljubno izbiramo, katero bomo nanesli na vodoravno os. Na sliki 2.8(a) stopinjske dni nanesemo na vodoravno os in porabo plina na navpično, ker so stopinjski dnevi obrazložitvena spremenljivka. Vreme je tisto, ki vpliva na porabo plina, ni poraba plina tista, ki bi pojasnjevala vreme.



Slika 2.8: Poraba zemeljskega plina v odvisnosti od stopinjskih dni. (a) Razsevni diagram. (b) Regresijska premica in njena uporaba pri napovedovanju.

Kot takrat, ko smo proučevali porazdelitve ene same spremenljivke, tudi tu pogledamo celostno sliko razsevnega diagrama in nato poiščemo presenetljiva odstopanja.

Celostno sliko razsevnega diagrama lahko opišemo z **obliko**, **smerjo** in **močjo** zveze.

Oblika zveze med stopinjskimi dnevi in porabo plina je jasna: točke težijo k premici. Celostno sliko lahko predstavimo s premico skozi točke diagrama. Na sliki 2.8(b) je prikazana taka premica. Ko se povečuje število stopinjskih dni, se povečuje tudi poraba plina. To je *smer* zveze. Točke v diagramu ležijo zelo blizu premice, zato je ta zveza precej *močna*. Število stopinjskih dni pojasni večino variacij v porabi plina. Pri šibkejši linearni zvezi bi bile točke bolj razpršene okoli premice. Razpršenost odseva učinke drugih dejavnikov, na primer uporabo plina za kuhanje ali pa izklop termostata, kadar gre družina na počitnice. Ti učinki so relativno majhni. Prav tako nimamo nobenih ubežnikov (točk, ki bi padle daleč izven splošne slike) ali drugih pomembnih odklonov.

2.9 Regresijske premice

Samo želi uporabiti svoje podatke, da bi napovedal, kolikšna bo poraba pri poljubni zunanji temperaturi (v stopinjskih dneh). To lahko naredi, če na razsevni diagram na sliki 2.8(a) nariše premico.

Regresijska premica je premica, ki opisuje, kako se odzivna spremenljivka y spreminja v odvisnosti od obrazložitvene spremenljivke x. Velikokrat uporabimo regresijsko premico za napoved vrednosti y pri neki dani vrednosti x.

Točke na sliki 2.8(a) ležijo tako blizu premice, da ni težko narisati regresijske premice na diagram, če uporabimo prozorno ravnilo. Na ta način dobimo premico na diagramu, ne pa tudi njene enačbe. Prav tako ni nobenega zagotovila, da je premica, ki jo narišemo po občutku, res najboljša za predvidevanje porabe. Obstajajo statistične metode, s katerimi iz podatkov dobimo enačbo najboljše premice (pri več različnih pomenih besede "najboljše"). Kmalu si bomo ogledali najbolj pogosto od teh metod, imenovano *regresija najmanjših kvadratov*. Premica na sliki 2.8(b) je regresijska premica najmanjših kvadratov za Samove podatke. Vsi računalniški programi za statistiko in veliko kalkulatorjev zna namesto nas izračunati premico najmanjših kvadratov, tako da nam je le-ta pogosto na voljo brez veliko dodatnega dela. Morali bi torej vedeti, kako uporabljamo take premice, četudi se ne naučimo, kako te premice iz podatkov tudi izračunamo.

Pri pisanju enačbe premice bo x obrazložitvena spremenljivka, ker jo nanašamo na vodoravno os, y pa odzivna spremenljivka. Vsaka premica ima enačbo oblike

$$y = a + bx.$$

Število b imenujemo *naklon* premice, ki pove, za koliko se spremeni y, ko se x poveča za 1. Naklon je običajno pomemben za statistika, ker nam pove hitrost, s katero se spreminja odziv y, ko x narašča. Število a je *začetna vrednost*, vrednost spremenljivke y pri $x = 0$.

Primer. (Pomen naklona in začetne vrednosti) Računalniški program nam pove, da je regresijska premica najmanjših kvadratov, ki jo dobimo iz Samovih podatkov,

$$y = 3,48 + 0,57x.$$

Naklon te premice je $b = 0,57$. To pomeni, da poraba plina naraste za 0,57 m^3 na dan za vsak dodani stopinjski dan. Začetna vrednost je $a = 3,48$. Ko ni stopinjskih dni (ko je torej povprečna temperatura 65°F ali več), bo poraba plina enaka 3,48 m^3 na dan. Naklon in začetna vrednost sta oceni, ki ju dobimo pri prilagajanju premice podatkom iz tabele 2.2. Ne pričakujemo, da bo vsak mesec z nič stopinjskimi dnevi povprečna poraba enaka natanko 3,48 m^3. Premica predstavlja le celostno sliko podatkov. ♦

Namen regresijske premice je napovedovanje vrednosti odzivne spremenljivke pri danih vrednostih obrazložitvene spremenljivke. Premico, ki jo narišemo v razsevni diagram, lahko uporabimo za postavljanje napovedi s pomočjo svinčnika in ravnila. Kadar pa poznamo enačbo premice, lahko vanjo enostavno vstavimo dano vrednost obrazložitvene spremenljivke.

Po namestitvi sončnih kolektorjev želi Samo izvedeti, koliko je s tem prihranil pri stroških ogrevanja. Ne more enostavno primerjati porabe pred in po namestitvi, ker zima pred namestitvijo ni bila nujno enako ostra kot tista po njej. Namesto tega lahko uporabi regresijsko premico, da predvidi, koliko plina bi porabil brez kolektorjev. Iz primerjave te napovedi z dejansko porabo bo lahko izračunal prihranek.

Primer. (Napovedovanje porabe plina) Tega februarja je bilo povprečje 35 stopinjskih dni na dan. Koliko plina bi Samo porabil brez kolektorjev? Slika 2.8(b) prikazuje uporabo regresijske premice za napovedovanje. Najprej poiščemo število 35 na vodoravni osi. Od tam gremo navzgor do regresijske premice in nato levo do osi, na kateri je poraba plina. Na ta način predvidimo, da bo poraba nekaj več kot 23 m^3 na dan. Bolj natančno oceno lahko dobimo z uporabo enačbe regresijske premice. Ta se glasi

$$y = 3,48 + 0,57x.$$

V tej enačbi je x število stopinjskih dni na dan v mesecu in y je predvidena poraba plina na dan v m^3. Naša predvidena poraba plina za mesec z $x = 35$ stopinjskimi dnevi bo

$$y = 3,48 + 0,57 \cdot 35 = 23,43.$$

Ta napoved skoraj gotovo ni popolnoma enaka porabi v mesecu s 35 stopinjskimi dnevi. Vendar pa podatki ležijo tako blizu premice, da smo lahko prepričani, da bo poraba zelo blizu 23,43 m^3 na dan. ♦

2.10 Korelacija

Razsevni diagram prikaže obliko, smer in moč zveze med dvema kvantitativnima spremenljivkama. Linearne zveze so pomembne, ker je premica preprosta in precej pogosta oblika. Pravimo, da je linearna zveza močna, če točke ležijo blizu premice, in šibka, če so točke širše raztresene okoli premice. Naše oči ne presodijo dobro, kako močna je zveza. Slediti moramo naši strategiji za analizo podatkov in uporabiti numerična merila, s katerim dopolnimo grafe. Uporabimo *korelacijo*.

Korelacija meri smer in moč linearne zveze med dvema kvantitativnima spremenljivkama. Običajno jo označimo z r.
Recimo, da imamo podatke o vrednosti spremenljivk x in y za n posameznikov. Vrednosti pri prvem posamezniku sta x_1 in y_1, pri drugem x_2 in y_2, in tako naprej. Povprečje in standardni odklon za prvo spremenljivko sta $\overline{x}$, s_x, za drugo pa $\overline{y}$, s_y. Korelacija r med x in y je

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right).$$

Ne pozabimo, da znak za vsoto $\sum$ pomeni "seštej te člene za vse posameznike". Formula za korelacijo r nam pomaga videti, kaj korelacija je, vendar pa v praksi za izračun uporabimo ustrezno programsko opremo ali kalkulator, ki poišče r za vnešene podatke o spremenljivkah x in y. Naloga 28 zahteva, da izračunamo korelacijo korak za korakom iz definicije. Pri tem nam pride prav tabela, podobna tisti, ki smo jo uporabili za računanje variance na strani 71.

Korelacija uporabi odklone spremenljivk x in y od njunih povprečij. Predznak r torej pokaže smer zveze med x in y. Višina in teža se na primer običajno spreminjata skupaj. Ljudje, ki so nadpovprečno veliki, so navadno tudi nadpovprečno težki. Ljudje, ki so manjši od povprečja, so običajno tudi lažji. Torej sta odklona od povprečja, ki ju zmnožimo, da dobimo posamičen člen iz vsote v formuli za r, večinoma oba pozitivna ali pa oba negativna. Produkti takih členov so večinoma pozitivni in zato je r pozitiven.

Bolj podrobnen pregled formule razkrije še več podrobnosti o korelaciji r. Za uspešno interpretacijo moramo vedeti naslednje:

(1) Korelacija ne razlikuje med obrazložitvenimi in odzivnimi spremenljivkami. V izračunu je vseeno, katero spremenljivko imenujemo x in katero y.

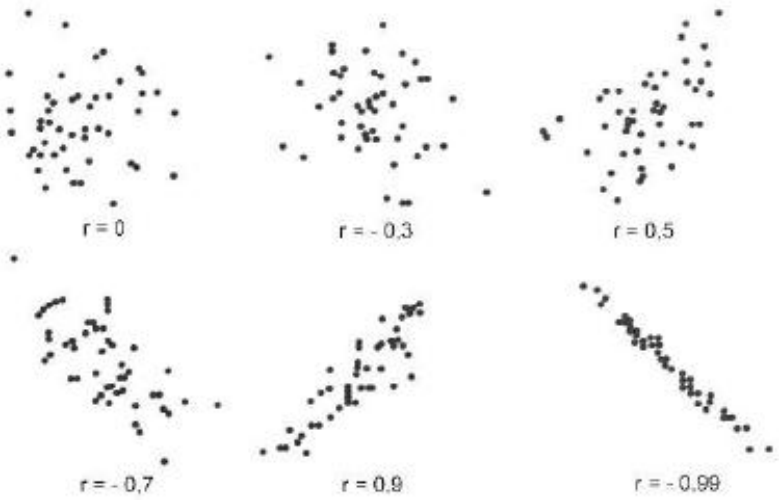
(2) Korelacija meri moč le za linearne zveze. Ne opisuje drugih zvez med spremenljivkami, ne glede na to, kako močne so.

(3) Predznak r pove smer zveze. Pozitivni r pomeni pozitivno zvezo: spremenljivki se gibljeta skupaj. Negativni r pomeni negativno zvezo: spremenljivki se gibljeta v nasprotnih smereh.

(4) Korelacija r je vedno število med -1 in 1. Vrednosti r blizu 0 pomenijo šibko linearno zvezo. Moč linearne zveze narašča, ko se r oddaljuje od 0 k -1 ali k 1. Vrednosti r blizu -1 ali 1 so znak, da točke ležijo skoraj na premici. Ekstremni vrednosti $r = -1$ in $r = 1$ se pojavita le v primeru, ko gre za popolno linearno zvezo, se pravi, kadar točke razsevnega diagrama ležijo natanko na neki premici. Razsevni diagrami na sliki 2.9 prikazujejo, kako vrednosti r blizu 1 ali -1 pomenijo močnejšo linearno zvezo.

(5) Korelacija r se ne spremeni, če spremenimo enote, v katerih merimo spremenljivki x in y. Če višino merimo v čevljih namesto v metrih in težo v funtih namesto v kilogramih, to ne spremeni korelacije med težo in višino. Korelacija r nima enot, je samo število.

(6) Kot na povprečje in na standardni odklon tudi na korelacijo močno vplivajo ubežniki.



Slika 2.9: Korelacija meri moč linearne zveze: Oblike, ki so bolj podobne premici, imajo korelacije bližje ± 1. V primerih na sliki je zaporedoma $r = 0$, $r = -0,3$, $r = 0,5$, $r = -0,7$, $r = 0,9$ in $r = -0,99$.

Slika 2.8 prikazuje zelo močno linearno zvezo med stopinjskimi dnevi in porabo zemeljskega plina. Korelacija je $r = 0.989$, kar je blizu $r = 1$, ki pripada popolni premici. Preveri to s kalkulatorjem tako, da si pomagaš s podatki iz tabele 2.2.

Pod žarometom

Florence Nightingale

Florence Nightingale (1820–1910) je zaslovela kot ustanoviteljica medicinskih sester in reformatorka zdravstvenega sistema. Kot glavna sestra britanske vojske v Krimski vojni med letoma 1854 in 1856 je prišla do zaključka, da so pomanjkanje higiene in bolezni ubili veliko število ranjenih vojakov. Njene reforme so zmanjšale smrtnost v njeni vojaški bolnišnici iz $42,7\%$ na $2,2\%$ in iz vojne se je vrnila slavna. Takoj je začela uspešen boj za reformo celotnega vojaškega sistema zdravstvene nege.

Eno od glavnih orožij, ki jih je Florence Nightingale uporabljala pri svojih bojih, so bili podatki. Poznala jih je, ker je reformirala tudi vodenje evidenc. Bila je pionir v uporabi grafov za predstavitev podatkov na slikovit način, ki so ga lahko razumeli tudi generali in člani parlamenta. Njeni domiselni grafi so mejnik v razvoju nove statistične znanosti. Menila je, da je statistika bistvena za razumevanje vseh socialnih vprašanj in poskušala je vpeljati študij statistike v visoko šolstvo.

2.11 Regresija najmanjših kvadratov

Kadar razsevni diagram kaže linearno zvezo med obrazložitveno spremenljivko x in odzivno spremenljivko y, želimo narisati premico, ki opisuje to zvezo. Točke bodo le redko ležale točno na premici, zato je naša naloga poiskati premico, ki se najbolje prilega tem točkam. Da bi to lahko naredili, moramo najprej povedati, kaj razumemo pod "premico, ki se najbolje prilega".

Recimo, da želimo uporabiti našo premico, da bi napovedali vrednosti y za dane vrednosti x, tako kot je to storil Samo, ko je iz stopinjskih dni napovedal porabo plina. Napako v naši napovedi merimo v navpični (y) smeri. Želimo torej, da bi bile navpične razdalje naših točk do iskane premice tako majhne, kot je le možno. Premica, ki se dobro prilega podatkom, ne leži v celoti nad ali pod vsemi točkami, zato bodo nekatere napake pozitivne in druge negativne. Njihovi kvadrati pa bodo seveda pozitivni. *Regresijska premica najmanjših kvadratov* je tista, za katero je

vsota kvadratov napak najmanjša možna.

Regresijska premica po metodi najmanjših kvadratov je premica, za katero je vsota kvadratov navpičnih razdalj od točk do premice najmanjša možna.

Ideja najmanjših kvadratov pove, v kakšnem smislu se premica najbolje prilega. Še vedno se moramo naučiti, kako to premico izračunamo iz podatkov. Če imamo n podatkov za spremenljivki x in y, kako se glasi enačba premice najmanjših kvadratov? Tule je rešitev tega matematičnega problema:

Dani so podatki o vrednostih obrazložitvene spremenljivke x in odzivne spremenljivke y za n posameznikov. Iz teh podatkov izračunamo $\overline{x}$ in $\overline{y}$, nato pa še standardna odklona s_x in s_y in korelacijo r. Regresijska premica najmanjših kvadratov je premica

$$y = a + bx$$

z **naklonom**

$$b = r\frac{s_y}{s_x}$$

in **začetno vrednostjo**

$$a = \overline{y} - b\overline{x}.$$

Ta enačba nam da vpogled v obnašanje regresijske premice najmanjših kvadratov, ker nam pokaže, da je le-ta povezana s povprečji, standardnimi odkloni in korelacijo spremenljivk x in y. V praksi ni potrebno najprej izračunati povprečij, standardnih odklonov in korelacije. Statistični programi ali kalkulatorji nam vrnejo naklon b in začetno vrednost a za vpisane podatke. S kalkulatorjem preveri, da je enačba regresijske premice najmanjših kvadratov iz Samovega primera porabe plina res $y = 3,48 + 0,57x$, kot smo trdili prej. Kalkulator bo pri tem vrnil večje število decimalnih mest za začetno vrednost in naklon.

Primer. (Ali težji ljudje porabijo več energije?) Stopnja metabolizma, hitrost, s katero telo porablja energijo, je pomembna pri proučevanju pridobivanja teže, diet in vadbe. Tabela 2.3 vsebuje podatke o pusti telesni teži in osnovni stopnji metabolizma za 12 žensk in 7 moških, ki so sodelovali pri študiji neke diete.

Pusta telesna teža, podana v kilogramih, je telesna teža brez maščob. Stopnjo metabolizma merimo v kalorijah, ki jih porabimo v 24 urah, istih kalorijah, s katerimi opisujemo, koliko energije vsebuje hrana. Raziskovalci verjamejo, da pusta telesna teža pomembno vpliva na stopnjo metabolizma.

Oseba	Spol	Teža (kg)	Stopnja (cal)	Oseba	Spol	Teža (kg)	Stopnja (cal)
1	M	62,0	1792	11	Ž	40,3	1189
2	M	62,9	1666	12	Ž	33,1	913
3	Ž	36,1	995	13	M	51,9	1460
4	Ž	54,6	1425	14	Ž	42,4	1124
5	Ž	48,5	1396	15	Ž	34,5	1052
6	Ž	42,0	1418	16	Ž	51,1	1347
7	M	47,4	1362	17	Ž	41,2	1204
8	Ž	50,6	1502	18	M	51,9	1867
9	Ž	42,0	1256	19	M	46,9	1439
10	M	48,7	1614				

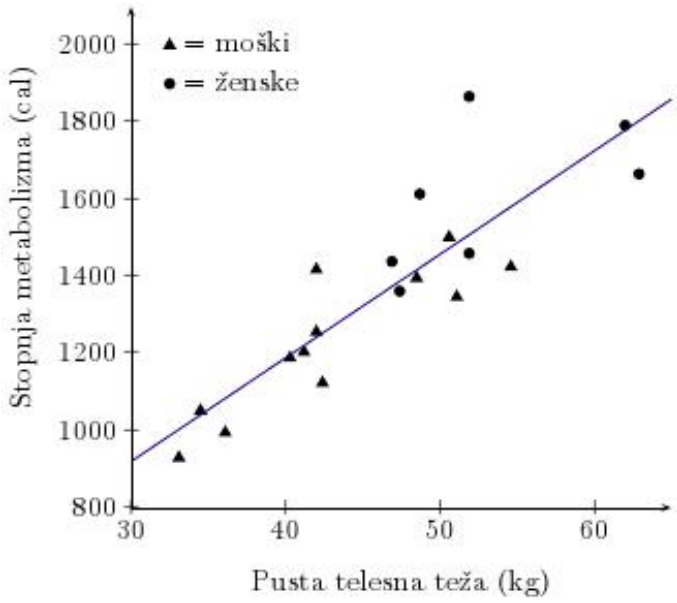
Tabela 2.3: Teža in stopnja metabolizma.

Slika 2.10 je razsevni diagram podatkov. Ker menimo, da telesna teža pomaga pojasniti stopnjo metabolizma, nanesemo težo na vodoravno os. Razsevnemu diagramu smo dodali še eno posebnost: dva različna simbola za označevanje točk nam pomagata razlikovati med moškimi in ženskami. To bo koristno, ker bomo kljub temu, da imajo ženske kot skupina nižjo težo kot moški, videli, da v obeh primerih velja podobna zveza. Računanje bomo zato izpeljali na vseh 19 primerkih skupaj.

Razsevni diagram kaže na srednje močno pozitivno linearno zvezo. Korelacija $r = 0,865$ opiše moč te zveze bolj natančno. Premica na diagramu je regresijska premica najmanjših kvadratov, s pomočjo katere napovemo stopnjo metabolizma iz puste telesne teže. Enačba te premice se glasi

$$y = 113,165 + 26,879x.$$

Naklon premice nam pove, da v povprečju osebki porabijo približno 27 kalorij na dan več za vsak dodatni kilogram telesne teže. Začetno vrednost $a = 113,165$ potrebujemo zato, da lahko narišemo premico, nima pa nobenega statističnega pomena. Telesna teža $x = 0$ ni možna, zato ne moremo govoriti o vrednosti stopnje metabolizma pri $x = 0$. ♦



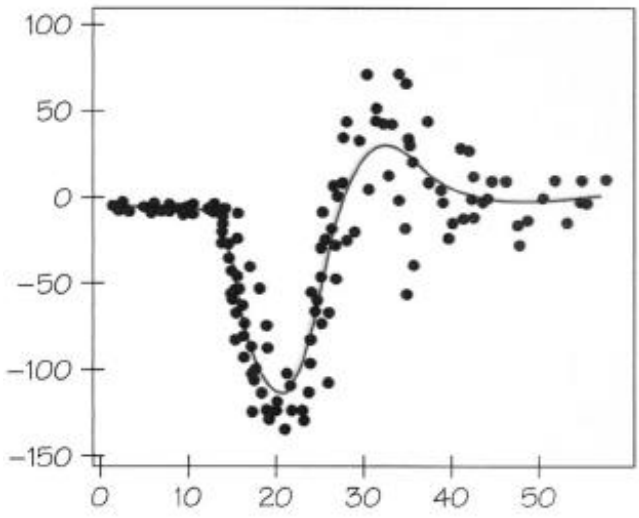
Slika 2.10: Razsevni diagram stopnje metabolizma glede na težo za 12 žensk in 7 moških. Za razlikovanje uporabimo različne simbole.

2.12 Sodobna analiza podatkov

Razsevni diagrami, korelacija in regresija so osnovna orodja za opisovanje zvez med dvema spremenljivkama. In če je zveza bolj zapletena in ne linearna? Kaj pa, če imamo več kot dve spremenljivki? Programi in računalniška grafika nam pomagajo prikazati in opisati zapletene zveze. Oglejmo si dva primera.

Primer. (Crash test motorja) Motor se zaleti v zid. Na srečo je voznik le lutka, ki ima v glavi vgrajeno napravo za merjenje pospeškov (sprememb hitrosti). Na sliki 2.11 je razsevni diagram pospeškov glave glede na čas v milisekundah. Pospeške merimo v večkratnikih gravitacijskega pospeška g. Motor se zidu približuje s konstantno hitrostjo (pospešek je blizu 0). Ko trešči v zid, lutkino glavo odnese naprej in jo silovito zavre (negativni pospešek doseže več kot 100 g), nato jo vrže nazaj (do 75 g), potem še malo niha in se ustavi.

Razsevni diagram ima jasno celostno podobo, vendar pa ne sledi preprostemu linearnemu pravilu. Še več, jasnost podobe variira, od precej močno opredeljene na levi do šibkejše (bolj razpršene) na desni. Statistični programi vključujejo *izglajevalec razsevnih diagramov*, ki odpravi to kompleksnost in nariše krivuljo, ki predstavlja splošno sliko. ♦

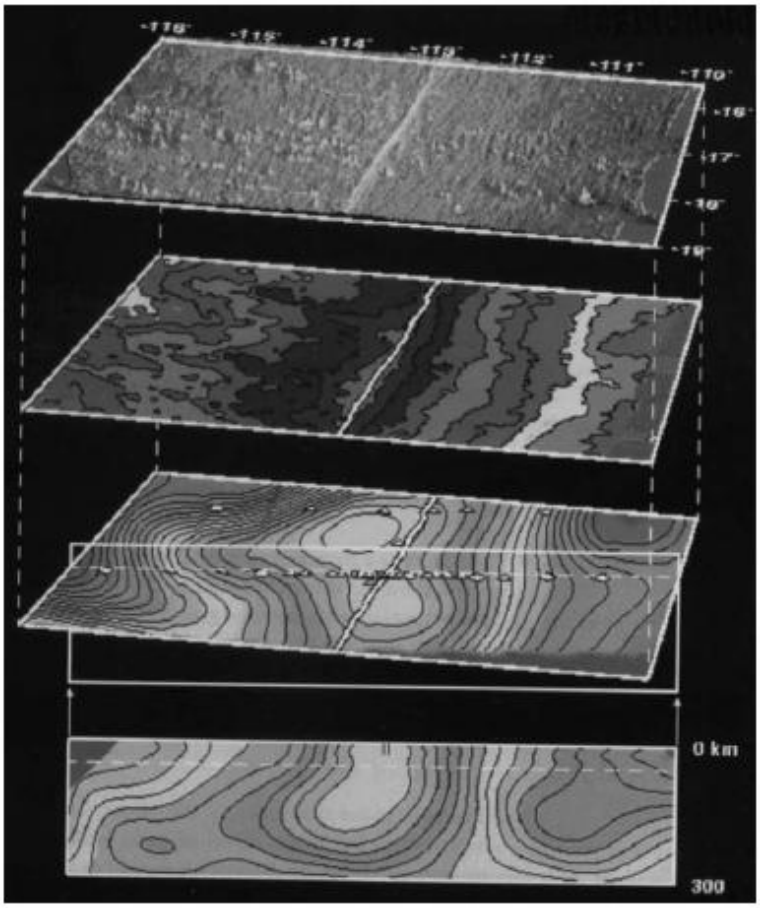


Slika 2.11: Odvisnost pospeška glave testne lutke od časa pri trku motorja z zidom. Krivulja je dobljena kot izglajevalec razsevnega diagrama.

Do zdaj smo si ogledali le diagrame z dvema spremenljivkama. Kaj pa, če želimo v istem diagramu prikazati še tretjo? Ker smo že porabili vodoravno in navpično os, nam preostane le še ena smer: pravokotno na ravnino lista. Tridimenzionalne grafe je težko jasno videti, razen če uporabimo barve ali gibanje (ali oboje), ki nam pomagajo prikazati perspektivo. Računalniška grafika lahko doda barve in gibanje, kar nam omogoča, da si ogledamo podatke za več spremenljivk naenkrat.

Primer. (Slike površja) Velike plošče, ki sestavljajo zemeljsko skorjo, lezejo narazen na grebenih sredi oceanov, kjer vroča magma (stopljene kamnine) privre iz globin. Znanstveniki proučujejo to širjenje morskega dna z združevanjem podatkov iz številnih virov, vključno z instrumenti, ki so nameščeni na morskem dnu dve milji pod gladino. Slika 2.12 je računalniška podoba podatkov iz študije v južnem Pacifiku.

Zgornja slika prikazuje topografijo oceanskega dna. Zemljepisna širina in dolžina označujeta položaj. Greben, ki ločuje dve plošči, poteka po sredini in na obeh straneh lahko opazimo male podvodne ognjenike. Drugi del prikazuje majhne variacije v gravitaciji, ki pomagajo ločevati med različnimi vrstami kamnin. Tretja slika doda podatke, ki jih dobimo s spremljanjem hitrosti potresnih valov, in prikazuje strukturo magme pod zemeljsko skorjo. Vsi trije diagrami so poravnani tako, da lahko znanstveniki vizualno primerjajo različne spremenljivke na različnih lokacijah za boljše razumevanje geoloških procesov, ki oblikujejo naš planet. ♦



Slika 2.12: Morsko dno v južnem Pacifiku blizu podvodnega grebena. Ta računalniška grafika prikazuje topografijo in meritve gravitacije in hitrosti potresnega vala, poleg tega pa še položaj na zemeljskem površju. (Vir: D. S. Scheirer, Brown University, *Science*, May 22, 1998)

2.13 Slovarček

asimetrična porazdelitev (ang. skewed distribution) Porazdelitev, pri kateri so na eni strani mediane vrednosti bistveno bolj oddaljene od mediane kot na drugi.

histogram (ang. histogram) Graf porazdelitve izidov (večkrat razdeljen v nekaj razredov) neke spremenljivke; višina vsakega stolpca je število opažanj, ki padejo v meje, določene z bazo tega stolpca; vsi stolpci naj bi bili enako široki.

korelacija (ang. correlation) Mera za smer in moč linearne zveze med dvema spre-

menljivkama; zavzame vrednosti med 0 (nobene linearne povezave) in ±1 (popolna linearna povezava).

kvartili (ang. quartiles) Prvi kvartil porazdelitve je točka, pod katero je 25% opaženih vrednosti, tretji kvartil je točka, pod katero je 75% vrednosti.

mediana (ang. median) Sredinska točka množice vrednosti; polovica vrednosti je manjših, polovica pa večjih od mediane.

odzivna in obrazložitvena spremenljivka (ang. response variable, explanatory variable) Odzivna spremenljivka meri izide študije, obrazložitvena služi pojasnjevanju opaženih izidov.

porazdelitev (ang. distribution) Slika izidov neke spremenljivke; porazdelitev opiše, katere vrednosti spremenljivka zavzame in kako pogosto se vsaka od vrednosti pojavi.

posamezniki (ang. individuals) Ljudje, živali ali stvari, ki jih opisujejo dani podatki.

povprečje ali srednja vrednost (ang. mean) Običajna aritmetična sredina; vsota vseh vrednosti, deljena s številom vrednosti.

povzetek s petimi števili (ang. five-number summary) Osnovni podatki o porazdelitvi vrednosti spremenljivke; sestavljajo ga mediana, prvi in tretji kvartil ter najmanjša in največja opažena vrednost.

razsevni diagram (ang. scatterplot) Graf vrednosti dveh spremenljivk kot množica točk v ravnini; na vodoravni koordinatni osi je obrazložitvena spremenljivka, na navpični pa odzivna.

regresijska premica (ang. regression line) Vsaka premica, ki opisuje, kako se odzivna spremenljivka y spreminja, ko spreminjamo obrazložitveno spremenljivko x; npr. premica po metodi najmanjših kvadratov.

regresijska premica najmanjših kvadratov (ang. least square regression line) Premica na razsevnem diagramu, za katero je vsota kvadratov navpičnih razdalj do točk, ki predstavljajo podatke, najmanjša; uporabimo jo lahko, da predvidimo vrednost odzivne spremenljivke y pri dani vrednosti obrazložitvene spremenljivke x.

simetrična porazdelitev (ang. symmetric distribution) Porazdelitev, katere histogram je približno zrcalno simetričen glede na mediano.

splošna analiza podatkov (ang. exploratory data analysis) Postopek pregleda podatkov v iskanju nepričakovanih vzorcev ali vplivov, v nasprotju z iskanjem odgovorov na specifična vprašanja.

spremenljivka (ang. variable) Vsaka izmerjena lastnost posameznika.

standardni odklon (ang. standard deviation) Mera za razpršenost porazdelitve okoli povprečja; kvadratni koren povprečja kvadratov razlik med podatki in povprečjem.

stebelni diagram (ang. stemplot) Prikaz porazdelitve spremenljivke, ki zadnje števke podatkov dodaja v ustrezne vrstice, sestavljene iz vseh ostalih števk.

škatla z brki (ang. boxplot) Graf, ki prikazuje povzetek s petimi števili; škatla predstavlja območje med obema kvartiloma, notranja črta označuje mediano; dve črti, ki segata iz škatle, se raztezata vse do minimalne in maksimalne izmerjene vrednosti.

ubežnik (ang. outlier) Točka, ki pade daleč izven splošnega vzorca v skupini podatkov.

varianca (ang. variance) Mera razpršenosti porazdelitve okoli povprečja; povprečje kvadratov razlik med podatki in povprečjem; kvadratni koren variance je standardni odklon.

2.14 Dodatna literatura

- Cleveland, William S. *The Elements of Graphing Data*, Wadsworth, Monterey, Calif., 1985. Podrobna študija najbolj učinkovitih elementarnih načinov grafičnih predstavitev podatkov, z veliko nasveti kako izboljšati preproste grafe.
- Moore, David S. *The Basic Practice of Statistics*, 2. izdaja, W. H. Freeman, New York, 1999. Prvi dve poglavji tega besedila vključujeta bolj podrobno obravnavo prikazovanja in opisovanja podatkov za eno ali dve spremenljivki. Snov tega poglavja je obravnavana bolj nadrobno, predstavljenih je veliko novih metod in interpretacij.

- Rossman, Allan J. *Workshop Statistics: Discovery with Data*, Springer-Verlag, New York, 1996. Čudovit vir praktičnih nalog, ki se osredotoča na opisovanje podatkov.

- Tufte, Edward R. *The Visual Display of Quantative Information*, Graphics Press, Cheshire, Conn.,1983. Lepo natisnjena knjiga z zgodovinskimi in sodobnimi grafi in predlogi za statistike in vizualne umetnike.

- Velleman, Paul F., David C. Hoaglin. Data analysis, *Perspectives on Contemporary Statistics*, Mathematical Association of America, Washington, D.C., 1992, str. 19–39. Esej, ki predpostavlja poznavanje osnovnih metod, opisanih v tem poglavju in Moorovi knjigi.

Spletne strani ne razlagajo postopkov izdelave histogramov ali razsevnih diagramov. Ponujajo pa veliko zanimivih dejanskih podatkov. Knjižnica podatkov

- *Data and Story Library*,
  `lib.stat.cmu.edu/DASL/`,

ima veliko podatkov in informacij, ki jih potrebujemo za njihovo uporabo. Slučajni splet,

- *Chance Web*,
  `www.dartmouth.edu/~chance/`,

objavlja podatke iz aktualnih novic, dostopen pa je tudi arhiv. Spletna revija *Journal of Statistics Education* ima članke o poučevanju statistike in arhiv podatkov. Najdemo jo na strani ameriškega statističnega združenja `www.amstat.org`.

2.15 Preverjanje znanja

(1) Tvoji bratranci so visoki 101, 91, 46, 96, 76 in 86 cm. Kateri so ubežniki?

(a) Le 101.

(b) Le 46.

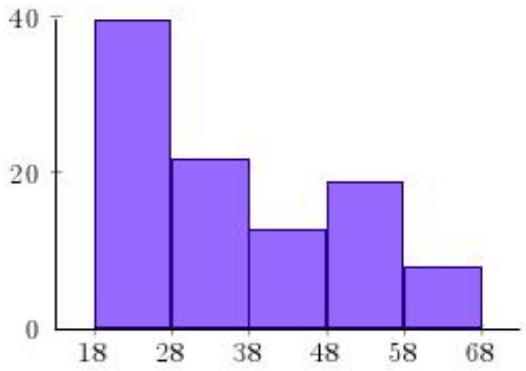
(c) Tako 101 kot 46.

(2) Spodaj je histogram, ki prikazuje starosti odraslih z neke zabave. Katera od trditev je pravilna?

(a) Histogram je približno simetričen.

(b) Histogram je desno asimetričen.

(c) Razred med 58 in 68 vsebuje 8 ubežnikov.



(3) Tukaj je sedem izmerjenih vrednosti: 4, 7, 5, 6, 5, 11, 4. Poišči mediano.

(a) 5

(b) 6

(c) 5,5

(4) Tukaj je sedem izmerjenih vrednosti: 4, 7, 5, 6, 5, 11, 4. Poišči povprečje.

(a) 5

(b) 6

(c) 6,6

(5) Povzetek s petimi števili vključuje

(a) povprečje in standardni odklon.

(b) mediano in povprečje.

(c) kvartile.

(6) Povprečje vrednosti 4, 5, 5, 7, 6, 6, 9 je 6. Koliko je standardni odklon?

(a) 2,67

(b) 1,63

(c) 1,51

(7) Dnevna poraba ledu y (v funtih) v zabaviščnem parku je povezana z maksimalno temperaturo x (v °F). Recimo, da je enačba regresijske premice najmanjših kvadratov $y = 50 + 20x$. Napovej porabo ledu za dan, ko je maksimalna temperatura 70°F.

(a) 1 funt

(b) 190 funtov

(c) 1450 funtov

2.16 Naloge

Veliko nalog zahteva uporabo kalkulatorja (ali programov), ki zna iz vnešenih podakov poiskati povprečje, standardni odklon, korelacijo, naklon in začetno vrednost regresijske premice najmanjših kvadratov.

Prikaz porazdelitev

(1) Spodaj je del podatkov, ki opisujejo porabo goriva za vozila letnika 1998 v miljah na galono. Kaj so pri teh podatkih posamezniki in kaj spremenljivke?

Znamka in model	Menjalnik	Število cilindrov	Mestna poraba	Zunajmestna poraba
⋮				
BMW 318I	Avtomatski	4	22	31
BMW 318I	Ročni	4	23	32
Buick Century	Avtomatski	6	20	29
Chevrolet Blazer	Avtomatski	6	16	20
⋮				

(2) Okoljevarstvena agencija zahteva, da proizvajalci avtomobilov za vsako vozilo navedejo porabo pri mestni vožnji in na daljših relacijah. V tabeli 2.4 so navedeni podatki o porabi na daljših relacijah (v miljah na galono) za 26 srednje velikih avtomobilov letnika 1998.

(a) Nariši histogram porabe za te avtomobile.

(b) Opiši glavne značilnosti (obliko, središče, razpon, ubežnike) porazdelitve porabe.

(c) Vlada uvede posebni davek na "požrešne" avtomobile. Kateri od navedenih po tvojem mnenju sodijo v to skupino?

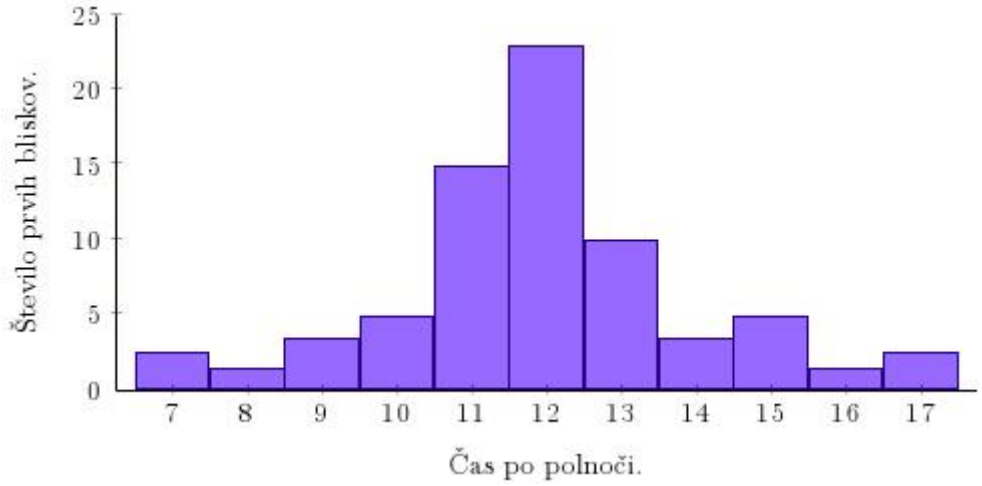
Model	Poraba	Model	Poraba
Acura 3.5RL	25	Lexus GS300	23
Audi A6 Quattro	26	Lexus LS400	25
Buick Century	29	Lincoln Mark VIII	26
Cadillac Catera	24	Mazda 626	33
Cadillac Eldorado	26	Mercedes-Benz E320	29
Chevrolet Lumina	29	Mercedes-Benz E420	26
Chrysler Cirrus	30	Mitsubishi Diamante	24
Dodge Stratus	28	Nissan Maxima	28
Ford Taurus	28	Oldsmobile Aurora	26
Honda Accord	29	Rolls-Royce Silver Spur	16
Hyundai Sonata	27	Saab 900S	25
Infiniti I30	28	Toyota Camry	25
Infiniti Q45	23	Volvo S70	25

Tabela 2.4: Poraba goriva na daljših relacijah za srednje velike avtomobile letnika 1998.

(3) Histogram na sliki 2.13 prikazuje podatke o urah, ob katerih so na posamezen dan prvič opazili bliske med neko študijo v Koloradu. Opiši to porazdelitev: Ali je približno simetrična ali izrazito asimetrična? Kje je središče? So prisotni ubežniki ali vrzeli?

(4) Skupni donos delnice je sprememba v tržni vrednosti plus morebitne dividende. Običajno skupni dobiček izrazimo kot odstotek začetne vrednosti. Na sliki 2.14 je histogram porazdelitve skupnih donosov za vseh 1528 delnic, s katerimi so poslovali na newyorški borzi v enem letu. Kot na sliki 2.3 gre za histogram odstotkov v vsakem razredu, ne pa števila delnic. (Vir: J. K. Ford, Diversification: How many stocks will suffice? *American Association of Individual Investors Journal* (Januar 1990): 14–16.)

(a) Opiši splošno obliko porazdelitve skupnih donosov.

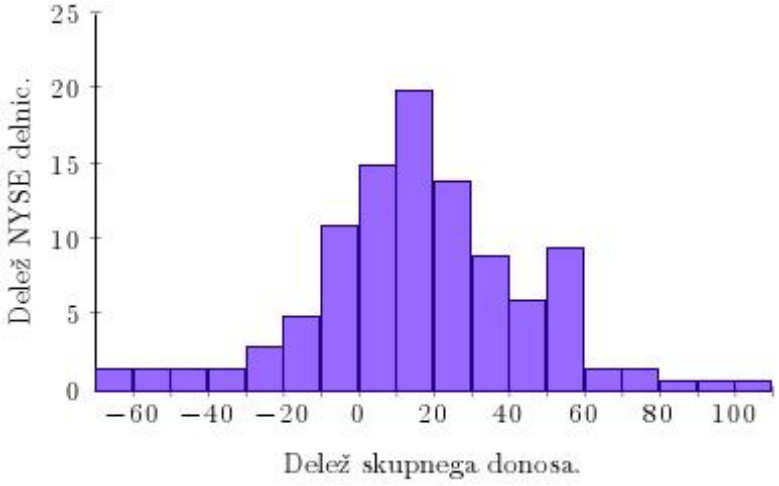


Slika 2.13: Razporeditev časov prvih pojavov bliskov za vsak dan opazovanj v zvezni državi Kolorado.

(b) Koliko približno je mediana te porazdelitve? (Se pravi, za katero vrednost velja, da je približno polovica donosov pod in polovica nad njo?)

(c) Približno kolikšna sta bila največji in najmanjši skupni donos? (To nam pove, kakšen je razpon porazdelitve.)

(d) Negativni skupni donos pomeni, da je lastnik delnice izgubil denar. Kolikšen je odstotek delnic, ki so prinesle izgube?



Slika 2.14: Razporeditev deleža skupnih donosov za vse standardne delnice newyorške borze v enem letu.

(5) Leta 1798 je angleški znanstvenik Henry Cavendish izmeril gostoto Zemlje v natančnem eksperimentu s torzijsko tehtnico. Spodaj je njegovih 29 zaporednih meritev iste količine (gostote Zemlje v primerjavi z gostoto vode), opravljenih z istim instrumentom. (Vir: S. M. Stigler, Do robust estimators work with real data? *Annals of Statistic*, 5(1977): 1055–1078.)

5,50	5,47	5,29	5,55	5,75	5,27
5,57	4,88	5,34	5,34	5,29	5,85
5,42	5,62	5,26	5,30	5,10	5,65
5,61	5,07	5,46	5,79	5,58	

(a) Izdelaj stebelni diagram.

(b) Opiši porazdelitev: Ali je približno simetrična ali izrazito asimetrična? Imamo vrzeli ali ubežnike?

(6) Raziskovalec ribištva je zbral naslednje podatke o dolžinah šestletnih samic belih krapov (v milimetrih):

217	230	220	221	225	223
219	217	225	228	234	222
231	222	220	222	222	223
225	214	221	233	227	234
223	225	253	220	213	224
235	283	210	218	235	231

(a) Izdelaj stebelni diagram.

(b) Izdelaj še histogram. Podatki ležijo med 210 in 283 mm. Razdeli jih v 5 razredov širine 15 mm, začenši z

$$210 \leq \text{dolžina} < 285.$$

(c) Opiši porazdelitev: Ali je približno simetrična ali izrazito asimetrična? Imamo vrzeli ali ubežnike?

Opisovanje porazdelitev

(7) Tabela 2.3 podaja puste telesne teže in stopnje metabolizma za 7 moških in 12 žensk. Primerjaj porazdelitvi telesne teže pri moških in pri ženskah s pomočjo povzetkov s petimi števili in z vzporednima škatlama z brki. Kaj pokažejo podatki?

(8) Vrnimo se k podatkom o porabi iz tabele 2.4.

(a) Zapiši povzetek s petimi števili.

(b) Pomagaj si s kalkulatorjem, da najdeš povprečje in standardni odklon.

(c) Odstrani Rolls-Roycea in ponovi izračune. Kateri od rezultatov se spremenijo in za koliko? Katero splošno dejstvo ilustrira ta primer?

(9) Tule so deleži glasov, ki so jih prejeli zmagoviti kandidati na volitvah med leti 1948 in 1996:

Leto	1948	1952	1956	1960	1964
Delež	49,6	55,1	57,4	49,7	61,1
Leto	1968	1972	1976	1980	1984
Delež	43,4	60,7	50,1	50,7	58,8
Leto	1988	1992	1996		
Delež	53,9	43,2	49,2		

(a) Napravi histogram.

(b) Kolikšna je mediana deležev glasov?

(c) Volitvam pravimo *plaz*, če je delež glasov, ki jih je dobil zmagovalec, nad tretjim kvartilom. Poišči tretji kvartil. Katere od volitev so bile plazovi?

(10) Nivo različnih substanc v krvi vpliva na naše zdravje. Spodaj so meritve nivoja fosfatov v krvi pacienta (v miligramih fosfata na deciliter krvi), ki so bile opravljene pri šestih zaporednih obiskih klinike:

5,6 5,2 4,6 4,9 5,7 6,4

Graf pri samo šestih podatkih ni zelo informativen, zato raje izračunamo povprečje in standardni odklon.

(a) Poišči povprečje po definiciji. Se pravi, seštej vseh šest podatkov in vsoto deli s 6.

(b) Poišči standardni odklon po definiciji. Za vsako vrednost izračunaj njen odklon od povprečja, jih kvadriraj in od tod izračunaj varianco in standardni odklon.

(c) Vnesi podatke v kalkulator in uporabi vgrajena programa za računanje povprečja in standardnega odklona, da dobiš $\overline{x}$ in s. Ali se rezultata ujemata s tvojimi izračuni?

(11) Nekateri ljudje pazijo na količino zaužitih kalorij. Revija *Consumer Reports* (Julij 1986, str. 366–367) je izmerila kalorije v 20 znamkah govejih hrenovk, 17 znamkah mesnih hrenovk in 17 znamkah piščančjih hrenovk. Tole so računalniški izpisi za vsako od treh vrst:

```
Povprečje = 156,8  Standardni odklon = 22,64
Min = 111  Max = 190  N=20
Mediana = 152,5  Kvartila = 140, 178,5

Povprečje = 158,7  Standardni odklon = 25,24
Min = 107  Max = 195  N=17
Mediana = 153  Kvartila = 139, 179

Povprečje = 122,5  Standardni odklon = 25,48
Min = 87  Max = 170  N=17
Mediana = 129  Kvartila = 102, 143
```

Uporabi te informacije, da narišeš vzporedne škatle z brki, ki bodo predstavljale količine kalorij v treh vrstah hrenovk. Na kratko primerjaj te porazdelitve. Ali uživanje perutninskih hrenovk zmanjša količino zaužitih kalorij v primerjavi z mesnimi ali govejimi hrenovkami?

(12) Ponovno si oglejmo podatke o dolžinah rib iz naloge 6.

(a) Poišči povzetek s petimi števili za to porazdelitev. Katere od dolžin se nahajajo v sredinskih 50% te porazdelitve?

(b) Ali po obliki porazdelitve pričakuješ, da je povprečje manjše od mediane, večje ali približno enako veliko? Izračunaj to povprečje in preveri svojo ugotovitev.

(c) Poišči standardni odklon. Na podlagi oblike porazdelitve sklepaj, ali sta $\overline{x}$ in s sprejemljivi meri za središče in razpon.

(13) Izdelaj povzetek s petimi števili za Cavendishove meritve gostote Zemlje iz naloge 5. Kako se simetrija porazdelitve kaže v tem povzetku?

(14) Povprečje 29 meritev iz naloge 5 je bila Cavendisheva najboljša ocena za gostoto Zemlje. Izračunaj to povprečje. Nato poišči še standardni odklon. (Zaradi simetrije lahko porazdelitev povzamemo z $\overline{x}$ in s.)

(15) Porazdelitev osebnih dohodkov v ZDA je močno desno asimetrična. Leta 1997 sta bila povprečje in mediana dohodkov zgornjega 1% Američanov 330 000 \$ in 675 000 \$. Katera od teh vrednosti je povprečje in katera mediana? Odgovor utemelji.

(16) Časopisni članek poroča, da je od 411 igralcev v registru državne košarkaške lige v februarju leta 1998 le 139 igralcev zaslužilo več od povprečne plače v ligi, ki je bila 2,36 milijona dolarjev. Ali je ta vrednost povprečje ali mediana višine plač igralcev? Kako to veš?

(17) Rezultati odraslih na Stanford-Binetovem inteligenčnem testu imajo povprečje 100 in standardni odklon 15. Kolikšna je varianca?

(18) Tole so podatki o številu "home runov", ki jih je dosegel Babe Ruth v svojih 15 letih pri ekipi *New York Yankees* med leti 1920 in 1934:

54	59	35	41	46	25	47	60
54	46	49	46	41	34	22	

Trenutni rekord števila "home runov" v eni sezoni velike lige pripada Marku McGuireu. Tole so rezultati, ki jih je Mark McGuire dosegel med letoma 1987 in 1998:

49	32	33	39	22	42	9	9
39	52	58	70				

Dvojni stebelni diagram nam pomaga pri primerjavi dveh porazdelitev. Stebla napišemo kot običajno, vendar tokrat narišemo eno navpično črto na levi in eno na desni. Na desni strani potem zapišemo liste, ki pripadajo Ruthu, na levi pa tiste, ki pripadajo McGuireu. Liste na vsakem steblu uredimo tako, da naraščajo od stebla navzven. Izdelaj dvojni stebelni diagram in na kratko primerjaj rezultate obeh igralcev. McGuire se je leta 1993 poškodoval, leta 1994 pa so igralci stavkali. Kako se ta dogodka odražata v rezultatih?

Prikaz zvez

(19) Morske krave so velika, krotka morska bitja, ki živijo ob obali Floride. Veliko jih ubijejo ali poškodujejo hitri motorni čolni. V tabeli 2.5 so podatki o številu registriranih čolnov (v tisočih) in številu morskih krav, ki so jih čolni ubili med leti 1977 in 1990.

Leto	Št. čolnov (v tisočih)	Št. ubitih morskih krav	Leto	Št. čolnov (v tisočih)	Št. ubitih morskih krav
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

Tabela 2.5: Smrti zaradi motornih čolnov za Florido, 1977-1990.

(a) Želimo raziskati zvezo med številom motornih čolnov in številom ubitih morskih krav. Katera od spremenljivk je obrazložitvena?

(b) Nariši razsevni diagram. Opiši smer, obliko in moč zveze. Ali opaziš kakšne ubežnike ali druge pomembne nepravilnosti?

(20) Kako se spreminja poraba goriva, ko se povečuje hitrost? V spodnji tabeli so zbrani podatki za britanski Ford Escort. Hitrost je merjena v kilometrih na uro, poraba pa v litrih na 100 kilometrov. (Vir: T. N. Lam, Estimating fuel consumption from engine size, *Journal of Transportation Engineering*, 111(1985): 339–357.)

(a) Nariši razsevni diagram. (Katera od spremenljivk je obrazložitvena?)

(b) Opiši vrsto zveze med spremenljivkama. Razloži, zakaj je smiselna.

(c) Kako bi opisal(a) smer te zveze?

(d) Ali je zveza razumno močna ali precej šibka? Odgovor utemelji.

Hitrost (km/h)	Poraba (l/100km)	Hitrost (km/h)	Poraba (l/100km)
10	21,00	90	7,57
20	13,00	100	8,27
30	10,00	110	9,03
40	8,00	120	9,87
50	7,00	130	10,79
60	5,90	140	11,77
70	6,30	150	12,83
80	6,95		

Regresijske premice

(21) Raziskovalci, ki proučujejo kisli dež, so izmerili kislost padavin v divjini Kolorada preko 150 zaporednih tednov. Kislost merijo v pH. Nižja pH vrednost pomeni večjo kislost. Raziskovalci so s časom opazili linearni vzorec. Poročali so, da se podatkom dobro prilega regresijska premica najmanjših kvadratov z enačbo

$$\text{pH} = 5,43 - (0,0053 \cdot \text{št. tednov}).$$

(Vir: W. M. Lewis in M. C. Grant, Acid precipitation in the western United States, *Science*, 207(1980): 176–177.)

(a) Nariši graf te premice. Razloži na preprost način, kaj nam premica pove o spreminjanju pH skozi čas.

(b) Iz premice razberi, kolikšna je bila vrednost pH na začetku opazovanj (tedni = 1) in na koncu (tedni = 150).

(c) Kolikšen je naklon regresijske premice? Pojasni, kaj nam ta naklon pove o hitrosti spreminjanja pH.

(22) Nadaljujmo z analizo podatkov o morskih kravah iz tabele 2.5. Tole so podatki za nadaljna štiri leta:

1991	716	53	1993	716	35
1992	716	38	1994	735	49

(a) Začni z razsevnim diagramov iz naloge 19. Približno koliko morskih krav bi bilo ubitih vsako leto, če bi se Florida odločila, da zamrzne število registracij motornih čolnov pri 716 000? V diagram približno vriši premico, ki bo dala napoved.

(b) Dodaj nove podatke v svoj diagram. Izkazalo se je, da je število registracij res obstalo na 716 000 za tri leta. Kako natančna je bila tvoja napoved?

(23) Recimo, da bi v daljni prihodnosti število registriranih čolnov na Floridi doseglo dva milijona. Podaljšaj premico iz prejšnje naloge in jo uporabi za napoved števila ubitih morskih krav. Pojasni, zakaj je ta napoved zelo nezanesljiva. (Uporabo premice, ki se prilega podatkom, za napoved odziva pri vrednosti spremenljivke x, ki leži zunaj območja, na katerega se nanašajo podatki, imenujemo *ekstrapolacija*. Napovedi, dobljene z ekstrapolacijo, so velikokrat nezanesljive.)

(24) Asfaltno cestišče se po izdelavi začne sušiti in s časom pridobiva trdnost. Inženirji uporabljajo regresijske premice, da predvidijo, kakšna bo trdnost po 28 dneh (ko bo sušenje končano) na podlagi meritev, ki jih izvedejo po 7 dneh. Naj bo x moč (v funtih na kvadratno inčo) po 7 dneh in y moč po 28 dneh. Iz enega dela meritev so ugotovili, da je enačba regresijske premice najmanjših kvadratov enaka

$$y = 1389 + 0,96x.$$

(a) Z besedami razloži, kaj nam pove naklon $0,96$ o sušenju asfalta.

(b) Neka nova merjenja po 7 dneh pokažejo, da je moč 3300 funtov na kvadratno inčo. Napovej moč tega materiala po 28 dneh.

Korelacija in regresija najmanjših kvadratov

(25) V nalogi 20 so podatki o porabi goriva v odvisnosti od hitrosti za manjši avto. Izračunaj korelacijo (pomagaj si s kalkulatorjem ali računalnikom). Pojasni, zakaj je r majhen, čeprav sta poraba in hitrost močno povezani.

(26) Poišči enačbo regresijske premice najmanjših kvadratov za podatke o morskih kravah iz tabele 2.5. S pomočjo enačbe napovej število smrti za leto, v katerem bo na Floridi registriranih 716 000 motornih čolnov. Primerjaj to napoved s svojo oceno iz naloge 22.

(27) Recimo, da bi se ženske vedno poročile z moškimi, ki so dve leti starejši od njih. Kolikšna bi bila v tem primeru korelacija med starostjo moža in žene? (Namig: Nariši razsevni diagram za različne starosti.)

(28) *Archaeopteryx* je izumrla zver, ki je imela perje kot ptice ter zobovje in dolgi rep kot plazilci. Znanih je le šest primerkov fosilov. Ker se ti primerki zelo razlikujejo v velikosti, so nekateri znanstveniki mnenja, da gre za različne vrste in ne za posamezne pripadnike iste vrste. Če fosili pripadajo isti vrsti in se razlikujejo v velikosti le zato, ker so eni mlajši od drugih, bi morala obstajati linearna zveza med dolžinami nekega para kosti za vse primerke. Ubežniki bi v tem primeru sugerirali, da gre za drugo vrsto. V naslednji tabeli so podatki o dolžini (v cm) kosti, imenovane *femur* (gre za eno od kosti noge) in dolžini kosti, imenovane *humerus* (kost zgornjega dela roke), za pet fosilov, pri katerih sta bili obe kosti ohranjeni. (Vir: M. A. Houck et al., Allometric scaling in the earliest fossil bird, Archaeopteryx lithographica, Science, 247(1990): 195–198.)

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

(a) Nariši razsevni diagram. Ali meniš, da pripada vseh pet primerkov isti vrsti?

(b) Po definiciji izračunaj korelacijo r. Se pravi, poišči povprečje in standardni odklon dolžin femurjev in dolžin humerusov. (Za računanje povprečij in standardnih odklonov uporabi kalkulator.) Nato izračunaj odklone od povprečja in uporabi formulo za r.

(c) Vnesi te podatke v kalkulator in uporabi vgrajeno funkcijo za izračun r. Prepričaj se, da dobiš isti rezultat kot v točki (b).

(29) Prehrambena industrija je prosila skupino 3368 ljudi, da ocenijo število kalorij v večjem številu pogostih vrst hrane. Tabela 2.6 prikazuje povprečja njihovih ocen in dejanska števila kalorij.

(a) Menimo, da število kalorij, ki jih določena vrsta hrane dejansko vsebuje, pomaga razložiti vrednosti, ki jih ljudje ugibajo. S tem v mislih nariši razsevni diagram za dane podatke.

(b) Poišči korelacijo r (pomagaj si s kalkulatorjem). S pomočjo razsevnega diagrama pojasni, zakaj je dobljeni r smiseln.

(c) Vsa ugibanja so večja od dejanskih vrednosti. Ali to dejstvo kakorkoli vpliva na korelacijo? Kako bi se r spremenil, če bi bile vse ocene za 100 kalorij višje?

(d) Ocene so veliko previsoke v primeru špagetov in tortice. Obkroži ustrezni točki na svojem razsevnem diagramu. Izračunaj r za ostalih osem vrst hrane, ti dve pa izpusti. Pojasni, zakaj se je r spremenil tako, kot se je.

Hrana	Uganjene kalorije	Dejanske kalorije
Polnomastno mleko	196	159
Špageti s paradižnikovo omako	394	163
Makaroni s sirom	350	269
Rezina pšeničnega kruha	117	61
Rezina belega kruha	136	76
Čokoladna rezina	364	260
Slani krekerji	74	12
Srednje veliko jabolko	107	80
Srednje velik krompir	160	88
Tortica s smetano	419	160

Tabela 2.6: Ocenjene in dejanske kalorije v desetih vrstah hrane.

(30) Nadaljujmo z analizo podatkov iz tabele 2.6.

(a) Pomagaj si s kalkulatorjem, da poiščeš regresijsko premico najmanjših kvadratov za primer uganjenih kalorij v odvisnosti od dejanskih kalorij. To naredi dvakrat, najprej za vseh 10 podatkov, nato pa še tako, da izpustiš špagete in tortico.

(b) Nariši obe premici na razsevni diagram iz prejšnje naloge. (Ena črta naj bo narisana črtkano, da ju lahko razločimo.) Ali ubežnika bistveno spremenita premico?

(31) Poišči enačbo regresijske premice najmanjših kvadratov za podatke o porabi goriva v odvisnosti od hitrosti iz naloge 20. Nariši razsevni diagram in vanj vriši še to premico. To je premica, ki se najbolje prilega podatkom (v smislu najmanjših kvadratov), vendar pa je ne bi uporabili za napovedovanje.

(32) Enačba regresijske premice za porabo plina y v odvisnosti od stopinjskih dni x se je glasila

$$y = 3,48 + 0,57x.$$

Vnesi podatke iz tabele 2.2 v svoj kalkulator.

(a) Uporabi funkcijo za računanje regresije na svojem kalkulatorju, da poiščeš enačbo regresijske premice najmanjših kvadratov.

(b) S kalkulatorjem poišči povprečji in standardna odklona spremenljivk x in y ter njuno korelacijo r. S pomočjo teh poišči naklon regresijske premice b in začetno vrednost a tako, da uporabiš ustrezno enačbo za regresijsko premico. Prepričaj se, da z (a) in (b) dobiš res premici iz primera. (Rezultati se lahko malo razlikujejo zaradi zaokroževanja.)

(33) Močna zveza med dvema spremenljivka *ne* pomeni vedno, da ena od spremenljivk povzroča spremembe druge. Nekdo ugotovi, "Obstaja močna pozitivna korelacija med številom gasilcev, ki gasijo požar, in škodo, ki jo ta požar povzroči. Če torej h gašenju pokličemo več gasilcev, bo škoda samo še večja." Pojasni, zakaj je takšno sklepanje napačno.

(34) Močna zveza med dvema spremenljivka *ne* pomeni vedno, da ena od spremenljivk povzroča spremembe druge. Raziskave kažejo, da obstaja pozitivna korelacija med velikostjo bolnišnice (merjeno s številom postelj x) in mediano števila dni y, ki jih pacienti preživijo v bolnišnici. Ali to pomeni, da si lahko skrajšamo število bolnišničnih dni, če za zdravljenje izberemo manjšo bolnišnico? Razloži.

(35) Sprememba merskih enot lahko močno spremeni izgled razsevnega diagrama. Oglej si naslednje podatke:

x	-4	-4	-3	3	4	4
y	0,5	-0,6	-0,5	0,5	0,5	-0,6

(a) Nariši osi x in y tako, da obe segata od -6 do 6. Vriši podatke na to sliko.

(b) Izračunaj vrednosti novih spremenljivk $x' = \frac{x}{10}$ in $y' = 10y$. Na isto sliko nariši še y' v odvisnosti od x' tako, da uporabiš za te točke drugačne simbole. Dobljena diagrama se po videzu zelo razlikujeta.

(c) Uporabi kalkulator, da izračunaš korelacijo med x in y. Nato izračunaj še korelacijo med x' in y'. Kako sta obe korelaciji povezani? Razloži, zakaj to ni presenetljivo.

(36) S pomočjo enačbe za regresijsko premico najmanjših kvadratov pokaži, da ta premica vedno poteka skozi točko $(\overline{x}, \overline{y})$. Se pravi, postavi $x = \overline{x}$ in pokaži, da iz enačbe za premico dobiš napoved $y = \overline{y}$.

Dodatne naloge

(37) V tabeli 2.7 so zbrani podatki o številu prebivalstva za države ZDA (v tisočih). Izdelaj stebelni diagram ali histogram s temi podatki. Na kratko opiši obliko, središče in razpon porazdelitve prebivalstva. Pri tem pazi, da podaš primerne številske vrednosti središča in razpona. Razloži, zakaj oblika porazdelitve ni presenetljiva. Ali se ti zdi, da so katere od držav ubežniki?

Država	Pop.	Država	Pop.	Država	Pop.	Država	Pop.
AL	4,273	IL	11,847	MT	879	RI	990
AK	607	IN	5,841	NE	1,652	SC	3,699
AZ	4,428	IA	2,852	NV	1,603	SD	732
AR	2,510	KS	2,572	NH	1,162	TN	5,320
CA	31,878	KY	3,884	NJ	7,988	TX	19,128
CO	3,823	LA	4,351	NM	1,713	UT	2,000
CT	3,274	ME	1,243	NY	18,185	VT	589
DE	725	MD	5,072	NC	7,323	VA	6,675
DC	543	MA	6,092	ND	644	WA	5,533
FL	14,400	MI	9,594	OH	11,173	WV	1,826
GA	7,353	MN	4,658	OK	3,301	WI	5,160
HI	1,184	MS	2,716	OR	3,204	WY	481
ID	1,189	MO	5,359	PA	12,056		

Tabela 2.7: Prebivalstvo ZDA (v tisočih).

(38) Zunanji igralec kluba New York Yankees Roger Maris je leta 1961 podrl rekord, ki ga je postavil Babe Ruth, in držal novi rekord vse do leta 1998, ko je Mark McGuire dosegel 70 "home runov". Tole so podatki o "home runih", ki jih je zadel Maris v svojih 10 letih v ameriški ligi:

14 28 16 39 61 33 23 26 8 13

Marisovih rekordnih 61 "home runov" je v tem primeru ubežnik.

(a) S pomočjo kalkulatorja poišči povprečje $\overline{x}$ in standardni odklon s.

(b) S pomočjo kalkulatorja izračunaj $\overline{x}$ in s za devet podatkov, ki ostanejo, ko odstraniš ubežnika. Kako to vpliva na vrednosti $\overline{x}$ in s?

(39) Običajni kriterij za iskanje potencialnih ubežnikov v množici podatkov izgleda takole:

I. Poišči kvartila Q_1 in Q_3 in *medkvartilni obseg*, $\text{IQR} = Q_3 - Q_1$. Medkvartilni obseg je razpon srednje polovice podatkov.

II. Neka vrednost je ubežnik, če leži več kot $1,5$ IQR nad tretjim ali pod prvim kvartilom.

Ali je po tem kriteriju Rolls-Royce iz tabele 2.4 potencialni ubežnik? Ali sta Aljaska in Florida iz tabele 2.1 ubežnika?

(40) Tabela 2.8 vsebuje podatke o času preživetja (v dneh) za 72 morskih prašičkov, ki so jih pri neki medicinski raziskavi okužili s *tubercle bacilli*[3]. Napravi histogram za te podatke. Ali je porazdelitev časa preživetja približno simetrična ali izrazito asimetrična? Ali bi bilo glede na obliko bolje uporabiti povzetek s petimi števili ali $\overline{x}$ in s za numerični opis porazdelitve? Izbrani opis tudi izračunaj.

43	45	53	56	56	57	58	66	67	73
74	79	80	80	81	81	81	82	83	83
84	88	89	91	91	92	92	97	99	99
100	100	101	102	102	102	103	104	107	108
109	113	114	118	121	123	126	128	137	138
139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511
522	598								

Tabela 2.8: Čas preživetja morskih prašičkov (v dneh).

(41) Poišči povprečje in mediano podatkov o preživetju morskih prašičkov iz tabele 2.8. S pomočjo splošne oblike porazdelitve pojasni zvezo med tema dvema merama središča.

[3]Povzročitelj tuberkuloze. (Op. prev.)

(42) Izbrati moraš štiri cela števila med 0 in 10, pri čemer lahko kakšno od števil izbereš večkrat.

(a) Izberi ta štiri števila tako, da bo standardni odklon kar najmanjši.

(b) Izberi štiri števila tako, da bo standardni odklon čim večji.

(c) Ali imaš pri (a) oz. (b) več možnih izbir? Pojasni.

(43) Poišči kakšno množico zanimivih podatkov na statističnem uradu ali v kakšnem poročilu (na primer stopnjo osipa v šolah ali pa bruto domači proizvod po državah). Izdelaj histogram teh podatkov ter opiši porazdelitev in morebitne ubežnike. Dodaj še numerični povzetek podatkov.

(44) Ameriški kolidži objavijo "povprečne" rezultate sprejemnih izpitov (SAT) bodočih brucev. Običajno kolidži želijo, da bi bilo to "povprečje" karseda visoko. V nekem članku v *New York Timesu* so ugotavljali, da "nekateri privatni kolidži, ki kupijo veliko najboljših študentov s štipendijami za nadarjene, raje uporabljajo povprečje, medtem ko imajo javni kolidži s prostim vpisom raje mediano." Uporabi svoje znanje o lastnostih povprečja in mediane ter tako pojasni te preference.

(45) Podaj primer majhne množice podatkov, za katero je povprečje večje od tretjega kvartila.

(46) Okrožnica nekega vzajemnega sklada pravi, "Dobro razpršen (*diverzificiran*) portfelj vsebuje vrednostne papirje z nizkimi korelacijami." V okrožnici je tudi tabela korelacij med donosi različnih vrst investicij. Na primer, korelacija med obveznicami in delnicami velikih podjetij je 0,50 in korelacija med obveznicami in delnicami manjših podjetij je 0,21.

(a) Rachel veliko investira v obveznice. Naložbe želi razpršiti z dodatnimi vlaganji, katerih donosi niso tesno povezani z donosi njenih obveznic. Ali naj izbere delnice velikih ali manjših podjetij? Pojasni.

(b) Če želi Rachel naložbo, ki bi naraščala, kadar bi donosi iz njenih obveznic padali, kakšno korelacijo mora poiskati?

(47) Nekateri ljudje mislijo, da obnašanje trga vrednostnih papirjev v januarju napoveduje, kako se bo trg obnašal preostali del leta. Naj bo obrazložitvena spremenljivka x delež sprememb v indeksu trga vrednostnih papirjev v mesecu januarju in naj bo odzivna spremenljivka y letna sprememba v indeksu.

Pričakujemo pozitivno korelacijo med x in y, ker spremembe v januarju prispevajo k celoletnim spremembam. Iz podatkov za leta 1960 do 1997 izračunamo:

$$\begin{array}{lll} \overline{x} = 1,75\% & s_x = 5,36\% & r = 0,596\% \\ \overline{y} = 9,07\% & s_y = 15,35\% & \end{array}$$

(a) Poišči enačbo premice najmanjših kvadratov za napoved celoletnih sprememb iz januarskih.

(b) Povprečna sprememba v januarju je $\overline{x} = 1,75\%$. Uporabi izračunano regresijsko premico za napoved spremembe indeksa za leto, ko je v januarju indeks narastel za $1,75\%$. Kaj opaziš? (Glej nalogo 36.)

(48) Kaže, da morda pitje zmernih količin rdečega vina zmanjša tveganje za kardiovaskularna obolenja. Tabela 2.9 vsebuje podatke o porabi rdečega vina in številu smrti zaradi kardiovaskularnih obolenj v 19 razvitih državah iz leta 1989. Porabo vina merimo v litrih alkohola na osebo, stopnjo smrti pa v številu smrti na 100 000 prebivalcev.

(a) Nariši razsevni diagram za te podatke tako, da bo razviden morebiten vpliv pitja vina na smrti zaradi kardiovaskularnih obolenj.

(b) Enačba ustrezne regresijske premice najmanjših kvadratov je

$$y = 260,56 - 22,969x.$$

Nariši to premico na razsevni diagram.

(c) S pomočjo regresijske premice predvidi smrtnost zaradi kardiovaskularnih obolenj v državi, v kateri letno porabijo 5 litrov alkohola na osebo.

(d) Ali prihajajo ti podatki iz opazovalne študije ali iz eksperimenta? Meniš, da so ti podatki dober razlog za prepričanje, da bi povečanje količine zaužitega alkohola v ZDA (na primer iz 1,2 na 5 litrov na osebo) zmanjšalo smrtnost zaradi kardiovaskularnih obolenj? Odgovor utemelji.

(49) Tabela 2.10 predstavlja štiri nabore podatkov, ki jih je pripravil statistik Frank Anscombe kot ilustracijo nevarnosti računanja brez predhodne grafične predstavitve podatkov.

(a) Brez risanja razsevnih diagramov poišči korelacijo in regresijsko premico najmanjših kvadratov za vsako od skupin. Kaj opaziš? Uporabi regresijsko premico za napoved vrednosti spremenljivke y pri $x = 10$.

Država	Alkohol (v litrih na prebivalca)	Stopnja smrti zaradi kardiovaskularnih obolenj
Avstralija	2,5	211
Avstrija	3,9	167
Belgija/Luks.	2,9	131
Kanada	2,4	191
Danska	2,9	220
Finska	0,8	297
Francija	9,1	71
Islandija	0,8	211
Irska	0,7	300
Italija	7,9	107
Nizozemska	1,8	167
Nova Zelandija	1,9	266
Norveška	0,8	227
Španija	6,5	86
Švedska	1,6	207
Švica	5,8	115
Združeno kraljevstvo	1,3	285
ZDA	1,2	199
Zahodna Nemčija	2,7	172

Tabela 2.9: Količina zaužitega vina in smrti zaradi kardiovaskularnih obolenj za izbrane države.

(b) Izdelaj razsevne diagrame za vsako od skupin in v vsakega dodaj pripadajočo regresijsko premico.

(c) V katerih od teh štirih primerov bi bilo smiselno uporabiti regresijsko premico za opis odvisnosti y od x? V vsakem od primerov svoj odgovor utemelji.

(50) Študija ukrepov za ravnanje z odplakami meri potrebe po kisiku pri razgrajevanju usedlin. Naj bo y logaritem potrebe po kisiku (v miligramih na minuto) in x skupni delež usedlin (v miligramih na liter odplak). Z 20 merjenji smo dobili podatke v spodnji tabeli.

(a) Iz podatkov izdelaj razsevni diagram. Ali obstaja približno linearna zveza? So prisotni ubežniki?

Skupina A

x	10	8	13	9	11	14	6	4	12	7	5
y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68

Skupina B

x	10	8	13	9	11	14	6	4	12	7	5
y	9,14	8,14	8,74	8,77	9,26	8,10	6,13	3,10	9,13	7,26	4,74

Skupina C

x	10	8	13	9	11	14	6	4	12	7	5
y	7,46	6,77	12,74	7,11	7,81	8,84	6,08	5,39	8,15	6,42	5,73

Skupina D

x	8	8	8	8	8	8	8	8	8	8	19
y	6,58	5,76	7,71	8,84	8,47	7,04	5,25	5,56	7,91	6,89	12,50

Tabela 2.10: Štiri skupine podatkov za proučevanje korelacije in regresije.

(b) Nariši na diagram premico, ki se na oko najbolje prilega podatkom. Uporabi premico za napoved logaritma potrebe po kisiku y pri $x = 4$.

x	7,2	7,8	7,1	6,4	6,4	5,1	5,9	5,3	5,0	5,0
y	1,56	0,90	0,75	0,72	0,31	0,36	0,11	0,11	-0,20	-0,15

x	4,8	4,4	4,3	3,7	3,9	3,6	4,4	3,3	2,9	2,8
y	0,00	0,00	-0,09	-0,22	-0,40	-0,15	-0,22	-0,40	-0,52	-0,05

(51) Multimedijska računalniška igrica vsebuje test veščine uporabljanja računalniške miške. Program nariše krog na slučajno izbranem položaju na zaslonu. Igralec se trudi z miško kar najhitreje klikniti kjerkoli v notranjosti kroga. Nov krog se pojavi takoj, ko je uporabnik kliknil prejšnjega. V tabeli 2.11 so podatki o rezultatih, ki jih je dosegel neki igralec, po 20 za vsako roko. *Razdalja* pomeni razdaljo miškinega kazalca od središča novega kroga v enotah, ki so odvisne od velikosti zaslona. *Čas* pomeni čas, ki ga je igralec potreboval za naslednji klik (v milisekundah).

(a) Sumimo, da je čas odvisen od razdalje. Napravi razsevni diagram časa v odvisnosti od razdalje, pri čemer za vsako roko uporabi drugačne simbole.

(b) Opiši obe zvezi. Ali se pozna, da je igralec desničar?

(c) Poišči regresijsko premico za vsako roko posebej. Nariši ti dve premici na diagram. Primerjaj korelaciji pri obeh rokah. Zakaj sta podobni, čeprav je eden od vzorcev precej bolj oster kot drugi?

Čas	Razdalja	Roka	Čas	Razdalja	Roka
115	190,70	desna	240	190,70	leva
96	138,52	desna	190	138,52	leva
110	165,08	desna	170	165,08	leva
100	126,19	desna	125	126,19	leva
111	163,19	desna	315	163,19	leva
101	305,66	desna	240	305,66	leva
111	176,15	desna	141	176,15	leva
106	162,78	desna	210	162,78	leva
96	147,87	desna	200	147,87	leva
96	271,46	desna	401	271,46	leva
95	40,25	desna	320	40,25	leva
96	24,76	desna	113	24,76	leva
96	104,80	desna	176	104,80	leva
106	136,80	desna	211	136,80	leva
100	308,60	desna	238	308,60	leva
113	279,80	desna	316	279,80	leva
123	125,51	desna	176	125,51	leva
111	329,80	desna	173	329,80	leva
95	51,66	desna	210	51,66	leva
108	201,95	desna	170	201,95	leva

Tabela 2.11: Odzivni čas v računalniški igrici.

2.17 Tehnološki kotiček

Računanje povprečja in standardnega odklona

Preglednice nas oskrbijo s preprostim načinom kopiranja formul. Ta posebnost nam pride prav pri računanju povprečja in standardnega odklona množice podatkov. Na

sliki 2.15 so podatki o 15 slučajnih metih kocke zapisani v stolpcu, označenim z "x". Za izračun povprečja teh vrednosti najprej vrednosti seštejemo z uporabo funkcije `=Sum(B2:B16)`. S pomočjo vsote iz `B18` lahko izračunamo povprečje tako, da jo delimo s številom podatkov n, torej s funkcijo `=B18/15`.

	A	B	C	D	E	F
1		x	povprečje	(x-povp.)	(x-povp.)^2	
2		2	3,6	-1,6	2,56	
3		6		2,4	5,76	
4		6		2,4	5,76	
5		1		-2,6	6,76	
6		1		-2,6	6,76	
7		1		-2,6	6,76	
8		5		1,4	1,96	
9		4		0,4	0,16	
10		6		2,4	5,76	
11		3		-0,6	0,36	
12		3		-0,6	0,36	
13		1		-2,6	6,76	
14		6		2,4	5,76	
15		5		1,4	1,96	
16		4		0,4	0,16	
17						
18	vsota	54			57,6	
19	povprečje	3,6				
20				varianca	4,114286	
21				st. odklon	2,02837	

Slika 2.15: Uporaba preglednice za računanje povprečja in standardnega odklona.

Če želimo uporabiti formulo za varianco, ki smo jo spoznali v tem poglavju, moramo izračunati razliko med vsakim od podatkov in povprečjem. Povprečje vpišemo v polje `C2` in razliko v polje `D2` s pomočjo formule `=B2-C2`. Kvadrat te razlike zapišemo v `E2` s formulo `=D2^2`. Ko smo napravili prvo vrstico teh treh stolpcev, dobimo preostale tako, da kopiramo in prilepimo te formule.

Nazadnje seštejemo vrednosti v zadnjem stolpcu in vsoto delimo z $n-1$. Če zapišemo

varianco v `E20`, lahko standardni odklon dobimo s formulo `Sqrt(E20)`.

Naloga 1. Generiraj rezultate 50 metov kocke z uporabo prave kocke ali pa z generatorjem naključnih števil iz programa za delo s preglednicami. Kot v zgornjem primeru izračunaj povprečje in standardni odklon teh podatkov.

Naloga 2. Uporabi ukaz `Sort`, da svoje podatke urediš od najmanjšega do največjega. S pomočjo te ureditve nato poišči mediano in kvartile.

Računanje premice najmanjših kvadratov

Premico najmanjših kvadratov lahko opišemo z njenim naklonom in začetno vrednostjo. S pomočjo formul, ki se nekoliko razlikujejo od tistih, ki smo jih spoznali v tem poglavju, lahko naklon in začetno vrednost izračunamo iz x, y, x^2 in xy. Preglednica na sliki 2.16 vsebuje stolpce s temi podatki. Formuli za izraza na poljih `C2` in `D2` sta `=A2*A2` in `=A2*B2`. S kopiranjem teh formul zaključi tretji in četrti stolpec.

	A	B	C	D	E
1	x	y	x^2	xy	
2	1	5	1	5	
3	6	8	36	48	
4	4	8	16	32	
5	4	7	16	28	
6	2	8	4	16	
7	4	10	16	40	
8	4	9	16	36	
9	1	3	1	3	
10	2	7	4	14	
11	1	5	1	5	
12					
13	29	70	99	227	10
14					
15	naklon	1,610738			
16	zač. vred.	2,328859			

Slika 2.16: Preglednica za računanje premice najmanjših kvadratov.

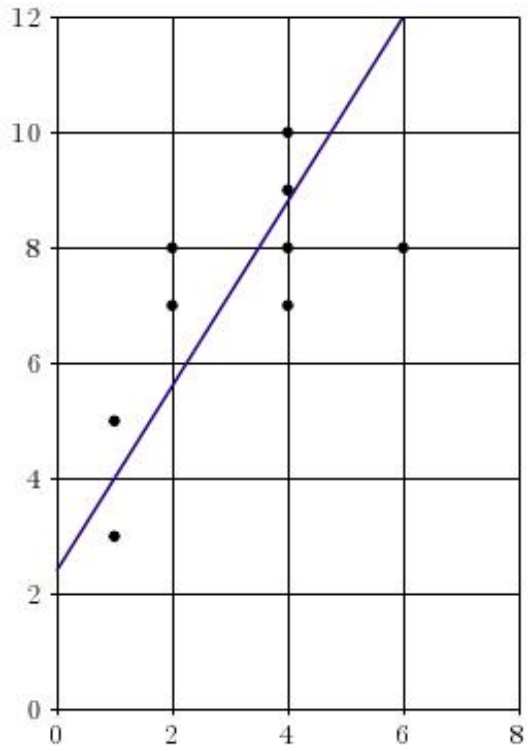
Nato uporabi ukaz `Sum` za izračun vsote v vsakem od stolpcev. Elementi 13. vrstice zdaj ustrezajo vrednostim x, y, x^2 in xy. Naklon dobimo s formulo

```
=(E13*D13-B13*A13)/(E13*C13-A13*A13).
```

Ker je v `E13` zapisano število podatkov, dobimo začetno vrednost po formuli

```
=B13/E13-C15*A13/E13.
```

Izberi podatke o x in y in nariši te točke s pomočjo ukaza za risanje razsevnega diagrama (`Scatterplot Graph`). Opaziš lahko, da podatki približno sledijo premici. Iz izračunanega naklona in začetne vrednosti razberi, da premica poteka skozi točki (0,2.3) in (6,12). Ta premica je prikazana na razsevnem diagramu na sliki 2.17.



Slika 2.17: Programi za delo s preglednicami znajo narisati tudi razsevne diagrame (*scatterplot*).

Naloga 3. Napravi naslednji eksperiment: vrzi črno in belo kocko. Naj bo x število pik na beli kocki in y vsota pik na obeh kockah. Naredi 20 ponovitev tega eksperimenta, nato pa nariši razsevni diagram dobljenih podatkov in izračunaj premico najmanjših kvadratov.

Naloga 4. Poišči v časopisu finančni poročili dveh različnih dni. Izberi 20 delnic in zapiši njihove vrednosti na prvi dan v stolpec x, vrednosti drugega dne pa v stolpec y. Nariši razsevni diagram in izračunaj premico najmanjših kvadratov za te podatke. Svoj model nato preveri tako, da izbereš 10 dodatnih delnic, poiščeš njihove vrednosti na prvi dan in uporabiš premico najmanjših kvadratov, da oceniš vrednosti delnic na drugi dan. Kako natančne so te ocene?

Raziskovanje

Kako se spreminjajo povprečje, mediana, kvartili, varianca in standardni odklon, ko število podatkov narašča? Simuliraj 20, 100 in 500 metov kocke in izračunaj ta števila za vsakega od teh primerov. Katera od teh števil ostajajo približno enaka? Katera se bistveno spremenijo? Zakaj?

2.18 Pisni projekti

(1) Del analize podatkov je oprezanje za neverjetnimi števili. Spodaj je del poročila o problemu počitniških jaht, ki onesnažujejo morje z metanjem odpadkov čez krov. Pojavilo se je v reviji *Condé Nast Traveler* junija 1992.

> Na sedemdnevnem križarjenju lahko srednje velika ladja (približno 1000 potnikov) nagrmadi 222 000 skodelic kave, 72 000 pločevink brezalkoholnih pijač, 40 000 pločevink in steklenic piva in 11 000 steklenic vina.

Ali so te številke verjetne? Napiši kratek esej, v katerem zagovarjaš svoje stališče. Vključi tudi par izračunov, ki podpirajo tvoje zaključke.

(2) Razmišljanje o številih zahteva več kot le sposobnost računanja. Ali se je dohodek Američanov v zadnjih desetletjih zmanjšal? Spodaj je nekaj podatkov, ki so se pojavili v debati na to temo. Po prilagoditvi podatkov inflaciji se je mediana dohodkov ameriških gospodinjstev zvišala iz 33 181 $ leta 1970 na 35 492 $ leta 1996. To je 7% porast v le 26 letih. Bruto domači proizvod pa se je po drugi strani povečal iz 12 070 $ v letu 1970 na 18 136 $ v letu 1996. To je 50% porast. Vsi ti podatki prihajajo z Urada za delavsko statistiko, torej so zanesljivi.

Napiši krajši esej, v katerem pojasniš to navidezno protislovje. Vpliv ekstremnih vrednosti na mediano in povprečje igra pomembno vlogo, to pa velja tudi za spremembe, do katerih je v tem obdobju prišlo v ameriških gospodinjstvih. (Gospodinjstvo sestavljajo ljudje, ki živijo skupaj na istem naslovu.)

(3) Mediji so polni dobrih in slabih grafov. Nekatere publikacije kot na primer *USA Today*, se še posebej pogosto poslužujejo grafov za prikaz podatkov. Poišči več grafov iz časopisov in revij (ne iz oglasov). Uporabi jih kot primere v kratkem eseju o jasnosti, točnosti in privlačnosti grafov v medijih. Informacije o tem, kaj so dobri grafi, lahko najdeš v knjigah Tufteja in Clevelanda, ki sta navedeni pod priporočenimi branji.

(4) Oboroženi s programsko opremo lahko začnemo raziskovati večje množice podatkov. Pojdi na `www.stat.purdue.edu/~dsmoore/data` in naloži datoteko `gpa.dat`. Ta datoteka vsebuje podatke o vseh 78 učencih sedmega razreda neke podeželske šole. Za vsakega učenca imamo pet podatkov: GPA, povprečno oceno, IQ, rezultat inteligenčnega testa, AGE, starost v letih, GENDER, spol, pri čemer 1 označuje ženski in 2 moški spol, in SC, rezultat psihološkega testa, ki meri "samopodobo".

Najprej si oglej porazdelitev GPA povprečij. Sestavi kratek opis porazdelitve, vključno z numeričnimi merami. Ali opaziš kakšne ubežnike ali druge nenavadne pojave? Nato analiziraj še zvezo med inteligenčnim količnikom in GPA. Ali učenci z višjim inteligenčnim količnikom dobivajo boljše ocene? Ali je zveza močna? Ali opaziš kakšne nenavadne točke?

Poglavje 3

Verjetnost: matematika naključij

Ste se kdaj vprašali, zakaj so igre na srečo, ki so za nekatere rekreacija ali pa droga, tako dober posel za igralnice? Vsak uspešen posel mora iz uslug, ki jih ponuja, kovati napovedljive dobičke. To velja tudi v primeru, ko so te usluge igre na srečo. Posamezni hazarderji lahko zmagajo ali pa izgubijo. Nikoli ne morejo vedeti, če se bo njihov obisk igralnice končal z dobičkom ali z izgubo. Igralnica pa ne kocka, pač pa dosledno dobiva in država lepo služi na račun loterij in drugih oblik iger na srečo.

Presenetljivo je, da lahko skupni rezultat več tisoč naključnih izidov poznamo s skoraj popolno gotovostjo. Igralnici ni potrebno obtežiti kock, označiti kart ali spremeniti kolesa rulete. Ve, da ji bo na dolgi rok vsak stavljeni evro prinesel približno pet centov dobička. Splača se ji torej osredotočiti na brezplačne predstave ali poceni avtobusne vozovnice, da bi privabili več gostov in tako povečali število stavljenega denarja. Posledica bo večji dobiček.

Igralnice niso edine, ki se okoriščajo z dejstvom, da so velikokratne ponovitve slučajnih izidov napovedljive. Na primer, čeprav zavarovalnica ne ve, *kateri* od njenih zavarovancev bodo umrli v prihodnjem letu, lahko precej natančno napove, *koliko* jih bo umrlo. Premije življenjskih zavarovanj postavi v skladu s tem znanjem, ravno tako kot igralnica določi glavne dobitke.

> Pojav je **slučajen**, če so posamezni izidi negotovi, vendar pa je na dolgi rok vzorec velikega števila posameznih izidov napovedljiv.

Za statistika *slučajen* ne pomeni *neurejen*. Za slučajnostjo je neke vrste red, ki se pokaže šele na dolgi rok, po velikem številu ponovitev. Veliko pojavov, naravnih in

tistih, ki so delo človeka, je slučajnih. Življenjska doba zavarovancev in barva las otrok sta primera naravne slučajnosti. Res, kvantna mehanika zagotavlja, da je na subatomskem nivoju v naravo vgrajena slučajnost. Teorija verjetnosti, matematični opis slučajnosti, je bistvenega pomena za prenekatero sodobno znanost.

Igre naključij so primeri slučajnosti, ki jo namenoma povzroči človek. Kocke v igralnicah so skrbno izdelane in izvrtane luknje, ki služijo označevanju pik, so zapolnjene z materialom, ki ima enako gostoto kot ostali del kocke. S tem je zagotovljeno, da ima stran s šestimi pikami enako težo kot nasprotna stran, na kateri je le ena pika. Tako je za vsako stran enako verjetno, da bo končala zgoraj. Vse verjetnosti in izplačila pri igrah s kockami temeljijo na tej skrbno načrtovani slučajnosti.

Tako statistiki kot upravniki igralnic se zanašajo na načrtovano slučajnost, čeprav statistiki uporabljajo tabele naključnih števil, ne pa kock in kart. Logika statističnih sklepov počiva na načrtovani slučajnosti in na matematiki verjetnosti, prav tisti, ki zagotavlja dobičke igralnicam in zavarovalnicam. Matematika naključij je tema tega poglavja.

3.1 Kaj je verjetnost?

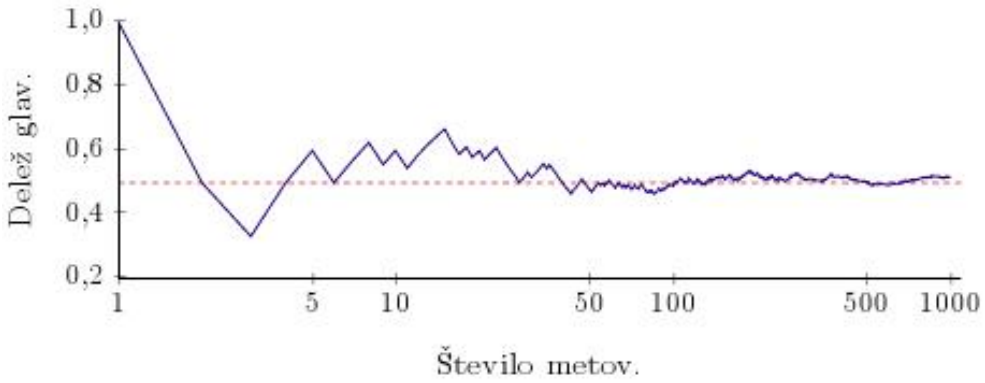
Matematika naključij, matematični opis slučajnosti, se imenuje *teorija verjetnosti*. Verjetnost opisuje napovedljive vzorce, ki na dolgi rok vladajo slučajnim izidom.

Primer. (Metanje kovanca) Ko vržemo kovanec, sta le dva možna izida: glava ali cifra. Na sliki 3.1 je prikazan rezultat 1000 metov kovanca. Za vsako število metov med 1 in 1000 smo narisali delež tistih metov, katerih rezultat je bila glava. Prvi met je bila glava, zato je začetni delež glav enak 1. Pri drugem metu je padla cifra, zato se je delež glav po dveh metih zmanjšal na 0,5. Tretji met je bila spet cifra, sledili pa sta ji dve glavi, zato je bil delež glav po petih metih enak $\frac{3}{5}$ ali 0,6.

Delež metov, pri katerih pade glava, se na začetku precej spreminja, vendar pa se ustali, ko število metov narašča. Nazadnje pride ta delež blizu 0,5 in tam obstane. Pravimo, da se glava pojavi z *verjetnostjo* 0,5. Ta verjetnost je prikazana na grafu z vodoravno črto. ♦

Verjetnost vsakega od izidov slučajnega pojava je delež števila pojavitev tega izida v dolgem zaporedju ponovitev.

Lahko bi posumili, da je verjetnost glave enaka 0,5, ker ima kovanec le dve strani.



Slika 3.1: Delež glav v odvisnosti od števila metov kovanca. Delež se sčasoma ustali pri verjetnosti za glavo.

Kot bomo videli v nalogah 1 in 2, take slutnje niso vedno pravilne. Ideja, ki se skriva za verjetnostjo, je izkustvo. Se pravi, verjetnost temelji na opazovanjih, ne pa na teoretiziranju. Verjetnost opisuje, kaj se zgodi po zelo velikem številu poskusov, in da jo določimo, moramo toliko poskusov res opazovati.

3.2 Verjetnostni modeli

Igralcem je že dolgo časa znano, da se meti kovanca, kart ali kock sčasoma ustalijo v točno določene vzorce. Verjetnostna matematika ima začetke v Franciji 17. stoletja, ko so hazarderji začeli prihajati k matematikom po nasvete (več v okvirju na strani 122). Ideja verjetnosti temelji na dejstvu, da lahko povprečni rezultat velikega števila slučajnih izidov poznamo z veliko gotovostjo. Vendar pa je definicija verjetnosti z izrazom "na dolgi rok" nejasna. Kdo ve, kaj "dolgi rok" je? Namesto tega matematično opišemo *kako se verjetnosti obnašajo*, pri čemer temeljimo na našem razumevanju deležev, ki se pojavljajo na dolgi rok. Da bi nadaljevali, si najprej zamislimo zelo preprost slučajni pojav, en sam met kovanca. Ko vržemo kovanec, ne vemo vnaprej, kakšen bo izid. Kaj pa vemo? Pripravljeni smo priznati, da bo izid bodisi glava bodisi cifra. Verjamemo, da se vsak od teh rezultatov pojavi z verjetnostjo $\frac{1}{2}$. Opis meta kovanca sestoji iz dveh delov:

- seznama vseh možnih izidov in
- verjetnosti za vsakega od teh izidov.

Tak opis je osnova vseh verjetnostnih modelov. Tule je slovarček besed, ki jih pri tem uporabljamo:

Vzorčni prostor S slučajnega pojava je množica vseh možnih izidov.
Dogodek je katerikoli izid ali množica izidov slučajnega pojava. Dogodek je torej podmnožica vzorčnega prostora.
Verjetnostni model je matematični opis slučajnega pojava, sestavljen iz dveh delov: vzorčnega prostora S in predpisa, ki dogodkom priredi verjetnosti.

Vzorčni prostor S je lahko zelo preprost ali pa zelo zapleten. Ko vržemo kovanec enkrat, sta le dva možna izida, glava ali cifra. Vzorčni prostor je torej $S = \{G, C\}$. Če izbiramo slučajni vzorec 1500 polnoletnih Američanov kot v Gallupovih raziskavah, pa vzorčni prostor vsebuje vse možne izbire 1500 izmed več kot 200 milijonov odraslih prebivalcev. Ta S je hudo velik. Vsak element vzorčnega prostora S je možni vzorec za Gallupovo raziskavo, od koder ime *vzorčni prostor*.

Primer. (Metanje kocke) Metanje dveh kock hkrati je običajni način za izgubljanje denarja v igralnicah. Ko vržemo dve kocki, je možnih 36 izidov, če med kockama razlikujemo. Ti možni izidi so prikazani na sliki 3.2. Tvorijo vzorčni prostor S. "Pade 5" je nek dogodek, označimo ga z A. Vsebuje 4 od 36 možnih izidov:

$A =$ {  }

Če so kocke dobro izdelane, izkušnje kažejo, da se vsak od 36 izidov s slike 3.2 pojavi enako pogosto. Razumen verjetnostni model torej pripiše vsakemu od izidov verjetnost $\frac{1}{36}$.

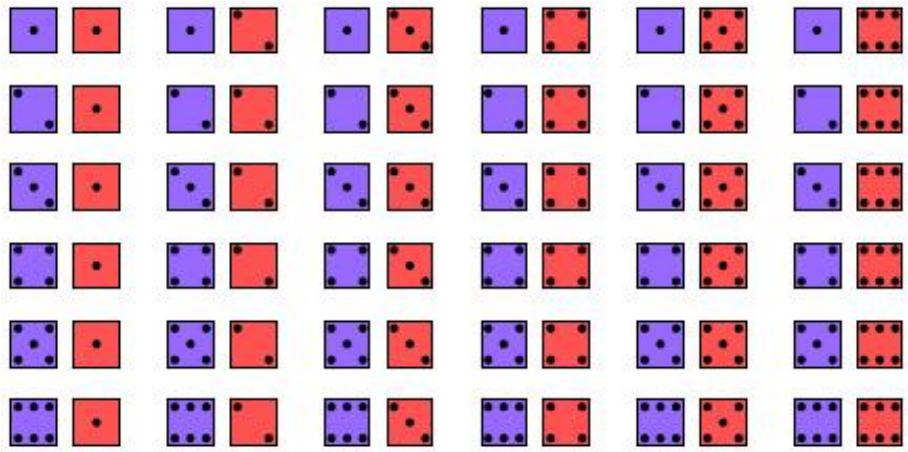
V igrah kot je *craps*[1] je pomembna samo *vsota* pik, ki jih vržemo. Spremenimo torej izide, ki nas zanimajo, takole: vržemo dve kocki in preštejemo število pik. V tem primeru je možnih le 11 izidov, od vsote 2, ki jo dobimo, če vržemo dve enici, do vsote 12, ki jo dobimo z dvema šesticama. Vzorčni prostor je

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

Če primerjamo ta S z vzorčnim prostorom na sliki 3.2, vidimo, da se S lahko spremeni, če spremenimo podrobnosti, ki jih pri nekem pojavu opazujemo. Izidi v

[1] Ameriška igra z dvema kockama. (Op. prev.)

tem novem vzorčnem prostoru *niso* več enako verjetni, ker lahko 7 dobimo na šest načinov, 2 ali 12 pa samo na enega. ♦



Slika 3.2: Možni izidi pri metu dveh kock.

3.3 Pravila verjetnosti

Načinov za predpisovanje verjetnosti je veliko, zato je smiselno začeti pri nekaj splošnih pravilih, ki se jim mora podrediti vsako prirejanje verjetnosti izidom. Ta pravila sledijo iz ideje, da je verjetnost "dolgoročni delež ponovitev, v katerih se dogodek zgodi".

(1) Verjetnost je vedno število med 0 in 1. Vsak delež je število med 0 in 1, zato je tudi verjetnost vedno število med 0 in 1. Dogodek z verjetnostjo 0 se ne zgodi nikoli (**nemogoč dogodek**), dogodek, ki ima verjetnost 1, pa se zgodi pri vsaki ponovitvi poskusa (**gotov dogodek**). Dogodek z verjetnostjo 0,5 se pri velikem številu ponovitev zgodi v polovici primerov.

(2) Vsi možni izidi morajo skupaj imeti verjetnost 1. Ker se mora pri vsaki ponovitvi pojaviti nek izid, mora biti vsota vseh verjetnosti natanko 1.

(3) Če dva dogodka nimata nobenih skupnih izidov (sta nezdružljiva ali disjunktna), je verjetnost, da se zgodi eden od njiju, enaka vsoti verjetnosti, da se zgodi prvi, in verjetnosti, da se zgodi drugi. Če se eden od dogodkov zgodi v 40% vseh poskusov, drugi v 25% in se nikoli ne

zgodita oba hkrati, potem se zgodi vsaj eden od njiju v 65% vseh poskusov, ker je $40\% + 25\% = 65\%$.

Uporabimo lahko matematični zapis, da pravila 1, 2 in 3 zapišemo krajše. Dogodke označimo z velikimi tiskanimi črkami z začetka abecede. Če je A nek dogodek, označimo njegovo verjetnost s $P(A)$. Spodaj so naša pravila za verjetnost v formalni obliki. Ko uporabljaš ta pravila, ne pozabi, da so samo še ena oblika intuitivno resničnih dejstev o dolgoročnih deležih.

Pravilo 1. Verjetnost $P(A)$ poljubnega dogodka A zadošča neenakosti $0 \leq P(A) \leq 1$.

Pravilo 2. Naj bo S vzorčni prostor verjetnostnega modela. Potem je $P(S) = 1$.

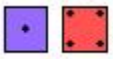
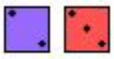
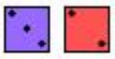
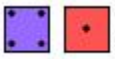
Pravilo 3. Dogodka A in B sta **disjunktna**, če nimata nobenih skupnih izidov in torej nikoli ne nastopita hkrati. Če sta A in B disjunktna, je

$$P(A \text{ ali } B) = P(A) + P(B).$$

Temu pravimo **pravilo vsote za disjunktne dogodke**.

Primer. (Verjetnosti pri metu kock) Na sliki 3.2 je prikazanih vseh 36 možnih izidov pri metu dveh kock. Za igralniške kocke je smiselno vsakemu od teh 36 izidov predpisati isto verjetnost. Ker mora biti verjetnost vseh 36 dogodkov skupaj enaka 1 (Pravilo 2), mora imeti vsak od izidov verjetnost $\frac{1}{36}$.

Kolikšna je verjetnost, da je vsota pik na obeh kockah enaka 5? Ker je ta dogodek sestavljen iz štirih možnih izidov, ki smo jih zapisali v prejšnjem primeru, nam pravilo vsote (Pravilo 3) pove, da je

$P(\text{pade } 5) = P($  $) + P($  $) + P($  $) + P($  $) =$

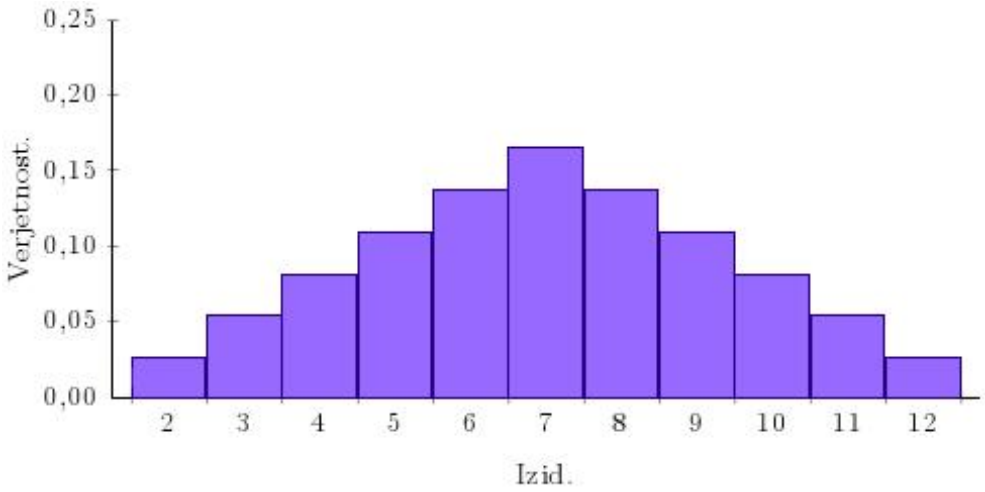
$$\begin{aligned} &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \\ &= \frac{4}{36} = 0,111. \end{aligned}$$

Kaj pa verjetnost, da dobimo 7? Na sliki 3.2 najdemo šest izidov, pri katerih je vsota pik enaka 7. Verjetnost za 7 je torej $\frac{6}{36}$, kar je približno 0,167. Na ta način nada-

ljuj z računanjem, da dobiš celoten verjetnostni model (vzorčni prostor in predpise verjetnosti) za met dveh kock, pri katerem opazujemo vsoto pik. Tole je rezultat:

Izid	2	3	4	5	6	7	8	9	10	11	12
Verjetnost	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Na sliki 3.3 je *verjetnostni histogram* tega verjetnostnega modela. Višina vsakega stolpca prikazuje verjetnost dogodka, ki ga ta stolpec predstavlja. Ker so višine ravno verjetnosti, se seštejejo v 1. Slika 3.3 je idealizirana podoba rezultatov velikega števila metov kock. Kot idealizacija je popolnoma simetrična.



Slika 3.3: Verjetnostni histogram za met dveh kock, ki prikazuje, kakšne so verjetnosti za posamezno vsoto pik.

Ta model vsakemu posameznemu izidu priredi neko verjetnost. Da bi poiskali verjetnost nekega dogodka, moramo sešteti verjetnosti izidov, ki ta dogodek sestavljajo. Na primer:

$$\begin{aligned} P(\text{izid je lih}) &= P(3)+P(5)+P(7)+P(9)+P(11) = \\ &= \frac{2}{36}+\frac{4}{36}+\frac{6}{36}+\frac{4}{36}+\frac{2}{36} = \\ &= \frac{18}{36}=\frac{1}{2}. \end{aligned}$$

♦

S tem primerom smo spoznali enega od načinov, kako dogodkom priredimo verjetnosti: predpišemo verjetnosti vsakemu od izidov, nato pa jih seštejemo, da dobimo

verjetnost dogodka. Da bi tako prirejanje zadoščalo pravilom verjetnosti, se morajo verjetnosti posameznih izidov sešteti v 1.

Verjetnostni model za končen vzorčni prostor podamo tako, da predpišemo verjetnost vsakemu posameznemu izidu. Te verjetnosti morajo biti števila med 0 in 1 in njihova vsota mora biti enaka 1. Verjetnost poljubnega dogodka je vsota verjetnosti izidov, ki ga sestavljajo.

Pod žarometom

Matematični Bernoulliji

Redke družine so prispevale k matematiki toliko, kot Bernoullijevi iz švicarskega Basla. Kar sedem Bernoullijev iz treh generacij med leti 1680 in 1800 je bilo odličnih matematikov. Pet izmed njih je pomagalo zgraditi novo matematično teorijo verjetnosti.

Jakob (1654–1705) in Johann (1667–1748) sta bila sina uspešnega švicarskega trgovca, vendar sta matematiko študirala proti volji svojega praktičnega očeta. Oba sta bila med najboljšimi matematiki svojega časa, vendar je bil Jakob tisti, ki se je osredotočil na verjetnost. Bil je prvi, ki je jasno videl idejo z dolgoročnimi povprečji kot način za merjenje slučajnosti.

Johannov sin Daniel (1700–1782) ter Jakobov in Johannov nečak Nicholas (1687–1759) sta se prav tako ukvarjala z verjetnostjo. Nicholas je opazil, da lahko z verjetnostjo pojasnimo vzorec v rojstvu dečkov in deklic. Kljub temu da se je tudi sam upiral očetovim željam, je Johann želel, da bi njegov sin Daniel postal trgovec ali zdravnik, a Daniela to ni odvrnilo in je vseeno postal še en matematik iz družine Bernoullijev. Na področju verjetnosti se je ukvarjal s pravičnim določanjem cen v igrah naključij, poleg tega pa je dokazal učinkovitost cepiva proti kozam.

Matematična družina Bernoullijev je, tako kot njihovi glasbeni sodobniki iz družine Bach, nenavaden primer nadarjenosti za določeno področje, ki se pokaže v zaporednih generacijah. Delo Bernoullijev je pomagalo verjetnosti, da je od svojega rojstva v svetu hazarda zrasla do spoštovanega orodja svetovne uporabnosti.

Primer. (Rangiranje srednješolcev) Naključno izbiramo študente in študentke prvih letnikov in jih vprašamo, kje so se v srednji šoli nahajali glede na učni uspeh. Tule so verjetnosti, ki jih dobimo iz deležev v velikem vzorcu študentov:

Razred	Zgornjih 20%	Drugih 20%	Tretjih 20%	Četrtih 20%	Spodnjih 20%
Verjetnost	0,41	0,23	0,29	0,06	0,01

Prepričaj se, da se te verjetnosti seštejejo v 1. Zdaj lahko izračunamo še naslednji dve verjetnosti:

$$\begin{aligned} P(\text{študent je v zgornjih 40 \%}) &= P(\text{zgornjih 20\%}) + P(\text{drugih 20\%}) = \\ &= 0,41 + 0,23 = 0,64, \end{aligned}$$

$$\begin{aligned} P(\text{študent ni v zgornjih 40 \%}) &= 0,23 + 0,29 + 0,06 + 0,01 = \\ &= 0,59. \end{aligned}$$

Ali veš, zakaj je verjetnost, da študent ni v zgornjih 20%, enaka 1 minus verjetnost, da izbrani študent je v zgornjih 20%? ♦

3.4 Enako verjetni izidi

Pri enostavnem slučajnem vzorčenju imajo vsi vzorci enake možnosti, da so izbrani. Pri metanju dveh pravičnih kock ima vseh 36 izidov enako verjetnost. Kadar je slučajnost produkt človekovega načrtovanja, se velikokrat zgodi, da so izidi iz vzorčnega prostora enako verjetni. Pravili 1 in 2 nam v tem primeru povesta, kolikšne so te verjetnosti.

Če je pri slučajnem pojavu možnih n izidov, ki so vsi enako verjetni, potem je verjetnost vsakega od izidov enaka $\frac{1}{n}$. Za poljubni dogodek A je njegova verjetnost enaka

$$P(A) = \frac{\text{število izidov v } A}{\text{število izidov v } S} = \frac{\text{število izidov v } A}{n}.$$

Primer. (Naključna števila) Zaporedna števila iz tabele naključnih števil 1.1 so bila generirana s pazljivim slučajenjem, zaradi katerega je vsak element z enako verjetnostjo katerikoli od desetih možnih. Ker mora biti celotna verjetnost enaka 1, je verjetnost vsakega od desetih izidov enaka $\frac{1}{10}$. Verjetnostni model je torej naslednji:

Izid	0	1	2	3	4	5	6	7	8	9
Verjetnost	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1

Na sliki 3.4 je ta model prikazan z verjetnostnim histogramom.

Verjetnost poljubnega dogodka poiščemo tako, da preštejemo, koliko izidov vsebuje. Naj bo na primer

$$A = \{\text{lih izid}\} = \{1, 3, 5, 7, 9\}$$

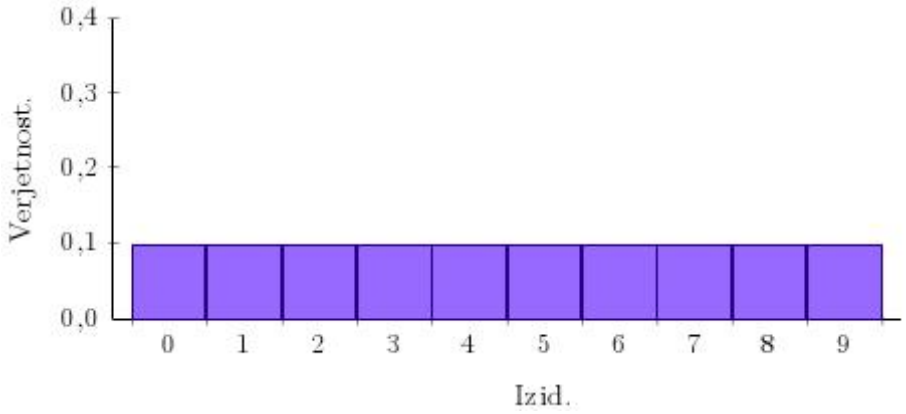
in

$$B = \{\text{izid je manjši ali enak } 3\} = \{0, 1, 2, 3\}.$$

Vidimo, da je $P(A) = 0,5$ in $P(B) = 0,4$. Dogodek [A ali B] vsebuje 7 izidov,

$$[A \text{ ali } B] = \{0, 1, 2, 3, 5, 7, 9\},$$

zato ima verjetnost 0,7. Ta verjetnost *ni* vsota verjetnosti $P(A)$ in $P(B)$, ker A in B *nista* disjunktna dogodka. Izida 1 in 3 pripadata tako A kot B. ♦



Slika 3.4: Verjetnostni histogram, ki prikazuje verjetnosti pri naključnem generiranju števke med 0 in 9.

Kadar so izidi enako verjetni, nas iskanje verjetnosti pripelje do študija metod za preštevanje, imenovanega **kombinatorika**. Kombinatorika je sama po sebi pomembno področje matematike. Za začetek si bomo ogledali metodo produkta, ki jo imenujemo tudi *osnovni princip štetja*.

Primer. (Identifikacijske kode) Nek računalniški sistem uporabnikom dodeljuje identifikacijske kode za prijavo tako, da naključno izbere tri črke (angleške abecede)[2]. Vse tričrkovne kode so torej enako verjetne. Kakšna je verjetnost, da koda, ki jo dobiš, ne vsebuje črke x?

Najprej preštejemo, koliko je vseh možnih besed. Na vsakem od treh mest se lahko pojavi katerakoli od 26 črk. Katerokoli od 26 črk na prvem mestu lahko kombiniramo

[2] Angleška abeceda ima 26 črk. (Op. prev.)

s katerokoli od 26 črk na drugem mestu, torej imamo $26 \cdot 26$ izbir. (To je res, ker je vrstni red črk pomemben, zato sta *ab* in *ba* različni izbiri.) Nazadnje lahko na tretjem mestu spet stoji katerakoli od 26 črk. Vseh možnih kod je torej

$$26 \cdot 26 \cdot 26 = 17\,576.$$

Zdaj pa preštejemo še besede, ki ne vsebujejo črke x. Take besede so sestavljene iz preostalih 25 črk, zato jih je skupaj

$$25 \cdot 25 \cdot 25 = 15\,625.$$

Verjetnost, da naša koda ne vsebuje črke x, je torej enaka

$$P(\text{ne vsebuje } x) = \frac{\text{število kod brez } x}{\text{število vseh kod}} = \frac{15\,625}{17\,576} = 0,889.$$

Recimo, da je računalnik programiran tako, da se izogne ponavljanju črk v identifikacijski kodi. Na prvem mestu se še vedno lahko pojavi katerakoli od 26 črk. Vendar pa je tokrat na drugem mestu dovoljenih le 25 preostalih črk, zato imamo skupaj $26 \cdot 25$ izbir za prvi dve črki kode. Pri katerikoli od teh izbir nam preostane še 24 črk za tretje mesto. Število različnih kod brez ponavljanja je

$$26 \cdot 25 \cdot 24 = 15\,600.$$

Pri kodah, v katerih se ne pojavi črka x, imamo na vsakem mestu še po eno možnost manj, zato jih je

$$25 \cdot 24 \cdot 23 = 13\,800.$$

Verjetnost, da naša koda ne vsebuje črke x, je torej

$$P(\text{ne vsebuje } x) = \frac{\text{število kod brez } x}{\text{število vseh kod}} = \frac{13\,800}{15\,600} = 0,885.$$

Če torej ne dovolimo ponavljanj, je verjetnost, da se bomo izognili črki x, nekoliko manjša. ♦

V prejšnjem primeru smo morali uporabiti dve od pravil preštevanja, ki jih velikokrat uporabljamo pri iskanju verjetnosti:

Pravilo preštevanja A. Dana je množica n različnih predmetov in k jih želimo urediti v vrsto. Isti predmet se lahko v razvrstitvi pojavi večkrat. Vseh možnih razvrstitev je

$$n \cdot n \cdot \ldots \cdot n = n^k.$$

Pravilo preštevanja B. Dana je množica n različnih predmetov. Spet jih želimo k postaviti v vrsto, vsak predmet pa se lahko pri tem pojavi največ enkrat. Število možnih ureditev je enako

$$n \cdot (n-1) \cdot \ldots \cdot (n-k+1).$$

V prejšnjem primeru je bil n (število črk, ki so bile na voljo) najprej enak 26, nato pa 25, k pa je bil enak 3 (število črk, ki jih moramo postaviti v vrsto, da dobimo kodo). Lažje je ta preštevanja vedno znova premisliti kot pa si zapomniti formule.

Primer. (Koliko razvrstitev?) Poroto, sestavljeno iz sedmih študentov, ki bodo sodili na debatnem turnirju, posedemo v vrsto s sedmimi stoli. Na koliko načinov lahko to storimo?

Ker na vsakem stolu sedi le en študent, seveda ne dovoljujemo ponavljanj. To situacijo torej opisuje pravilo B, pri čemer sta n in k oba enaka 7. Razmislek nadaljujemo takole: Katerikoli od sedmih študentov lahko sedi na prvem stolu, katerikoli od preostalih šestih lahko sede na drugi stol, in tako naprej. Število razporeditev je torej enako

$$7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040.$$

(To število navadno označimo 7! in preberemo "sedem fakulteta" ali pa "sedem faktoriela".) ♦

3.5 Srednja vrednost verjetnostnega modela

Recimo, da nam nekdo ponudi naslednji stavi, ki obe staneta enako: pri stavi A dobimo v primeru zmage 10 €, zmagamo pa z verjetnostjo $\frac{1}{2}$; pri stavi B dobimo 10000 €, vendar pa zmagamo z verjetnostjo $\frac{1}{10}$. Večina bi najverjetneje izbrala stavo B, čeprav so možnosti za zmago pri stavi A večje, ker pri B dobimo v primeru zmage veliko večji znesek. Bilo bi se neumno odločati med tema stavama zgolj na podlagi

verjetnosti za zmago. Znesek, ki ga pri tem lahko dobimo, je ravno tako pomemben. Kadar ima slučajni pojav številske izide, nas zanimajo tudi njihove vrednosti, ne pa zgolj verjetnosti.

Kakšen bo povprečni dobiček naših dveh stav po velikem številu iger? Spomnimo se, da so verjetnosti dolgoročni deleži iger, v katerih se pojavi vsak vsak od izidov. Pri stavi A dobimo 10 € v polovici primerov in 0 € v drugi polovici primerov. Povprečni dobiček je torej

$$\left(10\,€ \cdot \frac{1}{2}\right) + \left(0\,€ \cdot \frac{1}{2}\right) = 5\,€.$$

Po drugi strani pa pri stavi B dobimo 10 000 € pri $\frac{1}{10}$ vseh stav. Povprečni izkupiček je enak

$$\left(10\,000\,€ \cdot \frac{1}{10}\right) + \left(0\,€ \cdot \frac{9}{10}\right) = 1000\,€.$$

Če lahko stavimo velikokrat, se nam torej splača izbrati stavo B. Tole je splošna definicija "povprečnega izida", ki smo ga uporabili za primerjavo teh dveh stav:

Naj bodo možni izidi $s_1, s_2, \ldots, s_n$ iz vzorčnega prostora S števila in naj bo p_j verjetnost, da se zgodi s_j. **Srednja vrednost** μ tega verjetnostnega modela je

$$\mu = s_1p_1 + s_2p_2 + \ldots + s_np_n.$$

Spoznali smo že povprečje ali srednjo vrednost $\overline{x}$, povprečje n vrednosti, ki smo jih izmerili. Srednja vrednost μ pa opisuje verjetnostni model in ne neke množice podatkov. Lahko si μ predstavljamo kot teoretično povprečje, ki nam pove, kakšen je povprečni izid, ki ga lahko pričakujemo po veliko ponovitvah[3].

Primer. (Povprečna velikost gospodinjstev) Koliko članov šteje povprečno ameriško gospodinjstvo? V spodnji tabeli je prikazana porazdelitev velikosti gospodinjstev, ki jo je sestavil ameriški urad za štetje prebivalstva:

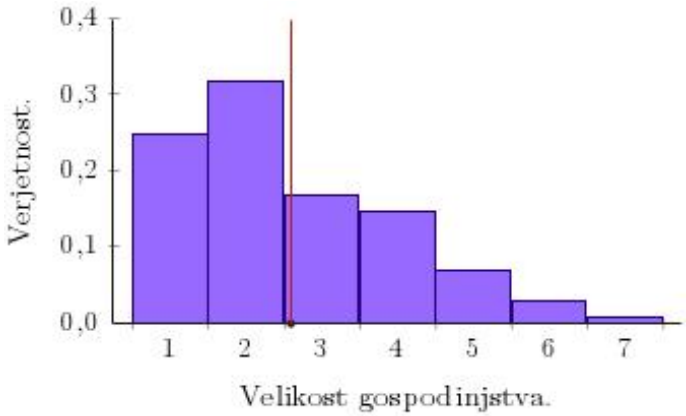
Število članov	1	2	3	4	5	6	7
Delež gospodinjstev	0,25	0,32	0,17	0,15	0,07	0,03	0,01

[3]V slovenščini se je za količino, ki jo ponavadi izračunamo po enaki formuli kot μ, uveljavil izraz *matematično upanje*. To poimenovanje lepo povzame bistvo: μ nam pove, na kakšen izkupiček lahko *upamo* po velikem številu stav. (Op. prev.)

Če naključno izberemo eno gospodinjstvo, njegova velikost sledi verjetnostnemu modelu, ki je povzet v zgornji tabeli. Srednja vrednost μ je povprečna velikost ameriškega gospodinjstva. Enaka je

$$\mu = 1 \cdot 0,25 + 2 \cdot 0,32 + 3 \cdot 0,17 + 4 \cdot 0,15 + 5 \cdot 0,07 + 6 \cdot 0,03 + 7 \cdot 0,01 = 2,6.$$

Na sliki 3.5 je verjetnostni histogram porazdelitve velikosti gospodinjstev, na katerem je označena tudi srednja vrednost $\mu = 2,6$.



Slika 3.5: Verjetnostni histogram, ki prikazuje verjetnosti za število članov naključno izbranega gospodinjstva. Povprečna velikost gospodinjstva je $\mu = 2,6$ človeka.

V tem primeru je srednja vrednost μ povprečna velikost ameriškega gospodinjstva. Če bi recimo izbrali slučajni vzorec 100 gospodinjstev in si zapisali velikosti, bi imenovali $\overline{x}$ povprečna velikost tega vzorca. Z drugim slučajnim vzorcem bi brez dvoma prišli do nekoliko drugačne vrednosti $\overline{x}$. Torej se $\overline{x}$ spreminja od vzorca do vzorca, μ, ki opisuje porazdelitev verjetnosti, pa je konstanten. ♦

Srednja vrednost μ je povprečni izid v dveh smislih. Definicija pravi, da je μ povprečje možnih izidov, uteženo z verjetnostmi. Izidom, ki so bolj verjetni, damo več teže. Pomembna lastnost verjetnosti, *zakon velikih števil*, pa pravi, da je μ povprečni izid še v enem smislu.

Oglejmo si poljuben slučajni pojav z numeričnimi izidi in končno srednjo vrednostjo μ. **Zakon velikih števil** pravi, da bo po velikem številu ponovitev pojava

- delež ponovitev, v katerih se pojavi vsak od izidov, vedno bližji verjetnosti tega izida in
- povprečje $\overline{x}$ vseh opaženih vrednosti vedno bližje μ.

Ta dejstva lahko povemo bolj natančno in jih matematično dokažemo. Zakon velikih števil naravno zaokroži idejo verjetnosti. Najprej smo opazili, da so nekateri pojavi sicer slučajni, vendar na dolgi rok kažejo pravilnosti. Nato smo uporabili idejo o dolgoročnih deležih kot motiv za vpeljavo osnovnih pravil verjetnosti. Ta pravila so matematične idealizacije, ki jih lahko uporabljamo, ne da bi pri tem verjetnost interpretirali kot delež pri velikem številu poskusov. Nazadnje pa nam zakon velikih števil pove, da se bo po velikem številu poskusov delež tistih, v katerih se pojavi določen izid, približeval verjetnosti tega izida.

Zakon velikih števil tudi pojasni, zakaj igralnice poslujejo tako dobro. Zmage (ali izgube) igralca v majhnem številu iger so negotove; to je navsezadnje tisto, kar naredi igre na srečo tako zanimive. Šele *na dolgi rok* lahko napovemo povprečni izid.

Igralnice pa igrajo igre več desettisočkrat. Igralnica lahko torej v nasprotju s posameznim igralcem računa na dolgoročno pravilnost, ki jo opisuje zakon velikih števil. Povprečni zaslužek igralnice po desettisoč igrah bo zelo blizu srednji vrednosti porazdelitve zaslužkov. Ni potrebno posebej omenjati, da ta srednja vrednost igralnici zagotavlja dobiček.

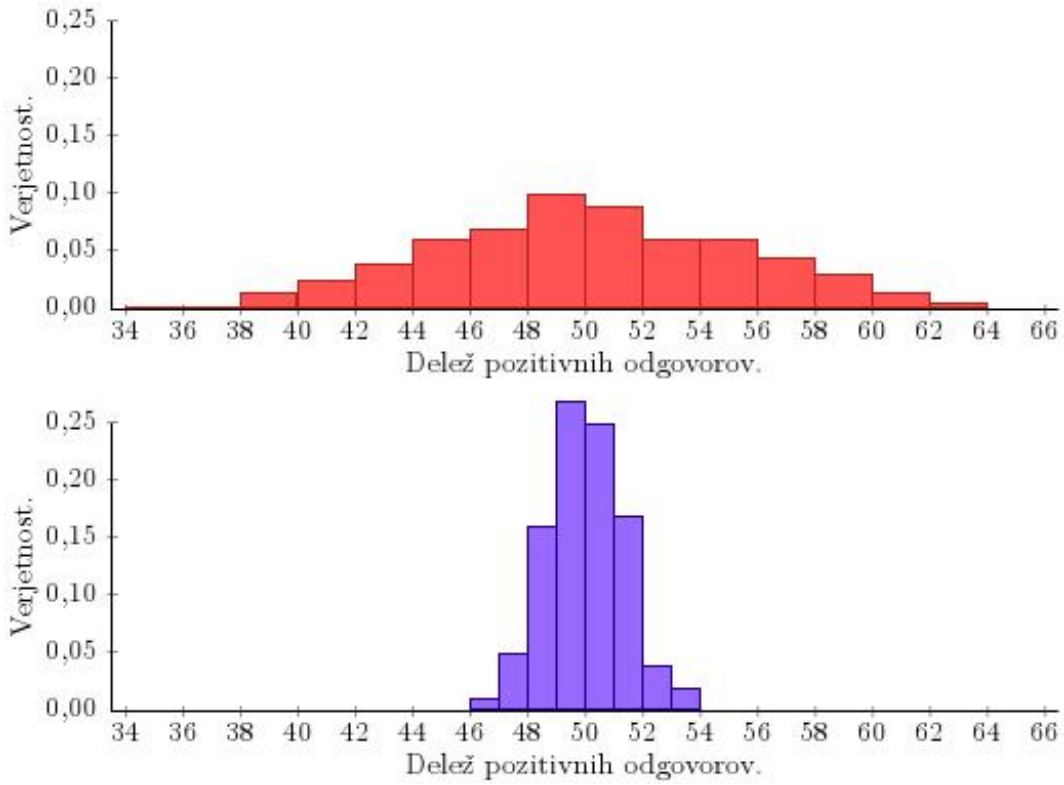
3.6 Vzorčne porazdelitve

Vzorčenje je na nek način zelo podobno kockanju. Oba se zanašata na načrtno uporabo slučajnosti. Verjetnost želimo uporabiti za opis rezultatov vzorčenja. Na prvi pogled se zdi naloga vse prej kot lahka. Recimo, da smo izbrali enostavni slučajni vzorec velikosti 100 izmed več kot 200 milijonov odraslih Američanov. Vsi možni vzorci so enako verjetni – to je definicija enostavnega slučajnega vzorčenja. Število možnih vzorcev je velikansko, zato iskanje verjetnosti s preštevanjem ne zveni prav privlačno. Obstajajo matematične bližnjice, obstaja pa še en način: namesto preštevanja dejansko izberemo veliko število vzorcev in opazujemo izide. V praksi programiramo računalnik, da posnema (formalni izraz za to je *simulirati*) izbiro velikega števila vzorcev. Poskusimo.

Primer. (Eksperiment z vzorčenjem) V poglavju o pridobivanju podatkov smo si ogledali Gallupovo raziskavo, v kateri so 1493 ljudi vprašali, če jih je ponoči strah zapustiti domove zaradi kriminala. To vprašanje postavimo enostavnemu slučajnemu vzorcu 100 ljudi in 48 jih odgovori pritrdilno. To predstavlja 48% vzorca. Izberemo še en vzorec in dobimo 50% pozitivnih odgovorov. Temu pojavu pravimo

vzorčna spremenljivost. Ko vedno znova izbiramo vzorce iz iste populacije, se bodo dobljeni rezultati razlikovali od vzorca do vzorca.

Na sliki 3.6(a) je prikazan histogram deleža pozitivnih odgovorov po izbiri 1000 enostavnih slučajnih vzorcev iz iste populacije. Opazimo značilno sliko, ki je karakteristika slučajnega vzorčenja. Zdaj lahko za opis te slike uporabimo jezik verjetnosti.



Slika 3.6: Vzorčni porazdelitvi, ki prikazujeta obnašanje deleža vzorca, ki odgovori pritrdilno na anketno vprašanje v primeru enostavnih slučajnih vzorcev iz iste populacije. (a) Vzorci velikosti 100. (b) Vzorci velikosti 1493.

Eden od razredov v histogramu pokriva območje

$$46\% < \text{delež pozitivnih odgovorov } \leq 48\%.$$

Natanko 70 od 1000 vzorcev je padlo v ta razred. Ker je za 1000 vzorcev potrebno veliko število ponovitev slučajnega vzorčenja, je delež vzorcev, ki pristanejo v tem razredu blizu verjetnosti tega razreda. Lahko torej ocenimo, da je verjetnost, da bo pozitivnih odgovorov več kot 46% in manj kot 48%, enaka $\frac{70}{1000} = 0,07$. Višina ustreznega stolpca na sliki 3.6(a) je 0,07. ♦

Statistiki imenujejo števila, ki jih izračunajo iz vzorcev, **statistike**. Delež pozitivnih odgovorov v našem vzorcu 100 ljudi je statistika. Histogram na sliki 3.6 (a) prikazuje vzorčno spremenljivost te statistike tako, da pripiše verjetnosti možnim vrednostim. Te verjetnosti sestavljajo *vzorčno porazdelitev* te statistike.

> **Vzorčna porazdelitev** statistike je porazdelitev vrednosti, ki jih zavzame statistika pri vseh možnih vzorcih iste velikosti iz iste populacije.

Če smo natančni, je vzorčna porazdelitev idealna slika, ki bi se pokazala, če bi pregledali vse možne vzorce velikosti 100 iz naše populacije. To idealno porazdelitev bi lahko prikazali z verjetnostnim histogramom. Porazdelitev, ki jo dobimo s fiksnim številom poskusov, na primer s 1000 poskusi na sliki 3.6 (a), je zgolj približek te vzorčne distribucije. V statistiki lahko z uporabo teorije verjetnosti dobimo točne vzorčne porazdelitve, ne da bi zares izbrali veliko vzorcev. Interpretacija vzorčne porazdelitve pa ostaja enaka, ne glede na to, ali jo dobimo z dejanskim vzorčenjem ali le s pomočjo matematične teorije verjetnosti.

Oglejmo si še drugi eksperiment z vzorčenjem. V Gallupovi raziskavi so anketirali 1493 ljudi in ne 100. Izberemo 1000 enostavnih slučajnih vzorcev 1493 ljudi. Za vsakega od teh vzorcev izračunamo delež pozitivnih odgovorov. Slika 3.6 (b) prikazuje porazdelitev dobljenih 1000 deležev, pri čemer uporabimo isto merilo kot pri sliki 3.6 (a). Dobimo vzorčno porazdelitev za to statistiko.

Primer. (Pregled vzorčnih porazdelitev) Uporabimo naša orodja za opisovanje porazdelitev na obeh vzorčnih porazdelitvah s slike 3.6. Ogledali si bomo *obliko*, *središče* in *razpon* teh dveh porazdelitev.

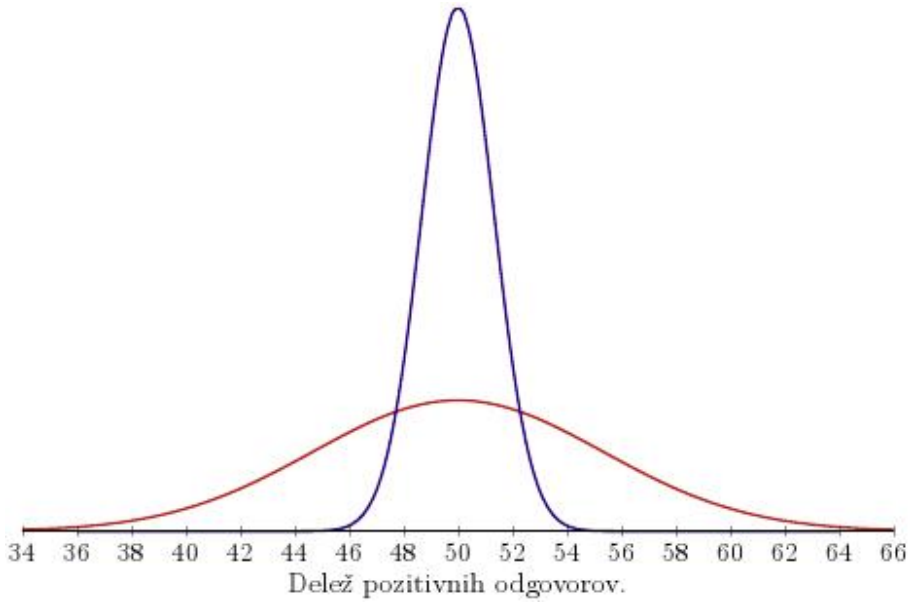
Obe porazdelitvi imata značilno zvončasto obliko. Sta precej simetrični, v središču imata en sam vrh. Ubežnikov, ki bi izstopali, ni. Središči obeh sta zelo blizu 50%. Pravzaprav sta srednji vrednosti 50,11 za vzorce velikosti 100 in 50,03 za vzorce velikosti 1493. Porazdelitev rezultatov za vzorce velikosti 1493 ima bistveno manjši razpon kot porazdelitev za vzorce velikosti 100, se pravi, histogram na sliki 3.6 (b) je višji in ožji od diagrama na sliki 3.6 (a). Standardni odklon rezultatov 1000 vzorcev je 4,986 za manjše in 1,289 za večje vzorce.

V resnici je populacija, iz katere smo izbirali vzorce, vsebovala natanko 50% ljudi, ki bi odgovorili pritrdilno. Središči vzorčnih porazdelitev sta zelo blizu 50%. Iz tega je razvidna odsotnost pristranskosti v enostavnih slučajnih vzorcih. Razpon rezultatov se manjša, ko izbiramo večje vzorce. Veliki vzorci torej običajno dajo rezultate, ki so blizu resnični vrednosti za populacijo. ♦

Iz našega eksperimenta z vzorčenji smo dobili približne vrednosti za verjetnosti (brez preštevanja), poleg tega pa smo se naučili nekaj o obnašanju vzorčnih porazdelitev, ko povečujemo velikost vzorca. Naš cilj je, da bi se naučili dovolj teorije verjetnosti, da bi lahko dobili rezultate, ki bi bili natančnejši od tistih, dobljenih z eksperimenti na vzorcih. V naslednjem razdelku bomo spoznali specifičen recept za računanje srednje vrednosti in standardnega odklona vzorčnih porazdelitev. Prvi korak je študij značilnih oblik teh porazdelitev. Imenujemo jih *normalne porazdelitve*.

3.7 Normalne porazdelitve

Čeprav se razlikujeta v spremenljivosti, sta histograma na slikah 3.6 (a) in (b) podobna. Oba sta simetrična s središči blizu 50%, na obeh straneh gladko padata, ubežnikov ni. Predstavimo obliko obeh histogramov z gladko krivuljo skozi vrhove vsakega od stolpcev. Če naredimo to zelo pazljivo z uporabo dejanskih verjetnosti posameznih izidov in ne ocen, ki smo jih dobili iz zgolj 1000 vzorcev, sta dobljeni krivulji precej blizu dvema članicama družine *normalnih krivulj*. Ti dve normalni krivulji sta prikazani na sliki 3.7.



Slika 3.7: Normalni krivulji, ki aproksimirata vzorčni distribuciji s slike 3.6. Višja krivulja pripada vzorcem velikosti 1493, nižja pa vzorcem velikosti 100. Ploščina pod vsako od krivulj je natanko 1.

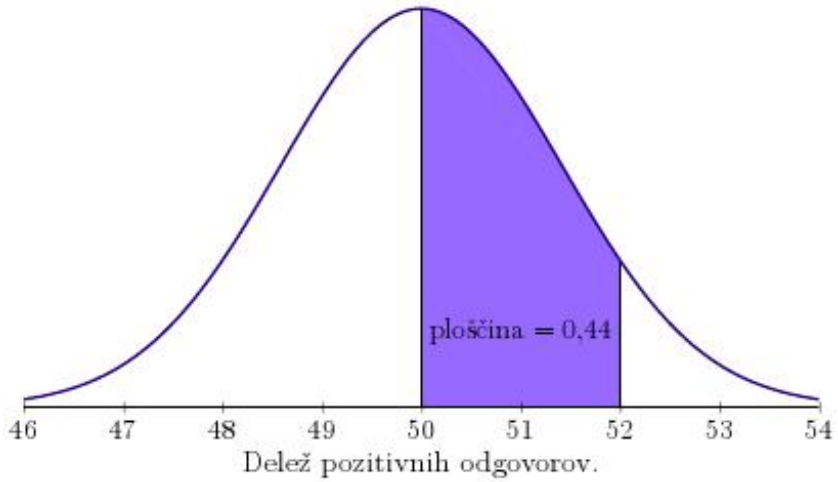
Normalne krivulje predstavljajo nov način za opisovanje verjetnosti. Prirejanje ver-

jetnosti posameznim vrednostim statistike lahko opišemo z verjetnostnim histogramom. Višina vsakega stolpca je verjetnost izidov, ki spadajo v ustrezni razred. Ker so vsi stolpci enake širine, so njihove ploščine (višina krat širina) sorazmerne z višinami. Normalne krivulje si lahko predstavljamo kot približke za verjetnostne histograme, v katerih je ploščina natančno enaka verjetnosti. Delo z normalnimi krivuljami je lažje kot delo s histogrami, ker stolpce zamenjamo z eno samo gladko krivuljo. Za normalne krivulje je značilno, da je skupna ploščina pod vsako krivuljo natanko 1, kar ustreza dejstvu, da je skupna verjetnost vseh možnih izidov ravno 1.

Normalna krivulja priredi verjetnosti izidom takole: verjetnost vsakega intervala je enaka ploščini pod normalno krivuljo nad tem intervalom. Celotna ploščina pod katerokoli normalno krivuljo je natančno 1.

Primer. (Verjetnost kot ploščina pod krivuljo) Na sliki 3.8 je še en primer normalne krivulje za vzorčno porazdelitev s slike 3.6 (b). Ta krivulja predpiše verjetnosti deležem pozitivnih odgovorov v enostavnih slučajnih vzorcih velikosti 1493 iz Gallupove raziskave.

Osenčeno območje je ploščina pod normalno krivuljo med 50% in 52%. Ta ploščina je enaka 0,44. Verjetnost, da bo med 50% in 52% ljudi v slučajno izbranem vzorcu odgovorilo pritrdilno, je torej 0,44. ♦



Slika 3.8: Verjetnost kot ploščina pod normalno krivuljo. Ploščina 0,44 je verjetnost, da leži izid med 50 in 52.

V naši prvi metodi za računanje verjetnosti smo najprej predpisali verjetnosti posameznim izidom, nato pa smo te vrednosti seštevali, da smo dobili verjetnost poljub-

nega dogodka. Verjetnost kot ploščina pod krivuljo pa je druga pomembna metoda za računanje verjetnosti. Lažje je, kadar imamo veliko posameznih izidov, ki so blizu skupaj. Krivulje različnih oblik opisujejo različno porazdeljene verjetnosti. Ukvarjali se bomo predvsem z normalnimi krivuljami, ker opisujejo verjetnost v številnih pomembnih situacijah. Kadar izidom priredimo verjetnosti s pomočjo normalne krivulje, govorimo o **normalni porazdelitvi verjetnosti**.

Sliki 3.6 in 3.7 nazorno prikazujeta, da je vzorčna porazdelitev deleža enostavnih slučajnih vzorcev blizu normalni porazdelitvi. To ni zgolj vprašanje umetniške presoje, je matematično dejstvo, ki ga je prvi dokazal Abraham DeMoivre leta 1718. Nekatere druge pogoste statistike, na primer povprečje $\overline{x}$ za velike vzorce, imajo prav tako vzorčne porazdelitve, ki so približno normalne. Normalna krivulja ne bo natančno opisala specifične množice izidov kot je na primer naša množica deležev 1000 vzorcev. Gre za idealizirano porazdelitev, ki je uporabna in daje dober približek dejanske porazdelitve izidov.

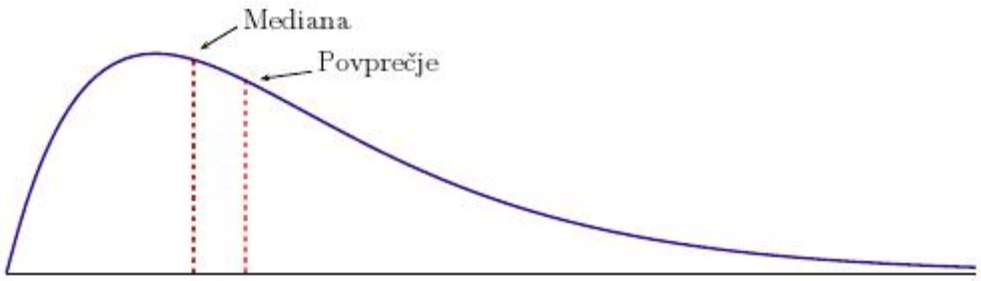
Obstaja tesna povezava med opisom predpisovanja verjetnosti numeričnim izidom in opisom množice podatkov. Za obe nalogi lahko uporabimo histograme. Prav tako lahko gladke krivulje, kakršne so na primer normalne krivulje, nadomestijo histograme tako v primeru velikih množic podatkov kot pri predpisovanju verjetnosti. Veliko množic podatkov lahko približno opišemo s pomočjo normalnih porazdelitev. Normalne porazdelitve si torej zaslužijo, da si jih podrobneje ogledamo.

3.8 Oblika normalnih krivulj

Normalno krivuljo je mogoče natančno podati z enačbo, vendar pa bomo mi zadovoljni že s slikami, kakršni sta sliki 3.7 in 3.8. Vse normalne krivulje so simetrične, zvončaste oblike in na obeh koncih hitro padejo proti nič. Sredina normalne krivulje je središče v več pomenih. Najprej je srednja vrednost μ, ki smo jo spoznali pri predpisovanju verjetnosti. Je tudi mediana, za katero velja, da polovica verjetnosti (polovica ploščine pod krivuljo) leži na eni in polovica na drugi strani. Kadar verjetnosti definiramo s pomočjo ploščin pod simetrično krivuljo, je srednja vrednost μ tudi mediana porazdelitve.

Povprečje in mediana asimetrične porazdelitve pa nista enaka. Primer na sliki 3.9 prikazuje *desno asimetrično porazdelitev*. Desni konec krivulje je veliko daljši od levega. Cene novih hiš so primer asimetrične porazdelitve: veliko je srednje dragih hiš, na desnem koncu pa nekaj izredno dragih graščin. Te graščine dvignejo srednjo

vrednost, povprečno ceno, zato je le-ta v teh primerih večja od mediane. Povprečna cena novih hiš v letu 1997 je bila 176 000 $, mediana pa le 146 000 $.



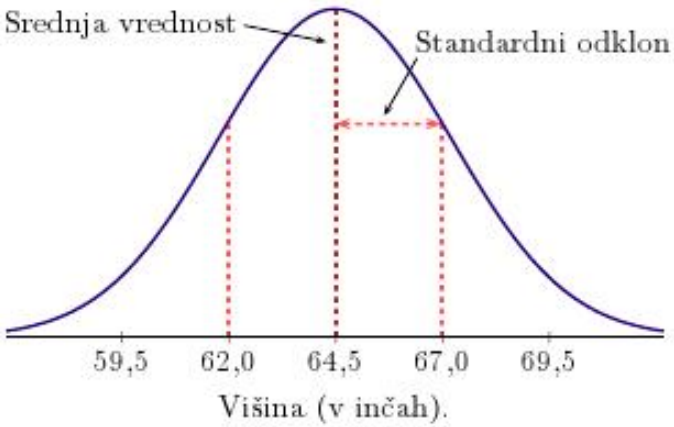
Slika 3.9: Povprečje asimetrične porazdelitve se nahaja bližje daljšemu koncu kot mediana.

Kot smo videli v prejšnjem poglavju, mora tudi najbolj površen opis podatkov o eni sami spremenljivki poleg podatka o središču vsebovati tudi informacijo o razponu. Kako pa je z razponom normalnih krivulj? *Normalne krivulje imajo posebno lastnost, da je njihov razpon mogoče natančno podati z enim samim številom, standardnim odklonom.* V prejšnjem poglavju smo se naučili, kako iz množice podatkov izračunamo standardni odklon. V primeru normalnih krivulj lahko standardni odklon (tako kot srednjo vrednost) preberemo direktno iz krivulje.

Srednja vrednost normalne porazdelitve leži v središču simetrije normalne krivulje.
Standardni odklon normalne porazdelitve poiščemo tako, da se s svinčnikom sprehodimo po krivulji od središča proti robu. Ko se oddaljujemo od središča, krivulja najprej vedno bolj strmo pada, nato pa se začne padanje upočasnjevati. Točki, v katerih se ukrivljenost krivulje spremeni, se nahajata en standardni odklon stran od središča.

Z nekaj vaje se naučimo precej natančno določati točke, v katerih se spremeni ukrivljenost. Slika 3.10 prikazuje porazdelitev višin američank med 18. in 24. letom. Krivulja je normalna in ima srednjo vrednost (in mediano) pri višini $\mu = 64,5$ inčev. Ukrivljenosti se spremenita pri 62 in pri 67 inčih. Standardni odklon porazdelitve je enak razdalji med katerokoli od teh sprememb in središčem, torej 2,5 inčev.

Običajna oznaka za standardni odklon verjetnostne porazdelitve je σ, grška črka sigma. Tako kot srednjo vrednost μ je tudi standardni odklon σ poljubne porazdelitve mogoče poiskati neposredno iz podanih verjetnosti. Prav tako tudi ločimo



Slika 3.10: Položaj srednje vrednosti in standardnega odklona pri normalni krivulji. Pri zgornji krivulji je $\mu = 64,5$ in $\sigma = 2,5$.

med standardnim odklonom s dane množice podatkov in σ, standardnim odklonom verjetnostne distribucije.

V prejšnjem poglavju smo velikokrat uporabili kvartile za opis razpona porazdelitve. Ker standardni odklon popolnoma opiše razpon poljubne normalne porazdelitve, nam tudi pove, kje se nahajata kvartila. Velja:

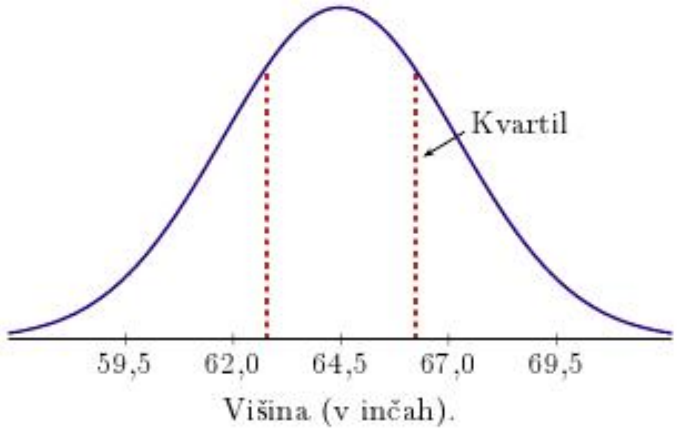
> Prvi kvartil vsake normalne porazdelitve se nahaja $0,67\sigma$ pod srednjo vrednostjo, tretji kvartil se nahaja $0,67\sigma$ nad srednjo vrednostjo.

Primer. (Porazdelitev višin) Na sliki 3.10 je prikazana porazdelitev višin deklet med 18. in 24. letom. Porazdelitev je približno normalna s središčem $\mu = 64,5$ inčev in standardnim odklonom $\sigma = 2.5$ inčev. Kvartila ležita $0,67\sigma$ ali

$$0,67 \cdot 2,5 = 1,7 \text{ inča}$$

od središča. Prvi kvartil je torej $64,5 - 1,7 = 62,8$ inčev, tretji kvartil pa je $64,5 + 1,7 = 66,2$ inčev. Kvartila sta označena na sliki 3.11. Med njima leži 50% vseh višin. ♦

Srednja vrednost in standardni odklon normalnih krivulj imata posebno lastnost: *normalna krivulja je popolnoma določena z μ in σ*. Podatka o središču in razponu nista dovolj, da bi natančno določili obliko večine porazdelitev podatkov, vendar pa zadostujeta, kadar gre za normalno porazdelitev. Če spremenimo srednjo vrednost, se normalna krivulja le premakne, če pa spremenimo razpon, se spremeni tudi njena



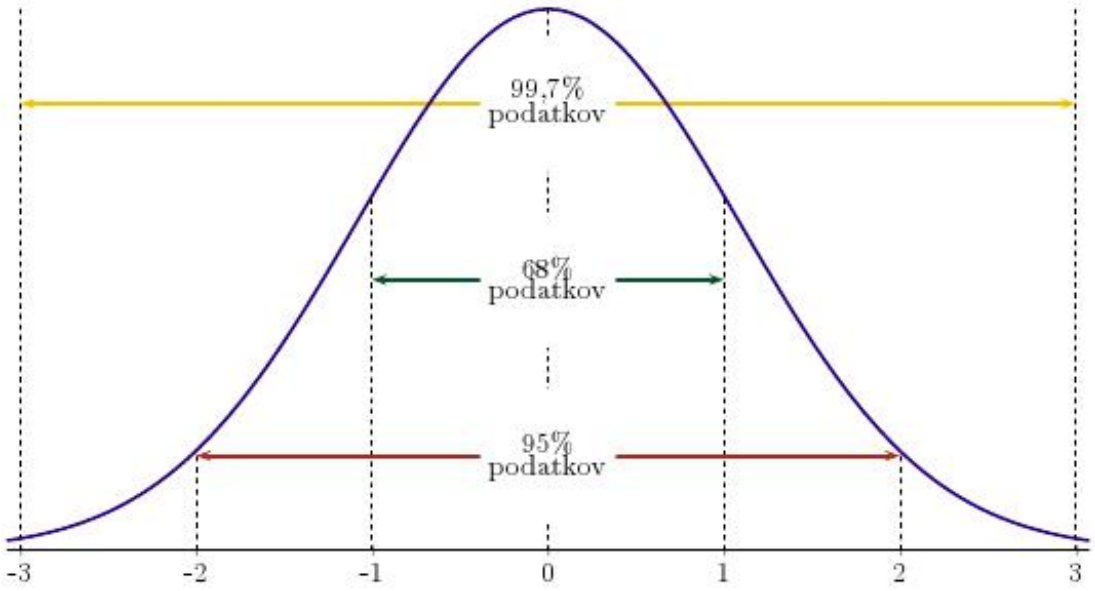
Slika 3.11: Kvartila normalne porazdelitve se nahajata na vsaki strani srednje vrednosti in sta od nje oddaljena za $0,67$ standardnega odklona. Za zgornjo krivuljo je $\mu = 64,5$ in $\sigma = 2,5$.

oblika. Normalna krivulja z manjšim razponom je višja in ožja (razpršenost je manjša) kot normalna krivulja z večjim standardnim odklonom. To lahko vidimo, če primerjamo krivulji na sliki 3.7. Obe imata isto srednjo vrednost, vendar ima krivulja, ki ustreza vzorcem velikosti 1493, manjši standardni odklon.

3.9 Pravilo 68-95-99,7

Vse normalne porazdelitve so enake, če nas zanima le, za koliko standardnih odklonov so podatki oddaljeni od središča. To je ena od posledic dejstva, da je normalna krivulja natančno določena s srednjo vrednostjo in standardnim odklonom. V posebnem so za vse normalne porazdelitve enake verjetnosti, da je neka vrednost oddaljena od središča za en, dva ali tri standardne odklone. Verjetnost, da se izid nahaja v razdalji enega standardnega odklona od središča, je 0,68. Za dva standardna odklona je ta verjetnost enaka 0,95. Pri treh standardnih odklonih je verjetnost že skoraj 1, natančneje, 0,997. Te številke lahko izračunamo iz enačb za normalne krivulje in ne veljajo za druge porazdelitve.

Slika 3.12 prikazuje te verjetnosti, izražene v odstotkih. Tej lastnosti pravimo *68–95–99,7 pravilo* za normalne porazdelitve.



Slika 3.12: Pravilo 68–95–99,7 za normalne porazdelitve.

Po **pravilu 68–95–99,7** za vsako normalno porazdelitev velja:

- 68% vrednosti leži v razdalji enega standardnega odklona od srednje vrednosti,
- 95% vrednosti leži v razdalji dveh standardnih odklonov od srednje vrednosti,
- 99,7% vrednosti leži v razdalji treh standardnih odklonov od srednje vrednosti.

Z uporabo treh števil iz pravila 68–95–99,7 lahko hitro izpeljemo koristne informacije o vsaki normalni porazdelitvi. Bolj podrobne informacije lahko dobimo iz tabel, ki navajajo ploščine pod normalnimi krivuljami, vendar pa bo za naše potrebe zadoščalo pravilo 68–95–99,7.

Primer. (Porazdelitve višin) Višine deklet med 18. in 24. letom so približno normalno porazdeljene s srednjo vrednostjo $\mu = 64,5$ inčev in standardnim odklonom $\sigma = 2,5$ inčev. En standardni odklon pod srednjo vrednostjo je $64.5 - 2.5 = 62$ inčev. Podobno je en standardni odklon nad srednjo vrednostjo 67 inčev. Pravilo 68–95–99,7 nam pove, da je približno 68% deklet visokih med 62 in 67 inči. Dva standardna odklona ustrezata 5 inčem, zato vemo, da je 95% deklet visokih med 59,5 in 69,5 inčev. Skoraj vse višine pa padejo v interval treh standardnih odklonov

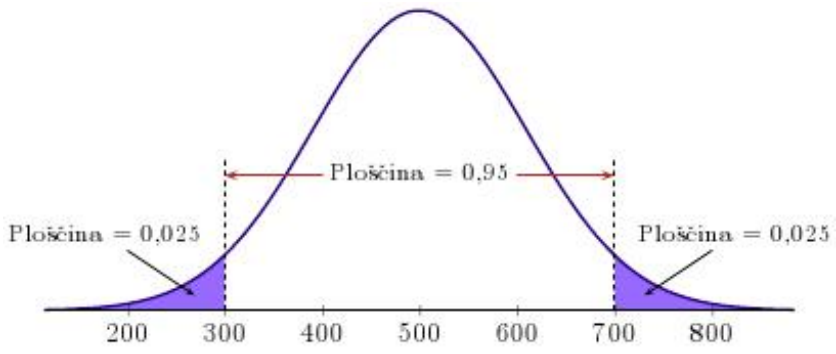
od srednje vrednosti, torej med 57 in 72 inči. Zelo malo deklet je višjih od 6 čevljev (72 inčev). ♦

Primer. (Rezultati sprejemnih izpitov) Porazdelitve rezultatov izpitov so približno normalne. Rezultati izpitov SAT so preračunani tako, da je povprečni rezultat približno $\mu = 500$ in je standardni odklon približno $\sigma = 100$. S pomočjo teh dejstev lahko odgovorimo na nekaj vprašanj o SAT rezultatih.

- *Koliko točk mora doseči kandidat, da se uvrsti v zgornjih 25%?*
  Tretji kvartil je $0,67 \cdot 100 = 67$ točk nad srednjo vrednostjo, zato so rezultati nad 567 v zgornjih 25%.

- *Kolikšen del rezultatov je med 200 in 800 točkami?*
  Rezultati med 200 in 800 točkami so največ tri standardne odklone oddaljeni od srednje vrednosti. Po pravilu 68–95–99,7 sklepamo, da leži $99,7\%$ rezultatov na tem intervalu. (V resnici sta 200 in 800 mejni vrednosti med doseženimi rezultati. Rezultate, ki so višji od 800 točk, zaokrožijo na 800.)

- *Kolikšen del rezultatov je nad 700 točkami?*
  Vrednost 700 leži dva standardna odklona nad srednjo vrednostjo. Po pravilu 68–95–99,7 sklepamo, da leži 95% vseh rezultatov med 300 in 700, torej jih 5% leži pod 300 in nad 700. Ker so normalne krivulje simetrične, leži polovica od teh 5% rezultatov nad 700. Če torej kandidat dobi več kot 700 točk, sodi v zgornjih 2,5% populacije.



Slika 3.13: Z uporabo pravila 68–95–99,7 poiščemo odstotek rezultatov sprejemnih izpitov SAT, višjih od 700. Za zgornjo krivuljo je $\mu = 500$ in $\sigma = 100$.

Graf normalne krivulje, na katerem označimo točke, ki so en, dva ali tri standardne odklone oddaljene od srednje vrednosti, nam lahko pomaga pri uporabi pravila 68–95–99,7. Na sliki 3.13 je prikazana porazdelitev rezultatov sprejemnih izpitov SAT.

Ploščini, ki nam pomagata pri določanju deleža rezultatov nad 700 točkami, sta osenčeni. ♦

3.10 Centralni limitni izrek

Pomembnost normalnih porazdelitev delno pojasnjuje ključna ugotovitev v teoriji verjetnosti, poznana pod imenom *centralni limitni izrek*. Izrek pravi, da porazdelitve slučajnih pojavov po velikem številu neodvisnih ponovitev težijo k normalnim. Centralni limitni izrek nam omogoča, da analiziramo slučajne pojave in predvidimo njihove izide, če imamo na voljo večje število podatkov.

Videli smo že, kako centralni limitni izrek deluje v našem primeru s slučajnim vzorčenjem. Vsak posameznik odgovori pritrdilno ali nikalno na anketno vprašanje. Možna sta le dva izida in nikjer ne opazimo kakšne normalne krivulje. Vendar pa delež pozitivnih odgovorov pri slučajnem vzorcu 100 ljudi že grobo sledi normalni porazdelitvi. Delež pozitivnih odgovorov si lahko predstavljamo kot povprečje pozitivnih odgovorov po množici 100 ljudi. Če v vzorec izberemo 1493 ljudi, delež pozitivnih odgovorov predstavlja povprečje po še večji skupini ljudi, zato je porazdelitev še bližje normalni.

V našem eksperimentu z vzorčenjem smo ugotovili, da so vzorci velikosti 1493 precej manj razpršeni kot vzorci velikosti 100. Razpon opišemo s standardnim odklonom normalne porazdelitve izidov. Centralni limitni izrek eksplicitno poveže velikost vzorca in standardni odklon. Tule je bolj natančna formulacija trditve:

Centralni limitni izrek pravi, da

- je vzorčno povprečje (ali delež vzorca) n ponovitev istega slučajnega pojava pri velikih n približno normalno porazdeljeno,
- srednja vrednost te normalne porazdelitve je enaka srednji vrednosti posamezne ponovitve poskusa in
- standardni odklon te normalne porazdelitve je standardni odklon posamezne ponovitve poskusa, deljen z $\sqrt{n}$.

Pozorni moramo biti tudi na dejstvo, da standardni odklon pada s kvadratnim korenom števila ponovitev poskusa, $\sqrt{n}$. To je res za vse vrednosti n, ne samo za tiste

dovolj velike n, pri katerih nam centralni limitni izrek pove, da je porazdelitev blizu normalni.

Primer. (Spremenljivost povprečij) Slučajno izberimo neko dekle iz primera o porazdelitvah višin. Pri ponavljanju takih izbir se dobljene višine spreminjajo, standardni odklon pri tem pa je $\sigma = 2,5$ inčev.

Zdaj pa izberimo pet deklet in izračunajmo povprečje njihovih višin $\overline{x}$. Povprečje $\overline{x}$ se prav tako spreminja, ko izberemo več takih skupin, standardni odklon pa je v tem primeru enak

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2,5}{\sqrt{5}} = \frac{2,5}{2,236} = 1,118 \text{ inčev.}$$

Oznaka $\sigma_{\overline{x}}$ nas opominja, da gre za standardni odklon porazdelitve $\overline{x}$, ne pa za standardni odklon ene same vrednosti. ♦

Da je spremenljivost povprečij $\overline{x}$ manjša od spremenljivosti ene spremenljivke, je pomembna ugotovitev. Na primer povprečje 25 vrednosti ($\sqrt{25} = 5$) iz iste populacije ima standardni odklon, ki je petkrat manjši od standardnega odklona spremenljivke. Standardni odklon povprečja 100 vrednosti ($\sqrt{100} = 10$) je le še $\frac{1}{10}$ standardnega odklona spremenljivke. Če želimo standardni odklon vzorčnega povprečja razpoloviti, moramo zaradi kvadratnega korena število vzorca kar štirikrat povečati, če velikost le podvojimo, ni dovolj.

3.11 Uporaba centralnega limitnega izreka

Centralni limitni izrek lahko uporabimo, da se prepričamo, kako dober posel so za igralnice igre na srečo. Oglejmo si le eno od številnih stav, ki jih ponujajo.

Primer. (Rdeče ali črno) Ameriško kolo pri ruleti ima 38 razdelkov, med katerimi je 18 črnih, 18 rdečih in 2 zelena. Ko kolo zavrtimo, je za vsak razdelek enako verjetno, da kroglica pristane v njem. Igralci lahko postavijo različne stave. Pri eni od najpreprostejših igralec izbere rdečo ali črno barvo. Recimo da stavi en evro na rdeče. Če kroglica pristane v rdeče obarvanem razdelku, igralec dobi poleg vloženega evra še enega. V nasprotnem primeru izgubi denar, ki ga je stavil. Pri tem je pomembno, da zelena razdelka pomenita zmago igralnice.

Odločimo se torej, da bomo stavili na rdeče. Možna sta le dva izida: zmagamo ali izgubimo. Zmagamo, če se kroglica ustavi v enem od 18 rdečih razdelkov, izgubimo pa, če konča v enem od 20 razdelkov črne ali zelene barve. Ker so kolesa v igralnicah

skrbno uravnotežena, so vsi razdelki enako verjetni in zato sta verjetnosti

$$\begin{aligned} P(\text{dobimo } 1\,€) &= \frac{18}{38}, \\ P(\text{izgubimo } 1\,€) &= \frac{20}{38}. \end{aligned}$$

Povprečni izid ene same stave na rdeče poiščemo na običajni način:

$$\mu = 1 \cdot \frac{18}{38} + (-1) \cdot \frac{20}{38} = -\frac{2}{38} = -0,053.$$

Zakon velikih števil nam pove, da je povprečje μ povprečni izid po velikem številu stav. Na dolgi rok torej igralci izgubimo (in igralnica zasluži) povprečno 5,3 centa na stavo. ♦

Kadar za mnenje vprašamo le eno osebo, ne vidimo nobene normalne krivulje. Situacija je enaka v primeru, ko igralec le enkrat stavi na rdeče pri ruleti. Vendar pa nam centralni limitni izrek zagotavlja, da bo povprečni izid velikega števila stav sledil porazdelitvi, ki bo blizu normalne. Recimo, da tekom večera postavimo 50 stav. Povprečni izid $\overline{x}$ je skupni dobiček (ali izguba), deljena s 50. Če zmagamo v 30 in zgubimo v 20 primerih, je skupni dobiček 10 €, povprečje pa $\overline{x} = 0,20\,€$. Če nadaljujemo z igranjem večer za večerom, se bo naš povprečni dobitek vsak dan spreminjal. Histogram teh vrednosti bo sledil normalni porazdelitvi. Na sliki 3.15 so prikazani rezultati velikega števila po 50 zaporednih stav. Preko njih narisana normalna krivulja, je porazdelitev, ki jo napoveduje centralni limitni izrek za ta primer.

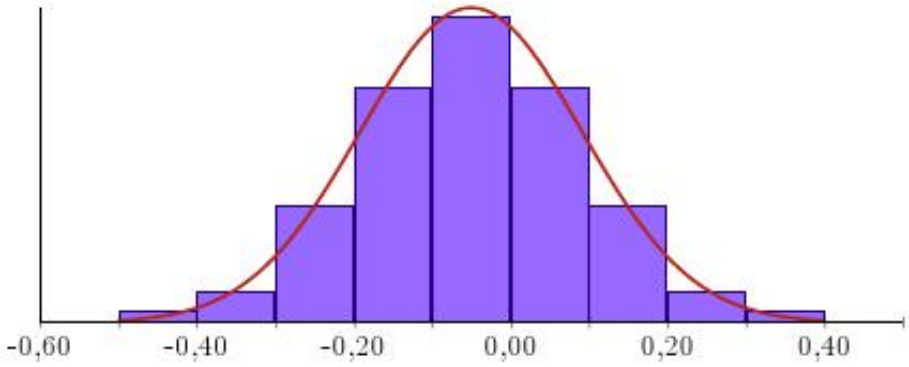
Slika 3.14: Igralec pri ruleti lahko izgubi ali zmaga, na dolgi rok pa zmaguje igralnica.

Vemo, da je srednja vrednost normalne porazdelitve s slike 3.15 enaka povprečni vrednosti ene same stave, -0,053. Kolikšen je standardni odklon? Vemo, da je enak $\frac{\sigma}{\sqrt{50}}$, kjer je σ standardni odklon porazdelitve posameznih stav. Posamične stave nimajo normalne porazdelitve, vendar pa je njihova porazdelitev tako preprosta, da lahko standardni odklon poiščemo z upoštevanjem ugotovitve, da je varianca povprečen kvadrat odklona od srednje vrednosti. Kot v primeru srednje vrednosti dobimo tu "povprečje" z uporabo verjetnosti posameznih izidov.

Recimo, da so možni izidi $s_1, s_2, \ldots, s_n$ iz vzorčnega prostora S števila in da je p_j verjetnost dogodka s_j. **Varianca** σ^2 tega verjetnostnega modela je

$$\sigma^2 = (s_1 - \mu)^2 p_1 + (s_2 - \mu)^2 p_2 + \ldots + (s_n - \mu)^2 p_n.$$

Standardni odklon σ je kvadratni koren iz variance.



Slika 3.15: Porazdelitev zmag pri ruleti ob večkratnem stavljenju na rdeče ali črno.

Primer. (Rdeče ali črno) Videli smo, da je povprečje pri stavi na rdeče (ali na črno) enako $\mu = -0,053$. Varianca in standardni odklon izidov ene same stave sta

$$\begin{aligned}
\sigma^2 &= (1 - (-0,053))^2 \cdot \frac{18}{38} + (-1 - (-0,053))^2 \cdot \frac{20}{38} = \\
&= 1,053^2 \cdot \frac{18}{38} + (-0,947)^2 \cdot \frac{20}{38} = \\
&= 0,9972, \\
\sigma &= \sqrt{0,9972} = 0,9986.
\end{aligned}$$

Standardni odklon povprečnega izida $\overline{x}$ pri 50 stavah je torej

$$\frac{\sigma}{\sqrt{n}} = \frac{0,9986}{\sqrt{50}} = 0,14.$$

Prepričaj se v pravilnost te trditve tako, da poiščeš na sliki 3.15 točki, kjer se krivulji spremeni ukrivljenost. ♦

Kaj bo torej izkusil kronični igralec, ki vsak večer odigra 50 iger? Skoraj vsi povprečni dnevni dobitki bodo ležali v območju treh standardnih odklonov od srednje vrednosti, torej med

$$-0,053 + 3 \cdot 0,14 = 0,367$$

in

$$-0,053 - 3 \cdot 0,14 = -0,473.$$

Skupni dobiček po 50 stavah bo torej ležal med

$$0,367 \cdot 50 = 18,35$$

in

$$-0,473 \cdot 50 = -23,65.$$

Državne loterije v ZDA

Igranje iger na srečo sega v preteklost vse do antike. V zgodnjih letih ZDA so bile pogoste tako javne kot zasebne loterije. Kasneje so za kakšno stoletje izginile, a se je igranje, upravljano s strani države, vrnilo leta 1964. V New Hampshireu so povzročili naval navdušenja, ko so ustanovili loterijo z namenom povečati državne dohodke brez povečanja davkov. Navdušenje se je hitro poleglo, ko so večje države posvojile to idejo, dokler ni večina držav sponzorirala svoje loterije. Državne loterije so hazarderstvo spremenile v sprejemljivo obliko zabave. V letu 1998 le tri države niso poznale nobene oblike legalnega hazarderstva.

Zveznim vladam je bila všeč zamisel, da bi svoje dohodke povečale brez zvišanja davkov. Igralnice pa po drugi strani zahtevajo visoke socialne žrtve v obliki povečanega kriminala in zasvojencev z igrami na srečo, ki zapravijo vse svoje in družinsko premoženje. Ena od študij, ki so jo izvedli na University of Illinois, je ocenila, da so s tem povzročeni stroški približno trikrat višji od zneska, ki ga v blagajne prispeva loterija.

Najbolj priljubljena igra na ameriških loterijah je Loto, pri katerem igralci izberejo (na primer) 6 števil izmed 49 in upajo, da jim je uspelo uganiti naključno izbrane zmagovalne številke. Možnosti za dobitek so neskončno majhne (6 991 908 proti 1 za šestico). Seveda pa so dobitki velikanski. Loterija je precej slaba stava, ker država izplača le približno polovico vsega stavljenega denarja. Veliko drugih iger na srečo je precej bolj radodarnih, na primer ruleta, o čemer se prepričamo tudi v tem poglavju.

Redni igralci Lota se ne morejo zanašati niti na centralni limitni izrek. Povprečni dobitek za 1€ je 50 centov, vendar pa to povprečje zajema peščico zmagovalcev in velikansko število poražencev. Variacija pri eni sami igri je velikanska in nobeno človeško izvedljivo število iger je ne more tako zmanjšati, da bi lahko prišli do kakšne uporabne napovedi. Pri tem nam ne pomaga niti $\sqrt{n}$. Edino, kar večina igralcev Lota dobi za svoj denar, je zadovoljstvo, ki ga občutijo, ko si predstavljajo, kako bodo obogateli. Lastnikom igralnic gre bolje. Eden izmed njih, Donald Trump, je rekel: "Še nikoli v življenju nisem igral iger na srečo. Hazarderji so ljudje, ki igrajo na avtomatih. Jaz sem raje lastnik teh igralnih avtomatov."

Igralec lahko torej dobi do $18,35$€, izgubi pa do $23,65$€. Igre na srečo so zanimive, ker so izidi tudi po celem večeru igranja še vedno negotovi. Mogoče je zapustiti preprogo kot zmagovalec. Vse je odvisno od sreče.

Igralnica pa je v povsem drugačnem položaju. Ne išče vznemirjenja, le konstanten dotok denarja. Kasino igra z vsako stranko, recimo 100 000 stav na rdeče ali črno na teden. Porazdelitev povprečnih zaslužkov njihovih strank pri 100 000 stavah je zelo blizu normalni in srednja vrednost je še vedno povprečni zaslužek ene stave, -0,053, torej izguba 5,3 centa pri stavi enega evra. Standardni odklon je veliko manjši, ko ga povprečimo po 100 000 stavah. Enak je

$$\frac{\sigma}{n} = \frac{0,9986}{\sqrt{100\,000}} = 0,003.$$

Takole izgleda razpon povprečnih rezultatov za igralnico po 100 000 stavah:

$$\begin{aligned} \text{Razpon} &= \text{Srednja vrednost} \pm 3 \text{ standardni odkloni} = \\ &= -0,053 \pm 3 \cdot 0,003 = \\ &= -0,053 \pm 0,009 = \\ &= -0,044 \text{ do } -0,062. \end{aligned}$$

Ker igralnica stavi velikokrat, postane standardni odklon povprečnega zaslužka zelo majhen. Ker je srednja vrednost negativna, bodo skoraj vsi izidi negativni. Izgube igralcev in dobitki igralnice so skoraj gotovo med 4,4 in 6,2 centa za vsak stavljeni evro.

Igralci, ki skupaj postavijo 100 000 stav, bodo izgubili denar. Verjeten obseg njihovih izgub je

$$-0,044 \cdot 100\,000 = -4400,$$
$$-0,062 \cdot 100\,000 = -6200.$$

Igralci bodo torej skoraj zagotovo izgubili – in igralnica zaslužila – med 4400 € in 6200 € pri teh 100 000 stavah. Še več, centralni limitni izrek nam pove, da je obseg možnih izidov vedno ožji, ko se število stav povečuje. To je razlog, da se igralnicam splača poslovati. Več denarja kot stavimo, bolj natančno lahko igralnica predvidi svoje zaslužke.

3.12 Slovarček

centralni limitni izrek (ang. central limit theorem) Povprečna vrednost velikega števila neodvisnih slučajnih izidov je približno normalno porazdeljena. Ko povprečimo n neodvisnih ponovitev istega slučajnega pojava, je povprečje porazdelitve izidov enako povprečnemu izidu enega samega poskusa, standardni odklon pa je sorazmeren z $\frac{1}{\sqrt{n}}$.

disjunktni dogodki (ang. disjoint events) Dogodki, ki nimajo nobenega skupnega izida.

dogodek (ang. event) Vsaka zbirka možnih izidov slučajnega pojava. Dogodek je podmnožica vzorčnega prostora.

enako verjetni izidi (ang. equally likely outcomes) Kadar je za vse možne izide slučajnega pojava verjetnost, da se zgodijo, enaka, govorimo o enako verjetnih izidih. Če je vseh izidov n, potem se vsak zgodi z verjetnostjo $\frac{1}{n}$.

kombinatorika (ang. combinatorics) Veja matematike, ki se ukvarja s prešteva-njem ureditev objektov.

normalne porazdelitve (ang. normal distributions) Družina verjetnostnih mode-lov, pri kateri je verjetnost dogodka določena s ploščino pod ustrezno normalno krivuljo. Normalne krivulje so simetrične in zvončaste oblike. Vsaka je na-tančno določena z μ in s standardnim odklonom σ.

pravilo 68–95–99,7 (ang. 68–95–99.7 rule) V vsaki normalni porazdelitvi leži 68% izidov znotraj enega standardnega odklona okoli srednje vrednosti; 95% izidov leži znotraj intervala s polmerom dveh standardnih odklonov okoli srednje vrednosti; 99.7% izidov leži manj kot tri standardne odklone stran od srednje vrednosti.

pravilo vsote za disjunktne dogodke (ang. addition rule for disjoint events) Če sta dva dogodka disjunktna, je verjetnost, da se zgodi vsaj eden od njiju, enaka vsoti verjetnosti, da se zgodi prvi, in verjetnosti, da se zgodi drugi.

slučajni pojav (ang. random phenomenon) Pojav je slučajen, če ne moremo pred-videti naslednjega izida, vseeno pa se vsak od izidov pojavlja v nekem točno določenem odstotku primerov v dolgem zaporedju ponovitev; ti deleži ponovi-tev so ravno verjetnosti posameznih izidov.

srednja vrednost verjetnostnega modela (ang. mean of a probability model) Povprečni izid slučajnega pojava z numeričnimi vrednostmi. Če imajo možni izidi $s_1, \ldots, s_n$ verjetnosti $p_1, \ldots, p_n$, potem je srednja vrednost enaka uteže-nemu povprečju izidov: $\mu = p_1 s_1 + \ldots + p_n s_n$.

standardni odklon normalne krivulje (ang. standard deviation of a normal curve) Standardni odklon σ normalne krivulje je razdalja med srednjo vre-dnostjo in prevojem na katerikoli strani.

standardni odklon verjetnostnega modela (ang. standard deviation of a probability model) Mera spremenljivosti verjetnostnega modela; če imajo možni izidi $s_1, \ldots, s_n$ verjetnosti $p_1, \ldots, p_n$, potem je varianca enaka

$$\sigma^2 = (s_1 - \mu)^2 p_1 + \ldots + (s_n - \mu)^2 p_n,$$

standardni odklon pa je σ, t.j. kvadratni koren iz variance.

statistika (ang. statistic) Število, ki ga izračunamo iz vzorca, na primer srednja vrednost vzorca; pri slučajnem vzorčenju se vrednosti statistik za različne vzorce razlikujejo.

verjetnost (ang. probability) Število med 0 in 1, ki pove, kako pogosto se pojavi nek dogodek pri veliko ponovitvah poskusa.

verjetnostni histogram (ang. probability histogram) Histogram, s katerim predstavimo verjetnostni model v primeru numeričnih izidov. Višina vsakega stolpca je verjetnost izida ali skupine izidov, ki jih ta stolpec predstavlja.

verjetnostni model (ang. probability model) Vzorčni prostor S skupaj s predpisom, ki vsakemu dogodku priredi njegovo verjetnost. Verjetnost $P(s)$, predpisana dogodku $s \in S$, je število med 0 in 1. Vsota verjetnosti vseh izidov mora biti enaka 1. Verjetnostni model lahko dogodkom priredi verjetnosti tudi kot ploščine pod neko krivuljo. V tem primeru mora biti celotna ploščina pod krivuljo enaka 1.

vzorčna porazdelitev (ang. sampling distribution) Porazdelitev vrednosti, ki jih zavzame statistika, ko izberemo veliko slučajnih vzorcev pod enakimi okoliščinami. Vzorčno porazdelitev sestavlja predpis, ki možnim vrednostim statistike priredi verjetnosti.

vzorčna spremenljivost (ang. sampling variability) naključna odstopanja v vrednosti statistike, ko iz iste populacije izbiramo več slučajnih vzorcev.

vzorčni prostor (ang. sample space) Množica vseh možnih izidov slučajnega pojava.

zakon velikih števil (ang. law of large numbers) Ko slučajni pojav velikokrat ponovimo, se povprečje $\overline{x}$ opaženih vrednosti približuje srednji vrednosti μ.

3.13 Dodatna literatura

- Mosteller, Frederick, Robert E. K. Rourke, George B. Thomas. *Probability with Statistical Applications*, Addison–Wesley, Reading, Mass., 1970. Obširna, a rahlo prefinjena obravnava osnov verjetnosti, ki zahteva le znanje srednješolske algebre.

- Olkin, Ingram, Leon J. Glesser, Cyrus Derman. *Probability Models and Applications*, 2. izdaja, Macmillan, New York, 1994. Ta knjiga izstopa s poudarjanjem uporabe verjetnosti pri resničnih pojavih in z odličnimi primeri modeliranja. Po nivoju je nekje med obema drugima.

- Snell, J. Laurie, *Introduction to Probability*, Random House, New York, 1988. Temelji na diferencialnem računu in je namenjena dodiplomskim študentom matematike. Tu jo priporočamo, ker vsebuje odlične primere in zgodovinske komentarje, še posebej pa zato, ker uporablja programe v BASICu, ki so vključeni v besedilo.

Na internetu lahko najdemo animirane simulacije, ki ilustrirajo pomembne ugotovitve o verjetnosti in vzorčnih porazdelitvah. Večina teh strani pripada šolam, univerzam ali statističnim uradom. Na žalost se pogosto spreminjajo. Brskanje po spletu za ključnimi besedami tipa "probability applet" bo najbrž vrnilo več rezultatov, kot jih imaš čas pregledati.

Že leta 1999 je David Lane z Rice University,

`http://www.ruf.rice.edu/~lane/hyperstat/index.html`,

objavil številne programčke in dober seznam s tem povezanih projektov.

3.14 Preverjanje znanja

(1) Vržemo kovanec in kocko. Nato pogledamo, koliko pik smo vrgli na kocki in ali je kovanec pokazal glavo ali cifro. Koliko izidov vsebuje vzorčni prostor?

(a) 6

(b) 8

(c) 12

(2) Vzorčni prostor vsebuje tri izide, A, B in C. Kateri od naslednjih primerov predstavlja pravilno prireditev verjetnosti tem izidom?

(a) $P(A) = 0{,}3$, $P(B) = 0{,}6$, $P(C) = 0{,}1$.

(b) $P(A) = 0{,}5$, $P(B) = 0{,}4$, $P(C) = 0{,}4$.

(c) $P(A) = 0{,}7$, $P(B) = -0{,}2$, $P(C) = 0{,}5$.

(3) V predalu so 3 črne in 7 modrih nogavic. Sežemo v predal in na slepo potegnemo iz njega eno nogavico. Kolikšna je verjetnost, da je ta nogavica črna?

(a) $\frac{3}{10}$

(b) $\frac{3}{7}$

(c) $\frac{1}{2}$

(4) Kovček ima kombinacijsko ključavnico, ki jo odpremo z izbiro pravega zaporedja treh števk med 0 in 9. Koliko je različnih možnih kod?

(a) 30

(b) 729

(c) 1000

(5) Vsaka srečka stane 2 €. Izmed 200 prodanih srečk bo ena zadela 100 € in ena 50 €. Kolikšna je povprečna vrednost ene igre?

(a) 0,75 €

(b) 1,25 €

(c) -1,25 €

(6) Življenjska doba baterije v uri je normalno porazdeljena s srednjo vrednostjo 2 leti in standardnim odklonom 0,75 leta. Kolikšna je verjetnost, da bo baterija zdržala manj kot 6 mesecev?

(a) 5%

(b) 2,5%

(c) manj kot 1%

(7) Življenjska doba baterije v uri je normalno porazdeljena s srednjo vrednostjo 2 leti in standardnim odklonom 0,75 leta. Recimo, da iz proizvodnje izberemo vzorec 16 baterij in jih preiskusimo. Kolikšen je standardni odklon $\sigma_{\overline{x}}$ povprečnega rezultata?

(a) 0,75 leta

(b) 0,1875 leta

(c) 0,0469 leta

3.15 Naloge

Kaj je verjetnost?

(1) Postavi kovanec na njegov rob in ga pridrži s kazalcem. Nato ga frcni z drugim kazalcem, da se nekajkrat zavrti, preden pade na podlago. Na osnovi 50 poskusov oceni, kolikšna je verjetnost, da pade glava.

(2) Morda se ti zdi očitno, da je verjetnost glave pri metu kovanca približno $\frac{1}{2}$, ker ima kovanec dve strani. Takšna prepričanja niso vedno točna. V prejšnji nalogi smo kovanec frcali, namesto da bi ga metali. Pri tem se je verjetnost, da pade glava, spremenila. Poskusimo še z eno različico. Postavi kovanec na rob na trdi, ravni podlagi. Nato udari po njej, da se kovanec prevrne. Kolikšna je verjetnost, da pade glava? Napravi vsaj 50 poskusov, da oceniš verjetnost.

(3) Odpri lokalni telefonski imenik na poljubni strani in si za prvih 100 telefonskih številk na tej strani zapiši, ali je zadnja števka liha ali soda. Koliko števk je bilo lihih? Kolikšna je približno verjetnost, da je zadnja števka telefonske številke liha?

(4) Tabelo naključnih števil 1.1 smo dobili s pomočjo slučajnega mehanizma, pri katerem je vsaka od števk z verjetnostjo 0,1 enaka 0. Kolikšen delež prvih 200 števil v tabeli predstavlja število 0? Ta delež je ocena za dejansko verjetnost, za katero v tem primeru vemo, da je enaka 0,1.

(5) V knjigi o pokru prebereš, da je verjetnost, da med petimi kartami dobiš tri enake $\frac{1}{50}$. Razloži, kaj to pomeni.

(6) Verjetnost je merilo za možnost, da se zgodi nek dogodek. Vsakemu od spodnjih dogodkov pripiši eno od verjetnosti, ki sledijo. (Verjetnost je običajno precej bolj natančna ocena možnosti kot pa besedni opis.)

$$0 \quad 0{,}01 \quad 0{,}3 \quad 0{,}6 \quad 0{,}99 \quad 1$$

(a) Ta dogodek je nemogoč. Nikoli se ne more zgoditi.

(b) Ta dogodek je gotov. Pojavil se bo pri vsaki ponovitvi slučajnega poskusa.

(c) Ta dogodek je zelo malo verjeten, vendar pa se bo v dolgem zaporedju ponovitev občasno pojavil.

(d) Ta dogodek se bo večkrat zgodil kot ne.

Verjetnostni modeli in pravila

V nalogah 7, 8 in 9 poišči smiseln vzorčni prostor za omenjene slučajne pojave. V nekaterih primerih je več možnih izbir.

(7) Desetkrat vržemo kovanec.

(a) Preštej število glav.

(b) Izračunaj, kolikšen delež vseh izidov predstavljajo glave.

(c) Ali se je pojavilo vsaj 5 glav ali ne?

(8) Pričakujemo novo leglo laboratorijskih podgan in prešteli bomo število novorojenih primerkov. (Ne vemo, kako velika so običano legla, lahko pa postaviš neko smiselno zgornjo mejo, če želiš.)

(9) Osebke v klinični študiji slučajno delimo v testno in kontrolno skupino. Za naslednjega si zabeležiš, v kateri skupini je, katerega spola je in ali kadi.

(10) Vsa človeška kri spada v eno od štirih krvnih skupin: 0, A, B ali AB, vendar pa je porazdelitev pri različnih rasah nekoliko različna. V spodnji tabeli je prikazana porazdelitev krvnih skupin pri slučajno izbranih Afroameričanih.

Krvna skupina	0	A	B	AB
Verjetnost	0,49	0,27	0,20	?

(a) Kolikšna je verjetnost, da spada kri v skupino AB? Zakaj?

(b) Maria ima krvno skupino B. Varno lahko sprejme kri skupin 0 in B. Kolikšna je verjetnost, da lahko naključno izbran Afroameričan Marii daruje kri?

(11) Če iz velike vreče bonbonov *M&M* slučajno izberemo en bonbon, bo ta ene izmed šestih možnih barv. Verjetnost vsake od barv je odvisna od deleža, ki ga med vsemi proizvedenimi bonboni predstavlja ta barva.

(a) V spodnji tabeli so verjetnosti vsake od barv pri slučajni izbiri navadnih *M&M* bonbonov. Kolikšna je verjetnost, da izberemo modri bonbon?

Barva	Rjava	Rdeča	Rumena	Zelena	Oranžna	Modra
Verjetnost	0,3	0,2	0,2	0,1	0,1	?

(b) Pri *M&M* bonbonih z arašidi so verjetnosti nekoliko drugačne. Tule so:

Barva	Rjava	Rdeča	Rumena	Zelena	Oranžna	Modra
Verjetnost	0,2	0,1	0,2	0,1	0,1	?

Kolikšna je verjetnost, da je naključno izbran arašidov *M&M* modre barve?

(c) Kolikšna je verjetnost, da je navaden *M&M* rdeč, rumen ali oranžen? Kolikšna je verjetnost, da je ene od teh barv *M&M* z arašidom?

(12) Lasvegaškega vedeževalca Zokija smo prosili za napoved izidov nekega prvenstva v košarki. Kot je dandanes v navadi, je Zoki v odgovor vpletel verjetnosti. Pravi, "Racmani imajo dvakrat tolikšne možnosti za zmago kot Carji. Za Strele je verjetnost zmage 0,1 in prav to velja tudi za Viharje, Carji imajo trikrat večje možnosti. Ostale ekipe nimajo niti najmanjše šanse." Ali je Zokijeva razporeditev verjetnosti za zmago med osem ekip, ki so udeležene v prvenstvu, prepričljiva? Odgovor utemelji.

(13) V ZDA pri vsakem primeru smrti zabeležijo vzrok (in sicer en sam). Podatki kažejo, da je z verjetnostjo 0,45 slučajno izbrani nesrečnež umrl zaradi kardiovaskularnih bolezni in z verjetnostjo 0,22 je bil vzrok rak. Kolikšna je verjetnost, da je smrt povzročil rak ali kardiovaskularna bolezen? S kolikšno verjetnostjo je bil vzrok kje drugje?

Enako verjetni izidi

(14) Abby, Deborah, Julie, Sam in Roberto delajo v istem podjetju v oddelku za stike z javnostjo. Njihov delodajalec mora izbrati dva izmed njih, ki se bosta udeležila konference v Parizu. Da bi se izognili pristranskosti, bodo izbiro izpeljali z vlečenjem listkov iz klobuka. (Gre torej za enostavni slučajni vzorec velikosti 2.)

(a) Naredi seznam vseh možnih izbir dveh imen izmed petih. To je vzorčni prostor.

(b) Ker izbiramo slučajno, so vse izbire enako verjetne. Kolikšna je verjetnost za vsako od izbir?

(c) Kolikšna je verjetnost, da bo izbrana Julie?

(d) Kolikšna je verjetnost, da ne bosta izbrana niti Sam niti Roberto?

(15) Nek mladi par si želi tri otroke. Skupaj je 8 možnih razporeditev deklic in dečkov. Dogovorimo se, da oznaka ŽŽM pomeni, da sta prva dva otroka deklici in tretji deček, in podobno za ostale možnosti. Vseh 8 razporeditev je (približno) enako verjetnih.

(a) Napravi seznam vseh razporeditev. Kolikšna je verjetnost vsake izmed njih?

(b) Na podlagi tega verjetnostnega modela izdelaj verjetnostni model za število deklic med tremi otroki. Nariši verjetnostni histogram.

(c) Uporabi model iz prejšnje točke in izračunaj verjetnost, da sta med tremi otroki vsaj dve deklici.

(d) Vrnimo se k modelu iz točke (a). Uporabi ta model za izračun verjetnosti, da sta med tremi otroki vsaj dve deklici. Primerjaj rezultat s tistim iz prejšnje točke.

(16) Računalnik dodeljuje uporabnikom tričrkovne identifikacijske kode za prijavo v sistem. Pri tem uporablja črke angleške abecede. Med samoglasnike štejemo a, e, i, o, u in y. Kolikšna je verjetnost, da v kodi, ki jo dobimo, ni samoglasnikov, če dovoljujemo ponovitve? Kaj pa, če se črke ne smejo ponavljati?

(17) Računalnik iz prejšnje naloge preprogramiramo tako, da so vse kode oblike soglasnik–samoglasnik–soglasnik. Oboje še vedno izbiramo naključno in soglasniki se lahko ponovijo. Kolikšna je verjetnost, da koda ne bo vsebovala črke x?

(18) Recimo, da lahko kode iz prejšnjih nalog vsebujejo tudi števila od 0 do 9, poleg tega pa dovolimo ponavljanja. Kolikšna je zdaj verjetnost, da koda ne vsebuje črke x? Kaj pa verjetnost, da ne vsebuje nobene številke?

(19) Osebne identifikacijske številke PIN so običajno štirimestne. Opaziš, da vsebuje večina tvojih PIN kod vsaj eno ničlo in se začneš spraševati, če proizvajalci kartic in naprav morda namenoma uporabljajo veliko ničel, da si je kode lažje zapomniti. Kolikšna je verjetnost, da vsebuje naključno izbrani PIN vsaj eno ničlo?

Srednja vrednost verjetnostnega modela

(20) Vržemo pošteno kocko. Izračunaj povprečno vrednost števila pik.

(21) Za verjetnostni model iz naloge 15 (a) izračunaj povprečno število deklic med tremi otroki.

(22) Profesor ekonomije ocenjuje svoje študente z ocenami od 0 do 4. Porazdelitev ocen je naslednja:

Razred	0	1	2	3	4
Verjetnost	0,10	0,15	0,30	0,30	0,15

Poišči povprečje (se pravi, srednjo vrednost) ocene pri tem predmetu. Izdelaj verjetnostni histogram za porazdelitev ocen in na njem označi izračunano srednjo vrednost.

(23) Ameriška državna loterija je predstavila igro *Izberi 3*, pri kateri ponujajo igralcu več možnih stav. Igralci izberejo trimestno število in stavijo 1 $. Loterija vsak večer oznani zmagovalno trimestno število, ki ga izbere slučajno. Če igralec ugane vsa tri števila v poljubnem vrstnem redu, dobi $83,33$ $, sicer pa izgubi vplačani dolar. Poišči povprečni dobitek. (Predpostavi, da igralec izbere število s tremi različnimi števkami.)

(24) V večini velikih ameriških mest je organizirana nekakšna ilegalna oblika lota, ki deluje na naslednji način: igralec izbere eno od 1000 trimestnih števil med 000 in 999 in plača posredniku en dolar, da lahko stavi na to število. Vsak dan slučajno izberejo eno trimestno število in če je igralec število uganil, dobi 600 $, sicer pa izgubi vplačani dolar. Kolikšen je povprečni dobiček? Janez že veliko let vsak dan stavi na eno število. Kaj pravi zakon velikih števil o Janezovem dobičku, medtem ko Janez veselo nadaljuje s stavami?

(25) Ameriška ruleta ima 38 žepkov, označenih z 0, 00 in s števili od 1 do 36. Ko zavrtimo kolo, kroglica z enako verjetnostjo pristane v kateremkoli od teh žepkov. Oznake žepkov so razporejene tudi po mizi, na kateri igralci postavljajo svoje stave. Eden od stolpcev vsebuje vse večkratnike števila 3, torej števila $3, 6, \ldots, 36$. Igralec položi 1 € na stolpec in prejme 3 €, če se kroglica ustavi na kateremkoli od teh števil.

(a) Kolikšna je verjetnost, da zmaga?

(b) Kolikšen je povprečni dobitek ene igre, če upoštevamo, da vsaka igra stane 1 €?

Vzorčne porazdelitve

(26) Ilustrirajmo idejo vzorčne porazdelitve z zelo majhnim vzorcem iz zelo majhne populacije. Populacija naj bo sestavljena iz rezultatov, ki jih je 10 študentov doseglo na izpitu:

Študent	0	1	2	3	4	5	6	7	8	9
Rezultat	82	62	80	58	72	73	65	66	74	62

Parameter, ki nas zanima, je povprečen rezultat μ. Vzorec je enostavni slučajni vzorec velikosti $n = 4$, ki ga izberemo iz naše populacije. Ker smo študente označili s številkami od 0 do 9, nam vsaka števka iz tabele naključnih števil 1.1 pomaga izbrati po enega študenta za vzorec.

(a) Izračunaj povprečje vseh 10 rezultatov. To je srednja vrednost populacije μ.

(b) Uporabi tabelo 1.1, da izbereš enostavni slučajni vzorec velikosti 4 iz te populacije. Izpiši ustrezne štiri rezultate in izračunaj povprečje $\overline{x}$. Ta statistika je ocena za μ.

(c) Desetkrat ponovi zgornji postopek in pri tem vsakič uporabi drug del tabele 1.1. Izdelaj histogram vseh 10 tako dobljenih vrednosti $\overline{x}$. Na ta način dobiš vzorčno porazdelitev $\overline{x}$. Ali je središče tvojega histograma blizu μ?

(27) V tabeli 2.8 so podani časi preživetja 72 morskih prašičkov, ki so bili del medicinskega eksperimenta. Teh 72 morskih prašičkov bo naša populacija.

(a) Nariši histogram vseh 72 časov preživetij. Ta populacija je močno desno asimetrična.

(b) Poišči povprečen čas preživetja. To je srednja vrednost populacije μ. Označi μ na x-osi svojega histograma.

(c) Označi člane populacije s števili med 01 in 72 in uporabi tabelo 1.1, da izbereš enostavni slučajni vzorec velikosti $n = 12$. Izračunaj povprečni čas preživetja $\overline{x}$ za ta vzorec. Označi vrednost $\overline{x}$ s točko na osi histograma iz (a).

(d) Izberi še štiri enostavne slučajne vzorce velikosti 12 s pomočjo različnih delov tabele 1.1. Za vsakega od vzorcev poišči $\overline{x}$ in ustrezno vrednost označi na histogramu iz točke (a). Ali bi bilo zelo nenavadno, če bi vseh pet izračunanih $\overline{x}$ ležalo na isti strani μ? Zakaj?

(e) Recimo, da bi iz te populacije izbrali veliko število enostavnih slučajnih vzorcev velikosti 12 in narisali histogram vrednosti $\overline{x}$. Kje bi po tvojem mnenju ležalo središče te vzorčne porazdelitve?

Normalne porazdelitve

(28) Porazdelitev višin odraslih ameriških moških je približno normalna s srednjo vrednostjo 69 inčev in standardnim odklonom 2,5 inčev. Nariši normalno krivuljo, na kateri sta pravilno označena srednja vrednost in standardni odklon. (Namig: Najprej nariši krivuljo nato pa oznake na vodoravni osi.)

(29) Z uporabo normalne porazdelitve, opisane v nalogi 28, in pravila 68–95–99,7 odgovori na naslednja vprašanja o višinah odraslih Američanov.

(a) Kolikšen delež moških je višjih od 74 inčev?

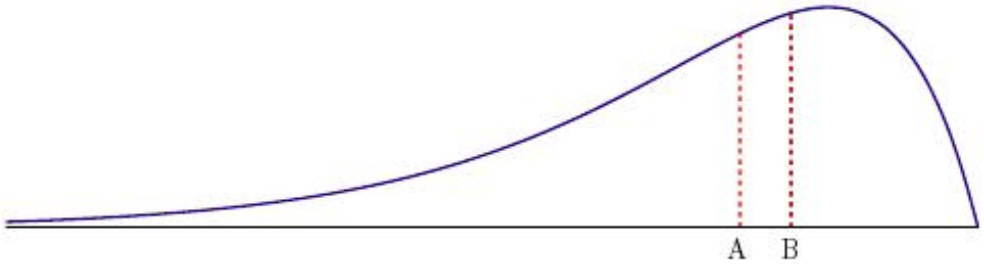
(b) Med katerima dvema višinama leži srednjih 95 % višin?

(c) Kolikšen delež moških je nižjih od 66,5 inčev?

(30) Kje so kvartili porazdelitve višin iz naloge 28?

(31) Število napak na kvadratni meter preproge se spreminja, srednja vrednost je 1,6, standardni odklon pa 1,2 napake na kvadratni meter. Porazdelitev ne more biti normalna, ker je število napak vedno celo. Inšpektor v vzorec izbere 200 kvadratnih metrov preprog, zabeleži število napak, ki jih je našel v vsakem kvadratnem metru, in izračuna $\overline{x}$, povprečno število napak na kvadratni meter v pregledanem vzorcu. Take preglede opravi večkrat. Na katerem intervalu leži srednjih 95% vseh $\overline{x}$?

(32) Na sliki 3.16 je prikazana nesimetrična verjetnostna porazdelitev. Srednja vrednost in mediana ne sovpadata. Katera od označenih črt je srednja vrednost porazdelitve in katera mediana? Odgovor utemelji.



Slika 3.16: Asimetrična porazdelitev.

(33) Rezultati standardnega inteligenčnega testa *Wechsler Adult Intelligence Scale* za skupino med 20. in 34. letom starosti so porazdeljeni približno normalno z $\mu = 110$ in $\sigma = 25$.

(a) Kolikšen delež ljudi iz te skupine doseže rezultate nad 110?

(b) Kolikšen delež teh ljudi ima rezultat, višji od 160?

(c) Poišči kvartila te porazdelitve. S preprostimi besedami razloži, kaj nam povesta ti dve števili.

(34) Vojska poroča, da je porazdelitev obsega glav vojakov približno normalna s srednjo vrednostjo 22,8 inčev in standardnim odklonom 1,1 inčev.

(a) Kolikšen delež vojakov ima obseg glave večji od 23,9 inčev?

(b) Vojska želi čelade pripraviti vnaprej in hoče, da bi se prilegale sredinskim 95 % vojakov. Preostalim vojakom bodo izdelali čelade posebej. Kateri obsegi so dovolj majhni ali pa dovolj veliki, da si bodo "prislužili" čelade po naročilu?

Centralni limitni izrek

(35) Rezultati sprejemnih izpitov SAT so približno normalno razporejeni s srednjo vrednostjo $\mu = 500$ in standardnim odklonom $\sigma = 100$.

(a) Naključno izberi enega kandidata. Kolikšna je verjetnost, da je rezultat izbranega kandidata večji od 500? Večji od 600?

(b) Izberi enostavni slučajni vzorec velikosti 4. Kolikšna je verjetnost, da je povprečje njihovih rezultatov večje od 500? Večje od 600?

(36) Juan v kemijskem laboratoriju izvede nekaj meritev in o rezultatih poroča v svojem laboratorijskem poročilu. Standardni odklon meritev, ki so jih dobili študenti, je $\sigma = 10$ mg. Juan ponovi meritve trikrat in izračuna povprečje $\overline{x}$ vseh treh rezultatov.

(a) Kolikšen je standardni odklon povprečja, ki ga je izračunal Juan? (Se pravi, če bi Juan nadaljeval z izvajanjem meritev in računal povprečja treh zaporednih rezultatov, kolikšen bi bil standardni odklon vseh tako dobljenih $\overline{x}$?)

(b) Kolikokrat mora Juan ponoviti meritve, da bi zmanjšal standardni odklon vrednosti $\overline{x}$ na 5? Razloži nekomu, ki ne zna statistike, zakaj je bolje poznati povprečje večih meritev kot pa rezultat ene same.

(37) Študentska organizacija namerava vprašati vzorec 50 študentov, če so opazili nove izobraževalne brošure o AIDSu. Zabeležili si bodo delež pozitivnih odgovorov. Njihov svetovalec za statistiko pravi, da bo standardni odklon tega deleža približno 7%. Kolikšen bi bil standardni odklon, če bi vzorec vseboval 100 študentov in ne 50?

(38) Kako velik vzorec bi morali izbrati v nalogi 37, da bi zmanjšali standardni odklon deleža pozitivnih odgovorov iz 7% na $3,5\%$? Pojasni nekomu, ki ne zna statistike, zakaj so večji vzorci pri raziskavah mnenja boljši kot manjši.

(39) V nalogi 25 smo poiskali povprečen dobitek pri stavi 1 € na stolpec pri ruleti. Seveda je povprečni dobitek negativen – na dolgi rok igralci izgubljajo in igralnica dobiva. Poišči standardni odklon za ta primer. Primerjaj svoj rezultat s tistimi, ki smo jih dobili v tem poglavju v primerih stave na rdeče ali črno. Ali je katera od teh stav boljša od drugih?

(40) V nalogi 24 smo poiskali povprečni dobitek v igri s števili. Poišči še standardni odklon za dobičke posamezne igre. Uporabi centralni limitni izrek, da podaš razpon (srednja vrednost $\pm$ 3 standardni odkloni) za dobiček, ki ga ima igralec v enem letu (po 365 igrah) in v 10 letih (po 3650 igrah).

Dodatne naloge

(41) Naključno izberi študenta in si zapiši, koliko denarja v bankovcih ima s seboj (ignoriraj kovance). Definiraj smiselni vzorčni prostor S za ta slučajni pojav. (Ne vemo, kakšna bi bila največja vrednost, ki bi jo lahko imel pri sebi študent, zato se moramo odločiti za neko smiselno vrednost, ko opisujemo vzorčni prostor.)

(42) Spodaj je porazdelitev zakonskega statusa Američank med 25. in 29. letom.

Izid	Nikoli poročena	Poročena	Vdovela	Ločena
Verjetnost	0,376	?	0,003	0,062

(a) Gre za resničen verjetnostni model. Kolikšna je verjetnost, da je ženska iz te skupine poročena?

(b) Pri tem modelu nima smisla govoriti o srednji vrednosti. Zakaj ne?

(43) Paket kart za bridge vsebuje 52 kart, po štiri vsake vrednosti: kralja, damo, fanta, desetko, devetko, ..., dvojko. Iz kupčka potegnemo eno karto in si zapišemo vrednost. Vsakemu od izidov pripiši verjetnost, ki bi jo imel, če bi bile karte pošteno premešane. Podaj še neko drugo možno porazdelitev verjetnosti (se pravi tako, ki upošteva pravila verjetnosti), ki se razlikuje od prve. Nato si izmisli še tretjo porazdelitev verjetnosti, ki pa ne bo legitimna. Pojasni, zakaj ni.

(44) Avtomobilske tablice v Indiani vsebujejo sedem znakov. Prva dva sta oznaka države, v kateri je avto registriran, tretja je črka in zadnje štiri so

številke. Kolikšna je verjetnost, da dobimo tablico, na kateri bodo zadnje štiri številke enake (na primer 7777)? Predpostavimo, da so črke in številke izbrane slučajno.

(45) Avtomobilske tablice na Havajih vsebujejo tri črke, ki jim sledijo še tri številke.

(a) Koliko je različnih možnih havajskih tablic?

(b) Ko obiščemo Honolulu, opazimo, da se vse tablice začenjajo z enim E, F, G ali H. Koliko tablic je možnih, če se lahko začnejo le s katero od teh štirih črk?

(c) Recimo, da bi država dovolila uporabo katerihkoli šestih črk ali številk v poljubnem vrstnem redu. Koliko različnih tablic bi bilo možnih v tem primeru?

(46) Opica se igra s pisalnim strojem in zadene črke i, m in t v naključnem vrstnem redu. Koliko tričrkovnih "besed" lahko sestavi? Katere od teh besed imajo v slovenščini nek pomen? Kolikšna je verjetnost, da je beseda, ki jo je opica sestavila, smiselna?

(47) Odločiš se obiskati svojega novega soseda. Veš, da so v družini štirje otroci, ne veš pa, koliko so stari in katerega spola so. Napravi seznam vseh možnih razporeditev deklic in dečkov, urejenih od najmanjšega do najstarejšega. Recimo MMŽŽ naj pomeni, da sta mlajša dva dečka, starejši dve pa deklici. Zakoni genetike pravijo, da so vse te razporeditve enako verjetne.

(a) Kolikšna je verjetnost, da je najstarejši otrok deklica?

(b) Kolikšna je verjetnost, da so v družini vsaj trije dečki?

(c) Kolikšna je verjetnost, da so vsaj trije od otrok istega spola?

(48) Pri neki študiji smo izbrali vzorec učencev petega razreda in si zapisali, koliko let šolanja so sčasoma dokončali. Na osnovi te študije lahko podamo naslednji verjetnostni model za leta šolanja, ki jih bo sčasoma končal slučajno izbrani petošolec:

Leta	4	5	6	7	8
Verjetnost	0,010	0,007	0,007	0,013	0,032
Leta	9	10	11	12	
Verjetnost	0,068	0,070	0,041	0,752	

(a) Prepričaj se, da ta verjetnostni model zadošča pravilom.

(b) Kateri izidi sestavljajo dogodek "učenec je končal vsaj eno leto srednje šole"? (Srednja šola se začne z devetim letom šolanja.) Kolikšna je verjetnost tega dogodka?

(c) Izračunaj povprečno število dokončanih let šolanja.

(49) V igralnicah je zelo popularna igra, imenovana Keno. Kroglice, označene s števili med 1 in 80, premešamo v posebni napravi, medtem ko lahko igralci stavijo tako, da označijo številke na kartici. Nato naključno izberemo 20 kroglic. Spodaj sta dva primera stav. Za vsakega napravi verjetnostni model izidov ter poišči srednjo vrednost in standardni odklon dobitkov. Ali je katera od stav boljša od druge?

(a) Pri stavi 1 € na "Označi eno!" dobimo 3 €, če je označeno število med 20 izbranimi, sicer pa izgubimo vloženi evro.

(b) Pri stavi 1 € na "Označi dve!" dobimo 12 €, če sta obe označeni števili med 20 izbranimi. Verjetnost tega dogodka je približno 0,06. Ali je bolje igrati "Označi eno!" ali "Označi dve!"?

(50) Obstaja preprost način za izdelavo verjetnostnega modela z danima srednjo vrednostjo μ in standardnim odklonom σ: izida sta samo dva, in sicer $\mu - \sigma$ ter $\mu + \sigma$, vsak pa se pojavi z verjetnostjo 0,5. S pomočjo definicij srednje vrednosti in variance verjetnostnega modela pokaži, da je res srednja vrednost takega modela enaka μ in standardni odklon σ.

(51) Psiholog Amos Tversky se je veliko ukvarjal s človeškim pojmovanjem slučajnosti. V njegovi osmrtnici (6. junij 1996) so v časopisu *New York Times* navajali naslednji primer:

(a) Tversky je osebke prosil, da izberejo med dvema zdravstvenima programoma, ki bi vplivala na 600 ljudi. V prvem bi imeli $\frac{1}{2}$ možnosti, da rešimo vseh 600 in prav toliko možnosti, da vseh 600 umre. V drugem bi zagotovo preživelo natanko 400 ljudi. Poišči povprečno število ljudi, ki jih reši prvi program.

(b) Tversky jim je nato ponudil drugačno izbiro. Pri prvem programu spet z enako verjetnostjo rešimo ali pa izgubimo vseh 600 ljudi, medtem ko pri drugem zagotovo izgubimo natanko 200 ljudi. V čem je razlika med to izbiro in tisto iz točke (a)?

(c) Pri izbiri iz točke (a) je večina vprašanih izbrala drugi program, pri izbiri iz (b) pa so izbrali prvega. Ali si vprašani pri odločanju pomagajo s srednjimi vrednostmi? Zakaj sta se po tvojem mnenju njihovi izbiri razlikovali?

(52) Spodnja tabela vsebuje rezultate 100 ponovitev izbire enostavnega slučajnega vzorca velikosti 200 iz velike množice ležajev, med katerimi je 10% takšnih, ki ne ustrezajo specifikacijam. Števila v tabeli so odstotki neustreznih ležajev v vsakem od vzorcev.

8,5	11,5	9,0	13,5	7,5	8,5	9,0	6,5	8,0	9,0
10,0	7,5	9,0	8,0	10,5	8,5	9,0	9,5	8,0	11,5
10,0	9,0	9,0	8,5	9,5	6,5	13,5	11,0	11,5	13,0
8,5	6,5	8,0	7,0	12,0	11,0	8,0	10,5	12,0	10,5
15,0	12,0	8,5	7,0	8,0	8,0	8,5	12,0	10,5	8,0
8,5	11,5	9,0	11,5	11,0	12,0	11,5	11,5	10,0	9,5
10,0	9,0	10,0	12,5	8,0	12,0	12,0	12,0	7,5	11,0
11,0	8,0	14,0	7,5	11,0	4,5	9,5	8,0	9,5	9,5
12,5	12,0	10,0	7,5	10,5	12,5	12,0	9,5	9,5	10,0
14,0	9,0	8,5	8,5	12,5	8,5	8,5	9,0	9,5	9,0

Podaj približno vzorčno porazdelitev za deleže vzorcev, in sicer tako, da si za vsak izid zapišeš delež vzorcev, pri katerih se je pojavil. Nariši histogram porazdelitve in opiši njegovo obliko. Ali je središče blizu 10%? Ali je porazdelitev približno simetrična? Ali je približno normalna? Iz porazdelitve poišči srednjo vrednost izidov. Ali je blizu 10%?

(53) Koncentracija aktivne sestavine v kapsulah je porazdeljena normalno z $\mu = 10\%$ in $\sigma = 0{,}2\%$.

(a) Kolikšna je povprečna koncentracija? Odgovor utemelji.

(b) Na katerem intervalu se nahaja srednjih 95% vseh koncentracij?

(c) Na katerem intervalu se nahajajo koncentracije srednje polovice vseh kapsul?

(54) Odgovori na naslednji vprašanji o kapsulah iz naloge 53:

(a) Pri kolikšnem odstotku kapsul je koncentracija aktivne sestavine višja od 10,4%?

(b) Pri kolikšnem odstotku kapsul je koncentracija aktivne sestavine višja od 10,6%?

(55) Dolžina nosečnosti pri človeku se spreminja po porazdelitvi, ki je približno normalna s srednjo vrednostjo 266 dni in standardnim odklonom 16 dni.

(a) Med kateri vrednosti pade srednjih 95% vseh nosečnosti?

(b) Kako dolgih je najkrajših 2,5% nosečnosti?

(56) *Decila* porazdelitve sta točki, pod katerima leži 10% (spodnji decil) oziroma 90% (zgornji decil) vseh vrednosti. Med spodnjim in zgornjim decilom leži 80% vseh podatkov. Pri normalnih porazdelitvah se nahajata na razdalji 1,28 standardnega odklona od srednje vrednosti. Koliko točk mora doseči študent, da sodi med zgornjih 10% porazdelitve rezultatov SAT (ki je normalna s srednjo vrednostjo 500 in standardnim odklonom 100)?

(57) S pomočjo podatkov iz nalog 55 in 56 ugotovi, kako dolgih je najkrajših 10% nosečnosti.

3.16 Tehnološki kotiček

Načrtovanje iger na srečo

Igre na srečo so načrtovane tako, da ponujajo mamljivo velik glavni dobitek, občasno izplačajo manjše nagrade in zagotavljajo majhen povprečen donos v korist ponudnika igre. Oglejmo si naslednjo igro: igralec vplača 1 €, nato pa potegne po eno karto iz vsakega od treh premešanih kupčkov 52 kart. Če so vse tri karte enake, zadene jackpot. Če se ujemata dve izmed kart, dobi manjšo nagrado. Od $52^3 = 140\,608$ možnih izbir treh kart v 52 primerih dobimo jackpot in v nadaljnjih 7956 primerih zadenemo manjšo nagrado.

Preglednica na sliki 3.17 opisuje situacijo, ko je jackpot vreden 1000 € in manjša nagrada 10 €. Verjetnosti, da dobimo nagrado, sta enaki $\frac{52}{52^3}$ in $\frac{7956}{52^3}$. Povprečno igralnica na vsako igro izplača 90 centov, igralec pa vloži 1 €. Igralnica torej zasluži 10 centov pri vsaki igri.

Naloga 1. Večina igralnic načrtuje igre tako, da zaslužijo približno 5 centov na vsak vloženi evro. Poišči takšne vrednosti za nagrade v zgornji igri, da bo zaslužek enak 5 centov na igro.

	A	B	C	D
1	nagrada	verjetnost	nagrada · verjetnost	
2	1 000,00€	0,000369822	0,36982	
3	10,00€	0,053440088	0,53440	
4				
5		skupno	0,90422	
6		cena	1,00€	
7		povp. dobitek	-0,09578	

Slika 3.17: Možna izbira dobitkov pri neki igri.

Naloga 2. Da bi naredila igro bolj vabljivo, igralnica doda še eno nagrado – če izvlečemo tri ase, dobimo super jackpot. Nastavi vrednosti nagrad tako, da bosta obe glavni nagradi veliki in bo igralnica vseeno zaslužila 5 centov za vsak vloženi evro.

Simulacija igre s preglednico

Predpostavimo, da so karte označene s števili od 1 do 52, pri čemer začnemo s štirimi asi. V preglednici na sliki 3.18 uporabimo funkcijo `=RandBetween(1,52)` da izberemo naključno število med 1 in 52 v vsakem od prvih treh stolpcev. V četrtem stolpcu preglednice program za nas preveri, če se izbrana tri števila ujemajo. Če želimo na primer preveriti, ali so števila iz `A1`, `B1` in `C1` enaka, uporabimo ukaz `=And(A1=B1,B1=C1)`. Ta funkcija vrne `ne`, kadar so ta tri števila različna. V petem stolpcu preverimo, če se vsaj dve od števil ujemata s pomočjo ukaza `=Or(A1=B1,A1=C1,B1=C1)`. S pomočjo preglednice lahko kopiramo prvo vrstico in jo prilepimo v naslednjih devetih ter tako izračunamo podatke še za preostale vrstice. Ta preglednica torej simulira 10 iger, vrednih 10 €.

Vsakič ko preglednico na novo evalviramo, so izbrana druga naključna števila in ustvarimo nove igre. V zadnjem stolpcu so shranjeni podatki o tem, kolikokrat sta se dve od izbranih števil ujemali. (Med to simulacijo nismo naleteli na trojico enakih števil.) Vsota števil iz zadnjega stolpca je 11. Če je vsaka manjša nagrada vredna 10 €, igralec dobi 110 €, vložil pa je 100 €.

Naloga 3. Simuliraj deset skupin po 100 iger kot je to prikazano v zgornjem primeru. Pri koliko od teh je bil izkupiček večji od vloženih 100 €?

	A	B	C	D	E	F	G
1	35	30	17	ne	ne	izidi:	2
2	14	46	48	ne	ne		1
3	10	28	20	ne	ne		2
4	11	29	36	ne	ne		1
5	29	26	16	ne	ne		1
6	25	9	6	ne	ne		0
7	21	50	33	ne	ne		2
8	8	41	41	ne	da		1
9	22	29	48	ne	ne		0
10	17	30	43	ne	ne		1
11							
12				vse 3	vsaj 2		

Slika 3.18: Simulacija igre s preglednico

Naloga 4. Ali je možno izbrati vrednosti za nagrade tako, da sta glavni nagradi še vedno visoki, povprečni zaslužek igralnice je 5 centov na igro in igralec iz prejšnjih simulacij v splošnem dobi povrnjeno večino svojih investicij? Poišči kompromisno zasnovo nagrad, pri kateri upoštevaš te želje.

Raziskovanje

Če igramo to igro velikokrat, bomo statistično gledano pri vsaki igri izgubili 5 centov. Ker pa so jackpoti zelo redki, bomo morali igrati zelo dolgo, preden bomo lahko zadeli in s tem poplačali svoje investicije. Kako dolgo bi po tvojem mnenju morali igrati, da bi zadeli jackpot?

3.17 Pisni projekti

(1) "Začetki matematične teorije verjetnosti segajo v Francijo sedemnajstega stoletja, ko so se kockarji začeli obračati na matematike po nasvete." Dva od teh matematikov sta bila Pierre de Fermat in Blaise Pascal. Poišči nekaj literature o začetkih verjetnosti in napiši kratek esej o vlogi Fermata in Pascala. (Dober vir je na primer Carl B. Boyer, *A History of Mathematics*, Wiley, New York,

1991.)

(2) Državne loterije so zelo pogoste v ZDA in drugih državah (glej besedilo v okvirčku). Napiši kratek esej, ki opisuje trenutno stanje loterij v Sloveniji. Koliko denarja dobijo? Kako ga porabijo? Kakšni so trendi ponujenih iger? Katere druge oblike igralništva dovoljuje država?

(3) Veliko ljudi je pretirano strah dogodkov, ki so sicer zelo malo verjetni. Verjetnost, da umremo v letalski nesreči, terorističnem napadu ali tornadu, je zelo majhna. Vendar pa pri javnem mnenju in v zasebnih odločitvah velikokrat ravnamo tako, kot da bi bila nevarnost takšna kot v primeru prometnih nesreč ali srčnega napada. Napiši kratek esej, v katerem opisuješ, kako ljudje ocenjujejo tveganje in kateri dejavniki poleg verjetnosti še vplivajo na njihova dejanja. Eden možnih virov je Richard J. Zeckhauser, W. Kip Vicusi, Risk within reason, *Science*, 4. maj 1990, str. 559–564.

Poglavje 4

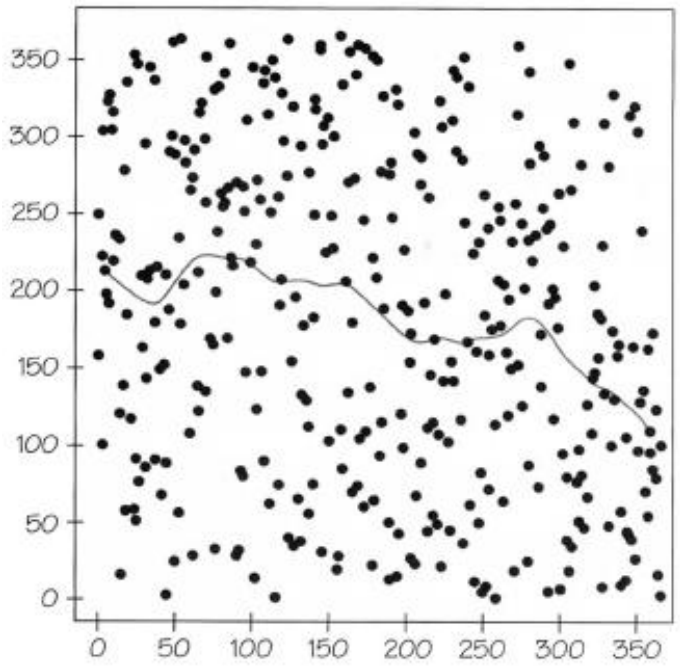
Statistično sklepanje

Sklepanje je proces, pri katerem pridemo do zaključkov na podlagi danih dokazov. Dokazi so lahko v mnogo različnih oblikah. V sojenju zaradi umora jih lahko predstavljajo izjave prič, posnetki telefonskih pogovorov, analize DNK iz vzorcev krvi in podobno. Pri statističnem sklepanju nam dokaze priskrbijo podatki. Neformalno statistično sklepanje velikokrat temelji na grafični predstavitvi podatkov. Formalno sklepanje, tema tega poglavja, uporablja verjetnost, da pove, do kakšne mere smo lahko prepričani, da so naši zaključki pravilni.

Primer. (Je bil nabor nepošten?) V času Vietnamske vojne so američani izbirali državljane za vojaško obveznost po sistemu loterije. Prvo so izvedli leta 1970. Vsi mediji so kmalu poročali, da je bila loterija pristranska proti tistim, ki so bili rojeni kasneje v letu. Na sliki 4.1 je razsevni diagram izbranih števil (manjša števila so bila izbrana prej) v odvisnosti od datuma rojstva, štetega od 1. januarja naprej. Krivuljo na diagramu smo dobili z izglajevalcem razsevnega diagrama. Vidimo, da kasneje v letu teži k nižjim številkam. Ta tendenca ni zelo močna, zato se lahko vprašamo, če je morda tak izid le naključen, ne pa rezultat sistematične pristranskosti v loteriji. Navsezadnje je pri vsaki loteriji nekaj slučajnega odstopanja od popolne poenotenosti.

Vendar pa izračun verjetnosti pokaže, da se tako močna odstopanja, kot je tisto na sliki 4.1, v zares slučajnih izborih pojavijo z verjetnostjo, manjšo od $\frac{1}{1000}$. Ta izračun je vse prepričal, da nabor ni bil naključen. Preiskava je pokazala, da so pri žrebanju slabo premešali kapsule, ki so vsebovale datume. ♦

V matematiki pridemo do zaključkov tako, da začnemo s hipotezo, nato pa z uporabo logičnega sklepanja pokažemo, da naši zaključki brezdvomno sledijo iz predpostavk.



Slika 4.1: Razsevni diagram številk, izbranih pri naboru (med 1 in 366) v odvisnosti od datuma rojstva (z začetkom pri 1. januarju). Splošni potek je opisan s pomočjo izglajevalca diagrama. Kaže, da je bila loterija pristranska v prid moškim, ki so bili rojeni v začetnem delu leta.

To je *deduktivno sklepanje* od hipoteze k posledicam. Statistični dokazi potekajo skoraj obratno. Če je bila loterija pristranska, pričakujemo, da bodo kasnejši datumi sistematično manj pogosto izbrani. V podatkih opazimo tak trend, zato dokazi govorijo v prid trditvi, da je bila loterija pristranska. To je *induktivno sklepanje* od posledic k hipotezi. Induktivno sklepanje ni dokaz. Manjša pogostost kasnejših datumov bi *lahko* bila naključna. *Statistično sklepanje uporabi verjetnost, da nam pove, kako močan je induktivni argument.* Težnja, ki jo opazimo v podatkih za leto 1970 se skoraj nikoli ne bi pojavila (verjetnost je manjša od 0,001) v slučajnem izboru. To močno govori v prid trditvi, da loterija ni bila zares naključna.

Primer. (Ali lahko zaupamo javnomnenjski raziskavi?) Kako lahko zaupamo rezultatom, ki jih dobimo s slučajnim vzorčenjem, če vemo, da bi z drugim vzorcem dobili drugačen rezultat? Pri Gallupovi raziskavi na 1493 ljudeh so ugotovili, da se 45% ljudi zaradi kriminala boji ponoči ven. V drugem slučajnem vzorcu bi izbrali nekih drugih 1493 ljudi in rezultat, ki bi ga dobili, bi bil različen od 45%. Kaj lahko kljub temu povemo o populaciji več kot 200 milijonov odraslih Američanov na osnovi Gallupovega vzorca?

Izračun verjetnosti nam pove, da pri 95% vseh Gallupovih vzorcev dobimo rezultat, ki se za manj kot tri odstotne točke razlikuje od resnične vrednosti za našo populacijo. Spoznamo torej, da je precej varno verjeti, da leži dejanska vrednost med 42% in 48%. Ker pa Gallupovi vzorci večkrat zgrešijo pravi rezultat za več kot 1%, ne moremo trditi, da leži resnični delež med 44% in 46%. ♦

V obeh primerih smo odgovor na vprašanje "Kaj bi se zgodilo, če bi to napravili večkrat?" dobili z izračunom verjetnosti. V velikem številu nepristranskih loterij bi le ena izmed tisoč pokazala tako močan trend, kot smo ga lahko opazili pri naboru. V velikem številu Gallupovih vzorcev bi jih 95% dalo rezultat, ki bi ležal $\pm 3\%$ od dejanske vrednosti. Takšne izjave so značilne za statistično sklepanje. Da bi ga razumeli, moramo najprej razumeti, kako pri tem uporabljamo verjetnost.

4.1 Ocenjevanje deleža populacije

Z uporabo poenostavljene verzije Gallupove raziskave o kriminalu bomo predstavili pomembno obliko statističnega sklepanja. Kot večina nacionalnih raziskav uporablja Gallupova anketa kompleksen večstopenjski načrt vzorca. Namesto tega si predstavljajmo, da izberemo enostavni slučajni vzorec 1500 odraslih in ugotovimo, da se jih med njimi 675 boji kriminala. **Delež vzorca**, ki ostane doma zaradi strahu pred kriminalom je enak

$$\hat{p} = \frac{675}{1500} = 0,45 = 45\%.$$

Delež vzorca bomo označili s $\hat{p}$ (kar preberemo "p streha"). Vedno bomo izražali delež vzorca v odstotkih.

Delež vzorca $\hat{p} = 45\%$ se nanaša na 1500 ljudi iz tega konkretnega vzorca. V resnici pa želimo poznati *delež populacije*, odstotek (imenujmo ga p) vseh odraslih Američanov, ki ponoči ostajajo doma zaradi strahu pred kriminalom. Da bi lahko razumno razpravljali o statističnem sklepanju, moramo jasno vedeti, katera števila opisujejo vzorce in katera populacije.

> Število, ki opisuje populacijo, imenujemo **parameter**. Število, ki ga izračunamo iz vzorca, imenujemo **statistika**.

Ni si težko zapomniti, da **p**arameter pripada **p**opulaciji, statistika pa torej spada

k vzorcu[1]. Pri statističnem sklepanju so parametri običajno neznani. Ne vemo, kolikšen delež p vseh odraslih zares ostaja doma zaradi strahu pred kriminalom. Za oceno neznanega parametra p uporabimo statistiko $\hat{p}$, ki smo jo dobili, ko smo dejansko anketirali ljudi iz vzorca. *Naš cilj ni zgolj oceniti p, temveč tudi povedati, kako natančna je ta ocena.* Da bi to izvedeli, se moramo vprašati "Kaj bi se zgodilo, če bi vzeli več vzorcev? Kako blizu neznanemu p bi običajno ležale ocene $\hat{p}$?"

Za odgovor na to vprašanje si natančneje ogledamo *porazdelitev vzorca* $\hat{p}$. To je porazdelitev vrednosti, ki jih zavzamejo deleži vzorcev, ko si ogledamo veliko število vzorcev iz iste populacije. V prejšnjih poglavjih smo že simulirali vzorčne porazdelitve. Zdaj želimo matematična dejstva. Tule so:

Izberimo enostavni slučajni vzorec velikosti n iz velike populacije, v kateri ima $p\%$ osebkov lastnost, ki nas zanima. Naj bo $\hat{p}$ delež osebkov v vzorcu, ki imajo to lastnost. Potem velja:

- Vzorčna porazdelitev $\hat{p}$ je *približno normalna* in je bližje normalni kadar je velikost vzorca n velika.
- *Srednja vrednost* vzorčne porazdelitve je natanko p.
- *Standardni odklon* vzorčne porazdelitve je

$$\sigma_{\hat{p}} = \sqrt{\frac{p(100-p)}{n}}.$$

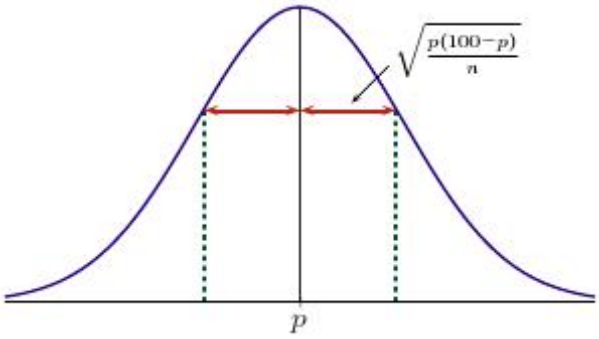
  S $\sigma_{\hat{p}}$ jo označimo zato, da ne pozabimo, da gre za standardni odklon, ki pripada porazdelitvi $\hat{p}$.

Na sliki 4.2 je prikazana ta vzorčna porazdelitev kot normalna krivulja. Središče (srednja vrednost) in razpon (standardni odklon) te krivulje vsebujeta pomembne statistične informacije.

Srednja vrednost krivulje je dejanski delež p vseh ljudi, ki se bojijo ponoči zapustiti domove zaradi kriminala. To nam pove, da $\hat{p}$ ni pristranska ali sistematično napačna cenilka za neznani p. Pri večkratnem vzorčenju bo dobljeni rezultat včasih višji in včasih nižji, povprečje velikega števila rezultatov, srednja vrednost vzorčne porazdelitve, pa bo pravilno. V praksi seveda ne poznamo vrednosti parametra p.

[1] V angleščini se ujemata tudi prvi črki izrazov za statistiko (statistic) in vzorec (sample). (Op. prev.)

Vemo pa, da se ne glede na to, kakšno vrednost ima p, opažene vrednosti statistike $\hat{p}$ kopičijo okoli p kot na sliki 4.2.



Slika 4.2: Vzorčna porazdelitev deleža vzorca $\hat{p}$. Je približno normalna s srednjo vrednostjo p in standardnim odklonom $\sqrt{\frac{p(100-p)}{n}}$.

Vendar pa nismo zadovoljni s tem, da imamo v povprečju prav. Dobra cenilka mora biti tudi ponovljiva, v smislu da daje skoraj enake odgovore pri večkratnem vzorčenju. Ponovljivost opišemo z razponom vzorčne porazdelitve, ki ga meri standardni odklon. Če bi vzorčenje ponovili velikokrat, tako da bi v vsaki ponovitvi poklicali naključno izbrane številke, bi vsakič dobili vrednost deleža vzorca $\hat{p}$ nekje vzdolž krivulje na sliki 4.2. Kako daleč od resničnega p ti rezultati ležijo, je odvisno od standardnega odklona $\sigma_{\hat{p}}$ te normalne krivulje. Standardni odklon se manjša, ko se velikost vzorca n povečuje. Tudi pri naših simulacijah iz prejšnjega poglavja (slika 3.6) je bilo tako. Zdaj natančno poznamo zvezo med n in standardnim odklonom. *Standardni odklon je odvisen od kvadratnega korena* $\sqrt{n}$. Če želimo razpon porazdelitve razpoloviti, moramo izbrati štirikrat večji vzorec.

Primer. (Vzorčna porazdelitev za raziskavo o kriminalu) Recimo, da se v resnici 40% odraslih boji ponoči zdoma zaradi kriminala. Se pravi, naj bo $p = 40\%$. Izberimo enostavni slučajni vzorec velikosti $n = 1500$. Pri večkratnem vzorčenju se bo delež vzorca $\hat{p}$ spreminjal v skladu z normalno porazdelitvijo s srednjo vrednostjo $p = 40\%$ in s standardnim odklonom

$$\sigma_{\hat{p}} = \sqrt{\frac{p(100-p)}{n}} = \sqrt{\frac{40 \cdot 60}{1500}} = \sqrt{1{,}6} = 1{,}265\%.$$

V praksi vrednosti parametra p ne poznamo. S kalkulatorjem izračunamo, da je standardni odklon porazdelitve $\hat{p}$ majhen, zato bo delež vzorca $\hat{p}$ ponavadi precej blizu p.

Zdaj pa predpostavimo, da je resnična vrednost za populacijo $p = 50\%$. Središče vzorčne porazdelitve se premakne na 50%. Standardni odklon se spremeni v

$$\sigma_{\hat{p}} = \sqrt{\frac{p(100-p)}{n}} = \sqrt{\frac{50 \cdot 50}{1500}} = \sqrt{1{,}67} = 1,29\%.$$

Standardni odklon $\sigma_{\hat{p}}$ se ne spremeni veliko, ko se spremeni p. Ko torej izbiramo vzorce istih velikosti iz različnih populacij, se središče vzorčne porazdelitve $\hat{p}$ premakne do parametra p, ki ustreza tej populaciji, razpon pa ostaja približno enak.

Velikost vzorca pomembno vpliva na razpon. Recimo, da bi izbrali vzorec velikosti $n = 375$ namesto 1500 iz populacije, za katero je $p = 40\%$. Srednja vrednost porazdelitve $\hat{p}$ je še vedno 40%, ker velikost vzorca ne spremeni središča vzorčne porazdelitve. Standardni odklon pa se poveča na

$$\sigma_{\hat{p}} = \sqrt{\frac{p(100-p)}{n}} = \sqrt{\frac{40 \cdot 60}{375}} = \sqrt{6{,}4} = 2,53\%.$$

Ker je nova velikost vzorca $n = 375$ le četrtina prvotnih 1500, je nov standardni odklon 2,53% dvakrat tolikšen kot prejšnji 1,265%. Tako na razpon vpliva $\sqrt{n}$. ♦

4.2 Intervali zaupanja

Pri našem anketiranju 1500 ljudi smo ugotovili, da je $\hat{p} = 45\%$. To je naša najboljša ocena za delež populacije p. Kako blizu resnične vrednosti p je najbrž naša ocena? Vemo, da se $\hat{p}$ spreminja v skladu z normalno porazdelitvijo. Pravilo 68–95–99,7 pravi, da $\hat{p}$ v 95% vseh vzorcev leži največ dva standardna odklona od p (srednje vrednosti vzorčne porazdelitve). Naša ocena, ki temelji na tem enem vzorcu, je torej najverjetneje od p oddaljena največ dva standardna odklona, se pravi kvečjemu

$$2\sigma_{\hat{p}} = 2\sqrt{\frac{p(100-p)}{1500}}.$$

Problem je v tem, da je ta standardni odklon odvisen od neznanega parametra p. Na srečo pa smo v zgornjem primeru videli, da se $\sigma_{\hat{p}}$ spreminja le zelo počasi, ko se spreminja p, če le ni p zelo blizu 0% ali pa 100%. Ker je $\hat{p}$ blizu p, enostavno zamenjamo v enačbi za standardni odklon neznani p s $\hat{p}$. Ker je tako dobljeni standardni odklon le ocena in ne poznamo natančne vrednosti, ga označimo s $s_{\hat{p}}$.

Primer. (Ocenjeni standardni odklon za raziskavo o kriminalu) Oceniti želimo standardni odklon deležev vzorca. Velikost vzorca je bila $n = 1500$ in za p

uporabimo oceno $\hat{p} = 45\%$, ki jo dobimo z anketo. Ocenjeni standardni odklon je

$$s_{\hat{p}} = \sqrt{\frac{45 \cdot 55}{1500}} = \sqrt{1{,}65} = 1,285\%.$$

♦

Zaključimo torej naslednje: v 95% vseh vzorcev bo delež vzorcev $\hat{p}$ kvečjemu za $2 \cdot 1,285\% = 2,6\%$ oddaljen od neznanega deleža populacije p. Izbrali smo en vzorec in dobili $\hat{p} = 45\%$. Od tod sklepamo, da p leži na intervalu

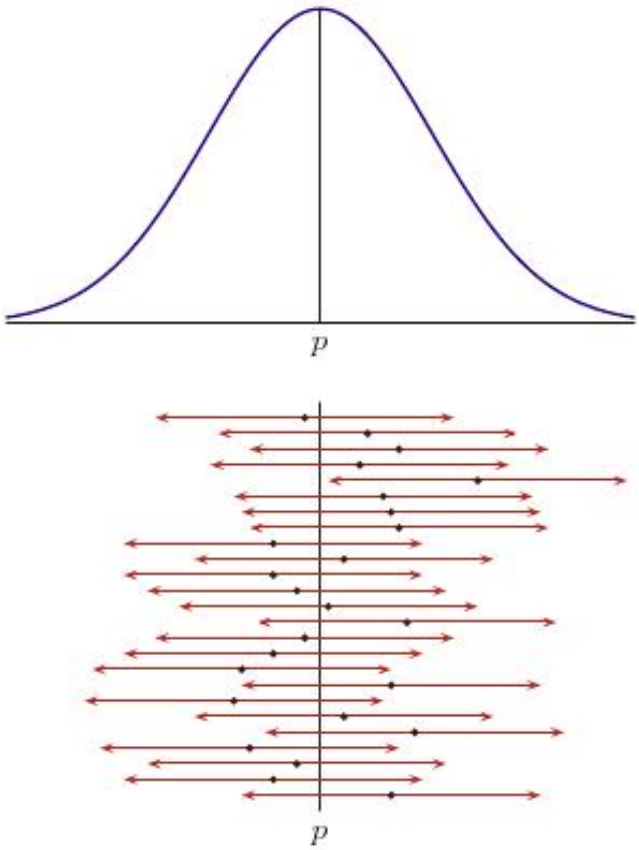
$$45\% \pm 2,6\%$$

oziroma med 42,4% in 47,6%. Pravimo, da je naše *zaupanje* v ta sklep 95%, ker smo dobili interval tako, da smo izračunali, kako blizu p ležijo deleži vzorcev v 95% vseh vzorcev. To je *95% interval zaupanja* za oceno neznanega deleža populacije.

Povedano po matematično, z verjetnostjo 0,95 je delež vzorca $\hat{p}$ oddaljen kvečjemu za ±2,6% od neznanega dejanskega deleža p vseh odraslih, ki jih je ponoči strah zapustiti domove zaradi kriminala. Slika 4.3 dodatno pojasnjuje to idejo. Normalna krivulja na zgornjem delu slike je vzorčna porazdelitev za $\hat{p}$. Ko izberemo več različnih vzorcev, se dejanske vrednosti $\hat{p}$ spreminjajo v skladu s to krivuljo. Na spodnjem delu slike so s pikami predstavljene vrednosti $\hat{p}$, ki smo jih dobili s 25 različnimi vzorci. Pri vsaki je narisan tudi interval zaupanja, ki sega 2,6% na vsako stran ustreznega $\hat{p}$. Dejanski delež populacije p je označen z navpično črto. Čeprav se intervali od vzorca do vzorca spreminjajo, vsi razen enega vsebujejo pravi p. Ko pravimo, da gre za 95% intervale zaupanja, želimo le povedati, da ti intervali vsebujejo p v 95% vseh vzorcev in ga zgrešijo le v 5% primerov. Pri tem je pomembno razumeti, da se teh 95% in 5% nanaša le na to, kaj bi se zgodilo, če bi izbirali take vzorce vedno znova in znova. Pri majhnem številu vzorcev je lahko število intervalov zaupanja, ki ne vsebujejo dejanskega p, nekoliko večje ali manjše od 5%. Na sliki 4.3 na primer le pri enem vzorcu od 25, torej v 4%, interval zaupanja ne vsebuje p.

Interval, ki ga dobimo iz podatkov o vzorcu z metodo, pri kateri za 95% vseh vzorcev dobimo interval, ki vsebuje dejanski parameter p dane populacije, se imenuje **95% interval zaupanja**.

Na sliki 4.3 lahko vidimo, da interval zaupanja vsakega vzorca vsebuje neznani parameter ali pa ne. Ne moremo vedeti, če je naš izbrani vzorec eden izmed 95% tistih, ki zadenejo parameter, ali pa med tistimi 5%, ki ga zgrešijo. Če za naš interval



Slika 4.3: Obnašanje 95% intervalov zaupanja pri večkratnem vzorčenju. Interval se spreminja od vzorca do vzorca, vendar pa na dolgi rok 95% vseh vzorcev porodi intervale, ki vsebujejo dejansko vrednost p.

45%±2,6% trdimo, da je 95% interval zaupanja, želimo reči "Ta interval smo dobili z metodo, ki v 95% primerov zadene pravi parameter."

Zdaj smo dosegli dvoje: ugotovili smo, kaj pomeni "95% zaupanje", poleg tega pa smo dejansko našli 95% interval zaupanja za ocenjevanje deleža populacije. Tule je recept:

Za enostavni slučajni vzorec velikosti n je **95% interval zaupanja za delež populacije** p enak

$$\hat{p} \pm 2s_{\hat{p}} = \hat{p} \pm 2\sqrt{\frac{\hat{p}(100-\hat{p})}{n}}.$$

Spomnimo se, da merimo tako p kot $\hat{p}$ v odstotkih. Ta recept je samo približno pravilen, vendar pa je precej natančen, ko je velikost vzorca n velika.

Primer. (Tvegano obnašanje v dobi AIDSa) Kako pogosto je obnašanje, pri katerem ljudje tvegajo okužbo? V nacionalni raziskavi so anketirali slučajni vzorec 2673 odraslih heteroseksualcev. Od teh jih je 170 v preteklem letu imelo več kot enega partnerja. Ustrezni delež vzorca je

$$\hat{p} = \frac{170}{2673} = 6,36\%.$$

S 95% stopnjo zaupanja je interval zaupanja za delež populacije p enak

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(100-\hat{p})}{n}} = 6,36 \pm 2\sqrt{\frac{6{,}36 \cdot 93{,}64}{2673}} = 6,36 \pm 0,94 = 5,42\% \text{ do } 7,30\%.$$

Kot ponavadi praktične težave lahko povzočijo dodatne napake, ki jih *ne* upoštevamo pri meji napake. Zelo verjetno je, da nekateri sodelujoči niso bili pripravljeni razkriti svojih resničnih navad. Rezultati raziskave torej najbrž podcenjujejo pogostost tveganega obnašanja, ne vemo pa, za koliko. ♦

Dolžina intervala zaupanja je odvisna od velikosti vzorca n: pri večjih vzorcih so intervali krajši. Niso pa intervali odvisni od velikosti populacije. To velja vse dokler je populacija veliko večja kot vzorec. Interval zaupanja iz zgornjega primera velja tako za vzorec 2673 ljudi iz mesta s 100 000 odraslimi prebivalci kot za vzorec iz države z 200 milijoni prebivalcev. Tisto, kar je pomembno, je, koliko ljudi anketiramo, ne pa kolikšen delež populacije kontaktiramo.

Vsak interval zaupanja ima dva bistvena podatka: interval sam in stopnjo zaupanja. Običajno ima obliko

$$\text{ocena } \pm \text{ meja napake.}$$

Ocena je vzorčna statistika, na primer $\hat{p}$, ki ocenjuje neznani parameter. Meja napake nam pove, kako natančna je ta ocena. V raziskavi o AIDSu je ocena 3,36%, meja napake pa je 0,94%.

Stopnja zaupanja nam pove, kako prepričani smo, da naš interval vsebuje dejansko vrednost parametra. Čeprav je pogosta stopnja zaupanja 95%, lahko zahtevamo večje zaupanje, na primer 99%, ali pa se zadovoljimo že samo z 90%. Naš 95% interval zaupanja je temeljil na srednjih 95% normalne porazdelitve. Za 99% stopnjo zaupanja zahtevamo srednjih 99% porazdelitve, zato je ustrezni interval zaupanja širši (ima večjo mejo napake). Podobno je 90% interval zaupanja krajši od 95% intervala zaupanja za iste podatke. Izbirati moramo med tem, kako natančno lahko določimo parameter (meja napake) in kako prepričani smo v ta rezultat.

Primer. (Razumevanje novic) V novicah se pogosto pojavljajo rezultati javnomnenjskih raziskav in drugih raziskav na vzorcih. Poročila velikokrat navajajo mejo

napake, redko pa omenijo stopnjo zaupanja. (O izjemi lahko prebereš v okvirju.) Medijsko poročilo o naši raziskavi o kriminalu bi izgledalo takole: "Raziskava je pokazala, da se 45% Američanov zaradi kriminala boji ponoči zapustiti domove. Meja napake te raziskave je ± 2,6 odstotne točke."

Poznati pa moramo tako mejo napake kot stopnjo zaupanja, ker večja stopnja zaupanja zahteva večjo mejo napake. V medijih velja nenapisano pravilo: skoraj vse javnomnenjske raziskave podajo mejo napake pri 95% stopnji zaupanja. Če torej v novici o raziskavi podajo mejo napake brez omembe stopnje zaupanja, lahko ponavadi predpostavimo, da je ta enaka 95%. ♦

Pod žarometom

Kako je potekala raziskava?

New York Times je 15. junija 1998 objavil članek z naslovom "Raziskava kaže: večina naklonjena Microsoftu", ki sta ga napisala Steve Lohr in Marjories Connelly. V času ko je država napadala računalniškega giganta z antitrust zakoni, je raziskava javnega mnenja pokazala, da ima 55% odraslih o Microsoftu pozitivno mnenje. Časopis je ob strani objavil tudi nekaj podrobnosti o tem, kako je potekala raziskava:

> Najnovejša raziskava, ki sta jo izvedla *New York Times* in *CBS News*, je temeljila na telefonski anketi, ki je bila opravljena med 7. in 9. junijem in je zajela 1126 odraslih po vseh ZDA.
>
> Vzorec telefonskih central je bil izbran slučajno s pomočjo računalnika s popolnega seznama več kot 42 000 aktivnih central po celi državi. Znotraj vsake centrale so nato slučajno izbrali telefonske številke, tako da so bili v vzorec vključeni tudi tisti, ki sicer niso v telefonskem imeniku. V vsakem gospodinjstvu je nato na vprašanja odgovarjala slučajno izbrana odrasla oseba.
>
> ... Teoretično se v 19 primerih od 20 rezultati raziskav, ki temeljijo na tovrstnih vzorcih, razlikujejo od dejanske vrednosti za največ tri odstotne točke.
>
> ... Poleg napak, ki nastajajo pri vzorčenju, se lahko pojavijo še druge, ki so povezane s praktičnimi omejitvami pri izvajanju vseh javnomnenjskih raziskav. Odstopanja v formulaciji vprašanj ali v vrstnem redu lahko na primer pripeljejo do drugačnih rezultatov.

Urad za delavsko statistiko pa se je odločil, da bo mesečno stopnjo nezaposlenosti podajal z 90% stopnjo zaupanja. Zaključki so temeljili na raziskavi, ki je vključevala

50 000 gospodinjstev, Urad pa pravi, da je meja napake $\pm 0,2\%$. Medtem ko naslovi časopisov oznanjajo 5,9% stopnjo nezaposlenosti, želi Urad povedati, da je z 90% zaupanjem med 5,7% in 6,1% delovno aktivnega prebivalstva nezaposlenega.

Raziskave mnenja imajo velikokrat meje napake okoli $\pm 3\%$. Bistveno manjša meja napake pri raziskavi o nezaposlenosti je posledica veliko večjega vzorca. Pri večjih vzorcih je meja napake manjša pri isti stopnji zaupanja. Vendar pa kvadratni koren iz n, ki se pojavi v računih, pove, da moramo vzeti štirikrat večji vzorec, če želimo mejo napake razpoloviti. Da bi dobili zelo majhno mejo napake, se pri zgornji raziskavi potrudijo in anketirajo 50 000 ljudi, medtem ko jih Gallupova raziskava kontaktira le 1500. Pri Gallupovi raziskavi si lahko privoščijo 3% napako, stopnja nezaposlenosti pa mora biti bolj natančno določena, ker na njej temelji veliko ekonomskih in političnih odločitev.

4.3 Ocenjevanje srednje vrednosti populacije

Statistikova škatla z orodjem vsebuje veliko različnih intervalov zaupanja, ki ustrezajo velikemu številu različnih parametrov populacije, ki bi jih želeli oceniti. Spoznali smo že interval zaupanja za oceno deleža populacije p. Zdaj pa želimo oceniti srednjo vrednost populacije. Za opis središča množice podatkov smo uporabili **vzorčno povprečje** $\overline{x}$. Zdaj bomo uporabili vzorčno povprečje $\overline{x}$ za oceno neznane srednje vrednosti μ celotne populacije, iz katere smo izbrali vzorec. Srednjo vrednost populacije označimo z μ, ki označuje tudi srednjo vrednost verjetnostne porazdelitve, ker je srednja vrednost populacije tudi srednja vrednost porazdelitve rezultata, ki ga dobimo pri slučajni izbiri enega posameznika iz populacije. Vzorčno povprečje $\overline{x}$ je statistika, ki se bo med večkratnim vzorčenjem spreminjala, medtem ko je srednja vrednost populacije μ parameter in ostaja konstantna. Na srečo je interval zaupanja za ocenjevanje μ precej podoben že znanemu intervalu zaupanja za ocenjevanje p, ker oba temeljita na normalni vzorčni porazdelitvi.

Primer. (NAEP) Namen ameriške raziskave NAEP (*National Assessment of Educational Progress*) je ugotavljati, kako napreduje izobraženost. Raziskava vključuje tudi kratek test kvantitativnega razumevanja, ki preverja v glavnem osnovno računanje in sposobnosti uporabe le-tega v realističnih problemih. Rezultati testov ležijo med 0 in 500 točkami. Oseba, ki dobi 233 točk, lahko na primer sešteje dva zneska, ki se pojavita na bančnem izpisku, nekdo, ki je dobil 325 točk je sposoben iz menija razbrati, koliko bo stalo kosilo, oseba s 375 točkami lahko pretvori ceno iz centov na

gram v dolarje na kilogram.

V zadnjem letu so v vzorec zajeli 840 moških med 21. in 25. letom starosti. Njihov povprečen rezultat je bil $\overline{x} = 272$. Teh 840 moških sestavlja enostavni slučajni vzorec populacije vseh mladih moških. Kaj lahko na osnovi tega vzorca povemo o povprečnem rezultatu μ populacije vseh 9,5 milijona moških med 21. in 25. letom starosti? ♦

Zakon velikih števil nam pove, da je vzorčno povprečje $\overline{x}$ pri velikih slučajnih vzorcih blizu neznani srednji vrednosti μ za dano populacijo. Ker je $\overline{x} = 272$, ugibamo, da je μ "nekje blizu 272". Da bi "nekje blizu 272" formulirali bolj natančno, se vprašamo: "Kako bi se spreminjalo vzorčno povprečje $\overline{x}$, če bi izbrali veliko vzorcev 840 oseb iz te populacije?" Na vprašanje nam odgovori *centralni limitni izrek* (str. 140).

Izberimo enostavni slučajni vzorec velikosti n iz velike populacije s srednjo vrednostjo μ in standardnim odklonom σ. Ko je velikost vzorca n velika, je **vzorčna porazdelitev vzorčnega povprečja** $\overline{x}$ približno normalna s srednjo vrednostjo μ in standardnim odklonom $\frac{\sigma}{\sqrt{n}}$.

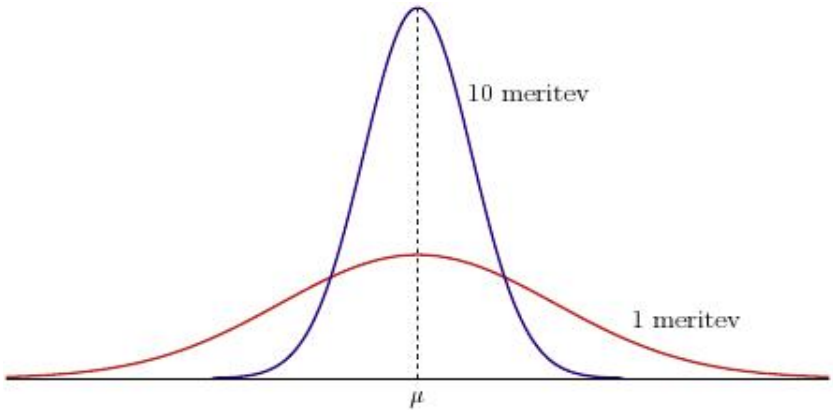
Tej že znani trditvi lahko dodamo še eno dejstvo. *Če je porazdelitev posameznikov v populaciji normalna, potem je vzorčna porazdelitev $\overline{x}$ popolnoma normalna.* To je res za vzorce poljubnih velikosti. Na sliki 4.4 je prikazana zveza med porazdelitvijo ene same vrednosti iz normalno porazdeljene populacije in porazdelitve povprečja več (v tem primeru 10) vrednosti. Povprečje večih vrednosti je manj spremenljivo kot posamezne vrednosti.

Če želimo uporabiti vzorčno porazdelitev $\overline{x}$, moramo za našo populacijo poznati standardni odklon σ. Iz preteklih izkušenj vemo, da je standardni odklon rezultatov NAEPa blizu $\sigma = 60$. Zdaj imamo informacije, ki jih potrebujemo za interval zaupanja srednje vrednosti.

Primer. (Ocena povprečnega NAEP rezultata) Normalna vzorčna porazdelitev za $\overline{x}$ ima srednjo vrednost, ki je enaka neznani srednji vrednosti populacije μ. Standardni odklon vzorčne porazdelitve je

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{60}{\sqrt{840}} = 2{,}1.$$

Pravilo 68–95–99,7 nam pove, da bo $\overline{x}$ v 95% vseh vzorcev največ za dva standardna odklona oddaljen od μ. Ta razdalja je $2 \cdot 2{,}1 = 4{,}2$ odstotne točke. V našem vzorcu



Slika 4.4: Vzorčna porazdelitev vzorčnega povprečja $\overline{x}$ iz enostavnega slučajnega vzorca 10 meritev v primerjavi s porazdelitvijo ene same meritve.

smo dobili $\overline{x} = 272$, torej s stopnjo zaupanja 95% trdimo, da leži srednja vrednost populacije μ na intervalu

$$272 \pm 4{,}2$$

oziroma med 267,8 in 276,2. ♦

Spodaj je recept, ki povzame naša dognanja. Interval zaupanja ima spet obliko

$$\text{ocena } \pm \text{ meja napake.}$$

Ocena je v tem primeru enaka vzorčnemu povprečju $\overline{x}$.

Naj bo μ neznana srednja vrednost populacije in σ znani standardni odklon. Iz te populacije izberemo enostavni slučajni vzorec velikosti n in izračunamo vzorčno povprečje $\overline{x}$. Potem je **95% interval zaupanja za srednjo vrednost populacije** μ enak

$$\overline{x} \pm 2\sigma_{\overline{x}} = \overline{x} \pm 2\frac{\sigma}{\sqrt{n}}.$$

Ta ocena velja v primeru, ko je populacija normalno porazdeljena, in je približno pravilna v drugih primerih, če so vzorci veliki.

Velikokrat v praksi ne poznamo vnaprej standardnega odklona σ za dano populacijo. V tem primeru moramo oceniti σ s standardnim odklonom s, ki ga izračunamo za naš vzorec. Če je vzorec velik, bo s blizu σ in zamenjava σ z s ne bo imela velikega vpliva na interval zaupanja.

Oglejmo si še en primer ocenjevanja srednje vrednosti populacije.

Primer. (Ocenjevanje količine prahu v rudnikih premoga) Povprečje več vrednosti je manj spremenljivo kot ena sama vrednost, zato je v primerih, ko je potrebna večja natančnost, dobro opazovati povprečje več meritev namesto ene same. Količino prahu v zraku v rudnikih premoga merimo tako, da izpostavimo filter zraku v rudniku, nato pa izmerimo količino prahu, ki se nanj nabere. Tehtanje ni povsem natančno. Večkratna tehtanja istega filtra se spreminjajo v skladu z normalno porazdelitvijo. Vrednosti, ki bi jih dobili s številnimi tehtanji, sestavljajo našo populacijo. Srednja vrednost μ te populacije je resnična teža (se pravi, da tehtanje ni pristransko). Standardni odklon populacije opisuje natančnost tehtanja. Vemo, da je $\sigma = 0{,}08$ mg. Vsak filter stehtamo trikrat in zapišemo povprečno težo.

Za enega od filtrov so izmerjene teže enake

$$123{,}1 \text{ mg} \qquad 122{,}5 \text{ mg} \qquad 123{,}7 \text{ mg}.$$

Kako bi poiskali 95% interval zaupanja za dejansko težo μ?

Najprej izračunamo vzorčno povprečje:

$$\overline{x} = \frac{123{,}1 + 122{,}5 + 123{,}7}{3} = \frac{369{,}3}{3} = 123{,}1 \text{ mg}.$$

Torej je 95% interval zaupanja enak

$$\overline{x} \pm 2\frac{\sigma}{\sqrt{n}} = 123{,}1 \pm 2\frac{0{,}08}{\sqrt{3}} = 123{,}1 \pm 0{,}09.$$

Smo torej 95% prepričani, da dejanska teža leži med 123,01 mg in 123,19 mg. ♦

4.4 Statistični nadzor procesov

Statistične metode uporabljamo pri zbiranju socialnih in ekonomskih informacij in pri raziskavah na številnih področjih. Večina primerov, ki smo si jih ogledali do sedaj, se je nanašala na ta dva načina uporabe statistike. Statistika pa prispeva tudi k izboljšanju kvalitete proizvedenih izdelkov. Skupaj z novo tehnologijo in novimi načini vodenja (na primer sodelovanjem z delavci in dobavitelji) so statistične ideje pomemben del bojev vsakega proizvajalca, ki želi tekmovati na globalnem tržišču. V tem razdelku si bomo ogledali preprosto a pomembno statistično orodje za kontroliranje in izboljšanje kakovosti, kontrolni diagram.

Primer. (Kontrola računalniških zaslonov) Proizvajalec računalniških zaslonov mora nadzorovati napetost na mreži tankih žic, ki ležijo pod površjem zaslona. Preveč napetosti bi poškodovalo mrežo, premalo bi povzročilo gubanje. Napetost merijo z električno napravo, ki izpisuje meritve v milivoltih (mV). Pravilna napetost je 275 mV. V proizvodnem procesu vedno prihaja do manjših odstopanj. Kadar proces poteka pravilno, je standardni odklon teh meritev $\sigma = 43mV$.

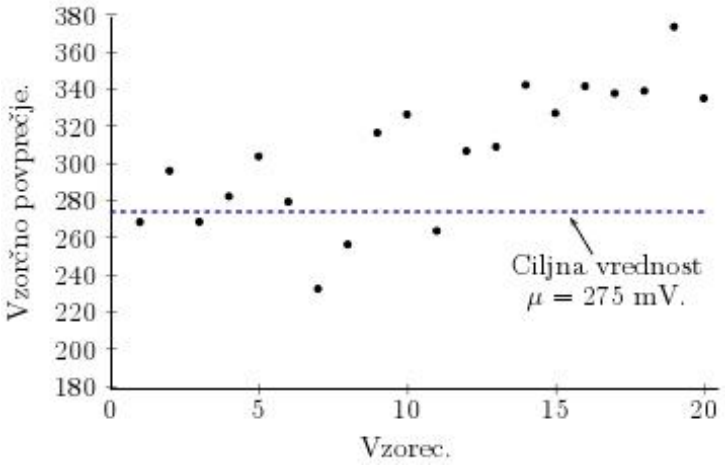
Tehnik vsako uro izmeri napetosti na vzorcu štirih zaslonov. Srednja vrednost $\overline{x}$ vsakega vzorca je ocena za povprečno napetost μ celotnega procesa v tistem trenutku. V tabeli 4.1 so prikazane vrednosti $\overline{x}$ za 20 zaporednih ur. Kako si lahko pomagamo s temi podatki, da bi obdržali proces stabilen? ♦

Vzorec	$\overline{x}$	Vzorec	$\overline{x}$
1	269,5	11	264,7
2	297,0	11	307,7
3	269,6	11	310,0
4	283,3	11	343,3
5	304,8	11	328,1
6	280,4	11	342,6
7	233,5	11	338,8
8	257,4	11	340,1
9	317,5	11	374,6
10	327,4	11	336,1

Tabela 4.1: Povprečja $\overline{x}$ za 20 vzorcev velikosti 4.

Graf podatkov v odvisnosti od časa nam bo pomagal videti, ali je bil proces moten. Na sliki 4.5 je diagram zaporednih povprečij meritev. Ker je ciljna vrednost procesa $\mu = 275$ mV, narišemo vzdolž diagrama *središčno črto* na tem nivoju. Povprečja iz kasnejših vzorcev vsa ležijo nad to črto in so dosledno višja od tistih iz prejšnjih vzorcev. Iz tega lahko sklepamo, da je srednja vrednost procesa μ dvignila stran od ciljne vrednosti 275 mV. Vendar pa morda ta dvig $\overline{x}$ predstavlja le naravno spremenljivost v procesu. Diagram moramo podpreti še z računom.

Pričakujemo, da bo $\overline{x}$ porazdeljen približno normalno. Ne samo da so meritve napetosti približno normalne, tudi centralni limitni izrek sugerira, da bodo vzorčna povprečja bližje normalni porazdelitvi kot posamezne meritve. Če ostane standar-



Slika 4.5: Graf vzorčnega povprečja meritev napetosti na računalniških zaslonih v odvisnosti od časa. Vodoravna črta je ciljna vrednost 275.

dni odklon pri $\sigma = 43$ mV, je standardni odklon povprečja $\overline{x}$ za štiri zaslone enak

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{43}{\sqrt{4}} = 21{,}5.$$

Dokler bo srednja vrednost enaka ciljni vrednosti $\mu = 275$ mV, nam bo pravilo 68–95–99,7 povedalo, da skoraj vse vrednosti $\overline{x}$ ležijo med

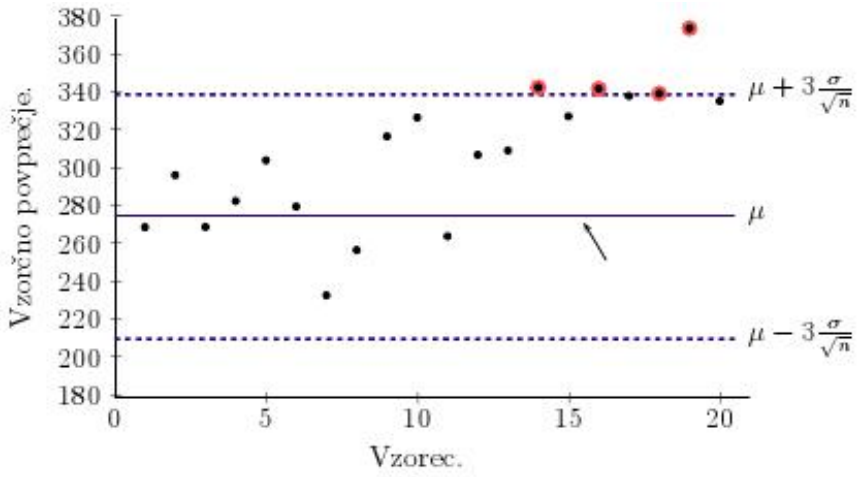
$$\mu - 3\sigma_{\overline{x}} = 275 - 3 \cdot 21{,}5 = 210{,}5 \text{ mV}$$

in

$$\mu + 3\sigma_{\overline{x}} = 275 + 3 \cdot 21{,}5 = 339{,}5 \text{ mV}.$$

Ti *kontrolni meji* smo vnesli v diagram kot črtkani črti.

Na sliki 4.6 je kontrolni diagram za meritve s slike 4.5. Štiri točke, ki so na diagramu obkrožene, ležijo nad zgornjo mejno vrednostjo. Ni zelo verjetno (verjetnost je manjša od 0,003), da bi kakšna od točk ležala nad to mejo, če bi μ in σ ostala nespremenjena. Te štiri točke so torej dober dokaz, da se je porazdelitev proizvodnega procesa spremenila. Zdi se, da se je srednja vrednost nekje pri 14. vzorcu premaknila navzgor. V praksi nadzorniki prežijo na takšne motnje v procesu takoj, ko opazijo prvo odstopanje, se pravi pri 14. vzorcu. Pomanjkanje nadzora bi lahko povzročili nov tehnik, nova serija mrež ali okvara na merilnem aparatu. Pojav meritev nad mejno vrednostjo nas opozori na spremembo, preden izdelamo veliko število pokvarjenih zaslonov. Spodaj je povzetek korakov za izdelavo kontrolnega diagrama, kakršen je na sliki 4.6.



Slika 4.6: Kontrolni diagram za $\overline{x}$ v primeru meritev napetosti na računalniških zaslonih. Črtkani kontrolni črti postavljata mejo pričakovanih odstopanj pri nemotenem procesu.

Za nadzorovanje stabilnosti procesa z danima standardoma μ in σ izdelamo **kontrolni diagram** $\overline{x}$ takole:

- Narišemo povprečja $\overline{x}$ rednih vzorcev velikosti n v odvisnosti od časa.
- Narišemo vodoravno *središčno črto* pri μ.
- Narišemo vodoravni kontrolni črti pri *mejnih vrednostih* $\mu \pm \frac{3\sigma}{\sqrt{n}}$.

Katerikoli $\overline{x}$, ki ne leži med obema kontrolnima črtama, je znak, da smo izgubili nadzor nad procesom.

Statistične ideje so temelji kontrolnih diagramov. Najprej se zavemo, da se v vseh procesih pojavljajo variacije. Naš cilj ni preprečiti te variacije, temveč le ločiti naravna odstopanja v procesu od tistih dodatnih, ki nas opozorijo na motnje. Poleg tega z uporabo normalne vzorčne porazdelitve za $\overline{x}$ in pravila 68–95–99,7 opišemo območje naravnih variacij. Nazadnje združimo to formalno sklepanje z diagramom podatkov, ki ga lahko uporabljajo tudi zaposleni v tovarnah, ki nimajo veliko statistične izobrazbe.

Zakaj vsako uro v vzorec izberemo le štiri zaslone? Namen statistične kontrole procesov ni preverjati funkcionalnosti zaslonov, te bodo podrobno pregledali, ko bodo dokončani. Cilj je spremljanje nameščanja mreže in pravočasna odprava morebitnih napak. Ni praktično preverjati vsakega zaslona na vsakem koraku proizvodnega procesa. Namesto tega nam statistične metode vzorčenja priskrbijo hiter in ekonomičen

način, da poskrbimo za gladko delovanje procesa. Kontrolni diagrami, ki temeljijo na vzorcih, zmanjšajo stroške, ki bi lahko nastali, če napak ne bi odkrili pravočasno, ker dopuščajo, da napake sproti odpravljamo. Tako ni potrebno popravljati ali zavreči končanih izdelkov.

V praksi ne iščemo le posameznih točk, ki padejo preko mej, ampak tudi druge sumljive vzorce.

Primer. (Signal "niza devetih") Eden od pogostih signalov, ki nas opozarjajo na napake, je *niz* devetih zaporednih točk nad ali pod središčno črto. Tak niz je zelo malo verjeten, če ostaja srednja vrednost procesa pri ciljni vrednosti, s pomočjo katere narišemo središčno črto – enaka je verjetnosti, da pri devetih zaporednih metih kovanca dobimo same glave. Tak niz torej napeljuje na možnost, da se je srednja vrednost odmaknila s središčne črte.

Na diagramu $\overline{x}$ s slike 4.6 nas signal niza devetih opozori na napako šele pri vzorcu številka 20. Signal, pri katerem smo pozorni na posamezne točke, ki prekoračijo katero od mejnih črt, nas opozori na napako že pri vzorcu 14. V tem primeru je bil signal z nizom devetih počasen pri odkrivanju napake. Kadar pa se srednja vrednost procesa le počasi odmika od ciljne vrednosti, nas bo signal niza devetih opozoril na napako veliko prej, preden bo katerakoli posamezna meritev padla iz dovoljenega območja. Zaradi tega pogosto uporabljamo oba postopka za odkrivanje napak istočasno. ♦

4.5 Nevarnosti analize podatkov

Statistični načrti zbiranja podatkov lahko, kot v primeru študije na zdravnikih, vključujejo eksperimente. Lahko pa tudi uporabljajo posebne postopke vzorčenja kot so na primer tisti, ki jih v ZDA uporabljajo pri raziskavah prebivalstva ali pri kontroli procesov. V obeh primerih se zanašamo na slučajenje in matematično teorijo verjetnosti za izračun vzorčnih porazdelitev. Iz teh lahko potem dobimo rezultate z znanimi stopnjami zaupanja. Vendar pa je formalno statistično sklepanje, ki se odraža v stopnjah zaupanja, sekundarnega pomena ob dobro zasnovanem zbiranju podatkov in vpogledu v njihovo obnašanje. Sklepanje je neuporabno v primeru prostovoljnih vzorcev in ne more odpraviti napak, ki nastanejo zaradi neodziva pri raziskavah na vzorcih. Še več, učinki *skritih spremenljivk* lahko celo iz navidezno jasnega sklepa napravijo zavajujočega. Kot smo videli v prvem poglavju, lahko pri dobro osnovanem eksperimentu nadzorujemo pomešanje skritih spremenljivk z

obrazložitvenimi. Kadar eksperiment ni možen, pa moramo včasih postati statistični detektivi. Oglejmo si primer. Čeprav je izmišljen, temelji na resnični situaciji.

Primer. (Katera bolnica je varnejša?) Z namenom pomagati potrošnikom, da bi lažje sprejemali odločitve, povezane z zdravstvom, vlada objavi podatke o usodah pacientov v posameznih bolnicah. Primerjati želiš bolnico A in bolnico B, ki sta v tvoji bližini. Tabela 4.2 predstavlja podatke o preživetju pacientov, ki so bili operirani v teh dveh bolnicah. Vključeni so vsi pacienti, ki so bili nedavno operirani, "preživeti" pa pomeni, da je pacient živel še vsaj 6 tednov po operaciji.

V bolnici A izgubijo 3% ($\frac{63}{2100}$) pacientov, v bolnici B pa le 2% ($\frac{16}{800}$). Zdi se, da je bolnica B boljša izbira, če potrebuješ operacijo. ♦

	Bolnica A	**Bolnica B**
Umrli	63	16
Preživeli	2037	784
Skupaj	2100	800

Tabela 4.2: Podatki o preživetju pacientov.

Tabela 4.2 je **dvosmerna tabela**. Kadar spremenljivke le razvrstijo osebke v kategorije (na primer umrl ali preživel), ne moremo narisati razsevnega diagrama, da bi prikazali zveze med njimi. Namesto tega prikažemo števila v dvosmerni tabeli in dobimo zvezo tako, da primerjamo deleže. Primerjava deležev pacientov, ki so umrli, nam pokaže, da je stopnja umrljivosti v bolnici B nižja. Poglejmo še bolj pozorno.

Primer. (Skrita spremenljivka udari) Niso vsi kirurški posegi enako resni. V nadaljevanju vladnega poročila so podatki o posegih, razdeljeni glede na predoperativno stanje pacientov. Opišemo ga kot "dobro" ali "slabo". Ti podrobnejši podatki so povzeti v tabeli 4.3. Prepričaj se, da so vnosi v prejšnji tabeli le vsote ustreznih števil iz te tabele.

Bolnica A premaga bolnico B pri pacientih, ki so bili v dobrem stanju: izgubili so jih le 1% ($\frac{6}{600}$), medtem ko je v bolnici B izguba znašala 1,3% ($\frac{8}{600}$). Prav tako bolnica A zmaga pri pacientih v slabem stanju z le 3,8% ($\frac{57}{1500}$) proti 4% ($\frac{8}{200}$) iz bolnišnice B. Torej je bolnica A varnejša izbira za paciente ne glede na njihovo stanje. Če se torej pripravljaš na operacijo, bi bilo pametneje izbrati bolnico A. ♦

Ta primer služi kot opozorilo k statističnim dokazom, še posebej v tistih primerih, ko podatkov ne dobimo z eksperimenti. Če ne upoštevamo stanja pacientov, se zdi

	Dobro stanje			Slabo stanje	
	Bolnica A	**Bolnica B**		**Bolnica A**	**Bolnica B**
Umrli	6	8	**Umrli**	57	8
Preživeli	594	592	**Preživeli**	1443	192
Skupaj	600	600	Skupaj	1500	200

Tabela 4.3: Podatki o preživetju pacientov glede na zdravstveno stanje.

bolnica B boljša izbira, čeprav se v resnici bolnica A odreže bolje ne glede na stanje. Kako lahko gre A bolje v obeh skupinah, pa vseeno slabše v skupnem seštevku? Oglejmo si še enkrat podatke. Bolnica A je zdravstveni center, ki pritegne resno bolne paciente iz celotne regije. Sprejeli so kar 1500 pacientov v slabem stanju. V bolnici B so imeli le 200 takih pacientov. Ker je za paciente v slabem stanju bolj verjetno, da bodo umrli, je bila stopnja umrljivosti pri bolnici A večja, kljub temu, da se je odrezala bolje kot B v vsakem posameznem razredu. Tabela 4.2 je zavajujoča, ker ne upošteva stanja pacientov. Statistično sklepanje, ki bi temeljilo le na podatkih iz tabele 4.2, bi bilo prav tako zavajujoče.

Tudi če podatke pridobimo previdno in jih pravilno analiziramo, ne moremo biti absolutno prepričani v pravilnost naših zaključkov. Vedno obstaja možnost, ne glede kako majhna, da nas bo slučajna izbira pripeljala do napačnih sklepov. Moč statističnih sklepov je v tem, da poznamo možnosti za napačne zaključke in jih lahko nadzorujemo s tem, da postavimo stopnjo zaupanja tako visoko, kot se nam zdi potrebno.

Statistika torej ne proizvaja dokazov. Vendar pa je v svetu, ki vedno zahteva resnico in v katerem je večina dokazov nezanesljivih, statistika največkrat najboljša možna izbira.

4.6 Slovarček

delež vzorca (ang. sample proportion) Delež $\hat{p}$ tistih elementov iz vzorca, ki imajo neko lastnost. V primeru enostavnega slučajnega vzorca ga uporabimo kot oceno za ustrezni delež p celotne populacije, iz katere je bil izbran vzorec.

dvosmerna tabela (ang. two-way table) Tabela, ki navaja izide glede na dve spremeljivki (npr. bolniki, razporejeni glede na zdravstveno stanje in glede na

uspešnost zdravljenja).

interval zaupanja (ang. confidence interval) Interval, ki ga izračunamo iz vzorca s pomočjo metode, ki z znano verjetnostjo poda interval, ki vsebuje neznani parameter. To verjetnost imenujemo *stopnja zaupanja*; običajno imajo intervali zaupanja obliko

$$\text{ocena} \pm \text{meja napake}.$$

kontrolni diagram (ang. control chart) Graf, ki prikazuje vrednost statistike pri zaporednih vzorcih (npr. en vzorec na uro ali en vzorec vsako izmeno). Graf vsebuje tudi sredinsko črto pri ciljni vrednosti procesnega parametra in kontrolni mejni vrednosti, ki jih statistika običajno ne prestopi, razen če se proces oddalji od ciljne vrednosti. Namen kontrolnega diagrama je spremljanje procesa v času in odkrivanje neobičajnih vplivov na proces.

parameter (ang. parameter) Število, ki opisuje populacijo. Običajno je cilj statističnega sklepanja oceniti neznani parameter ali pa določiti njegovo vrednost.

statistika (ang. statistic) Število, ki opisuje vzorec. Izračunamo jo lahko iz vzorca in pri tem ne potrebujemo nobenih neznanih parametrov, ki opisujejo populacijo.

vzorčno povprečje (ang. sample mean) Srednja vrednost (aritmetično povprečje) $\overline{x}$ vseh vrednosti v vzorcu. V primeru enostavnega slučajnega vzorca ga uporabimo kot oceno za neznano srednjo vrednost μ za populacijo, iz katere je bil izbran vzorec.

4.7 Dodatna literatura

- David S. Moore, *The Basic Practice of Statistics*, 2. izdaja, Freeman, New York, 1999. Šesto poglavje te knjige predstavlja podrobnosti statističnega sklepanja. V sedmem in osmem poglavju avtor obravnava praktično uporabo sklepanja pri ocenjevanju srednjih vrednosti in deležev.

- Lincoln E. Moses, *The Reasoning of statistical interference,* Perspectives on Contemporary Statistics*, Mathematical Association of America, Washington, D.C., 1992, str. 107–122.* Ta esej o naravi sklepanja zahteva nekaj znanja verjetnosti in pred tem je priporočljivo prebrati zgoraj omenjeno knjigo.

Zanimivo in poučno je videti, kako vodilne agencije opisujejo natančnost svojih raziskav. Celotna tiskovna poročila, ki jih izdajajo pri agenciji Gallup, so dostopna na www.gallup.com. Harrisova raziskava je dosegljiva na www.louisharris.com pod "*Harris Poll this week*". Mesečna poročila o stopnji nezaposlenosti shranjuje Urad za statistiko dela na stats.bls.gov. Nacionalno združenje raziskovalnih agencij NCPP ima na svoji strani www.ncpp.org nekaj precej zanimivega materiala (*Principles of Disclosure, 20 Questions for Journalists*).

4.8 Preverjanje znanja

(1) Slučajni vzorec 10 vreč sladkorja ima povprečno težo 4,9 funta, kar je manj kot povprečna teža vseh proizvedenih vreč sladkorja, ki znaša 5,05 funta. V tem primeru je 4,9

(a) statistika.

(b) parameter.

(c) vzorec.

(2) Da bi ugotovili, koliko je zanimanja za novi park, anketiramo 300 okoliških prebivalcev. Med njimi jih je 135 naklonjenih parku. Kolikšen je delež vzorca $\hat{p}$?

(a) 0,45%

(b) 40,5%

(c) 45%

(3) Enostavni slučajni vzorec 400 prebivalcev nekega mesta povprašamo za mnenje in 144 jih je za novi gasilski dom. Kolikšen je približno standardni odklon deleža vzorca $\hat{p}$ za ta primer?

(a) 6%

(b) 2,4%

(c) približno 1%

(4) Tovarna izdeluje vreče sladkorja s povprečno težo 5,05 funta in standardnim odklonom 0,05 funta. Slučajno izberemo štiri vreče in izračunamo njihovo povprečno težo. Kolikšen je standardni odklon vzorčne porazdelitve povprečne teže?

(a) 0,0125 funta

(b) 0,025 funta

(c) 0,1 funta

(5) S slučajnim vzorcem velikosti 2000, ki ga izberemo med prebivalci Spodnjega Kašlja, ugotovimo, da jih 45 še nikoli ni imelo noric. Poišči 95% interval zaupanja za dejanski delež meščanov, ki še nikoli niso imeli noric.

(a) 2,25%± 0,3316%

(b) 2,25%± 0,6632%

(c) 2,25%± 0,2199%

(6) Vreče sladkorja imajo povprečno težo 5,05 funta in standardni odklon 0,05 funta. Tovarna uporablja 95% kontrolni limiti v kontrolnem diagramu proizvodnega procesa. Izberemo slučajni vzorec 9 vreč in jih stehtamo. Povprečna teža je 5,09 funta. Katera od naslednjih trditev je pravilna?

(a) Povprečna teža vzorca je ušla nadzoru.

(b) Povprečna teža vzorca je še pod nadzorom.

(c) Nimamo dovolj informacij.

(7) Spodaj je dvosmerna tabela, ki prikazuje sprejem gostov v privatni klub. Poišči delež sprejetih moških.

	Moški	**Ženske**
Sprejeti	20	20
Zavrnjeni	30	10

(a) 20%

(b) 25%

(c) 40%

4.9 Naloge

Ocenjevanje deleža populacije

Za vsako od debelo natisnjenih števil v nalogah 1–3 povej, ali gre za *parameter* ali *statistiko*.

(1) Za slučajni vzorec študentk določimo povprečno višino **65** inčev, kar je več kot povprečna višina vseh Američank, ki znaša **63** inčev.

(2) Raziskovalec izvede slučajeni primerjalni eksperiment na mladih podganah, da bi ugotovil vpliv toksične snovi v hrani. Kontrolna skupina dobi nekontaminirano hrano, eksperimentalna skupina pa dobi v hrani 2500 delcev na milijon toksične snovi. Po osmih tednih je povprečna pridobljena teža v kontrolni skupini **335** g, v eksperimentalni pa **289** g.

(3) Podjetje za telefonsko oglaševanje iz Los Angelesa uporablja napravo, ki slučajno izbira telefonske številke iz mesta. Od prvih 100 klicanih jih **48**% ni bilo v imeniku. To ni presenetljivo, ker vemo, da **52**% vseh telefonskih številk ni v imeniku.

(4) Tonya želi oceniti, kolikšen delež študentov v njenem študentskem domu je zadovoljnih s hrano v študentski menzi. Ankerira enostavni slučajni vzorec 50 od 680 študentov, ki stanujejo v tem domu. Od tega jih 14 meni, da je hrana dobra.

(a) Opiši populacijo in z besedami povej, kaj je parameter p.

(b) Podaj (v %) vrednost statistike $\hat{p}$, ki ocenjuje p.

(c) Recimo, da je v resnici 25% študentov zadovoljnih s hrano iz menze. Kolikšna sta srednja vrednost in standardni odklon vzorčne porazdelitve $\hat{p}$?

(5) PTC je spojina, ki ima za nekatere močan grenak okus, drugim pa se zdi brez okusa. Sposobnost zaznavanja PTC je dedna. Približno 75% Italijanov lahko zazna PTC. Oceniti želiš delež Američanov, pri katerih je vsaj eden od starih staršev iz Italije in ki lahko okusijo PTC. Predpostavi, da ocena o 75% deležu Italijanov drži za to populacijo in testiraj 500 ljudi. Skiciraj normalno krivuljo, ki prikazuje, kako se bo delež $\hat{p}$ vseh tistih, ki okusijo PTC, spreminjal pri večkratnem vzorčenju.

(6) V eni od držav na ameriškem srednjem zahodu ima 84% gospodinjstev za praznike ob koncu leta božično drevesce. V raziskavi slučajni vzorec 400 gospodinjstev vprašamo "Ste letos postavili božično drevesce?" Kakšna je vzorčna porazdelitev deleža pozitivnih odgovorov?

(7) Standardni odklon $\sigma_{\hat{p}}$ vzorčnega deleža $\hat{p}$ se spreminja skupaj z dejansko vrednostjo deleža populacije p. Na srečo se ne spreminja veliko, če p ni blizu

0% ali 100%. Recimo, da je velikost vzorca $n = 1500$. Izračunaj $\sigma_{\hat{p}}$ za $p =$ 30%, 40%, 50%, 60% in 70%. Nato izračunaj $\sigma_{\hat{p}}$ za $p = 0\%$, 10% in 20%. Na katerem delu se $\sigma_{\hat{p}}$ spreminja najhitreje, ko se spreminja p? Nariši graf $\sigma_{\hat{p}}$ v odvisnosti od p.

Intervali zaupanja

(8) Poročilo o raziskavi na vzorcu 1500 odraslih pravi: "S stopnjo zaupanja 95% med 27% in 33% vseh Američanov meni, da so droge največji problem v javnih šolah." Razloži nekomu, ki ne zna statistike, kaj v tem primeru pomeni "s stopnjo zaupanja 95%".

(9) Pri Gallupovi raziskavi so vprašali slučajni vzorec 1005 odraslih, če podpirajo dvojezične javne šole ali pa menijo, da bi se morali učenci, ki ne govorijo angleško, vključiti v angleške šole in se naučiti jezika na ta način. Pri tem je bilo 63% za enojezične šole. V tisku so objavili, da je bila meja napake te raziskave 3%. Razloži, kaj to pomeni.

(10) Recimo, da so v raziskavi iz prejšnje naloge uporabili enostavni slučajni vzorec velikosti 1005 in ugotovili, da je 60% vprašanih za enojezične šole. Podaj 95% interval zaupanja za delež odraslih, ki bi odgovorili, da so za enojezične šole.

(11) V preteklem letu je 73% vprašanih študentov prvega letnika v nacionalni raziskavi odgovorilo, da je "dobra finančna preskrbljenost" zanje pomemben cilj. Državna univerza je ugotovila, da je v enostavnem slučajnem vzorcu 200 študentov prvega letnika 132 študentov označilo ta cilj kot pomemben. Podaj 95% interval zaupanja za delež prvih letnikov na univerzi, za katere je to pomemben cilj.

(12) V ZDA razmišljajo, da bi dodatno omejili število vozil, ki lahko vstopijo v Yellowstonski nacionalni park. Da bi ocenili odziv javnosti, vprašajo enostavni slučajni vzorec 150 obiskovalcev, če podpirajo takšne ukrepe. Od teh jih 89% odgovori pritrdilno. Podaj 95% interval zaupanja za delež obiskovalcev parka, ki so za omejitev. Ali lahko s 95% stopnjo zaupanja trdiš, da jih je več kot polovica za omejitev? Odgovor utemelji.

(13) *New York Times* in *CBS News* sta izvedla nacionalno raziskavo na 1048 slučajno izbranih najstnikih med 13. in 17. letom. Med njimi jih je 692 imelo

televizijski sprejemnik v svoji sobi in 189 jih je kot svoj najljubši program navedlo *Fox*. Predpostavljali bomo, da je šlo za enostavni slučajni vzorec.

(a) Podaj 95% interval zaupanja za delež vseh ljudi te starosti, ki imajo v svoji sobi TV sprejemnik, in za tiste, ki najraje gledajo *Fox*.

(b) Časopisni članek pravi, "Teoretično se v 19 primerih od 20 rezultati ne razlikujejo za več kot 3 odstotne točke od vrednosti, ki bi jo dobili, če bi anketirali vse ameriške najstnike." Pojasni, kako se tvoji izračuni ujemajo s to trditvijo.

(14) V raziskavo o enakopravnosti žensk v ZDA, ki jo je izvedla televizijska hiša MSNBC, je bilo vključenih 1019 odraslih oseb. Časopisni članek, ki je poročal o raziskavi, je navajal, "Meja napake pri rezultatih je 3 odstotne točke."

(a) Skupno je 54% vzorca (550 od 1019 ljudi) odgovorilo, da se je na tem področju naredilo dovolj. Poišči 95% interval zaupanja za delež odraslih, ki bi odgovorili pritrdilno. Ali je poročilo o meji napake približno pravilno? (Predpostavi, da je šlo za enostavni slučajni vzorec.)

(b) Časopisni članek je trdil, da 65% moških in le 43% žensk meni, da se je na tem področju storilo dovolj. Pojasni, zakaj nimamo dovolj informacij, da bi lahko podali intervale zaupanja posebej za moške in za ženske.

(c) Ali bi bila meja napake pri 95% intervalu zaupanja za ženske večja, manjša ali enaka 0,03? Zakaj? Navedbe časopisa o meji napake so očitno nekoliko zavajujoče.

Naloge 15–18 temeljijo na naslednji situaciji: Poročilo pravi, da so pri nacionalni raziskavi na 1500 slučajno izbranih odraslih ugotovili, da jih je 43% mnenja, da jim bo šlo naslednje leto slabše. V nadaljevanju poročila je bilo zapisano, da je meja napake 3 odstotne točke s 95% stopnjo zaupanja.

(15) Kateri od naslednjih virov napak so vključeni v mejo napake iz poročila?

(a) Pri raziskavi so slučajno izbirali telefonske številke, zato so pri tem izpustili ljudi, ki niso imeli telefona.

(b) Neodziv: pri nekaterih od teh številk se ni nihče oglasil ali pa je klicani zavrnil sodelovanje.

(c) Slučajne variacije pri slučajnem izbiranju telefonskih številk.

(16) Ali bi bila pri 90% intervalu zaupanja na osnovi rezultatov te ankete meja napake večja, manjša ali enaka 3 odstotnim točkam? Zakaj?

(17) Recimo, da bi pri raziskavi anketirali 1000 ljudi namesto 1500 (in spet ugotovili, da jih je 43% mnenja, da jim bo šlo prihodnje leto slabše). Ali bi bila meja napake za 95% interval zaupanja večja, manjša ali enaka 3 odstotnim točkam? Zakaj?

(18) Recimo, da smo rezultat 43% dobili s podobno metodo enostavnega slučajnega vzorčenja za vse odrasle v državi New York (z 18 milijoni prebivalcev) ne pa za celotne ZDA (z 270 milijoni prebivalcev). Ali bi bila meja napake 95% intervala zaupanja večja, manjša ali enaka 3 odstotnim točkam? Zakaj?

Ocenjevanje srednje vrednosti populacije

(19) Pošiljka mehanskih delov ima kritično dimenzijo, ki je normalno porazdeljena s srednjo vrednostjo 12 cm in standardnim odklonom 0,01 cm. Sprejemna skupina premeri slučajni vzorec 25 delov. Kakšna je vzorčna porazdelitev vzorčnega povprečja $\overline{x}$ kritičnih dimenzij za te dele?

(20) Rezultati študentov na ACT sprejemnih testih v preteklem letu so bili normalno porazdeljeni s srednjo vrednostjo $\mu = 18{,}6$ in standardnim odklonom $\sigma = 5{,}9$.

(a) Kateri interval vsebuje srednjih 95% rezultatov?

(b) Izračunamo povprečje 25 slučajno izbranih rezultatov. Kateri interval vsebuje srednjih 95% povprečij $\overline{x}$?

(21) Napake pri natančnih merjenjih so velikokrat normalno porazdeljene. Izkušnje kažejo, da se napake pri kontrolnih metodah spreminjajo, ko meritev ponavljamo, v skladu z normalno porazdelitvijo s srednjo vrednostjo 0 (se pravi, da postopek ne bo sistematično precenil ali podcenil dejanske razdalje) in standardnim odklonom 0,03 m. Geodet ponovi vsako meritev trikrat in za končno vrednost uporabi povprečje treh meritev. Napaka pri tej vrednosti je povprečna napaka $\overline{x}$ teh treh zaporednih meritev.

(a) Kakšna je porazdelitev povprečne napake $\overline{x}$, ko geodet izmeri veliko razdalj?

(b) Med katerima vrednostima leži 95% napak?

(22) Pri študiji poklicnih poti upravnikov hotelov so poslali vprašalnike enostavnemu slučajnemu vzorcu 160 hotelov iz velikih ameriških hotelskih verig. Prejeli so 114 odgovorov. Povprečen čas, v katerem je teh 114 upravnikov delalo pri svojem trenutnem podjetju, je bil $\overline{x} = 11{,}78$ let. Ne poznamo standardnega odklona populacije σ, vzorčni standardni odklon pa je enak $s = 3{,}2$ leta. Ker je vzorec velik, je s blizu σ. Podaj 95% interval zaupanja za povprečno število let, ki so jih upravniki velikih verig preživeli pri svojem trenutnem podjetju.

(23) Pri laboratorijski tehtnici je standardni odklon $\sigma = 0{,}001$ g pri večkratnih tehtanjih. Predpostavi, da so meritve pri večkratnih tehtanjih normalno porazdeljene s srednjo vrednostjo, ki je enaka dejanski teži tehtanega predmeta. Pri treh tehtanjih nekega primerka smo dobili

3,412 3,414 3,415

Podaj 95% interval zaupanja za dejansko težo primerka. Kolikšna sta ocena in meja napake pri tem intervalu?

(24) V spodnji tabeli so rezultati IQ testa 31 učenk 7. razreda iz neke ameriške šole.

114	100	104	89	102	91
114	114	103	105	108	130
120	132	111	128	118	119
86	72	111	103	74	112
107	103	98	96	112	112
93					

(a) Pričakujemo, da bo porazdelitev rezultatov blizu normalni. Napravi stebelni diagram porazdelitve teh 31 rezultatov. Ali na diagramu opaziš kakšne ubežnike, izrazito asimetričnost ali druga nenavadna odstopanja?

(b) Obravnavaj teh 31 deklic kot enostavni slučajni vzorec vseh sedmošolk iz tega okrožja. Predpostavi, da je standardni odklon rezultatov IQ testa v tej populaciji enak $\sigma = 15$. Podaj 95% interval zaupanja za povprečni rezultat v tej populaciji.

(c) V resnici pripadajo ti rezultati vsem sedmošolkam ene od šol iz tega okrožja. Natančno razloži, zakaj se ne moremo zanesti na interval zaupanja iz točke (b).

(25) Poišči mejo napake za 95% interval zaupanja iz naloge 23, če tehtamo vsak primerek dvanajstkrat in ne le trikrat. Prepričaj se, da je tvoj rezultat dvakrat manjši od meje napake iz naloge 23. Razloži, zakaj lahko že brez računanja ugotovimo, da bo nova meja napake dvakrat manjša.

(26) Pri NAEP testu je sodelovalo tudi 1077 žensk med 21. in 25. letom starosti. Njihov povprečni rezultat je bil 275. Predpostavi, da je standardni odklon posameznih rezultatov enak $\sigma = 60$.

(a) Podaj 95% interval zaupanja za povprečni rezultat μ v populaciji vseh žensk med 21. in 25. letom.

(b) Predpostavi, da so dobili enak rezultat, $\overline{x} = 275$, za vzorec 250 žensk. Podaj 95% interval zaupanja za srednjo vrednost populacije v tem primeru.

(c) Predpostavi, da so pri vzorcu 4000 žensk dobili vzorčno povprečje $\overline{x} = 275$ in še enkrat izračunaj 95% interval zaupanja za μ.

(d) Kolikšne so meje napake za vzorce velikosti 250, 1077 in 4000? Kako večja velikost vzorca vpliva na mejo napake za interval zaupanja?

(27) Naprava za predelovanje mleka spremlja število bakterij na mililiter v surovem mleku, ki ga sprejmejo v predelavo. Za slučajni vzorec 10 enomililiterskih primerkov so podatki zbrani v spodnji tabeli.

5370	4890	5100	4500	5260
5150	4900	4760	4700	4870

Predpostavimo, da je število bakterij normalno porazdeljeno in je standardni odklon $\sigma = 265$ na ml. Podaj 95% interval zaupanja za povprečno število bakterij na ml v mleku tega proizvajalca.

(28) V radijski oddaji povabijo poslušalce k razpravi o predlaganem povišanju plač mestnih svetnikov. "Kako visoke bi morale biti po vašem mnenju plače mestnih svetnikov? Pokličite nas in povejte svoje mnenje." Skupno pokliče 958 ljudi. Povprečna predlagana plača je $\overline{x} = 8740\,€$ na leto in standardni odklon

odgovorov je $s = 1125$ €. Pri velikih vzorcih kakršen je ta je s zelo blizu neznanemu standardnemu odklonu za populacijo, σ. Radijska postaja izračuna 95% interval zaupanja za povprečno plačo μ in ugotovi, da bi bilo povprečje predlogov vseh meščanov med 8667 € in 8813 €.

(a) Pokaži, da je zgornji izračun pravilen.

(b) Kljub temu dobljeni rezultat ne velja za celotno populacijo. Pojasni, zakaj.

Statistični nadzor procesov

Pri obravnavanju kontrolnih diagramov uporabljaj hkrati signal "ene zunanje točke" in signal "niza devetih".

(29) Proizvajalec avtomobilskih klimatskih naprav vsako uro proizvodnje pregleda vzorec štirih termostatov. Termostati so nastavljeni na $75°F$ in nato vstavljeni v komoro, v kateri postopoma zvišujejo temperaturo. Zabeležijo temperaturo, pri kateri termostat vključi klimatsko napravo. Standard za srednjo vrednost je $\mu = 75°$. Pretekle izkušnje kažejo, da se odzivna temperatura pravilno nastavljenih termostatov spreminja s $\sigma = 0{,}5°$. Povprečno odzivno temperaturo $\overline{x}$ za vsakega od vzorcev vnašajo v kontrolni diagram za $\overline{x}$. Izračunaj središčno črto in mejni vrednosti za ta diagram.

(30) Širina reže, ki jo izreže frezalni stroj, je pomembna za pravilno delovanje hidravličnih sistemov pri velikih traktorjih. Proizvajalec vsako uro preverja širino z vzorcem petih zaporednih izdelkov. Povprečno širino reže za vsak vzorec vnese na kontrolni diagram. Ciljna vrednost je $\mu = 0{,}8759$ inča. Frezalni stroj pri pravilni nastavitvi izdeluje reže s povprečno širino enako ciljni vrednosti in standardnim odklonom $\sigma = 0,0012$ inča. Določi središčno črto in obe mejni vrednosti za kontrolni diagram $\overline{x}$.

(31) Laboratoriji imajo ponavadi kontrolni diagram za proces merjenja, ki temelji na večkratnih meritvah standardnega primerka. Tvoja zadolžitev je kontrolni diagram za tehtnico iz naloge 23. V rednih časovnih intervalih moraš trikrat zapored stehtati standardno 5-gramsko utež. Določi sredinsko črto in mejni vrednosti za ta kontrolni diagram.

(32) Premer potisnega ležaja v električnem motorju naj bi bil 2.205 cm. Kadar je proizvodni proces pravilno nastavljen, izdelujemo ležaje s povprečnim premerom 2,2050 cm in standardnim odklonom 0,0010 cm. Vsako uro izmerimo vzorec pet zaporedoma izdelanih ležajev. V spodnji tabeli so vzorčna povprečja $\overline{x}$ za 12 ur.

Napravi kontrolni diagram $\overline{x}$ za diameter potisnega ležaja. Uporabi oba možna signala, da oceniš, če je proces pod kontrolo. Kdaj bi morali ukrepati, da bi popravili potek procesa?

Ura	1	2	3	4
$\overline{x}$	2,2047	2,2047	2,2050	2,2049
Ura	5	6	7	8
$\overline{x}$	2,2053	2,2043	2,2036	2,2042
Ura	9	10	11	12
$\overline{x}$	2,2038	2,2045	2,2026	2,2040

Naloge 33–35 se nanašajo na naslednje podatke: Proizvajalec zdravil oblikuje tablete s pomočjo kompresije granul, ki vsebujejo aktivno sestavino in razna polnila. Za vsako skupino tablet izmerijo trdoto vzorca in tako nadzorujejo proces kompresije. Ciljni vrednosti za trdoto sta $\mu = 11{,}5$ in $\sigma = 0{,}2$. V tabeli 4.4 so tri skupine podatkov, od katerih vsaka predstavlja povprečje $\overline{x}$ za 20 zaporednih vzorcev $n = 4$ tablet.

(33) Napravi kontrolni diagram $\overline{x}$ za skupino A iz tabele 4.4. Srednja vrednost procesa μ se je med zbiranjem teh podatkov nenadoma prestavila k novi vrednosti. Ali je na diagramu opazno uhajanje izpod nadzora? Kako? Pri katerem od vzorcev se je spremenila srednja vrednost?

(34) Srednja vrednost procesa μ se je ustalila pri ciljni vrednosti $\mu = 11{,}5$ med zbiranjem podatkov iz skupine B. Vzorčna povprečja se spreminjajo, ampak le v okviru pričakovanih slučajnih odstopanj. Izdelaj kontrolni diagram $\overline{x}$ in ga komentiraj.

(35) Skupina podatkov C iz tabele 4.4 ilustrira učinek enakomernega premikanja srednje vrednosti populacije. Proces je stabilen pri prvih desetih vzorcih, nato pa se začne srednja vrednost μ vztrajno dvigati. Napravi kontrolni diagram

Vzorec	Skupina A	Skupina B	Skupina C
1	11,602	11,627	11,495
2	11,547	11,613	11,475
3	11,312	11,493	11,465
4	11,449	11,602	11,497
5	11,401	11,360	11,573
6	11,608	11,374	11,563
7	11,471	11,592	11,321
8	11,453	11,458	11,533
9	11,446	11,552	11,486
10	11,522	11,463	11,502
11	11,664	11,383	11,534
12	11,823	11,715	11,624
13	11,629	11,485	11,629
14	11,602	11,509	11,575
15	11,756	11,429	11,730
16	11,707	11,477	11,680
17	11,612	11,570	11,729
18	11,628	11,623	11,704
19	11,603	11,472	12,052
20	11,816	11,531	11,905

Tabela 4.4: Tri skupine $\overline{x}$ za 20 vzorcev velikosti 4.

$\overline{x}$ za te podatke. Ali je razvidno kakšno uhajanje izpod nadzora? Ali je na diagramu opazno plezanje μ navzgor?

Nevarnosti analize podatkov

(36) Kako na valjenje jajc vodnega pitona vpliva temperatura kačjega gnezda? Raziskovalci so razdelili novo znesena jajca v tri skupine glede na temperaturo: vroče, nevtralno in mrzlo. V prvi skupini so posnemali toploto, ki jo običajno priskrbi samica pitona, pri tretji skupini pa so simulirali odsotnost samice. V spodnji tabeli so podatki o številih jajc in njihovih usodah. (Vir: R. Shine, T. R. L. Madsen, M. J. Elphick, P. S. Harlow, The influence of nest temperatures and maternal brooding on hatchling phenotypes in water pythons, *Ecology*, 78

(1997): 1713–1721.)

	Št. jajc	Izvaljeni
Hladno	27	16
Nevtralno	56	38
Vroče	104	75

(a) Izdelaj dvosmerno tabelo temperature v odvisnosti od izida (izleglo ali ne).

(b) Za vsako skupino izračunaj delež jajc, iz katerih so se izlegli pitoni. Raziskovalci so pričakovali, da bo mrzlo okolje zmanjšalo število izvaljenih jajc. Ali podatki podpirajo njihova pričakovanja?

(37) V študiji vpliva kadilskih navad staršev na navade srednješolcev so raziskovalci anketirali dijake iz osmih srednjih šol v Arizoni. Rezultati so povzeti v spodnji tabeli. (Vir: S. V. Zagona, ed., *Studies and Issues in Smoking Behavior*, University of Arizona Press, Tucson, 1967, str. 157–180.)

	Študent kadi	Študent ne kadi
Oba starša kadita	400	1380
Eden od staršev kadi	416	1823
Nobeden od staršev ne kadi	188	1168

Opiši povezavo med kadilskimi navadami staršev in njihovih otrok, tako da izračunaš različne deleže in jih primerjaš. Nato povzemi rezultate še s preprostimi besedami.

(38) Strelno orožje je drugi najpogostejši vzrok smrti iz nezdravstvenih razlogov (prvi so motorna vozila). V tabeli so zbrani podatki za Milwaukee, Wisconsin, med leti 1990 in 1994. Primerjati želimo vrste strelnega orožja, ki so bile uporabljene v umorih in v samomorih. Predpostavljamo, da bodo šibrovke in lovske puške bolj pogosto uporabljene pri samomorih, ker jih ima veliko ljudi doma za lov. Kaj pravi o tem spodnja tabela? (Vir: S. W. Hargarten et al., Characteristics of firearms involved in fatalities, *Journal of American Medical Association*, 275 (1996): 42–45.)

	Umori	**Samomori**
Pištola	468	124
Šibrovka	28	22
Lovska puška	15	24
Nedoločeno	13	5
Skupaj	524	175

(39) V tabeli so podatki o točnih letih in zamudah dveh letalskih družb na petih različnih letališčih v obdobju enega meseca. V medijih pogosto poročajo o skupnem deležu točnih letov. Zaradi skritih spremenljivk so lahko taka poročila zavajujoča. (Vir: A. Barnett, How numbers can trick you, *Technology Review*, oktober 1994, str. 38–45.)

	Alaska Airlines		**America West**	
	Točen	Zamujal	Točen	Zamujal
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1841	305	201	61

(a) Kolikšen delež vseh letov družbe Alaska Airlines predstavljajo tisti z zamudami? Kolikšen je ta delež pri družbi America West? To so vrednosti, ki jih običajno navajajo v medijih.

(b) Poišči delež letov z zamudami za družbo Alaska Airlines na vsakem od petih letališč. Ponovi izračune še za družbo America West.

(c) America West se odreže slabše na čisto vsakem od petih letališč, vendar jih gre v celoti bolje. Zveni nemogoče. S pomočjo podatkov razloži, zakaj pride do tega. (Vzroki tičijo v vremenskih pogojih na letališčih v Phoenixu in v Seattlu.)

(40) Kaže, da je to, ali se obsojenemu morilcu izreče smrtna kazen ali ne, odvisno od rase žrtve. Spodaj so podatki o 326 primerih, v katerih so bili obtoženi umora spoznani za krive. (Vir: M. Radelet, Racial characteristics and imposition of the death penalty, *American Sociological Review*, 46 (1981): 918–927.)

Beli obtoženec	**Smrtna kazen**	
	Da	Ne
Bela žrtev	19	132
Črna žrtev	0	9

Črni obtoženec	**Smrtna kazen**	
	Da	Ne
Bela žrtev	11	52
Črna žrtev	6	97

(a) S pomočjo zgornjih podatkov sestavi dvosmerno tabelo rase obtoženega (bela ali črna) v primerjavi s smrtno kaznijo (da ali ne).

(b) Pokaži, da velja Simpsonov paradoks: skupno je na smrt obsojenih več belcev kot črncev, vendar pa je pri obeh rasah žrtev odstotek obsojenih črncev višji.

(c) S pomočjo podatkov razloži vzroke za paradoks tako, da te bo lahko razumel tudi sodnik.

Dodatne naloge

(41) Študent prebere, da je 95% interval zaupanja povprečnega rezultata NAEP testov za moške med 21. in 25. letom starosti med 267,8 do 276,2. Ko ga vprašamo, kaj to pomeni, pravi, da "95% moških iz te starostne skupine doseže rezultate med 267,8 in 276,2". Ali ima prav? Odgovor utemelji.

(42) V tehnološkem kotičku *New York Timesa* so 29. maja 1995 poročali o raziskavi o uporabnikih interneta, ki je pokazala, da je med njimi kar dvakrat več moških kot žensk. To je bilo presenetljivo, saj je ena prejšnjih raziskav pokazala, da je bilo moških kar devetkrat več. V nadaljevanju članka najdemo naslednje:

> Natančne raziskave so vključevale več kot 13 000 organizacij, kjer uporabljajo internet. Prejeli so 1468 uporabnih odgovorov, g. Quaterman pa je povedal, da je meja napake 2,8 odstotka pri stopnji zaupanja 95%.

(a) Kolikšna je stopnja odziva te raziskave? (Se pravi, kolikšen delež načrtovanega vzorca je odgovoril na vprašalnik?)

(b) Ali je po tvojem mnenju majhna meja napake dobro merilo natančnosti rezultatov te raziskave? Odgovor utemelji.

(43) Anketa *New York Timesa* je zajela 1025 slučajno izbranih žensk iz ZDA (vključno z Aljasko in Havaji). Med vprašanimi jih je bilo 47% mnenja, da nimajo dovolj časa zase.

(a) Zapisali so, da je meja napake ±3 odstotne točke pri 95% stopnji zaupanja. Poišči 95% interval zaupanja za delež odraslih žensk, ki menijo, da nimajo dovolj časa zase.

(b) Razloži nekomu, ki ne zna statistike, zakaj ne moremo trditi, da 47% odraslih žensk nima časa zase.

(c) Nazadnje jasno razloži, kaj pomeni "pri 95% stopnji zaupanja".

(44) Kadar je statistika, ki ocenjuje neznani parameter, normalno porazdeljena, ima 95% interval zaupanja za parameter obliko ocena $\pm 2\sigma_{\text{ocena}}$. Pri kompleksno načrtovanih vzorcih zahtevajo ocene in standardni odkloni zamotane izračune. Vendar pa lahko v primerih, ko sta dana ocena in standardni odklon, izračunamo interval zaupanja za μ, ne da bi pri tem poznali formule, ki so pripeljale do teh vrednosti.

Poročilo, ki temelji na redni študiji prebivalstva, ocenjuje, da je bila stopnja nezaposlenosti v ZDA junija 1998 4,7%. (To pomeni, da je bilo 4,7% civilistov, starejših od 16 let, ki so iskali zaposlitev, nezaposlenih.) Poročilo navaja tudi, da je bil standardni odklon približno 0,11%. Vzorec, ki ga uporabljajo pri teh raziskavah, je kompleksen večstopenjski vzorec približno 50 000 gospodinjstev. Vzorčna porazdelitev ocenjene stopnje nezaposlenosti je približno normalna. Podaj 95% interval zaupanja za stopnjo nezaposlenosti v tej populaciji.

(45) Enostavni slučajni vzorec študentov na univerzi Upper Wabash Tech so vprašali, če se strinjajo z omejitvijo vpisa pri bolj obleganih predmetih, katere namen je obdržati visoko kakovost predavanj. Študentska organizacija sumi, da tak načrt ne bo všeč brucem, ki še niso vpisali glavnih predmetov. V tabeli so podani odgovori, ki so jih dobili.

	Za	**Proti**
Bruc	40	160
Absolvent	80	20

(a) Podaj 95% interval zaupanja za delež brucev, ki podpirajo omejitev.

(b) Podaj 95% interval zaupanja za delež absolventov, ki podpirajo omejitev.

(46) Žveplove spojine povzročajo neprijeten vonj v vinu, zato želijo vinarji poznati *zaznavni prag*, najmanjšo koncentracijo, ki jo človek še lahko zazna. Zaznavni prag za dimetil sulfid (DMS) pri poklicnih poskuševalcih vina je približno 25

mikrogramov na liter vina (μg/l). Nevajeni nosovi potrošnikov pa so ponavadi manj občutljivi. Tole so podatki o zaznavnem pragu za 10 študentov:

31	31	43	36	23
34	32	30	20	24

Predpostavi, da je standardni odklon zaznavnega pragu pri nevajenih ljudeh $\sigma = 7$ μg/l.

(a) Izdelaj stebelni diagram in se tako prepričaj, da je porazdelitev približno simetrična in brez ubežnikov. (Večje število podatkov potrdi, da ni nobenih sistematičnih odstopanj od normalnosti.)

(b) Podaj 95% interval zaupanja za povprečni zaznavni prag DMS med študenti.

(c) Ali lahko z gotovostjo trdiš, da je povprečni zaznavni prag pri študentih višji kot 25 μg/l? Zakaj?

(47) V ZDA so zakladne menice varna naložba, seveda pa nas zanima, kolikšni so dobički. V tabeli so zbrani podatki o skupnem donosu menic (v %) med leti 1970 in 1996.

Leto	1970	1971	1972	1973
Dobiček	6,45	4,37	4,17	7,20
Leto	1974	1975	1976	1977
Dobiček	8,00	5,89	5,06	5,43
Leto	1978	1979	1980	1981
Dobiček	7,46	10,56	12,18	14,71
Leto	1982	1983	1984	1985
Dobiček	10,84	8,98	9,89	7,65
Leto	1986	1987	1988	1989
Dobiček	6,10	5,89	6,95	8,43
Leto	1990	1991	1992	1993
Dobiček	7,72	5,46	3,50	3,04
Leto	1994	1995	1996	
Dobiček	4,37	5,60	5,13	

(a) Napravi histogram teh podatkov. Stolpci naj bodo široki dve odstotni točki. Kakšno odstopanje od normalnosti opaziš? Zaradi centralnega limitnega izreka pa lahko kljub temu obravnavamo porazdelitev $\overline{x}$ kot normalno.

(b) Predpostavimo, da lahko obravnavamo podatke za teh 27 let kot slučajni vzorec donosov zakladnih menic. Podaj 95% interval zaupanja za dolgoročni povprečni donos. (Predpostavi, da poznaš standardni odklon donosov, $\sigma = 2{,}75\%$.)

(c) Stopnja inflacije v tem času je bila povprečno 5,5%. Ali smo lahko prepričani, da je povprečni donos zakladnih menic večji od 5,5%? Zakaj?

(48) Uporabili smo vzorčno porazdelitev $\hat{p}$ in pravilo 68–95–99,7, da smo poiskali 95% interval zaupanja za delež populacije p.

(a) Natančno razloži, zakaj je

$$\hat{p} \pm \sqrt{\frac{\hat{p}(100 - \hat{p})}{n}}$$

68% interval zaupanja za p.

(b) Izpelji formulo za 99,7% interval zaupanja za p.

(49) Uporabi nalogo 48 (a) in podatke iz naloge 12 ter podaj 68% interval zaupanja za delež obiskovalcev Yellowstonskega parka, ki podpirajo omejitev števila vozil v parku. Primerjaj dolžino 68% intervala z dolžino 95% intervala iz naloge 12 in pojasni razliko.

(50) Uporabi nalogo 48 (b) in podatke iz naloge 9 ter podaj 99,7% interval zaupanja za delež vseh odraslih, ki podpirajo enojezične šole. Primerjaj dolžino dobljenega intervala z dolžino 95% intervala zaupanja iz naloge 10. Kaj je razlog za razliko v dolžinah?

(51) Zgornji in spodnji decil vsake normalne porazdelitve se nahajata na razdalji 1,28 standardnega odklona pod in nad srednjo vrednostjo. (Spodnji decil je točka, pod katero je verjetnost 10%, zgornji decil pa tista, pod katero je verjetnost 90%.)

(a) Uporabi to dejstvo ter izpelji formulo za 80% interval zaupanja za delež populacije p na osnovi deleža vzorca $\hat{p}$, ki je točen za velike velikosti vzorca n.

(b) Podaj 80% interval zaupanja za delež obiskovalcev iz naloge 12, ki podpirajo omejitve.

(52) Zgornji in spodnji decil vsake normalne porazdelitve se nahajata na razdalji 1,28 standardnega odklona pod in nad srednjo vrednostjo.

(a) Podaj formulo za 80% interval zaupanja za srednjo vrednost μ normalne populacije s pomočjo vzorčnega povprečja $\overline{x}$ enostavnega slučajnega vzorca velikosti n.

(b) Podaj 80% interval zaupanja za povprečen IQ iz naloge 24.

(53) Stopnja donosa delnic se spreminja iz meseca v mesec. Lahko si pomagamo s kontrolnim diagramom, da vidimo, če je vzorec spreminjanja v času stabilen ali pa so morda obdobja, v katerih so delnice posebej nestabilne v primerjavi s svojim lastnim dolgoročnim vzorcem. Spodaj so povprečne mesečne stopnje donosov (v %) Wal-Martovih delnic za 38 šestmesečnih obdobij, urejene od leve proti desni po vrsticah. Podatki zajemajo obdobje od januarja 1973 do decembra 1990.

-11,78	1,68	7,88	-11,01	19,72
1,66	1,13	2,70	-0,60	5,91
2,95	0,05	1,88	6,46	2,25
8,05	4,10	2,08	4,02	11,36
8,13	0,12	1,30	-1,27	6,59
3,01	8,68	-1,52	6,64	-3,20
2,92	0,63	3,49	2,94	5,86
-0,25	6,02	5,87		

Ta števila obravnavamo kot povprečja $\overline{x}$ iz 38 vzorcev, od katerih je vsak vključeval 6 vrednosti. Narišemo jih na kontrolni diagram $\overline{x}$.

(a) Za središčno črto uporabi povprečje vseh $\overline{x}$, ki ga ponavadi v kontroli kakovosti označijo z $\overline{\overline{x}}$. To je isto kot povprečje 228 posameznih mesečnih donosov.

(b) Standardni odklon 228 mesečnih donosov vključuje tako dolgoročne kot kratkoročne trende in je v splošnem prevelik za uspešno kontrolo. Namesto tega povprečimo standardne odklone 38 vzorcev in dobimo $\overline{s} = 9450$. Predpostavi, da je $\overline{\overline{x}}$ približek za μ in $\overline{s}$ približek za σ. Določi mejni vrednosti za kontrolni diagram $\overline{x}$.

(c) Izdelaj kontrolni diagram. Ali opaziš kakšno zunanjo točko ali pa niz devetih? Ali so bila v teh 19 letih obdobja, ko so donosi ušli izpod kontrole?

(54) Spodaj so dane vsote po vrsticah in stolpcih dvosmerne tabele z dvema vrsticama in dvema stolpcema.

a	b	50
c	d	50
60	40	100

Poišči dva *različna* nabora števil a, b, c in d, ki bi dala enake vsote po vrsticah in stolpcih. Ta primer nam pokaže, da ne moremo razbrati zveze med dvema spremenljivkama le iz posameznih porazdelitev.

(55) Nedavne študije so pokazale, da so prejšnja poročila podcenila tveganje, ki ga predstavlja prekomerna telesna teža. Do napake je prišlo, ker so spregledali skrite spremenljivke. Tako na primer kajenje hkrati zmanjšuje težo in vodi v zgodnjo smrt. Ilustriraj nevarnosti, ki jih predstavljajo skrite spremenljivke, s poenostavljeno verzijo zgornje situacije. Napravi tabelo, ki bo vključevala prekomerno težo (da ali ne), zgodnjo smrt (da ali ne) in kadilske navade (da ali ne), poleg tega pa bo iz nje razvidno naslednje:

- Kadilci in nekadilci s prekomerno telesno težo umrejo prej kot tisti, ki niso predebeli.
- Ko kadilce in nekadilce združimo v dvosmerno tabelo s podatki o teži in umrljivosti, vidimo, da tisti z normalno težo umrejo prej.

4.10 Tehnološki kotiček

Testiranje generatorja naključnih števil

Velikokrat smo se že zanašali na generator naključnih števil v programu za delo s preglednicami kot nadomestek za pošteni kovanec, pošteno kocko ali dobro premešan kupček kart. Ali so bile te simulacije res poštene in nepristranske? Na to vprašanje lahko odgovorimo z uporabo vzorcev in intervalov zaupanja.

Met kovanca simuliramo tako, da izberemo slučajno število med 0 in 1, kar storimo z ukazom `=RandBetween(0,1)`. Če predpostavimo, da je program pošten, potem je verjetnost, da izbere 1, enaka $p = 0{,}5$. Če izberemo 10 vzorcev, je v 95% primerov dobljeni $\hat{p}$ v 95% intervalu zaupanja, $0{,}5 \pm 0{,}316$, torej med 0,184 in 0,816. To pomeni, da bi med 2 in 8 vzorcev moralo biti vsaj v 95% primerov enakih 1.

Naloga 1. Izvedi eksperiment, ki smo ga pravkar opisali. Izračunaj vsoto 10 števil, naključno generiranih z ukazom `=RandBetween(0,1)`. Ponovi ta eksperiment skupno dvajsetkrat. Kolikokrat je število enic padlo iz sprejemljivih mej?

Naloga 2. Če izberemo večje vzorce, se velikost 95% intervala zaupanja zmanjša. Povečaj velikost vzorca na 100 in izračunaj pripadajoči interval zaupanja. Nato izračunaj vsoto 100 naključnih števil, dobljenih z zgoraj navedenim ukazom. Ponovi ta eksperiment dvajsetkrat. Kako pogosto je število enic padlo iz sprejemljivih mej?

Ocena μ z uporabo vzorčnega povprečja $\overline{x}$

Velikokrat uporabljamo vzorčno povprečje $\overline{x}$ za oceno dejanske srednje vrednosti populacije μ. V preglednici na sliki 4.7 je naštetih 5 tež vzorcev (v gramih) za nek izdelek. Vsaka posamezna teža je natančna s standardnim odklonom $\sigma = 0{,}05$ g. Na osnovi teh vrednosti izračunamo vzorčno povprečje in vzorčni standardni odklon $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$. Krajišči 95% intervala zaupanja za srednjo vrednost dobimo tako, da prištejemo in odštejemo $\sigma_{\overline{x}}$ vzorčnemu povprečju.

	A	B	C	D
1	38,85		sigma	0,05
2	38,76		vzorčno povp.	38,804
3	38,80		n	5
4	38,83		sigma/sqrt(n)	0,022361
5	38,78			
6			interval min	38,75928
7	194,02	vsota	interval max	38,84872

Slika 4.7: Izračun intervala zaupanja s pomočjo preglednice.

Naloga 3. Napravi tabelo, podobno tisti na sliki 4.7, za naslednjih pet podatkov:

38,82 38,76 38,75 38,75 38,81

Naloga 4. Uporabi funkcijo `= 25,5 + RandBetween(5,25)/50`, da generiraš nove "meritve". Standardni odklon je približno $\sigma = 0{,}1$. Napravi pet vzorcev in izračunaj 95% interval zaupanja za "dejansko" težo.

Raziskovanje

Recimo, da moraš po pogodbi zagotoviti, da je vzorčna varianca $\sigma_{\overline{x}}$ manjša od 0,01. Vsako tehtanje te stane nekaj časa in denarja, vendar pa so bolj natančne tehtnice ponavadi bolj drage. Kako poiskati kompromis med stroški, ki jih predstavlja večje število tehtanj, in ceno bolj natančne tehtnice? Zamisli si scenarij, v katerem se lotiš tega problema.

4.11 Pisni projekti

(1) Kako agencije opisujejo natančnost svojih raziskav? Pod žarometom je opisano, kako so to naredili pri New York Timesu. Pojdi na spletno stran agencije Gallup (`www.gallup.com`) in preberi objavo kakšne nedavne ankete. Nato pojdi na stran Urada za statistiko dela (`stats.bls.gov/newsrels.htm`) in poišči najnovejše podatke o zaposlenosti (*Employment Situation*) v razdelku *Employment & Unemployment*.

V obeh objavah poišči tisti del, ki opisuje metode vzorčenja, mejo napake rezultata in vire napake, ki v to mejo niso vključni. V drugem primeru se posveti samo podatkom iz raziskave prebivalstva (*Current Population Survey*). Napiši kratko primerjavo obeh primerov. Zakaj je po tvojem mnenju Gallup precej bolj redkobeseden kot Urad za statistiko dela?

(2) Meja napake za raziskavo na vzorcu upošteva naključne variacije, ki so posledica slučajnega vzorčenja. V praksi pa so lahko rezultati napačni iz drugih razlogov. Nekaterih izbranih ni mogoče kontaktirati, drugi se zlažejo ali pa se ne spomnijo, formulacija vprašanj vpliva na odgovore. To so "praktične ovire", ki jih omenjamo pod žarometom.

Napiši kratko razpravo o najbolj pomebnih praktičnih ovirah, na katere naletimo pri anketiranju in drugih raziskavah na človeški populaciji. Nekaj materiala je v *Statistics: Concepts and Controversies* (pod priporočeno literaturo

v prvem poglavju) in v članku P.E. Converse, M.W. Traugott, Assessing the accuracy of polls and surveys, *Science*, 234 (1986): 1094–1098.

(3) Intervalov zaupanja ne uporabljamo le v situacijah, ki smo jih opisali v tem poglavju. Eden od primerov je ocenjevanje razlike med dvema deležema populacije, p_1 in p_2. Na primer, p_1 bi lahko bil delež žensk in p_2 delež moških, ki ponoči ostajajo doma, ker jih je strah kriminala. Preberi poročilo o intervalih zaupanja za $p_1 - p_2$ v knjigah o statističnih metodah, na primer v razdelku 8.2 Moorove *Basic Practice of Statistics*. Nato natančno razloži, kako dobimo ta novi interval s pomočjo enakega sklepanja kot tiste, ki smo jih že spoznali: poišči oceno, opazi, da je vzorčna porazdelitev ocene približno normalna, ugotovi, kakšen je standardni odklon te normalne porazdelitve in se za dva standardna odklona odmakni od ocene, da dobiš 95% interval zaupanja.

Stvarno kazalo