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		

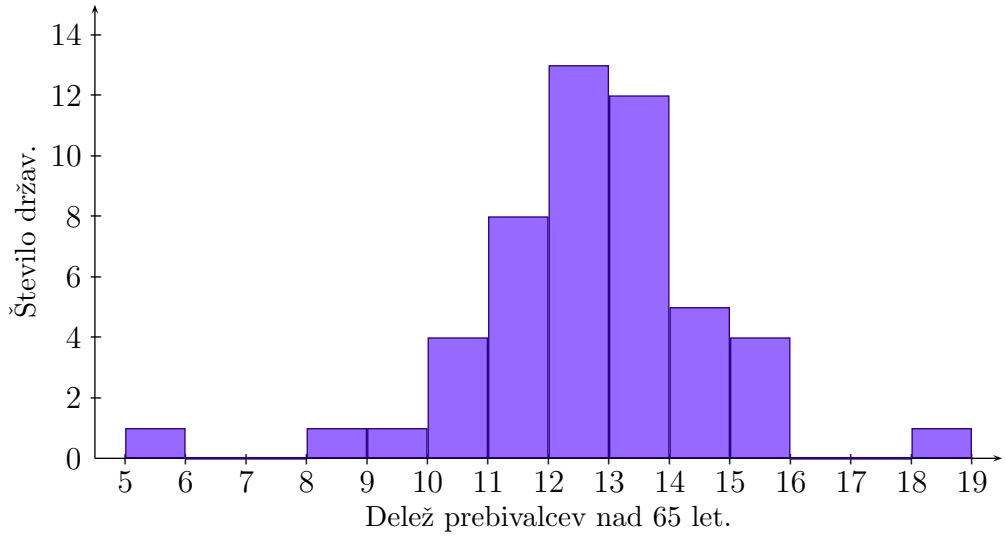
(3) Narišemo histogram. Najprej na vodoravni osi označimo merilo za spremenljivko, katere porazdelitev rišemo. V našem primeru je to "delež prebivalcev nad 65 let". Označimo vrednosti med 5 in 19, ker smo na tem intervalu izbrali naše razrede. Na navpični osi je merilo za števila, ki smo jih dobili s 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

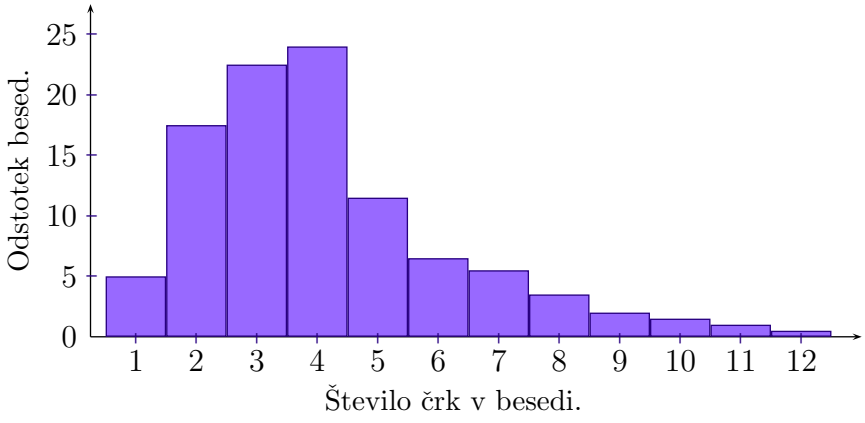
namesto 50 vnesli 5,0. Drugi ubežniki pa opozarjajo na posebne lastnosti nekaterih opažanj. Na Floridi, kjer živi veliko upokojencev, je velik del populacije starejši od 65 let, in na Aljaski, ki je čisto na severu, jih je malo. ♦

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.

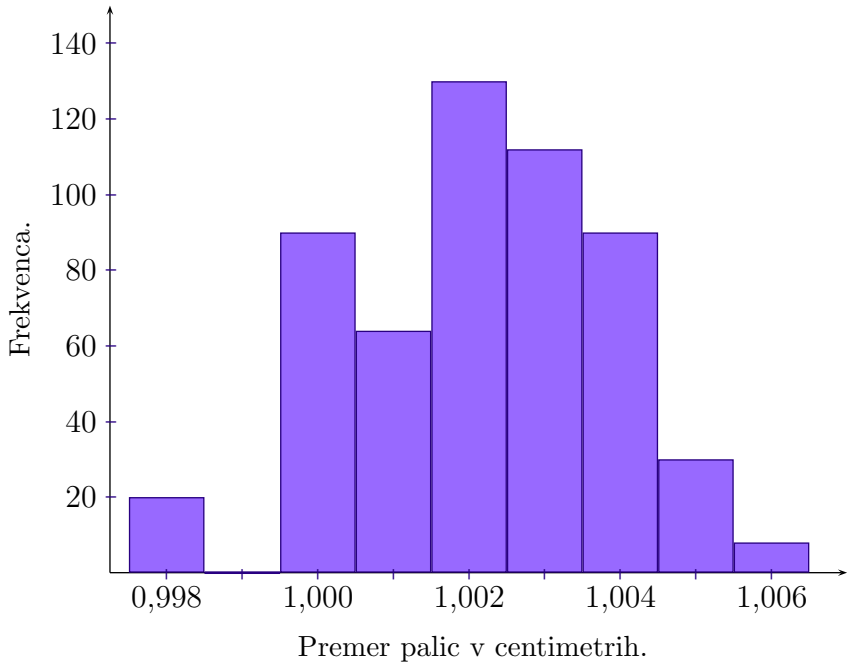
Primer. (Shakespearove besede) Slika 2.3 prikazuje porazdelitev dolžin besed, ki se pojavijo v Shakespearovih dramah. Tudi ta porazdelitev ima en sam vrh, vendar pa je desno asimetrična. Veliko je torej kratkih besed, takih s tremi ali štirimi črkami, daljših, takih z 10, 11 ali 12 črkami, pa je zelo 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 25 25 27 28 31 33 **34** 36 37 44 50 59 85 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

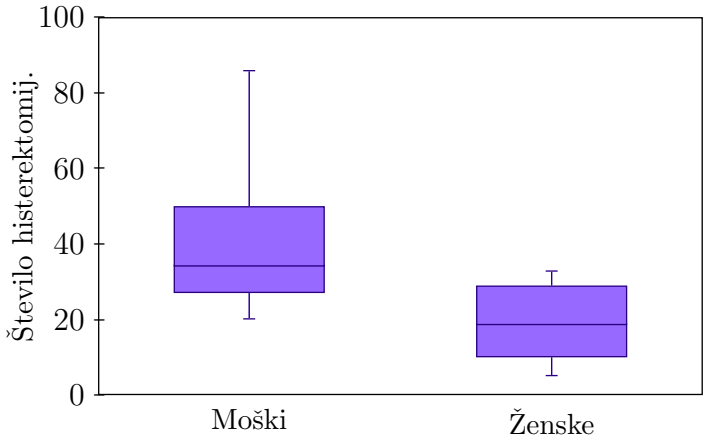
20 27 34 50 86

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

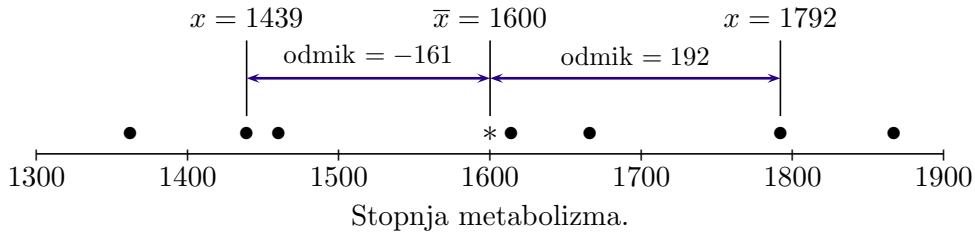
Čeprav je povzetek s petimi števili najbolj splošno uporaben numerični opis porazdelitve, ni najbolj pogost. To mesto pripada kombinaciji povprečja in *standardnega odmika*. Povprečje je (tako kot mediana) merilo za središče. Standardni odklon pa je (tako kot kvartili in ekstremi iz povzetka s petimi 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 razprostirajo 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. ♦

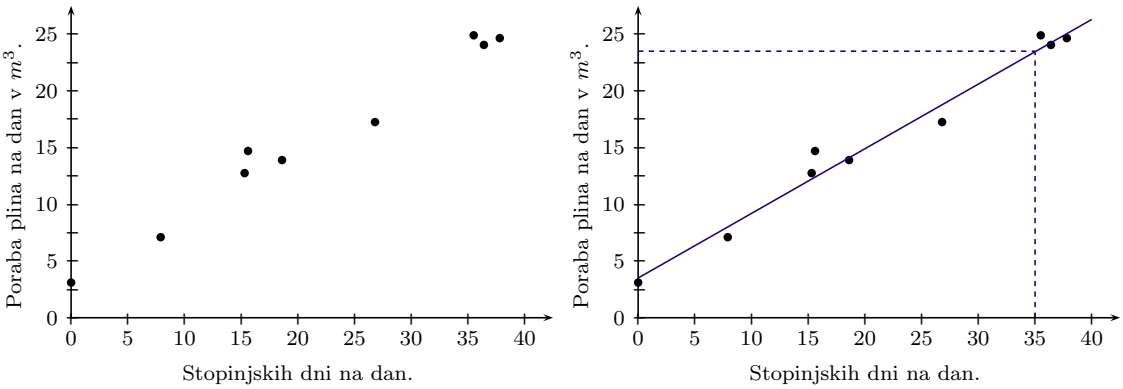
Razsevni diagram pokaže povezavo med dvema numeričnima spremenljivkama, ki ju izmerimo na istih posameznikih. Vrednosti ene od spremenljivk se pojavijo na vodoravni osi, vrednosti druge pa na navpični. Vsak posameznik iz množice podatkov je predstavljen s točko na diagramu, ki jo določata koordinati,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

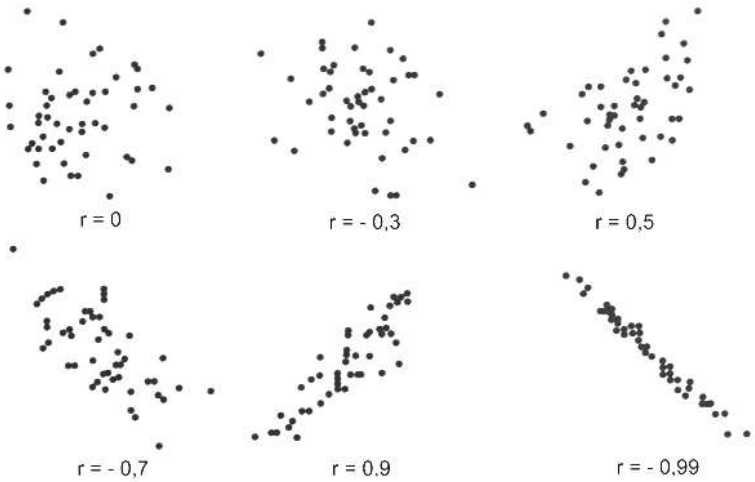
(2) Korelacija meri moč le za linearne zveze. Ne opisuje drugih zvez med spremenljivkami, ne glede na to, kako močne so.

(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				

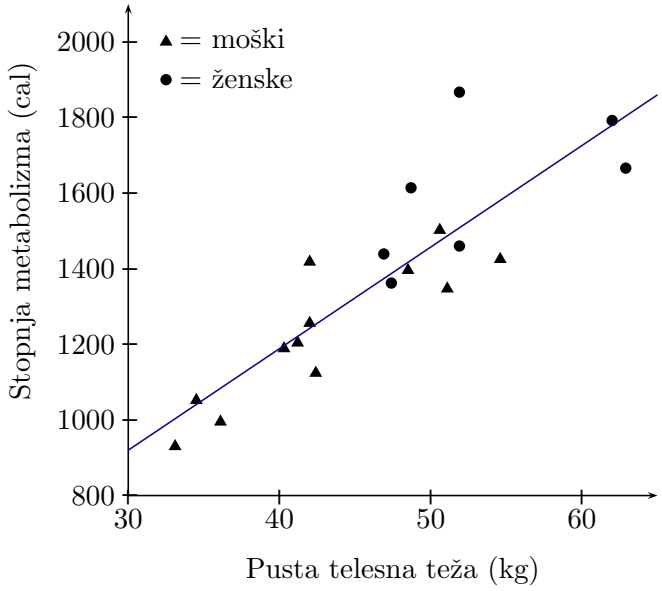
Tabela 2.3: Teža in stopnja metabolizma.

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 $x = 0$. ♦



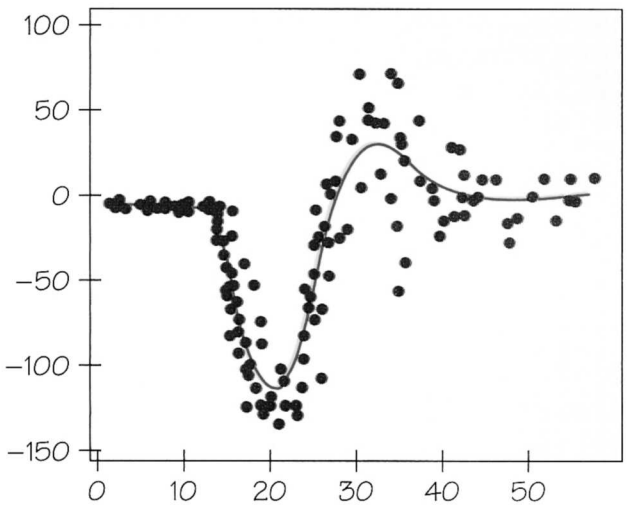
Slika 2.10: Razsevni diagram stopnje metabolizma glede na težo za 12 žensk in 7 moških. Za razlikovanje uporabimo različne simbole.

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. ♦

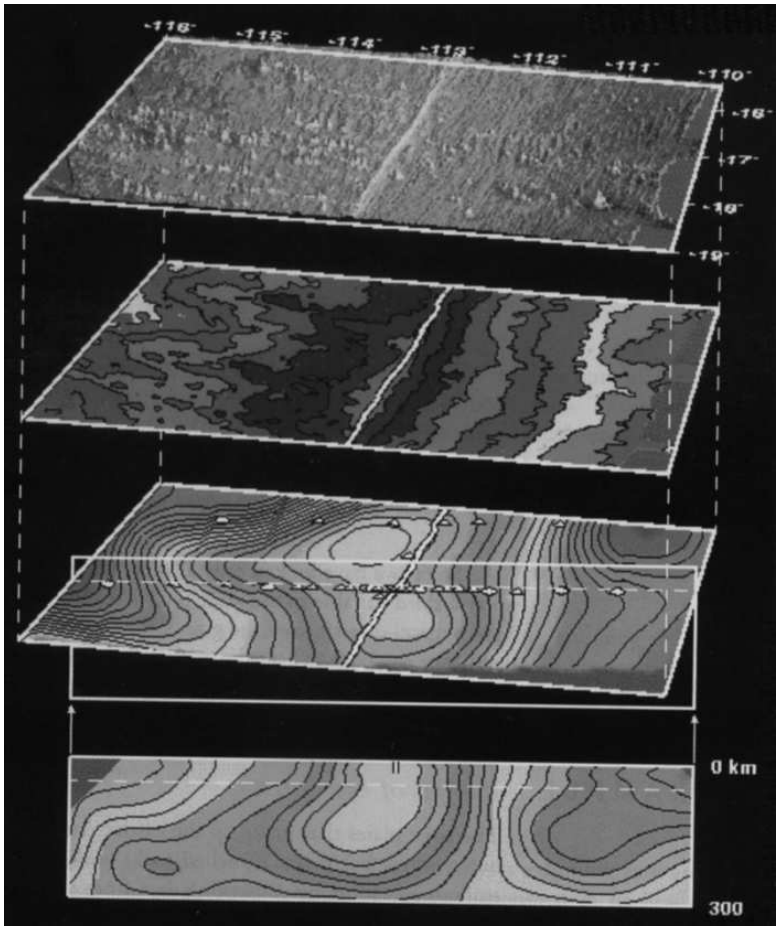


Slika 2.11: Odvisnost pospeška glave testne lutke od časa pri trku motorja z zidom. Krivulja je dobljena kot izglajevalec razsevnega diagrama.

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

spremenljivkama; 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.

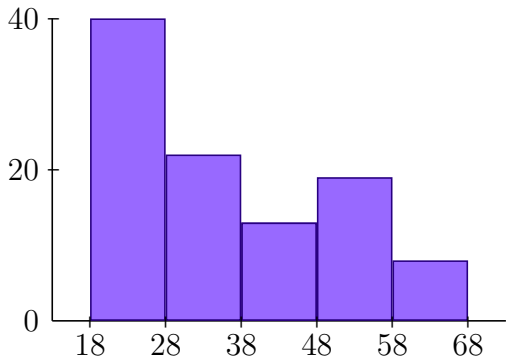
(c) Tako 101 kot 46.

(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?

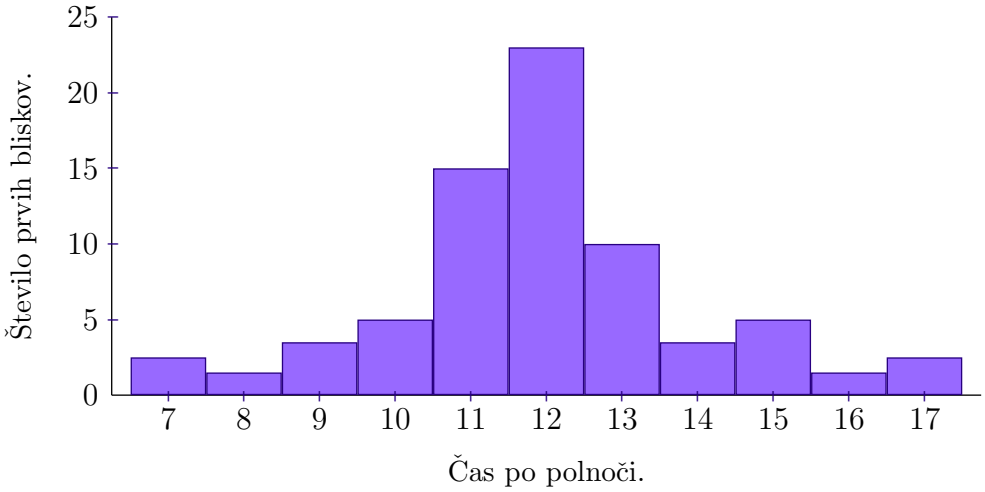
Model	Poraba	Model	Poraba
Acura 3.5RL	25	Lexus GS300	23
Audi A6 Quattro	26	Lexus LS400	25
Buick Century	29	Lincoln Mark VIII	26
Cadillac Catera	24	Mazda 626	33
Cadillac Eldorado	26	Mercedes-Benz E320	29
Chevrolet Lumina	29	Mercedes-Benz E420	26
Chrysler Cirrus	30	Mitsubishi Diamante	24
Dodge Stratus	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.

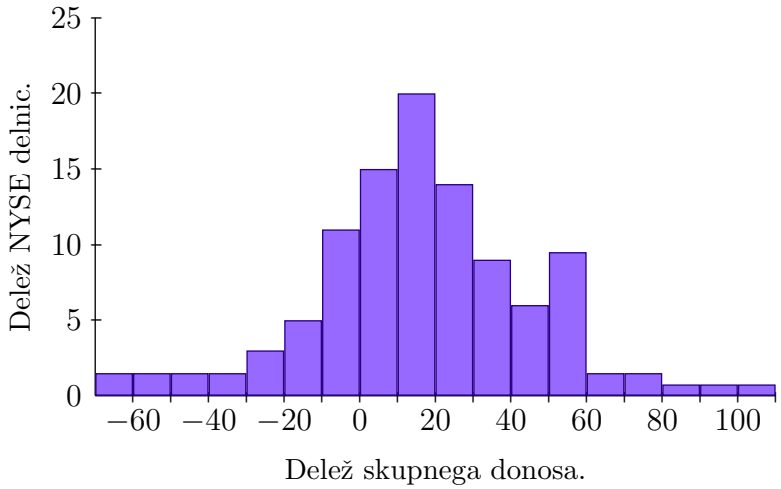


Slika 2.13: Razporeditev časov prvih pojavov bliskov za vsak dan opazovanj v zvezni državi Kolorado.

(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 za-

porednih 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.

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.

(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 $x = 10$.

(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

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.

dobili podatke v spodnji tabeli.

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

(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.

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.

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

	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			

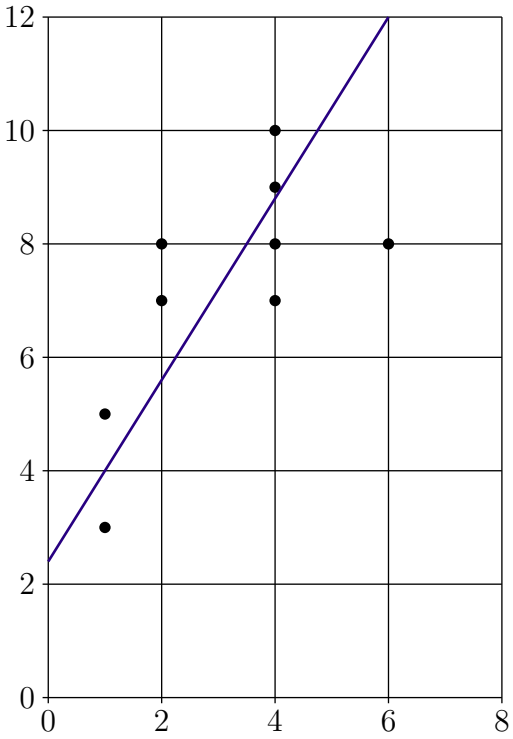
Slika 2.16: Preglednica za računanje premice najmanjših kvadratov.

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?



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

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?