Poglavje 1

Statistike

V živinoreji spremljamo neko lastnost - spremenljivko x_i in jo izmerimo. Tako dobimo niz podatkov, ki ga imenujemo tudi vzorec **Z** (1.1). Če smo živali naključno izbirali, je vzorec naključni.

$$Z = (x_1,\ x_2,\ \ldots x_n) \qquad [1.1]$$

Opravili smo n meritev dnevne količine mleka pri 1000 kravah. Ko želimo podatke predstaviti, je povsem neprimerno, da bi navajali vse meritve, tudi takrat, ko je vzorec manjši. Podatke moramo primerno predelati, da iz njih potegnemo najpomembnejše informacije. Izračunane vrednosti imenujemo ***statistike***. Če želimo pri tem poudariti, da se nanašajo na vzorec, jih poimenujemo ***vzorčne statistike***. V tem poglavju bomo obravnavali tri skupine statistik: srednje vrednosti, mere razpršenosti in mere povezanosti. Omeniti moramo tudi nekatere statistike, ki jih pravzaprav poiščemo in ne računamo. To so najmanjša (minimum) in največja vrednost (maksimum).

1.1 Srednje vrednosti

Vrednosti spremenljivk se med enotami razlikujejo. Nekatere vrednosti so pogostejše, druge pa manj verjetne, vse pa so bolj ali manj podobne "osrednji vrednosti".

Za srednjo vrednost imamo več statistik. V živinoreji so pogoste aritmetična sredina, mediana in modus. Srečamo lahko tudi geometrično sredino, le redko pa harmonično sredino. Srednje vrednosti sodijo med najpomembnejše statistike in praviloma veliko povedo o vzorcu. Če je vzorec slučajen, lahko zaključke posplošimo tudi na populacijo. Pri srednjih vrednostih, zlasti pri povprečju, obstaja velika nevarnost, da jih uporabimo tudi takrat, ko jih ne bi smeli. So tudi priročne za izračun. S srednjimi vrednostmi dobro opišemo populacijo, izgubimo pa informacije, ki so značilne za manjše skupine v vzorcu ali celo za posamezne meritve. Tako se ne smemo prehitro zadovoljiti z njimi, nadalje moramo iskati povezavo med meritvami, spreminjanje meritev s časom, proučevati različnost med njimi.

1.1.1 Aritmetična sredina

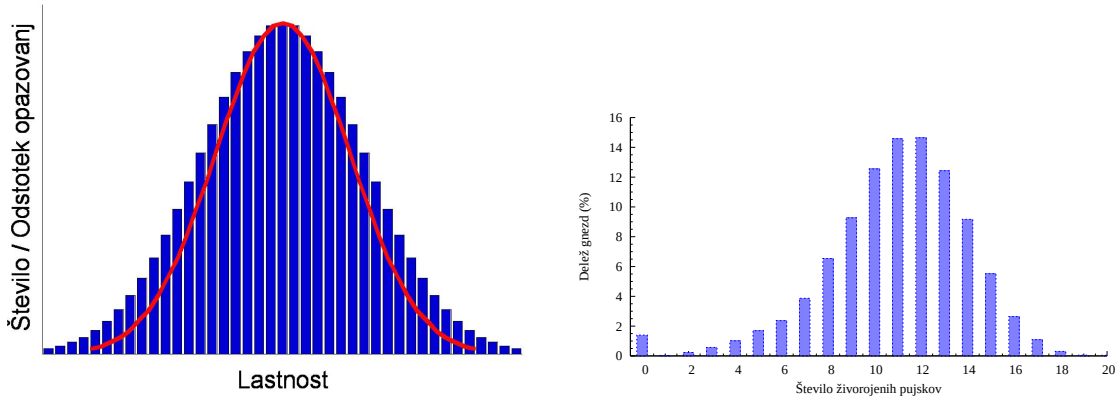
1.1.1.1 Navadna aritmetična sredina

Aritmetična sredina je poznana tudi kot povprečje. **Povprečje** (enačba 1.2) dobimo tako, da seštejemo vrednosti spremenljivk x_i in jih delimo s številom vrednosti (n). Vsota odklonov od povrečja je vedno enaka 0.

$$\overline{X} = \frac{1}{n} * \sum_{i=1}^{n} x_i \qquad [1.2]$$

Aritmetično sredino pri populacijah bomo označevali z μ, pri vzorcu pa z $\overline{X}$.

Povprečje bomo omenjali pri več porazdelitvah, prav poseben pomen pa ima pri normalni ali Gaussovi porazdelitvi, kjer predstavlja njen lokacijski parameter (slika 1.1) in je tudi najpogostejša vrednost pri porazdelitvi. Porazdelitev podatkov smo narisali s stolpci modre barve, ki predstavljajo dovolj majhne intervale, z rdečo črto pa smo vrisali normalno porazdelitev. Na vajah boste preverjali, ali sta teoretična



Slika 1.1: Aritmetična sredina pri normalni (levo) in diskretni (desno) porazdelitvi

porazdelitev in porazdelitev podatkov dovolj blizu, da lahko predpostavimo normalno porazdelitev. Presoje ne smemo opravljati subjektivno, saj smo različno natančni: eni premalo, drugi preveč. Tako se bomo raje naučili postopka presoje.

Povprečje lahko uporabljamo tudi pri diskretnih porazdelitvah (desno na sliki 1.1), ki so približno simetrične in imajo dovolj razredov. Kot primer navajamo število živorojenih pujskov v gnezdu. Porazdelitev ni popolnoma simetrična, delež gnezd s samo mrtvorojenimi pujski nekoliko odstopa, saj bi pričakovali takih gnezd manj kot z 1 živorojenim pujskom. To je lahko zaradi prezgodnjih porodov, ko so pujski slabotni, svinja pa lahko prasi celo v skupinskem boksu, kjer porod motijo ostale svinje. Te nepravilnosti so zanemarljive, povprečje v tem primeru dovolj dobro ocenjuje vrh porazdelitve.

Povprečje ni vedno primerno. Kadar so populacije neenovite, heterogene oziroma asimetrične, preverimo smiselnost uporabe mediane oziroma modusa.

1.1.1.2 Tehtana aritmetična sredina

Aritmetična ali navadna sredina je primerna, kadar so meritve (x_i) pri spremenljivki enakovredne, so enako natančno merjene. Če pa imajo opazovanja različno težo, pa niso enakovredna in jih moramo tehtati. Najpogostejši je primer, ko smo prišli do povprečij različnih podskupin populacije, ki so bila izračunana iz različnega števila opazovanj. V tem primeru moramo poznati vrednost in število ali pogostnost.

Če poznamo frekvenco podatkov (f_i, porazdelitve), moramo razmišljati o tehtanem povprečju (enačba 1.3).

$$\overline{X} = \sum_{i=1}^{r} f_i x_i \qquad [1.3]$$

Kadar poznamo število opazovanj, lahko izračunamo frekvenco ali pa uporabimo enačbo (1.4).

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{r} n_i x_i \qquad [1.4]$$

PRIMER : Povprečna ocena Za primer vzemimo dva študenta. Oba sta pri izpitih dobila samo dve oceni: 10 in 6. Pri tem je prvi dosegel 20 desetic in eno šestico, drugi pa 20 šestic in eno desetico. Izračunajmo navadno in tehtano aritmetično sredino!
Navadna aritmetična sredina je pri obeh študentih enaka: $(10 + 6)/2 = 8$.
Tehtana aritmetična sredina pri prvem študentu je: $(20 * 10 + 1 * 6)/21 = 9.81$.
Tehtana aritmetična sredina pri drugem študentu je: $(1 * 10 + 20 * 6)/21 = 6.19$.
Katera povprečna ocena se vam zdi pravilna?

PRIMER : Velikost gnezda pri prašičih
Na kmetiji je prasilo 5 mladic po 9.40 živorojenih pujskov v gnezdu in 20 starih svinj po 11.20 živorojenih pujskov v gnezdu. Izračunajte povprečno velikost gnezda na kmetiji! Katero povprečje je primernejše?
Izračun:

$$\overline{X} = \frac{1}{5+20}(5 * 9.40 + 20 * 11.20) = \frac{271}{25} = 10.84$$

PRIMER : Tehtano povprečje - frekvenca
Trije kmetje so prodali 520 prašičev v klavnico. Prvi je prispeval 25 % živali, drugi 40 % in tretji pa 35 %. Prašiči so v povprečju tehtali 105 kg pri prvem kmetu, 110 kg pri drugem in 107 kg pri tretjem. Izračunajte povprečno maso celotne skupine!
Izračun:

$$\overline{X} = 0.25 * 105 + 0.40 * 110 + 0.35 * 107 = 107.7$$

1.1.2 Geometrijska sredina

Pri izračunavanju povprečij iz verižnih indeksov, koeficientov rasti in stopenj rasti raje uporabimo geometrijsko sredino (enačba 1.5). Imenujemo jo tudi povprečna proporcionalna vrednost. Uporabljamo jo samo, kadar so vrednosti vseh spremenljivk x_i pozitivne.

$$\overline{X_g} = \sqrt[n]{\prod_{i=1}^{n} x_i} \qquad [1.5]$$

Tudi geometrijsko povprečje leži med najmanjšo in največjo spremenljivko x_i. Kadar spremenljivke x_i logaritmiramo, je antilogaritem aritmetične sredine logaritmiranih vrednosti enak geometrijski sredini nespremenjenih spremenljivk. To je uporabno, če so vrednosti na originalni skali zelo različne.

$$\overline{X_{ln}} = ln\left(\sqrt[n]{\prod_{i=1}^{n} x_i}\right) = \frac{1}{n}\sum \ln(x_i) \qquad [1.6]$$

1.1.3 Harmonična sredina

Harmonična sredina je inverzna vrednost povprečja inverznih vrednosti spremenljivke x_i.

$$\overline{X_h} = n * \left(\sum \frac{1}{x_i}\right)^{-1} \qquad [1.7]$$

Na harmonično sredino ne naletimo prav pogosto. V živinoreji je najbolj poznan primer pri izračunu povprečja efektivnih velikosti populacij po generacijah. Ta parameter boste obravnavali pri kvantitativni genetiki in predstavlja oceno števila nesorodnih živali v populaciji. V zadnjem času se ta značilnost populacije pogosto uporablja pri presoji ogroženosti populacije in genetske raznovrstnosti. Harmonično sredino izračunavamo tudi pri številu somatskih celic.

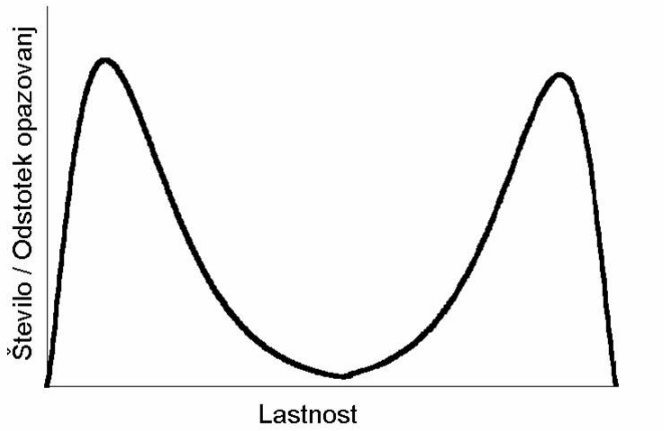
1.1.4 Mediana

Mediana ali **centralna vrednost** je tista vrednost spremenljivke, ki razdeli meritve tako, da je enako število meritev večjih in manjših od nje. Določanje mediane je enostavno, če so podatki razvrščeni oziroma rangirani po vrednosti. Če je število enot liho, dobi mediana vrednost srednje enote. V primeru sodega števila opazovanj pa je mediana povprečje srednjega para meritev. Mediana je neobčutljiva na posamezne vrednosti spremenljivk, dokler spremenjena vrednost ostane na isti strani mediane. Mediana pove o podatkih manj kot povprečje, je pa lahko primerna, če porazdelitev ni simetrična.

PRIMER : Določitev mediane
Bolezen pri 7 obolelih živalih traja 6, 6, 7, 7, 8, 29 in 35 dni. Povprečje znaša 14 dni, vendar pa je to predvsem zaradi 2 živali, ki se dolgo nista pozdravili. Mediana je 7 dni in nekako bolje opiše porazdelitev kot povprečje.

1.1.5 Modus

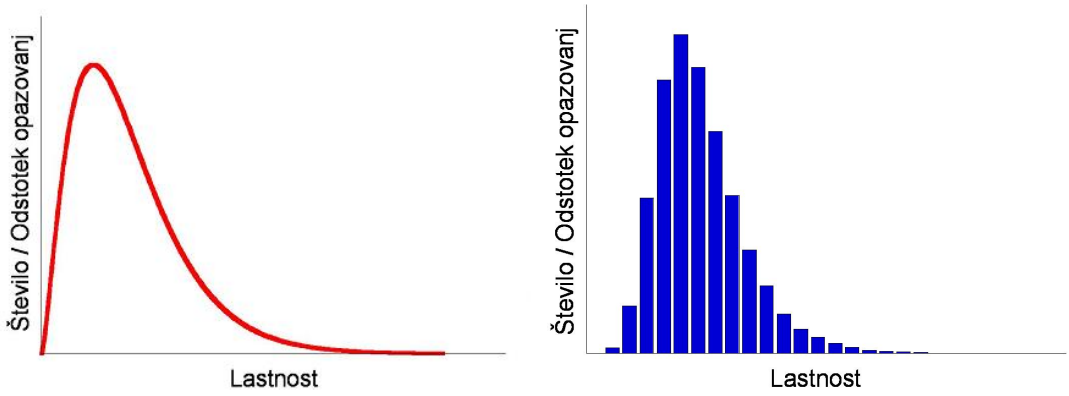
Modus je srednja vrednost, ki je enaka najpogostejši vrednosti. Ugotovimo ga lahko pri sorazmeroma velikem vzorcu, medtem ko so lahko pri manjših vzorcih vse vrednosti redke. Porazdelitve imajo lahko več modusov (vrhov) in lahko govorimo o unimodalnih, bimodalnih in polimodalnih porazdelitvah (1.2). V živinoreji je kar nekaj primerov porazdelitev z več modusi.



Slika 1.2: Bimodalna porazdelitev

Modus predstavlja srednjo vrednost bolje kot povprečje na selekcioniranih vzorcih. Primeren je tudi za heterogene ali asimetrične porazdelitve (slika 1.3). Kot primer prikazujemo debelino hrbtne slanine pri prašičih. Zaradi selekcije na mesnatost in urejenih tehnologij je slanina s kožo tanka. Ker pa je navzdol omejena (koža ima debelino 3 mm), je strma ob y-osi in bolj položna s podaljšanim repom na levi strani. Porazdelitev označimo kot desno asimetrično. Desno asimetrične so tiste, ki imajo na desni strani modus.

Modus lahko prav tako uporabimo pri diskretnih spremenljivkah (desno na sliki 1.3). Za primer lahko vzamemo porazdelitev dobe od odstavitve do pojava estrusa pri odstavljenih svinjah, dobe od telitve do pojava estrusa kravah, porazdelitev dobe od pripusta do pregonitve. Pojavljajo se tudi pri spremljanju dnevnega ritma posameznih živali, zlasti tistih aktivnosti, ki so vezane na prehranjevanje.



Slika 1.3: Asimetrični porazdelitvi (levo zvezna, desno diskretna)

1.2 Mere razpršenosti

Pri tem predmetu bomo mere razpršenosti zelo izpostavili, ker imajo veliko vlogo pri številnih rejskih opravilih. V prireji razpršenost ni zaželjena. Prirejo je veliko lažje organizirati, če so potrebe živali enake, na trgu so izenačeni proizvodi praviloma bolje plačani. Razpršenost je pomembna tudi pri odbiri plemenskih živali, kjer zagotavlja možnosti genetskega napredka. Če so vse živali enake, potem populacije ne moremo ne poslabšati in ne izboljšati. Spoznali boste tudi, da je raznolikost pomembna, da populacija ohrani prilagodljivost in se tako ohrani.

Živinorejci se preveč pogosto zadovoljimo le s povprečnimi vrednostmi in zato pogosto izgubimo veliko zaslužka. Naslednji primer je izmišljen in poenostavljen tako , da bomo hitreje računali, vendar pa so podobni primeri v živinoreji pogosti. Mnogo končnih produktov uvrščamo v tržne razrede, ki jih plačujejo po plačilnih maskah, ki nagrajujejo dobro kakovost in kaznujejo manj željene razrede. Najvišja cena ni nujno na sredini ali pri ekstremu, vendar izenačenost ali razpršenost opazovanj pri proizvodnih lastnostih vpliva na rezultat.

PRIMER : Pomen rapršenosti
Trije rejci A, B in C proizvajajo za trg isti proizvod (preglednica 1.1), ki se uvršča v osem kakovostnih razredov. Po ceni vidimo, da sta najbolj zaželjena kakovostna razreda 4 in 5, manj cenjeni pa so tako nižji kot višji razredi. Porazdelitev izdelkov pri rejcih po posameznih razdelkih je precej različna. Rejec A nima najslabših razredov, manj pogosta sta tudi razreda 2 in 7, odtopa pa tudi pri srednjih, bolje ocenjenih razredih. Največjo razpršenost ima rejec C, ki ima vse razrede enakomerno zastopane.
V tej nalogi je primerno tehtano povprečje. Izračunajmo tehtana povprečja za kakovostne razrede pri vseh treh rejcih!

$$2 * 0.050 + 3 * 0.100 + 4 * 0.350 + 5 * 0.350 + 6 * 0.100 + 7 * 0.050 = 4.5$$

$$1 * 0.035 + 2 * 0.065 + 3 * 0.100 + 4 * 0.300 + 5 * 0.300 + 6 * 0.100 + 7 * 0.065 + 8 * 0.035 = 4.5$$

$$1 * 0.125 + 2 * 0.125 + 3 * 0.125 + 4 * 0.125 + 5 * 0.125 + 6 * 0.125 + 7 * 0.125 + 8 * 0.125 = 4.5$$

Povprečna vrednost za kakovostni razred je 4.5 pri vseh treh rejcih, vendar pa ne bodo enako zaslužili. Izračunajmo povprečno ceno za enoto izdelka!

$0.6 * 0.050 + 0.8 * 0.100 + 1.0 * 0.350 + 1.0 * 0.350 + 0.8 * 0.100 + 0.6 * 0.050 = 0.920$

$2 * (0.3 * 0.035 + 0.6 * 0.065 + 0.8 * 0.100 + 1.0 * 0.300) = 0.859$

$.125 * (0.30 + 0.60 + 0.80 + 1.00 + 1.00 + 0.80 + 0.60 + 0.30) = 0.495$

Povprečna cena proizvoda med rejci je različna. Prvi pri vsakem iztrži pri vsakem kosu skoraj še enkrat višjo ceno, drugi rejec pa prejme 6.6 % manj.

Tabela 1.1: Porazdelitev proizvodov po kakovostnih razredih pri treh rejcih

		Rejec		
Razred	Cena	A	B	C
1	0.30		0.015	0.125
2	0.60	0.050	0.085	0.125
3	0.80	0.100	0.100	0.125
4	1.00	0.350	0.300	0.125
5	1.00	0.350	0.300	0.125
6	0.80	0.100	0.100	0.125
7	0.60	0.050	0.085	0.125
8	0.30		0.015	0.125

1.2.1 Povprečni odklon

Najprej moramo definirati odklon, ki ga bomo pogosto imenovali tudi napaka ali ostanek. Označili ga bomo s črko e, indeks pa označuje, kateri meritvi pripada. Odklon je odstopanje meritve od povprečja spremenljivke (enačba 1.8).

$$e_i = (x_i - \overline{x}) \tag{1.8}$$

Sedaj pa si izračunajmo povprečni odklon!

$$\overline{O} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = \tag{1.9}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i) - \frac{1}{n}\sum_{i=1}^{n}(\overline{x}) = \tag{1.10}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i) - \frac{1}{n}n\overline{x} = \overline{x} - \overline{x} = 0 \tag{1.11}$$

Ugotovili smo, da je povprečni odklon ali pričakovana vrednost odklona **vedno** enaka 0 (enačba 1.11) in **ni primerna statistika za razpršenost**.

Povprečna vrednost ali pričakovane vrednosti so vedno izbrane tako, da je vsota odklonov enaka 0. Tako je povprečje odklonov pri vseh poskusih že vnaprej znano in ne kaže na nobeno razliko.

1.2.2 Absolutni odklon

Absolutni odklon je povprečje absolutnih vrednosti odklonov (enačba 1.12). Vsi odkloni so pozitivni, vsi odkloni k absolutnemu odklonu enako prispevajo, z drugimi besedami imajo enako težo.

$$AO = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \qquad [1.12]$$

1.2.3 Varianca

Kako natančno smo (lahko) izvedli poskus, opisuje mera za razpršenost - **varianca vzorca** (enačba 1.13). Tudi to ni parameter - varianca populacije, je samo ocena, morda še to slaba, kadar imamo le nekaj opazovanj. Doprinos odklonov je različen in sicer imajo pri varianci večji pomen večja odstopanja meritev od pričakovane vrednosti.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \qquad [1.13]$$

Definicija: Varianca je povprečni kvadratni odklon od pričakovane (primerjalne) vrednosti.

Varianco bomo označili na dva načina. V prvem primeru bomo uporabili σ^2 - grško črko sigma in potenco 2, ki poudarja, da smo odklone kvadrirali. Drugi način pa je označen po zgledu funkcije$var\,(x)$: kratica var prihaja iz besede **var**ianca, v oklepaju pa obvezno navedemo spremenljivko. Prvi način uporabimo, ko mislimo na vrednosti varianc, drugo pa v primeru, ko z variancami "računamo". Kadar pri prvi obliki želimo poudariti, kateri spremenljivki varianca pripada, navedemo njeno oznako v indeksu. Oznaka σ_x^2 torej pomeni varianco za spremenljivko x.

Če je pričakovana vrednost za celotno populacijo enaka, je zgornji izračun dober. V živinoreji pa se bomo srečali s primeri, ko je pričakovana vrednost posameznih podskupin različna (enačba 1.14). Takrat bomo imeli več povprečij ($\overline{x_i}$), odklone bomo izračunali ločeno za vsako poskupino in jih seštevali skupaj. Vsoto bomo delili z razliko med številom opazovanj (n) in številom izračunanih povprečij (p).

$$\sigma^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i}\right)^2}{n-p} \qquad [1.14]$$

Kadar imamo več skupin in kvantitativne odvisne spremenljivke, raje govorimo o pričakovani vrednosti ($E\,(x_i)$), v imenovalcu pa tudi odštejemo število p, ki predstavlja število ocenjenih parametrov. Enačbo 1.15 smo tu navedli samo zato, da so enačbe nekje skupaj in jih lahko primerjamo. Podrobneje jo bomo obrazložili kasneje.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - E\,(x_i))^2}{n-p} \qquad [1.15]$$

Sedaj se vrnimo k enačbi 1.13 in jo malo preuredimo (enačba 1.16). Enačbo bomo poenostavili samo toliko, da bomo pri oznaki za vsoto ($\sum_{i=1}^{n}$) opustili informacijo, da indeks i zavzema vse vrednosti od 1 do n.

$$var\,(x) = \frac{\sum\left(x_i - \overline{X}\right)\left(x_i - \overline{X}\right)}{n-1} = \qquad [1.16]$$

Sedaj pomnožimo člena znotraj vsote iz enačbe 1.16 in dobimo enačbo 1.17.

$$\frac{\sum \left(x_i x_i - \overline{X} x_i - x_i \overline{X} + \overline{X}\overline{X}\right)}{n-1} = \quad [1.17]$$

Ker sta srednja člena enaka (enačba1.18), ju lahko seštejemo in razstavimo vsoto (enačba 1.19).

$$\overline{X} x_i = x_i \overline{X} \quad [1.18]$$

$$= \frac{\sum \left(x_i^2\right) - 2\sum \left(\overline{X} x_i\right) + \sum \left(\overline{X}^2\right)}{n-1} = \quad [1.19]$$

Povprečje in njegov kvadrat sedaj lahko izpostavimo pred znak za vsoto (enačba 1.20), ker ostaneta pri vseh členih pač vrednosti obeh statistik enaki, konstantni.

$$= \frac{\sum \left(x_i^2\right) - 2\overline{X}\sum (x_i) + \overline{X}^2 \sum (1)}{n-1} = \quad [1.20]$$

Če le malo preuredimo enačbo za izračun povprečja, vidimo, da je vsota vseh spremenljivk enaka zmnožku števila opazovanj in povprečja spremenljivk (enačba 1.21).

$$\sum (x_i) = n\overline{X} \quad [1.21]$$

Enačba 1.22 pa pove, da takrat, ko seštejemo toliko enk, kot imamo opazovanj, dobimo število opazovanj (n).

$$\sum (1) = n \quad [1.22]$$

Ko uporabimo zadnji dve enačbi, lahko zapišemo varianco z enačbo **??**.

$$= \frac{\sum \left(x_i^2\right) - 2n\overline{X}^2 + n\overline{X}^2}{n-1} = \quad [1.23]$$

Zadnja dva člena sta spet enaka, zato ju lahko seštejemo in dobimo enačbo 1.24.

$$= \frac{\sum \left(x_i^2\right) - n\overline{X}^2}{n-1} = var(x) \quad [1.24]$$

V zadnjem členu enačbe 1.24 lahko povprečje zamenjamo s formulo 1.2 za izračun povprečja, kot je to prikazano v enačbi 1.25.

$$= \frac{\sum \left(x_i^2\right) - n\left(\frac{1}{2}\sum (x_i)\right)^2}{n-1} = \quad [1.25]$$

S preureditvijo dobimo še enačbo 1.26 za izračun variance.

$$= \frac{\sum \left(x_i^2\right) - \frac{1}{n}\sum^2 (x_i)}{n-1} = var(x) \quad [1.26]$$

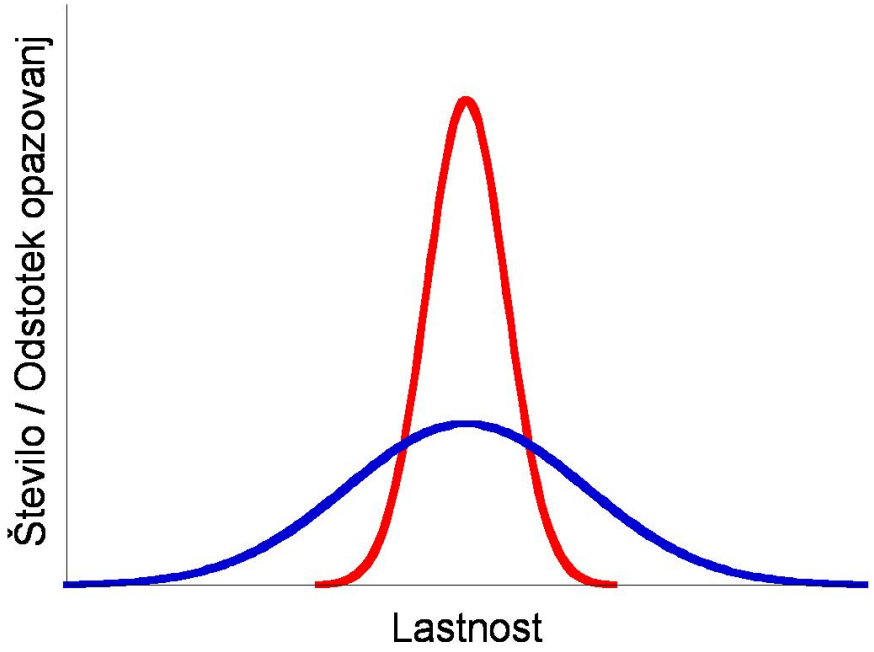
Z enačbami 1.13 , 1.24 in 1.26 lahko izračunamo varianco in bomo dobili isti rezultat. Prva enačba je osnovna definicija, ni najprimernejša za izračun variance. Ker z odštevanjem povprečja lahko dobivamo vrednosti z decimalkami, jih želimo zaokrožiti in smo pri izračunu morda nismo več dovolj natančni. Enačbi 1.24 in 1.26 pa sta bolj primerni tako pri "peš" računanju kot pri uporabi računalnika. Seveda imamo najraje, če nam programski paket, s katerim analiziramo podatke, ponudi kar funkcijo in sam izbere primerno enačbo. Kljub temu pa bomo živinorejci te enačbe znali na pamet!

1.2.4 Standardni odklon

Standardni odklon oziroma standardna deviacija je pozitivna vrednost kvadratnega korena iz variance.

$$\sigma = \sqrt{\sigma^2} \qquad [1.27]$$

Standardni odklon si lažje predstavljamo kot varianco, pri raznih izračunih pa je primernejša varianca, saj se izognemo kvadratnemu koreni v enačbah. V grafu 1.4 imamo dve normalni porazdelitvi. Pri modri porazdelitvi sta varianca in standardni odklon večja kot pri rdeči. Modra porazdelitev je nižja in širša, rdeča pa ožja in višja. Če imamo na y-osi delež opazovanj, je površina pod porazdelitveno funkcijo vedno enaka 1.



Slika 1.4: Normalna porazdellitev z veliko (modra krivulja) in malo (rdeča krivulja) razpršenostjo

1.2.5 Standardna napaka ocene

Povprečje je ocenjeno zanesljivo - z majhno **standardno napako** ocene (enačba 1.28), če smo poskus izvedli v nadzorovanih pogojih, opravili meritve natančno in v zadostnem številu.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad [1.28]$$

Izpeljimo zgornjo enačbo! Najprej moramo ugotoviti varianco za povprečje (enačba 1.29). Pa poskusimo. Namesto povprečja vstavimo enačbo za izračun povprečja. Konstanto $\frac{1}{n}$ lahko izpostavimo, vendar jo moramo pri tem kvadrirati.

$$\sigma^2_{\bar{x}} = var\,(\overline{x}) = var\left(\frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n}\right) = \qquad [1.29]$$

Ostane nam varianca vsote (enačba 1.30). Ker so meritve x_i neodvisne, so vse kovariance enake 0 in tako odpadejo.

$$= \frac{1}{n^2} var\,(x_1 + x_2 + \cdots + x_i + \cdots + x_n) = \qquad [1.30]$$

Tako nam ostanejo le členi z variancami, ki jih lahko zapišemo tudi kot vsoto (enačba 1.31). Meritve smo opravili z enako natančnostjo, zato je varianca pri vseh meritvah enaka. Označimo jo z σ^2. Meritev je bilo n, zato lahko izpeljemo enačbo do konca.

$$= \frac{1}{n^2}\sum_{i=1}^{n} var\,(x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n}{n^2}\sigma^2 = \frac{1}{n}\sigma^2 \qquad [1.31]$$

Da dobimo enačbo 1.28, moramo končni rezultat še koreniti.

Standardno napako ocene bomo izvrednotili tudi pri drugih ocenah sistematskih vplivov, vendar pa nam bodo v tem primeru v pomoč elementi v matriki koeficientov oziroma njeni inverzi. O standarni napaki bomo še pogosto govorili, zato se je velja zapomniti.

1.2.6 Koeficient variabilnosti

Koeficient variabilnosti je tudi mera za variabilnost, kjer primerjamo standardni odklon s povprečno vrednostjo (enačba 1.32). Vrednosti navajamo v odstotkih. V starejši literaturi je ta statistika pogosto uporabljena, sedaj pa se je izogibamo predvsem v tabelah in grafih. Še vedno pa je lahko dobrodošla statistika pri interpretaciji rezultatov.

$$KV = \frac{\sigma}{\overline{x}} * 100 \qquad [1.32]$$

Lahko pa dobimo tudi "čudne vrednosti", ko je koeficient povsem neuporaben. To se zgodi, ko je povprečje blizu 0 ali pa izredno veliko v primerjavi s standardnim odklonom.

1.2.7 Kvantili

Kvantili (slika **??**) so način prikazovanja porazdelitev. Porazdelitev je razdeljena na štiri predele, ki predstavljajo 25 % podatke. Dva pravokotnika, ki sta nakazana na dveh straneh mediane, prikazujeta prvi dve četrtini podatkov. Ker na naši sliki nista enako velika, pomeni, da porazdelitev ni simetrična. Tudi ročaja predstavljata vsak zase po četrtino podatkov. Vrednosti, ki po statističnih merilih ne sodijo v to porazdelitev, predstavljajo **osamelce** in so označeni s krogci. Povprečje je prikazano s križcem. Grafi so primerni, ko želimo v grobem primerjati porazdelitev velikega števila spremenljivk. Praviloma imamo tudi manjše število opazovanj in bi bolj natančne porazdelitve nič doprinesle k razumevanju rezultatov.

1.3 Mere podobnosti

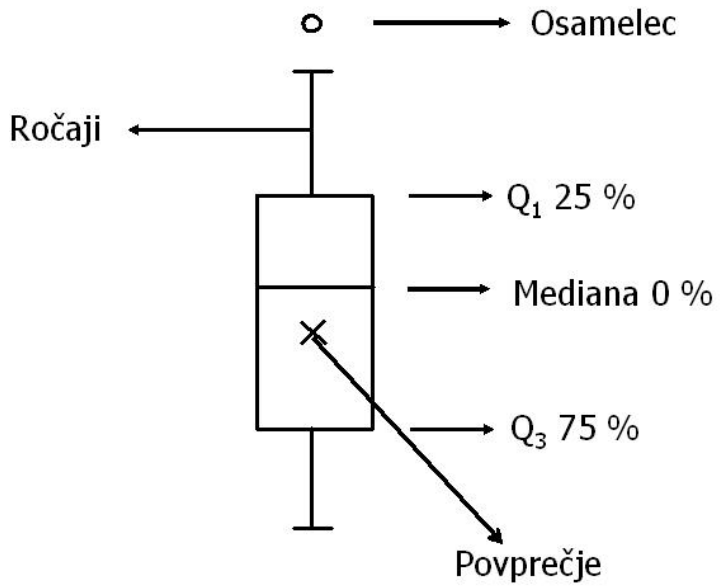
1.3.1 Kovarianca

Definicija: Kovarianca je povprečni produkt odklonov obeh spremenljivk (enačba **??**).

$$\sigma_{xy} = cov\,(x,\, y) = \frac{\sum \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{n-1} \qquad [1.33]$$

Tudi za kovarianco uporabljamo dve oznaki. Pri prvi oznaki (σ_{xy}) uporabimo malo grško črko sigma, v indeksu pa kot produkt navedemo obe spremenljivki, v našem primeru sta to spremenljivki $x$in y. Ta oznaka je primerna, kadar predstavljamo vrednosti. Druga oznaka ($cov\,(x,\, y)$) v obliki je primernejša, kadar s kovariancami in variancami računamo. V tem izrazu moramo obvezno navesti v oklepaju spremenljivki, med katerima računamo varianco. Zaloga vrednosti za kovarianco je obsežna, kovarianca ima lahko vrednost od $-\infty$do ∞. Spremenljivki sta neodvisni, kadar je kovarianca enaka 0, povezava med spremenljivkama obstaja, kadar je kovarianca pozitivna (enačba 1.34) ali negativna (enačba 1.35) .

$$\sigma_{xy} > 0 \qquad [1.34]$$



Slika 1.5: Kvantili z ročaji

$$\sigma_{xy} < 0 \qquad [1.35]$$

Preuredimo enačbo 1.33 tako, da pomnožimo člena znotraj vsote.

$$\frac{\sum \left(x_i y_i - \overline{X} y_i - x_i \overline{Y} + \overline{XY}\right)}{n-1} = \qquad [1.36]$$

Končni rezultat je isti, če najprej izračunamo vrednosti izrazov v oklepaju in jih potem seštejemo, ali pa seštejemo vse člene posamezno in jih potem prišejemo ali odštejemo. Torej lahko enačbo 1.36 razstavimo.

$$= \frac{\sum (x_i y_i) - \sum \left(\overline{X} y_i\right) - \sum \left(x_i \overline{Y}\right) + \sum \left(\overline{XY}\right)}{n-1} = \qquad [1.37]$$

Tako kot smo to storili pri variancah, lahko iz členov v 1.37 izpostavimo povprečja $\overline{X}$ in $\overline{Y}$ pred vsoto.

$$= \frac{\sum (x_i y_i) - \overline{X} \sum (y_i) - \overline{Y} \sum (x_i) + \overline{XY} \sum (1)}{n-1} = \qquad [1.38]$$

V nadaljevanju bomo uporabili naslednji dve enačbi.

$$\sum (y_i) = n\overline{Y} \qquad [1.39]$$

$$\sum (1) = n \qquad [1.40]$$

Ko v enačbi 1.38 zamenjamo levi strani enačb 1.39 in 1.40 z desnima stranema, dobimo enačbo 1.41.

$$= \frac{\sum (x_i y_i) - n\overline{XY} - n\overline{XY} + n\overline{XY}}{n-1} = \qquad [1.41]$$

Zadnji trije členi so enaki, zato jih lahko seštejemo (enačba 1.42).

$$= \frac{\sum (x_i y_i) - n\overline{XY}}{n-1} = \sigma_{xy} \qquad [1.42]$$

Tako kot pri varianci lahko povprečja iz enačbe 1.42 zamenjamo s formulo 1.2za izračun povprečja, kot je to prikazano v enačbi 1.43.

$$= \frac{\sum (x_i y_i) - \frac{1}{n} \sum (x_i) \sum (y_i)}{n-1} = \sigma_{xy} \qquad [1.43]$$

Tudi pri kovarianci smo torej izpostavili enačbe 1.33, 1.42 in 1.43, ki se jih bomo zapomnili, ne glede na to,

1.3.2 Korelacija

Korelacija je mera za podobnost dveh kvantitativnih spremenljivk x in y. Parameter, ki opisuje korelacijo, je korelacijski koeficient (r_{xz}). Izračunamo ga kot razmerje med kovarianco in geometrijskim povprečjem varianc obeh spremenljivk (enačba 1.44). Predstavlja standardizirano kovarianco in ima zalogo vrednosti od -1 do 1.

$$r_{xy} = corr\,(x,\, y) = \frac{cov\,(x,\, y)}{\sqrt{var\,(x)\, var\,(y)}} \qquad [1.44]$$

Korelacija bo višja, če se kovarianca približuje geometrijskemu povprečju variac obeh spremenljivk. Kadar je vrednost enaka nič, korelacije ni. To se zgodi samo takrat, ko je kovarianca med spremenljivkama enaka nič.

S preureditvijo enačbe 1.44 lahko dobimo še eno enačbo za izračun korelacije. Enačba 1.45 je primerna, kadar želimo korelacijo izvrednotiti direktno iz podatkov (npr. pri programiranju).

$$= \frac{\sum \left(x_i - \overline{X}\right) \left(y_i - \overline{Y}\right)}{\sqrt{\sum \left(x_i - \overline{X}\right)^2 \sum \left(y_i - \overline{Y}\right)^2}} \qquad [1.45]$$

Povezanost spremenljivk je močna, kadar je vrednost korelacije blizu ena, tako na pozitivni kot negativni strani. Kadar pa je vrednost korelacije zelo blizu nič ali pa celo nič, pa je povezava med spremenljivkama šibka ali pa je sploh ni. V živinoreji pogosto označimo korelacijo glede na pričakovano vrednost. Če je absolutna vredenost ocene višja od pričakovane, potem jo že označimo za močno povezavo, čeprav morda vrednosti niso prav visoke. Kadar pa je absolutna vrednost manjša od pričakovane, pa tako že govorimo o šibki oz. nizki korelaciji.

Korelacijo se lahko naučimo tudi presojati iz grafov. Kadar se pri povečevanju spremenljivke na osi x, povečuje tudi spremenljivka na osi y, je korelacija pozitivna. Lahko pa se pri povečevanju prve spremenljivke, druga zmanjšuje. Takrat je korelacija negativna. Poseben primer, ko pa povezave ne moremo ne uganiti in ne dokazati z izračuni, pa je takrat, ko med spremenljivkama ni povezave. To se zgodi, kadar so točke okrog premice porazdeljene skoraj v krogu, ali pa ima premica smerni koeficient enak 0 (leži vzporedno z x-osjo) ali ∞ (leži vzporedno z y-osjo). Vlogi spremenljivk sta lahko zamenjani (enačba 1.46), a se vrednost korelacije ne spremeni.

$$corr\,(x,\, y) = corr\,(y,\, x) \qquad [1.46]$$

Na grafičnih prikazih korelacija opisuje razpršenost spremenljivk okrog vrisane linearne premice. Kadar so podatki razpršeni tesno ob premici, je korelacija blizu 1 ali -1. Kadar se oblak meritev oblikuje v elipso, so korelacije močnejše, če je elipsa bolj sploščena ob premici, in šibkejša, če se elipsa približuje krogu. Morda bo lažje, če povežete premico z enko in krog z ničlo.

1.3.3 Regresija

Regresijo bomo razumeli kot funkcijo spremenljivke x, ki najbolje pojasni spremenljivko y. Pri linearni regresiji predpostavljamo, da spremenljivko y lahko dobro napovemo z linearno enačbo. V statistiki enačbe pišemo nekoliko drugače kot v matematiki (enačba 1.47). Tako pri polinomih začenjamo s konstanto (μ), nadaljujejo z linearnim členom ($b_{xy}x_i$) in členi iz višjimi potencami, ki jih v tej enačbi nimamo. Konstanta μ predstavlja presečišče premice z y-osjo, b_{xy} pa predstavlja smerni koeficient premice, ki ga v statistiki poimenujemo regresijski koeficient.

$$y_i = \mu + b_{xy}x_i \qquad [1.47]$$

V enačbi 1.47 smo neodvisne spremenljivke označili z x_i, kjer indeks i predstavlja števec spremenljivk. Indeks zavzema vrednosti od 1 (prva spremenljivka) do števila meritev n (zadnja meritev). Če poznamo konstanto a in smerni koeficient b_{xz}, lahko iz odvisne spremenljivke izračunamo odvisno spremenljivko y_i.

PRIMER : Izračun odvisne spremenljivke
Debelina hrbtne slanine pri pitancih se z rastjo povečuje. Pri analizi podatkov na intervalu od 90 do 110 kg smo ugotovili, da se debelina hrbtne slanine povečuje za 0.10 mm/kg telesna mase. Pri 90 kg so prašiči imeli v povprečju 12 mm slanine. Izračunajmo pričakovano debelino hrbtne slanine pri masi 90, 95, 100, 105 in 110 kg.
Nastavimo linearno enačbo! V tem primeru bomo y-os premaknili na mesto, kjer živali tehtajo 90 kg.

$$y_i = 12\,mm + 0.10\,\frac{mm}{kg} * (x_i - 90\,kg) \qquad [1.48]$$

Izračunane vrednosti odvisnih spremenljivk so prikazane v tabeli 1.2.

Tabela 1.2: Izračun odvisne spremenljivke

Neodvisna spremenljivka (kg)	90	95	100	105	110
Odvisna spremenljivka (mm)	12.0	12.5	13.0	13.5	14.0

Sedaj smo se naučili uporabljati regresijski koeficient, pa ga poskusimo še izračunati. Pri matematiki smo se naučili, da za smerni koeficient potrebujemo dve točki na premici, npr. $T_1\,(x_1,\,y_1)$ in $T_2\,(x_2,\,y_2)$. Smerni koeficient premice (enačba 1.49) izračunamo iz razmerja med spremembama pri odvisni in neodvisni spremenljivki.

$$b_{xy} = \frac{(y_2 - y_1)}{(x_2 - x_1)} \qquad [1.49]$$

PRIMER : Izračun smernega koeficienta premice
Indeks plemenske vrednosti (slika 1.6) se y leti povečuje. Za vsako leto imamo dve ali tri vrednosti, ki so prikazane s točkami. Vrisali smo premico, ki naj bi ponazorila povprečne letne spremembe. Na premici smo izbrali dve točki in sicer $T_1\,(84,\,105)$ in $T_2\,(94,\,142)$. Prva izbrana točka T_1 se slučajno ujema z eno od meritev, ki leži tudi na premici. Druga točka T_2 pa ne predstavlja meritve. Sedaj lahko izračunamo smerni koeficient (enačba 1.50).

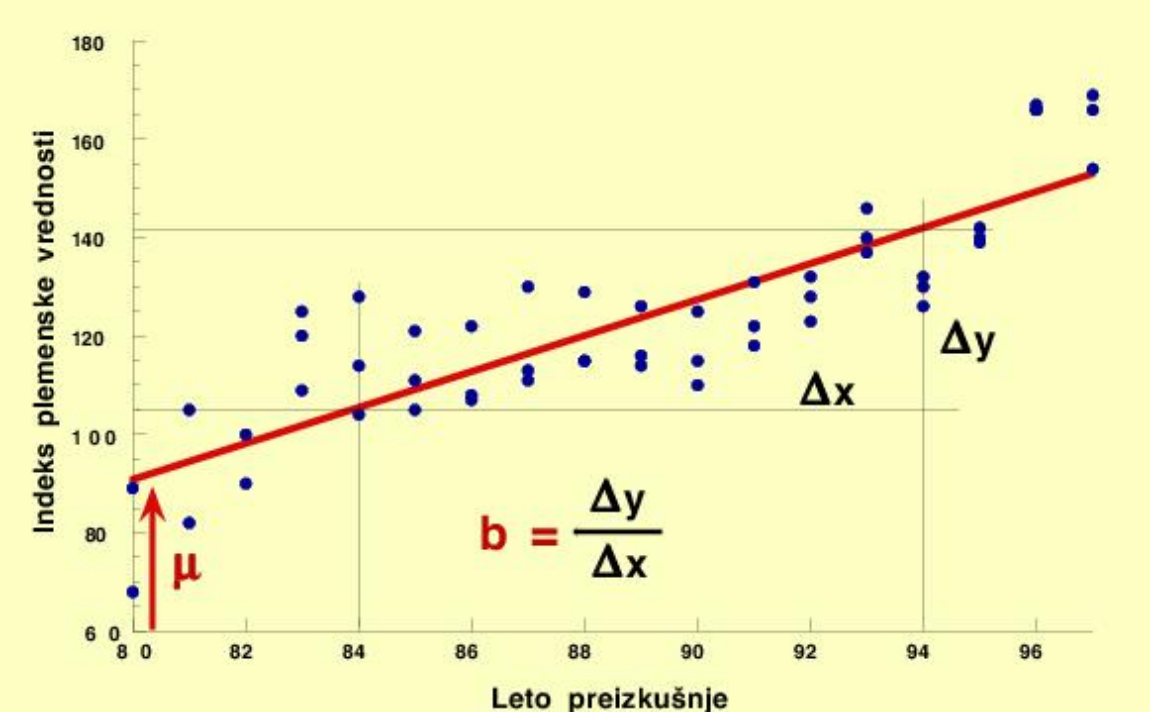
$$b_{xy} = \frac{(142 - 105)}{(94 - 84)} = 3.7 \quad [1.50]$$

Indeks plemenske vrednosti je naraščal za 3.7 enot na leto. Presečišče premice z y-osjo, ki ga v biometriji najraje označimo z μ, lahko preberemo iz slike ali izračunamo (enačba 1.51).

$$\mu = 105 + 3.7 * (80 - 84) = 90.2 \quad [1.51]$$

Sedaj si lahko napišemo enačbo premice. Tudi v tem primeru smo y-os premaknili v točko, kjer ima neodvisna spremenljivka vrednost 80.

$$y_i = 90.2 + 3.7\,(x_i - 80) \quad [1.52]$$



Slika 1.6: Izračun smernega koeficienta premice

Vrednosti za konstanto μ in smerni koeficient b_{xy} smo prebrali iz vrisane premice, za katero pa ne vemo, kako dobro je bila vrisana. Imeti moramo bolj natančno metodo, s katero bomo neposredno iz meritev izračunali parametre - iskani neznanki. Sedaj se bomo enačbo 1.51 kar naučili, kasneje jo bomo tudi izpeljali. Ko bomo smerni koeficient vrisane premice izračunali iz podatkov, ga bomo imenovali tudi regresijski koeficient.

$$b_{xy} = \frac{cov\,(x,\, y)}{\sqrt{var\,(x)}} \quad [1.53]$$

$$= \frac{\sum \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum \left(x_i - \overline{X}\right)^2}} = \quad [1.54]$$

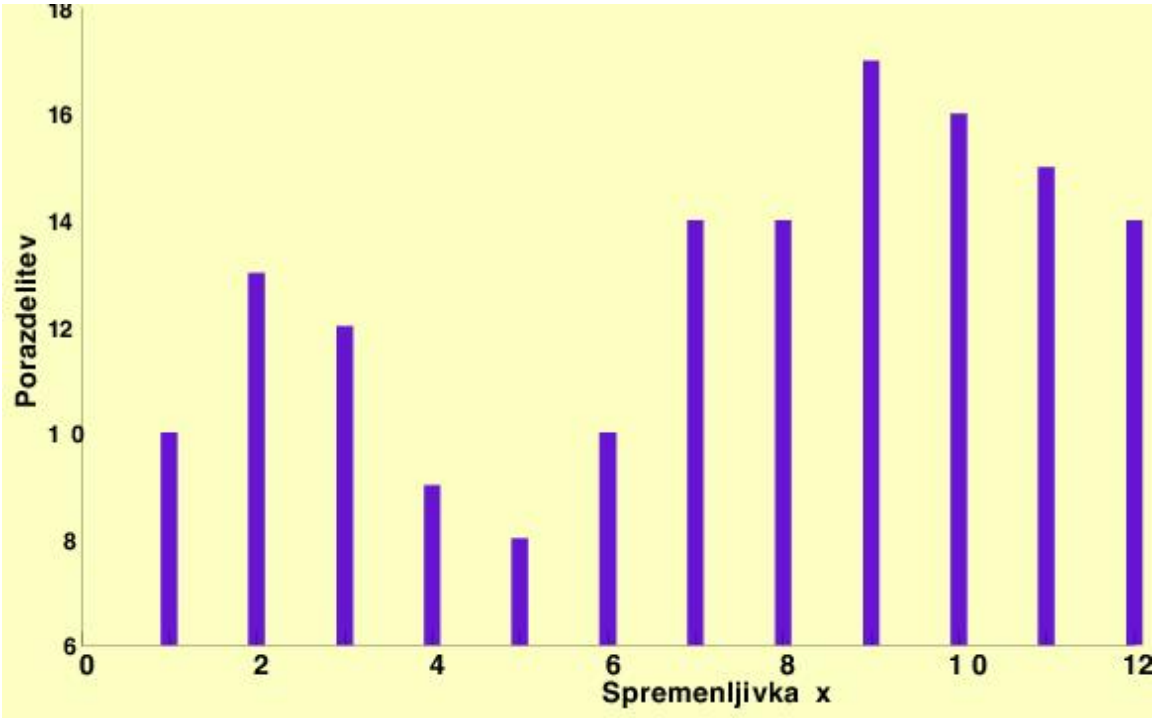
$$= r_{xy}\sigma_y \quad [1.55]$$

Vrednosti za regresijski koeficient je lahko od $-\infty$ do ∞.

1.4 Vaje

Naloga 1:

Določite mere srednje vrednosti na naslednjem grafu in komentirajte rezultat!



Slika 1.7: Diskretna bimodalna porazdelitev

Naloga 2:

V poskusu v dveh čredah smo preiskušali dopolnilno krmo za breje ovce. Zanimala nas je rast jagnjet v času od rojstva do odstavitve. Pri tem smo za vsako jagnje zabeležili: čredo, krmo ovce v času brejosti, število jagnjet v gnezdu, spol, zaporedno jagnjitev kot jagnjitev mladice (ml) ali stare ovce (so), očeta, mater ter starost ob odstavitvi (dni). Jagnjeta smo stehtali ob rojstvu ter ob odstavitvi (v kg, na 100 g natančno). Izračunajte osnovne statistike (pomagajte si z alinejami pod preglednico s podatki) in jih uredite v preglednici!

Izračunajte:

- povprečne vrednosti za rojstno in odstavitveno maso ter starost ob odstavitvi po rejcih in celotni vzorec
- mediano in modus za vse tri spremenljivke v celotnem vzorcu
- varianco in standardni odklon pri obeh rejcih in celotni vzorec
- standardne napake za povprečja iz prve alineje
- kovarianco med rojstno in odstavitveno maso ter med starostjo in odstavitvena maso
- regresijski in korelacijski koeficient med rojstno in odstavitveno maso ter starostjo in odstavitveno maso

Tabela 1.3: Podatki iz preizkusa dopolnilne krme za breje ovce

Čreda	Spol jag.	Zap. jagn.	Jagnje	Rojstna masa	Starost ob odst.	Masa ob odst.
1	m	ml	1	3.2	120	25.2
1	m	so	2	3.5	115	30.6
1	z	so	3	3.8	115	28.7
2	z	ml	4	3.6	125	24.3
2	z	so	5	4.3	130	34.8
2	m	ml	6	2.8	118	22.5
2	z	so	7	3.7	124	29.2
1	m	ml	8	3.4	115	26.4
1	z	so	9	3.9	114	27.3
1	z	so	10	4.1	114	28.7
1	m	ml	11	3.6	132	27.8

Naloga 3:

V prvi skupini je bilo 49 prašičev. Skupaj so tehtali 4983.3 kg, vsota kvadratnih odklonov pa znaša 168.0 kg^2. V drugi skupini je bilo 81 prašičev s povprečno maso 101.7 kg. Predpostavimo, da so v obeh poskusih ostali pogoji konstantni. Izračunajte osnovne statistike in jih uredite v preglednici!

Izračunajte:

- povprečne vrednosti za obe skupini in celotno populacijo
- variance in standardne odklone za obe skupini in celotno populacijo
- minimalno in maksimalno vrednost , če so podatki porazdeljeni normalno
- standardne napake ocen za vsa tri povprečja
- ocenite razliko med prvo in drugo skupino

Naloga 4:

Pri proizvodnji sira preverjajo kakovost na štirih kontrolnih točkah. Pri prvi kontroli so izločili 9.4 % sumljivih sirov. Na drugi kontroli je bilo potrebno iz proizvodnje umakniti 4.1 % sira, pri tretji samo še 1.5 % in pri zadnji je bilo pokvarjenih še 0.4 % sirov. Kolikšen delež sirov je bilo pravilno zorjenih in koliko izločenih do konca?

Naloga 5:

Pri pitanju prašičev na kmetiji smo proučevali izgube od rojstva do konca pitanja. Rejec nam je posredoval izgube za posamezne faze prireje. Pri sesnih pujskih je imel 17.3 % izgub, od odstavitve do preselitve v predpitanje 8.5 %, v predpitanju 3.4 % in v pitanju 1.4 %. Kolikšen delež prašičev je kmet dopital in koliko je bilo izgub do konca pitanja?

Naloga 6:

V dveh loviščih smo tehtali najdeno rogovje jelenov. Izračunajte osnovne statistike in jih uredite v preglednici!

Izračunajte:

- povprečne vrednosti za maso rogovja po loviščih in celotni vzorec
- mediano in modus za maso rogovja v celotnem vzorcu
- varianco in standardni odklon pri obeh rejcih in celotni vzorec
- standardne napake za povprečja iz prve alineje

Tabela 1.5: Poskus z jeleni v dveh loviščih - revirjih

Revir	Jelen	Masa (kg)	Revir	Jelen	Masa (kg)
1	1	13.0	2	9	10.6
1	2	10.3	2	10	10.2
1	3	12.0	2	11	13.7
1	4	13.3	2	12	12.7
1	5	14.0	2	13	14.1
1	6	12.4	2	14	14.2
1	7	14.2	2	15	12.0
1	8	13.5	2	16	10.4