STATISTIKA

1. OPISNA STATISTIKA

1.1. Imenske spremenljivke

Imenska spremenljivka ima zalogo vrednosti, ki jo predstavljajo imena (kategorije) a_1, a_2, ..., a_M. Imenu a_j pripadajočo pogostost (frekvenco, torej število enot vzorca oz. populacije s tem imenom) označimo z n_j. Zbrane podatke pregledno prikažemo v obliki **frekvenčne tabele**

imena	a_1	a_2	...	a_j	...	a_M
frekvence	n_1	n_2	...	n_j	...	n_M

.

Pri tem je vsota frekvenc enaka številu n zajetih enot (to je velikost vzorca oz. populacije). Velja

$$n_1 + n_2 + \ldots n_M = \sum_{j=1}^{j=M} n_j = n \ .$$

Če frekvence n_j delimo z n, dobimo relativne frekvence (deleže vzorca oz. populacije) $r_j = n_j/n$. Dobljene vrednosti lahko izrazimo v odstotkih, $p_j = 100 r_j$.

Primer: 200 dijakov srednje šole so vprašali, kakšne knjige najraje berejo. Tabela prikazuje frekvence, relativne frekvence in odstotke

zvrsti	romani	potopisi	pesmi	življenjepisi	drugo	ni odg.
frekvence	40	30	20	10	50	50
deleži	0,2	0,15	0,1	0,05	0,25	0,25
v %	20	15	10	5	25	25

.

Dobljene rezultate prikažemo grafično:

a) s **stolpčnim diagramom**, ki ga sestavljajo stolpci (iste širine), pri tem so njihove višine **sorazmerne** frekvencam (in tudi deležem) navedenih imen. Pri tem ni smiselno imeti prevelikega števila stolpcev. Če so vrednosti kakšnega stolpca nesorazmerno velike v primerjavi z ostalimi, ga lahko "prekinemo", a moramo na to jasno opozoriti, v pomoč lahko tudi navedemo frekvenco (oz. delež).

b) z **ležečim** (vrstičnim) diagramom: stolpce nadomestijo po vrsticah "ležeči" pravokotniki.

c) s **strukturnim krogom** (kolačem): središčni koti so sorazmerni frekvenci (deležu) posameznih imen.

Če je imenska spremenljivka *ordinalne* narave (vrednosti je mogoče primerjati in razvrstiti po kakšnem kriteriju), je potrebno to urejenost upoštevati pri vrstnem redu stolpcev v tabeli, stolpičev v stolpčnem diagramu in izsekov v strukturnem krogu.

Vrednost (kategorijo, razred), ki ima največjo frekvenco, imenujemo **modus** (modalna vrednost).

Kadar imamo opraviti z več opisnimi spremenljivkami na isti populaciji oz. vzorcu, je težje dobiti jasno sliko o zbranih podatkih. V primeru dveh spremenljivk najlepše prikažemo frekvence v **kontingenčni tabeli**.

Recimo, da imamo lastnosti A in B , vrednosti prve naj bodo $a_1, \ldots, a_r$ in imena za drugo $b_1, \ldots, b_s$. Število elementov, ki imajo hkrati lastnost A enako a_j in lastnost B enako b_k, naj bo $n_{j,k}$ ali, kadar to ne povzroča zmede, n_{jk}. To število imenujemo **celična** frekvenca. Frekvenca $n_{j.}$ imena a_j je tedaj enaka vsoti vseh frekvenc n_{jk}, ko k preteče vrednosti od 1 do s. Podobno je pogostost $n_{.k}$ imena b_k enaka vsoti frekvenc n_{jk}, ko j preteče vse vrednosti od 1 do r. Te **robne** (ali s tujko *marginalne*) frekvence zapišemo v robnem stolpcu oz. robni vrstici tabele. Spet je vsota vseh celičnih frekvenc enaka številu n vseh v proučevanje zajetih enot. Programski paketi za statistične obdelave podatkov omogočajo prikaze **razširjenih** kontingenčnih tabel, kjer so v celicah vključeni tudi deleži glede na vrstico, na stolpec ali na celotno tabelo, pogosto so izraženi v odstotkih. Formalni zapis tabele pri $r = 2$ in $s = 4$ je

$A \setminus B$	b_1	b_2	b_3	b_4	
a_1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
a_2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

Primer: 200 učencev šole razvrstimo po barvi las in oči. Spremenljivka *Lasje* ima vrednosti: *svetli, rjavi* in *črni*, spremenljivka *Oči* pa vrednosti *modre, rjave, drugo*. Rezultati so podani v spodnji tabeli, kjer so v prvi vrstici podani še deleži glede na vrstico, glede na stolpec in glede na celoto.

Lasje \ *Oči*	modre	rjave	drugo	
svetli	30	20	10	60
	0,5	0,333	0,167	1
	0,375	0,25	0,25	
	0,15	0,1	0,05	0,3
rjavi	40	40	20	100
črni	10	20	10	40
	80	80	40	200

Grafično lahko prikažemo porazdelitev celičnih frekvenc v obliki **diagrama strukturnih stolpcev**, kjer so stolpci razvrščeni glede na vrednosti prve spremenljivke, notranje pa so strukturirani glede na vrednosti druge spremenljivke (vlogi seveda lahko zamenjamo). Pri tem si lahko še dodatno pomagamo z barvami ali drugačnimi polnili.

Možen je tudi prikaz z več strukturnimi krogi, kjer krogi pripadajo vrednostim prve spremenljivke, v izseke pa so krogi razdeljeni v skladu s celičnimi frekvencami, torej glede na drugo spremenljivko. Ploščine krogov morajo biti sorazmerne frekvencam prve spremenljivke (polmeri so torej sorazmerni kvadratnim korenom iz frekvenc).

Včasih srečamo tudi prikaze s **piktogrami**, kjer za frekvence uporabljamo primerne grafične simbole: npr. za proizvodnjo pralnih strojev različnih blagovnih znamk za vsako firmo uporabimo pralni stroj. Pri tem mora biti s frekvenco sorazmerna **ploščina** uporabljenega lika!

1. 2. Številske spremenljivke

Števila kot vrednosti takšnih spremenljivk prinašajo naravno urejenost. Zato lahko podatke razvrstimo po velikosti (v naraščajočem ali, manj pogosto, v padajočem vrstnem redu) v **ranžirno vrsto**. Če so podatki števila $x_1, \ldots, x_n$, zapišemo te iste vrednosti (npr. v naraščajočem redu)

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, kjer velja $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Tu smo z indeksom v oklepaju povedali **rang** ustreznega elementa, torej njegovo mesto v tej razporeditvi. Če imamo podatke 21, 32, 27, 31 in 28, se ranžirna vrsta glasi $21, 27, 28, 31, 32$. Vrednost 28 ima pri tem rang 3. kadar imamo več enakih vrednosti, jih seveda v vrsto zapišemo vse, pripišemo pa jim vsem enak povprečni rang. Recimo, da so se na mesta od sedmega do vključno dvanajstega uvrstile vrednosti $23, 24, 24, 24, 24, 26$. Tedaj vsem štirim vrednostim 24 pripada rang $9,5$, saj je $(8 + 9 + 10 + 11)/4 = 9,5$.

Razvrstitev podatkov omogoča enostaven **prikaz** osnovne oblike porazdelitve podatkov **s petimi števili**: najmanjšo vrednostjo (min), prvim kvartilom Q_1, **mediano** me (ali Me ali Q_2), tretjim kvartilom Q_3 in največjo vrednostjo (max). **Kvartili** Q_1, Q_2 in Q_3 so števila, ki celoten nabor vrednosti razdelijo (približno) na četrtine:

- prvi kvartil je meja med spodnjo četrtino urejenih podatkov in ostalimi tremi četrtinami,
- drugi kvartil ali mediana je vrednost, ki po velikosti deli podatke v spodnjo in zgornjo polovico,
- tretji kvartil je meja med zgornjo četrtino urejenih podatkov in ostalimi tremi četrtinami.

Pri lihem številu podatkov ($n = 2k+1$) je mediana vrednost $x_{(k+1)}$, npr. pri 21 podatkih je mediana vrednost z rangom 11; pri sodem številu podatkov ($n = 2k$) pa je mediana aritmetična sredina vrednosti $x_{(k)}$ in $x_{(k+1)}$, torej
$me = (x_{(k)} + x_{(k+1)})/2$.

Torej je pri sto podatkih mediana na sredini med vrednostima z rangoma 50 in 51.

Analogno določimo ostala kvartila. Obstajajo seveda formule, ki omogočajo izračune kvartilov, in so obvezni sestavni del programske opreme za delo s podatki.

Primer:
za 20 podatkov je $Q_1 = (x_{(5)} + x_{(6)})/2$, $me = (x_{(10)} + x_{(11)})/2$ in $Q_3 = (x_{(15)} + x_{(16)})/2$.

Kvartili omogočajo prikaz strukture porazdelitve podatkov v obliki "**škatle z ročaji**" (box and whiskers plot, boxplot na kratko). Škatla sega od Q_1 do Q_3 in je predeljena v dva dela z mediano. Na obeh straneh škatli dodamo ročaja, daljici, ki segata od škatle do najmanjšega (največjega) podatka, ki ne presega oddaljenosti $3kr/2$ od škatle, kjer je $kr = Q_3 - Q_1$ **kvartilni razmik**. Le-ta nam pove velikost intervala, na katerem je zbrana osrednja polovica podatkov. Predstavlja eno izmed mer za razpršenost (raztros) podatkov. Iz nje izpeljemo tudi **kvartilni odklon** (semiinterkvartilni razmik) ko, ki je enak $kr/2$. Če obstajajo podatki, ki so od škatle oddaljeni za več od $3kr/2 = 3ko$, vse te vrednosti, jih imenujemo **osamelci** (outliers). V grafičnih prikazih jih lahko označimo, npr. z majhnimi krožci (kot točke) in morda opremimo z mestom ustrezne enote v podatkih. Običajno preverimo, ali so vrednosti resnične, saj so tolikšna odstopanja izjemna in morda tudi posledica napak (pri merjenju, ob vnosu podatkov, ...). Takšen **strukturni prikaz s kvartili** je ugoden za primerjavo porazdelitev statistične spremenljivke na različnih (pod)populacijah.

Včasih imamo opraviti s številskimi spremenljivkami, ki imajo končno zalogo vrednosti, npr. $\{a_1, a_2, \ldots, a_M\}$, kjer je število M različnih vrednosti majhno. Tedaj zbrane podatke urejeno

prikažemo tako, da podamo ob vrednostih a_j tudi število elementov n_j v naboru podatkov, ki imajo to vrednost (frekvenco, kot v primeru imenskih spremenljivk). Podatke združimo v **frekvenčni tabeli**:

vrednosti	a_1	a_2	$\ldots$	a_j	$\ldots$	a_M
frekvence	n_1	n_2	$\ldots$	n_j	$\ldots$	n_M

.

Na osnovi tega seveda lahko narišemo tudi (**frekvenčni) histogram** (stolpci nad vrednostmi so enako široki, njihove višine so sorazmerne frekvencam) ali pa **frekvenčni poligon** (z daljicami povežemo točke, ki jih postavimo nad pripadajočo vrednostjo v višini sorazmerno s frekvenco). Podobno kot pri imenskih spremenljivkah tudi tu imenujemo **modus** vsako vrednost, ki ima v primerjavi s sosednjimi največjo frekvenco. Če je takšna vrednost ena sama ali dve sosednji, govorimo o **unimodalni** porazdelitvi (frekvenčni poligon ima en maksimum). Če sta takšni vrednosti dve in dovolj razmaknjeni, pravimo, da je porazdelitev **bimodalna**, če jih je več, pa **polimodalna**. Pogosto tiči razlog za več modusov v različni porazdelitvi opazovanega statističnega znaka na podmnožicah osnovne populacije ("plasti"). Npr., raven kakšnega hormona pri odraslih osebah je lahko odvisna od spola.

Številskim spremenljivkam lahko priredimo tudi druge številske mere, s katerimi opisujemo (merimo) nekatere pomembne lastnosti porazdelitve. Najvažnejši sta povprečje in standardni odklon (njegov kvadrat je varianca), včasih pa nas zanimata tudi asimetričnost in kurtozis (koničavost, nasprotje sploščenosti).

Povprečje ali **povprečna vrednost** nabora številskih podatkov je njihova aritmetična sredina. Če imamo za spremenljivko (količino) X nabor vrednosti x_1, x_2,..., x_N, označimo povprečje z oznako $\overline{x}$ ali, če želimo poudariti odvisnost od števila podatkov, $\overline{x_n}$. Velja:

$$\overline{x} = (x_1 + x_2 + \ldots + x_n)/n = \frac{1}{n} \sum\nolimits_{j=1}^{j=n} x_j \ .$$

Če je različnih vrednosti spremenljivke malo in se torej vrednosti v naboru podatkov ponavljajo, podamo podatke s frekvenčno tabelo, v kateri so a_k, $1 \leq k \leq M$, možne vrednosti, n_k pa pripadajoče frekvence. Tedaj računanje povprečja poenostavimo v formulo

$$\overline{x} = (n_1 a_1 + n_2 a_2 + \ldots n_M a_M)/n = \frac{1}{n} \sum\nolimits_{k=1}^{k=M} n_k a_k \ .$$

Zaradi različnih razlogov (veliko različnih vrednosti, zveznost spremenljivke, poenostavljanje prikazov,...) vrednosti pogosto "grupiramo", razdelimo v disjunktne intervale in ob tem pogosto ohranimo le števila podatkov (frekvence) v teh intervalih ter seveda meje intervalov. Za vse nadaljnje izračune uporabljamo osiromašeno informacijo, saj različne vrednosti v intervalu nadomestimo z eno samo: ali je to sredina razreda, ali število, ki je za interval tipično ali vnaprej izbrano po nekem kriteriju. Če to vrednost označimo z a_k za k-ti interval, lahko za približek povprečja uporabimo kar zgornjo formulo. V primeru, da imamo podatke v elektronski obliki, uporabimo ustrezno možnost v menijih programa in dobimo izračunano povprečje.

Kot mediana je tudi povprečje mera za "središče" podatkov. Kako dobro jih opiše, pa nam pove **standardni odklon** (ali deviacija) s. Iz nabora podatkov ga izračunamo po formuli

$$s = \sqrt{s^2}, \text{ kjer je } s^2 = \frac{1}{n} \sum\nolimits_{k=1}^{k=n} (x_k - \overline{x})^2$$

oziroma $s^2 = \frac{1}{n} \sum\nolimits_{k=1}^{k=M} (a_k - \overline{x})^2 n_k$ v primeru ponovljenih vrednosti.

Tako povprečje kot standardni odklon merimo z isto enoto kot spremenljivko X, **varianco** ali **disperzijo** s^2 nabora podatkov pa s kvadrirano enoto. Nadaljnje lastnosti pridejo včasih prav pri računanju:

1. $\overline{(x+c)} = \overline{x} + c$, $\overline{(cx)} = c.\overline{x}$;
2. $s^2 = \frac{1}{n}\sum_{k=1}^{k=n} x_k^2 - \overline{x}^2 = \frac{1}{n}\sum_{k=1}^{k=M} a_k^2 n_k - \overline{x}^2$.

Povprečje odklonov od povprečja, $\frac{1}{n}\sum_{k=1}^{k=n}(x_k - \overline{x})$, je enako 0. Včasih pri opisu podatkov izračunamo tudi **povprečni absolutni odklon** od povprečja $\frac{1}{n}\sum_{k=1}^{k=n}|x_k - \overline{x}|$. Podobno imamo tudi mero za razpršenost pozitivnih spremenljivk, imenovano **variacijski koeficient** $vk = s/\overline{x}$, ki pa je brez enote, torej je enak za isti nabor podatkov, podan v različnih merskih sistemih. Smiselno ga je uporabljati pri vrednostih, ki so dokaj ozko zbrane okoli pozitivnega povprečja.

Če je neočutljivost na spremembe nekaj podatkov dobra lastnost mediane in kvartilov, je dobra lastnost povprečja in standardnega odklona ta, da upošteva vse vrednosti.

Slednje velja tudi za **koeficient asimetrije** podatkov a, (tudi: asimetrija ali asimetričnost) ki ga izračunamo s pomočjo kubov odklonov od povprečja,

$$a = \frac{1}{n}\sum_{k=1}^{k=n}(x_k - \overline{x})^3/s^3 = \frac{1}{n}\sum_{k=1}^{k=M} n_k(a_k - \overline{x})^3/s^3 \ ,$$

in **kurtozis**, ki je mera za sploščenost oz. koničavost nabora podatkov

$$k = \frac{1}{n}\sum_{k=1}^{k=n}(x_k - \overline{x})^4/s^4 - 3 = \frac{1}{n}\sum_{k=1}^{k=M} n_k(a_k - \overline{x})^4/s^4 - 3 \ .$$

Obe količini sta brez enote, torej primerni za primerjavo porazdelitve različnih spremenljivk. Asimetričnost je pozitivna v primeru znatnih pozitivnih odklonov ("rep" na desni), kurtozis pa je pozitivna za porazdelitve, ki so bolj koničaste pri povprečju kot normalne porazdelitve, in negativna, kadar je porazdelitev sploščena. Če je porazdelitev podatkov asimetrična v desno ($a > 0$), imamo pri enimodalni porazdelitvi naslednji vrstni red od leve proti desni: modus, mediana, povprečje. Pri $a < 0$ se vrstni red obrne.

Primer:
Podatke $0, 0, 1, 1, 1, 2, 3, 7, 10, 15$ lahko prikazemo graficno v obliki "histograma":

	x														
x	x														
x	x	x	x				x			x					x
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Vidimo, da ima porazdelitev "rep" na desni. Štetje pokaže, da je modus 1, mediana $me = 1,5$, račun pa razkrije povprečje $\overline{x} = 4$ in varianco $s^2 = 19,8$. Tako je standardni odklon $s = \sqrt{19,8} \simeq 4,45$, variacijski koeficient $vk = \sqrt{19,8}/4 \simeq 1,11$ in koeficient asimetričnosti $a = 166,3/\sqrt{19,8^3} \simeq 1,89$. Kvartila sta $Q_1 = 0,5$ in $Q_3 = 8,5$, zato je kvartilni razmik $kr = 8$.

Če te podatke grupiramo tako, da strnemo po tri sosednje vrednosti, dobimo frekvenčno tabelo

sredina	1	4	7	10	13	16
frekvence	6	1	1	1		1

.

Približek povprečja je tedaj $4,3$ in približek za s kar $\sqrt{24,21} \simeq 4,92$.