STATISTIKA

Populacija in vzorec

Populacija je ime za množico, ki jo preučujemo. Elementi populacije so *statistične enote*. Populacija je lahko zelo številčna množica, tako da je statistična raziskava na njej predraga ali celo neizvedljiva. Zato izvedemo raziskavo na manjši podmnožici, ki je po značilnostih kar najbolj podobna osnovni množici.

Taki podmnožici rečemo *vzorec*. Vzorec je *reprezentativen*, če je po značilnostih kar najbolj podoben osnovni množici. Moč vzorca *imenujemo* numerus in ga označimo z N. Ponavadi vzorec izberemo naključno - torej imajo vsi elementi osnovne množice enako možnost, da so izbrani. V tem primeru rečemo, da je vzorec *slučajen*.

Na podlagi vzorca naredimo statistično oceno, ki velja za celotno populacijo. Stopnja natančnosti oziroma verjetnost trditve za celotno populacijo je odvisna od vzorca in metode raziskave. Lahko se zgodi, da pride do velikih odstopanj.

Podatki

Vrednost ali lastnost statistične enote, ki jo proučujemo, imenujemo *statistični znak, spremenljivka* ali *podatek*. Podatki, ki jih ponavadi zbiramo so lahko:

- številski (numerični)

  Numerični podatki so lahko:
  - celoštevilski (diskretni);
  - ali pa lahko zavzamajo poljubno vrednost na nekem intervalu in so zato zvezni.

- stvarni (opisni, atributni, nenumerični)

Podatki, pridobljeni v posmezni raziskavi, so največkrat neurejeni in nepregledni, zato jih moramo urediti oz. narediti pregledne. To storimo na naslednja načina:

- če je podatkov **malo**, jih **uredimo po velikosti**;
- če je podatkov **veliko**, jih **združimo v skupine**.

Kako ravnati z njimi, je precej odvisno od samih podatkov. Ločimo:

- zvezne podatke (*v danem intervalu zavzamejo poljubno vrednost*),
- diskretne podatke (*jih dobimo s štetjem*)

Če se določeni podatki ponavljajo, potem število ponovitev neka podatka imenujemo frekvenca.

Če je diskretnih podatkov zelo veliko ali če so podatki zvezni, jih združujemo v skupine ali frekvenčne razrede. Celoten razpon podatkov zajamemo z določenim številom frekvenčnih razredov, ki so ponavadi enako široki, ni pa nujno. Veljata naslednji oznaki:

$\boldsymbol{x}_{\max}$ predstavlja podatek z največjo vrednostjo;

$\boldsymbol{x}_{\min}$ predstavlja podatek z najmanjšo vrednostjo;

Grupiranje podatkov

Podatke grupiramo s pomočjo frekvenčnih razredov. **Grupiranja se lotimo tako, da podatke razdelimo v frekvenčne razrede**. Velikost razreda določimo glede na potrebe konkretnega primera in ga izračunamo na nasednji način:

Širina frekvenčnega razreda d_k je razlika med zgornjo mejo razreda z_k in spodnjo mejo razreda s_k:

$$\boldsymbol{d_k = z_k - s_k}$$

Enako širokim razredom določimo njihovo širino tako, da celoten razpon podatkov $x_{\max} - x_{\min}$ delimo z ustreznim naravnim številom (odvisno od primera).

Ko enkrat določimo širino frekvenčnega razreda, lahko določimo še sredino posameznega frevenčnega razreda:

Sredina frekvenčnega razreda predstavlja število, ki je srednja vrednost celotnega razreda:

$$x_k = \frac{z_k + s_k}{2}$$

Frekvenčna porazdelitev

Podatke iz histograma s številkami ponavadi predstavimo še s tabelo, ki ji rečemo frekvenčna porazdelitev. Vsaki vrednosti ali vsaki skupini vrednosti

Frekvenčna porazdelitev je razvrstitev statističnih enot populacije po vrednosti spremenljivke v pripadajoče razrede. **Frekvenca razreda** pa je število enot v posameznem razredu. Frekvenco razreda j označimo kot f_j.

pripišemo število posameznih enot s to vrednostjo ali frekvenco.

Zanimive so tudi vrednosti, ki povedo, kolikšen **delež celote** pomeni posamezna vrednost statističnega znaka oz. število podatkov v frekvenčnem razredu. Kvocient teh dveh števil je relativna frekvenca f' podatka.

Relativna frekvenca razreda j je delež enot populacije, ki je razvrščen v razred j:

$$f'_j = \frac{f_j}{N}$$

pri čemer je f_j frekvenca razreda j, N pa število vseh podatkov.

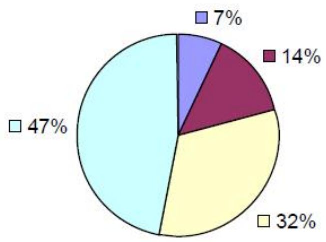
Največkrat ga podajamo v odstotkih:

Poznamo pa še **kumulativno frekvenco F**, ki pove koliko podatkov je zavzelo manjšo vrednost od zgornje meje frekvenčnega razreda.

GRAFIČNO PRIKAZOVANJE PODATKOV

Krožni diagram ali strukturirani krog

Pri majhnem številu frekvenčnih razredov je najbolje uporabiti *krožni diagram* ali *strukturirani krog*. Celota v tem primeru pomeni 360 stopinj, središčne kote, ki pripadajo posameznim vrednostim, pa dobimo s sklepnim računom.

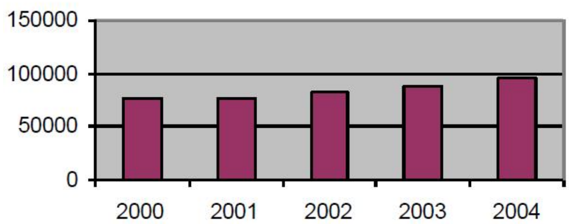


Stolpični diagram

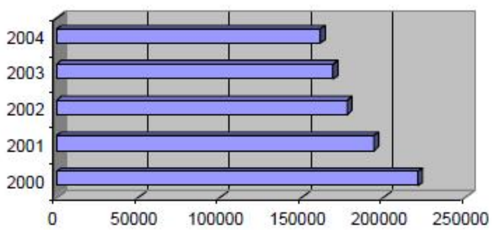
Če so podatki razvrščeni v veliko frekvenčnih razredov lahko dosežejo veliko različnih diskretnih vrednosti. Zato je najbolje uporabiti *stolpični diagram*.

Stolpični diagrami so lahko:

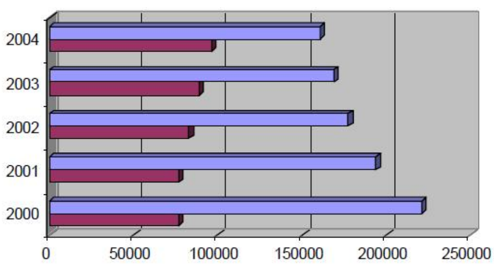
- pokončni



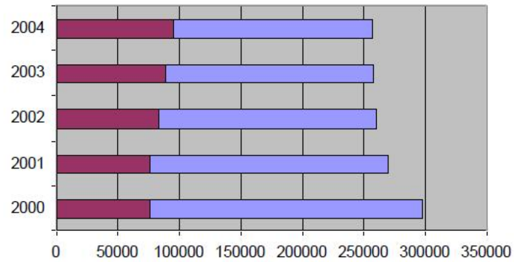
- ležeči



- sestavljeni



- strukturirani



Katero obliko izberemo je odvisno od vrste podatkov in od tega, ali želimo z diagramom primerjati isto vrsto podatkov različnega izvora ali podatke v različnih časovnih obdobjih.

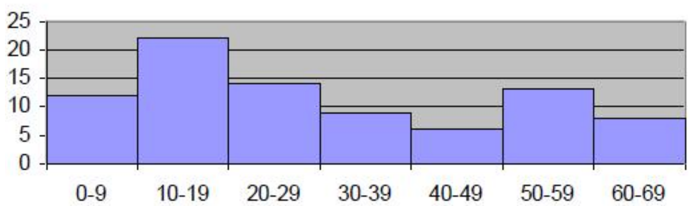
S sestavljenimi stolpčnimi diagrami lahko predstavimo primerjavo med podatki iste vrste različnega izvora tako, da rišemo stolpce drugega ob drugem ali pa jih postavljamo drugega na drugega.

Pomembna lastnost strukturiranih diagramov (strukturirani krog ali strukturirani stolpčni diagram) je prikaz deležev oz. odstotkovna sestava glede na celoto. Zato morajo vsi stolpci strukturiranega stolpčnega diagrama pomeniti 100%.

Strukturni stolpični diagram pa lahko oblikujemo tudi tako, da stolpce, ki predstavljajo deleže, postavimo drugega poleg drugega. Takemu diagramu pravimo sestavljen stolpčni diagram.

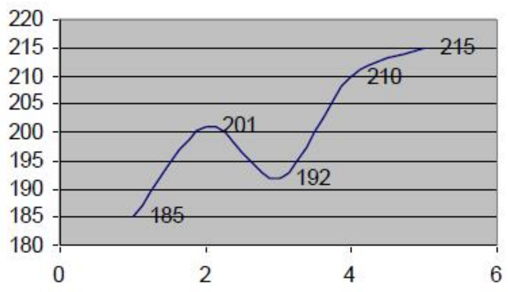
Histogram

Pri tem ni pomembno, ali so širine frekvenčnih razredov enake. Meje razredov narišemo na vodoravni osi, frekvence posameznih razredov pa na navpični osi. Tako nastanejo pravokotniki, ki so drug ob drugem, ploščina posameznega pravokotnika pa je sorazmerna frekvenci tistega razreda.



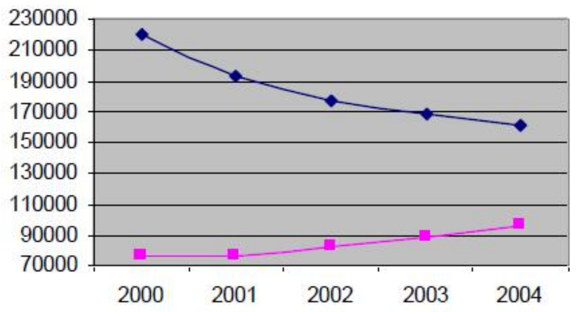
Linijski diagram

Podatki so lahko zvezni ali grupirani. Za grupirane podatke dobimo krivuljo tako, da povezujemo vrednosti sredin frekvenčnih razredov, za zvezne podatke pa krivulja predstavlja funkcijsko odvisnost podatka od časa.



Frekvenčni poligon

Frekvenčni poligon je linijski diagram. Na abscisno os narišemo sredine razredov, nad njimi pa točke v višini frekvenc razredov. Točke povežemo z daljicami.



SREDNJE VREDNOSTI

Aritmetična sredina

Aritmetični sredini pravimo tudi povprečna vrednost. Povprečna vrednost je [vsota vseh statističnih spremenljivk] deljeno s [številom vseh vrednosti]:

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

V primeru, da se vrednosti statistične spremenljive ponavljajo (vrednosti x_1 se pojavijo k_1 −krat, vrednosti x_2 se pojavijo k_2 −krat, itd..), srednjo vrednost izračunamo po enačbi, ki nam računanje nekoliko pospeši in olajša:

$$\overline{x} = \frac{k_1 x_1 + k_2 x_2 + \ldots + k_n x_n}{n}$$

Aritmetična sredina grupiranih podatkov

V primeru, da smo podatke grupirali, aritmetično sredino izračunamo tako:
Najprej za vsak frekvenčni razred izračunamo njemu lastno aritmetično sredino.
Nato pa izračunamo povprečje vseh aritmetičnih sredin frekvenčnih razredov.

**Aritmetična sredina ni najboljša srednja vrednost, v kolikor kakšen podatek zelo izstopa od drugih podatkov, saj lahko ta en podatek nepravično zelo popači rezultat.*

Modus

Modus ali gostiščnica (oznaka: **Mo**) je tista vrednost, ki se v množici vseh vrednosti **najpogosteje ponavlja**.

(OPOMBA: *ob majhnem številu podatkov modus ni najbolj zanesljiva informacija o srednji vrednosti)*

Modus grupiranih podatkov

V primeru, da imamo podatke grupirane, namesto modusa poiščemo modalni razred: to je tisti razred, ki ima največjo frekvenčno gostoto.

Mediana

Mediana ali središčnica (oznaka: **Me**) je tista vrednost statistične spremenljivke, pri kateri je **polovica vrednosti manjših ali enakih, druga polovica pa večjih od nje**.

Da to vrednost ugotovimo, moramo podatke razvrstiti po velikosti. Nato se mediana nahaja ravno na sredini, torej je na zaporednem mestu, ki ga dobimo tako:

Rang mediane: $\frac{\boldsymbol{n+1}}{\mathbf{2}}$ (PAZI: to NI vrednost mediane)

Na tem zaporednem mestu preberemo vrednost mediane, če je število podatkov liho.

V primeru sodega števila podatkov ugotovimo, da ni pravega srednjega podatka. Za vrednost mediane v tem primeru vzamemo aritmetično sredino srednjih dveh podatkov.

Razpršenost podatkov

Da bi lahko informacijo o povprečni vrednosti bolje razumeli (in ne bili zavedeni), vpeljemo novo mero - *razpršenost podatkov*. Razpršenost podatkov si lahko predstavljamo kot vrednost, ki nam pove, **kako blizu so podatki povprečni vrednosti** oz. **kako so porazdeljeni**:

- tem manjša je razpršenost, tem bolje odraža povprečna vrednost tipično vrednost naključnega podatka;
- tem večja je razpršenost, tem slabše odraža povprečna vrednost tipično vrednost naključnega podatka.

Razpršenost podatkov lahko izračunamo z variacijskim razmikom, varianco in standardnim odklonom, ki so najpomembnejše mere za to.

Variacijski razmik

To je razlika med največjo in najmanjšo vrednostjo med podatki. Oznaka: R

$$R = x_{\max} - x_{\min}$$

Varianca in standardni odklon

Varianca je *najpomembnejša* mera za razpršenost podatkov. Pokaže nam stopnjo odstopanja podatkov od povprečja (aritmetične sredine).

Do variance pa pridemo z naslednim razmislekom: če bi sešteli vsa odstopanja od aritmetične sredine - vrednosti večje od povprečja imajo pozitivno odstopanje, vrednosti manjše od povprečja pa negativno - bi se odstopanja seštela natanko v nič. Da se vrednosti ne bi seštele v nič, se znebimo negativnih vrednosti tako, da odstopanja kvadriramo in jih seštejemo:

$$\sigma^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \dots + (x_n - \overline{x})^2}{n}$$

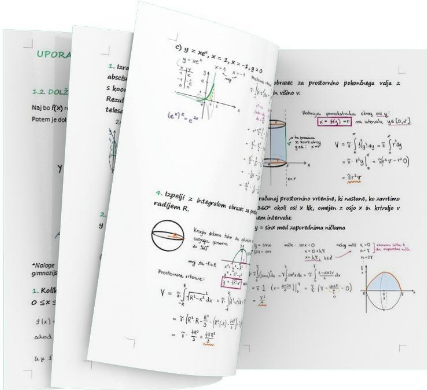
Večja kot je varianca, bolj so podatki razpršeni. A varianca nam podaja mero v kvadrirani obliki. Želimo podatek, ki bo primerljiv s povprečno vrednostjo, zato varianco korenimo; korenjeno varianco imenujemo **standardni odklon**:

$$\sigma = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \dots + (x_n - \overline{x})^2}{n}}$$

ZBIRKE 1000+ REŠENIH NALOG, KI SO ŽE NA VOLJO:

- **ODVOD** (pravila za računanje, uporaba odvoda) z vso teorijo in formulami: **230+ nalog**

- **INTEGRAL** (pravila, določeni in nedoločeni integral, metode integriranja, ploščine in prostornine) z vso teorijo in formulami: **170+ nalog**



- **ZAPOREDJA** (lastnosti zaporedij, limita in konvergenca, aritmetično zaporedje, geometrijsko zaporedje in vrsta, obrestni račun): **280+ nalog**

- **KOMBINATORIKA IN VERJETNOST** (osnovni izrek kombinatorike, permutacije, variacije, kombinacije, binomski izrek, verjetnost): **250+ nalog**

- **REŠENE MATURITETNE POLE** (18 pol; spomladanski rok 2004-2021)

Za dostop do rešenih nalog preveri [TUKAJ.]()