

STATISTIKA



sŠolsko leto 2013/2014

1. OSNOVE STATISTIKE

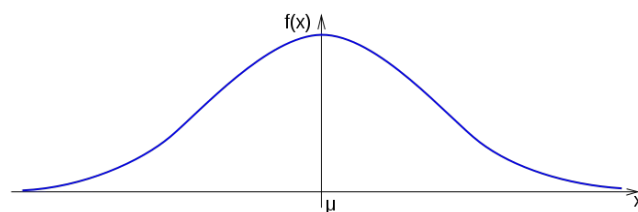
Statistika je znanost in večina o razvoju znanja z uporabo podatkov. Njena osnova je matematična statistika, ki je veja uporabne matematike. Beseda statistika izvira iz latinske besede **status**, ki pomeni stanje, pa tudi država. Izvor besede nakazuje, da so bile prve statistične raziskave ljudska štetja ter gospodarska in demografska vprašanja države.

Do približno 15. stol. Se je statistika uporabljala večinoma za štetje prebivalstva in s tem naborni sistem ter za določanje davkov. Čez čas so statistiki spoznali, da je uporaba te znanosti lahko mnogo širša. Prvi ki je to naredil je bil Anglež John Graunt (1620-1674), lastnik tekstilnih tovarn, ki se je v prostem času ukvarjal s študijem socialnih problemov. To znanost so prvič resno uporabili med epidemijo kolere v Londonu leta 1854, ko je doktor John Snow (1813-1858) statistično dokazal, da sta za epidemijo kriva predvsem slaba kanalizacija in vodovod.

V začetku 18. stol. je prišlo do pomembnega odkritja. Znanstveniki, predvsem astronomi so pri svojem delu opravljali množične meritve iste količine, so pri študiju porazdelitve napak pri ponavljajočih se meritvah ugotovili, da izmerjene vrednosti oblikujejo značilno zvonasto krivuljo oz. da so normalno porazdeljene. Enačbo krivulje je zapisal Abraham de Moivre (1667-1754). Nemški matematik C. F. Gauss (1777-1855) po katerem se imenuje krivulja, je razložil normalne porazdelitve.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-M)^2}{\sigma^2}}$$

Slika 1: Formula normalne porazdelitve



Slika 2: Gaussova krivulja

Uporabnost normalne porazdelitve se je hitro pokazala tudi zunaj matematike in naravoslovnih znanosti. Za nadaljnji razvoj statistike v 20. stol. so bili najzaslužnejši Karl Pearson, William Gosset in Sir Ronald Aylmer Fischer.



Slika 3: Karl Pearson



Slika 4: William Gosset



Slika 5: Sir Ronald Aylmer Fischer

Statistika je danes eno od najpomembnejših orodij za raziskovanje v fiziki, meteorologiji, biologiji, kemiji, agronomiji, sociologiji, psihologiji, ekonomiji, politologiji, komunikologiji, kineziologiji ...

2. OSNOVNI POJMI STATISTIKE

Populacija: množica, ki jo proučujemo

Statistične enote: elementi populacije

Vzorec: podmnožica, katere elementi predstavljajo največjo možno mero značilnosti celotne množice

Reprezentativen vzorec: elementi iz osnovne množice so izbrani tako, da so v njih zajete vse značilnosti osnovne množice

Slučajen vzorec: vsi elementi osnovne množice imajo enako možnost, da bodo izbrani

Numerus: moč vzorca

Statistični znak (spremenljivka, podatek): vrednost ali lastnost statistične enote, ki jo proučujemo.

Statistični znaki so:

Številski (numerični)

- celoštevilski (diskretni) in zvezni

Stvarni (opisni, atributivni)

Parameter: tipična vrednost, s katero opišemo celotno množico podatkov naenkrat

3. PODATKI

Vrednost ali lastnost statistične enote, ki jo proučujemo, imenujemo **statistični znak**, **spremenljivka** ali **podatek**. Podatki, ki jih zbiramo, so po navadi **številski**: višina, teža, starost ... Velikokrat nas tudi zanima spol, rojstni kraj, izobrazba ... Ti podatki pa so **stvarni (opisni, atributivni)**.

Pri zbiranju podatkov moramo upoštevati določena pravila:

1. Najprej preverimo izvor podatkov.
2. Preverimo, ali so izsledki raziskave pred objavo pregledali in preverili neodvisni strokovnjaki.
3. Poizvemo, kdo je zbiral podatke.
4. Ne zaupamo slepo končnim izsledkom, dobljenim s povezavami med podatki.

4. UREJANJE IN GRUPIRANJE PODATKOV

Podatki, pridobljeni v posamezni raziskavi, so največkrat neurejeni in nepregledni, zato jih moramo urediti oziroma narediti pregledne. Če jih je malo, jih uredimo po velikosti, drugače jih združujemo v skupine. Kako ravnamo z njimi, je precej odvisno od samih podatkov; ali so diskretni ali zvezni.

Frekvenca: posamezno število diskretnih statističnih enot iste vrednosti tega statističnega znaka.

Frekvenčni razred: če je diskretnih podatkov zelo veliko ali če so podatki zvezni, jih združimo v frekvenčni razred.

Širina frekvenčnega razreda: d_k je razlika med zgornjo in spodnjo mejo razreda.
 $d_k = z_k - s_k$ (z_k – zgornja meja razreda, s_k – spodnja meja razreda)

Sredina frekvenčnega razreda: $x_k = \frac{z_k + s_k}{2}$
(x_k – sredina frekvenčnega razreda, s_k – spodnja meja razreda, z_k – zgornja meja razreda)

Kumulativna frekvenca: pove, koliko podatkov je doseglo manjšo vrednost od zgornje meje frekvenčnega razreda. To je način grupiranja, ko združujemo frekvence ali frekvenčne razrede od spodaj navzgor oziroma ko »kopičimo« podatke.

4. GRAFIČNO PRIKAZOVANJE PODATKOV

Podatke prikazujemo z različnimi tabelami, vendar jih lahko veliko učinkoviteje prikažemo z **grafikoni**. Izbira vrste grafikona je odvisna od vrste podatkov in števila frekvenčnih razredov pri posameznem primeru.

Vrste grafikonov:

– **Strukturni krog** ali **krožni diagram** je najbolje uporabiti pri majhnem številu frekvenčnih razredov.

– **Stolpčni diagram** uporabimo, če so podatki razvrščeni v veliko frekvenčnih razredov ali lahko dosežejo veliko različnih diskretnih vrednosti.

Vrste stolpčnih diagramov:

- **pokončni**
- **ležeči**
- **sestavljani**
- **strukturni**

– **Frekvenčni poligon** ali **linijski diagram** uporabljamo za opisovanje postopnega spreminjanja vrednosti nekega podatka, največkrat skozi daljše časovno obdobje.

– **Piktogram** ali **slikovni diagram** ponazarjanje podatkov s sličicami (vsaka sličica pomeni določeno število).

5. SREDNJA VREDNOST

Običajno množico podatkov grupiramo in grafično ponazorimo, vendar pa obstaja tudi taka tipična vrednost, s katero lahko opišemo celotno množico podatkov naenkrat.

Taki tipični vrednosti rečemo **parameter**. Parametrov je več vrst, najpomembnejši je **povprečna vrednost** ali **povprečje**.

Tudi srednjih vrednosti je več vrst:

Aritmetična sredina: je kvocient vsote vseh vrednosti statistične spremenljivke s številom teh vrednosti.

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

Modus (gostiščnica): je vrednost podatka, ki se v množici vseh vrednosti najpogosteje ponavlja. Če se zgodi, da se v neki množici dve vrednosti enako mnogokrat pojavita, potem rečemo, da je porazdelitev vrednosti **bimodalna**. Za grupirane podatke namesto modusa poiščemo **modalni razred**: to je tisti razred, ki ima največjo frekvenčno gostoto.