Osnove strojnega učenja in podatkovnega rudarjenja

Statistika: osnovni pojmi in prijemi

izr. prof. Branko Kavšek

Vsebina

osnove
porazdelitve
verjetnost
vzorci

O statistiki …

Definicija (SSKJ):

1. številčni podatki o množičnih pojavih, prikazani navadno v tabelah, grafikonih, …
2. veda o metodah zbiranja in analize podatkov o množičnih pojavih
3. kazalec, ki se izračuna iz vzorčnih podatkov

Primeri “statističnih” izjav

- Najmočnejši izmerjeni potresni sunek je meril 9,2 po Richterjevi lestvici.
- Verjetnost, da je morilec moški, je vsaj 10-krat večja kot za žensko.
- Vsak osmi Južnoafričan je okužen z virusom HIV.
- Leta 2020 bo na vsakega novorojenčka 15 ljudi, starejših od 64 let.

Statistika torej …

- … vključuje matematične izračune
- … se opira na številke

Vendar je pomembno tudi …

- … kako številke izbiramo
- … kako interpretiramo zaključke

Poglejmo na (sledečih treh) primerih →

1. primer

“Statistična” ugotovitev:

Zaradi nove reklamne akcije sladoleda XYZ konec maja meseca se je prodaja le-tega v sledečih treh mesecih povečala za 30%.

Prodaja sladoleda v poletnih mesecih (junij, julij, avgust) tipično naraste ne glede na reklamo.

“Zgodovinski efekt” – interpretacija rezultata glede na eno spremenljivko, ko je zanj odgovorna neka druga spremenljivka (v zgornjem primeru čas).

2. primer

“Statistična” ugotovitev:

Več kot je cerkva v mestu, več je kriminala. Torej: cerkve vodijo v kriminal.

Tako povečanje števila cerkva kot kriminala lahko pojasnimo z večanjem populacije v mestu – v večjih mestih je več cerkva in tudi kriminala.

“Efekt tretje spremenljivke” – napačno predpostavljamo, da obstaja povezava med dvema spremenljivkama, ko v bistvu na obe vpliva tretja spremenljivka.

3. primer

“Statistična” ugotovitev:

V letošnjem letu je 75% več medrasnih porok kot pred 25 leti.

Kaj, če je bilo pred 25 leti 1% medrasnih porok, letos pa jih je 1,75% (75% več) Ali to res kaže na dramatičen porast? Kaj pa fluktuacije v vmesnem obdobju?

Pomanjkanje podatkov – preprosto nimamo dovolj podatkov, da bi lahko nekaj z gotovostjo zatrdili.

Zakaj je poznavanje statistike pomembno?

- Vsak dan se srečujemo s “statističnimi” izjavami, podobnimi tistim iz prejšnjih prosojnic
  - Nekaterim je verjeti
  - Nekatere so lahko zavajajoče
- Poznavanje statistike nam omogoča ločevati med gornjima tipoma izjav
- **Uvod v metode podatkovnega rudarjenja**

Osnovni pojmi in definicije

- Opisne statistike
- Sklepne statistike (inferential statistics)
  - Izbira vzorcev
- Spremenljivke
- (Per)centili
- Merjenje
  - Kakšno mero izbrati?
  - Osnove zbiranja podatkov
- (verjetnostne) Porazdelitve
- Linearne transformacije

Opisne statistike

- Opisujejo podatke, ki jih imamo na voljo
- Ne “sklepajo” na osnovi teh podatkov

- **Opisna statistika**:

  Zanimivo, da Američani več plačujejo ljudi, ki skrbijo za njihove zobe in noge, kot pa tiste, ki jih varujejo in vzgajajo njihove otroke.

  (je Slovenija kaj boljša?)

- **Primer** – tabela povprečnega letnega zaslužka ljudi v ZDA po poklicih za leto 1999:

$ 112.760	pediatri
$ 106.130	zobozdravniki
$ 100.090	podiatri
$ 76.140	fiziki
$ 53.410	arhitekti
$ 49.720	psihologi
$ 47.910	hostese
$ 39.560	učitelji v osnovnih šolah
$ 38.710	policaji
$ 18.980	aranžerji cvetja

Sklepne statistike

- Iz lastnosti **vzorca** sklepamo na lastnosti celotne **populacije**
  - Kako izbrati “pameten” / naključen vzorec?
  - Kaj je to pristranskost (bias) vzorca?

Kako izbiramo vzorce? – vzorčenje

Pravilo:

pristranskost vzorca (*sample **bias***)

Vzorec naj bo **reprezentativen** = naj čimbolje odraža lastnosti populacije + pozor na **velikost** vzorca!

- Vrste vzorčenj:
  - (preprosto) naključno vzorčenje
  - napredna vzorčenja:
    - naključna dodelitev (*random assignment*)
    - stratificirano vzorčenje (*stratified sampling*)

Vzorčenje – primeri (1)

- Naključno vzorčenje:
  - vsak osebek (primer) iz populacije mora imeti **enako verjetnost**, da ga bomo izbrali (v vzorec)
  - izbor enega osebka ne sme vplivati na izbor ostalih = **neodvisnost**

Primer:

Izmed vseh Slovencev, starih med 19 in 35 let anketiramo samo tiste, katerih priimek se začne na črko “Z” in še to vsakega stotega od teh.

V čem je problem?

Vzorčenje – primeri (2)

- Velikost vzorca:
  - majhni vzorci so lahko **nereprezentativni** = ne predstavljajo pravilno lastnosti populacije

Primer:

Na podlagi 10-ih metov “poštenega” kovanca sklepamo o verjetnostih, da pade cifra oz. grb.

V čem je problem?

Vzorčenje – primeri (3)

- Naključna dodelitev:
  - dejanske populacije ni;
    opravka imamo s **hipotetično populacijo**
  - vzorec iz hipotetične populacije naključno razdelimo v 2 ali več skupin = osebke iz vzorca **naključno dodeljujemo** skupinam

Primer:

Pri testiranju učinka zdravil izbrani vzorec ljudi razdelimo v dve skupini. Eni (kontrolni) skupini dajemo t.i. *placebo*, drugi pa pravo zdravilo. Opazujemo ali prihaja do razlik med skupinama.

V čem je lahko problem?

Vzorčenje – primeri (4)

- Stratificirano vzorčenje:
  - vzorčimo v skupinah (**slojih – *stratus*-ih**) glede na določeno lastnost populacije

Primer:

V košari je 1000 žog (populacija), od katerih (vemo, da) je **70% rdečih**, **20% zelenih** in **10% modrih**. Lastnost, ki jo bomo uporabili za stratificiranje vzorca je torej ***barva***.

Kako naj izberemo vzorec 10-ih žog iz zgornje populacije, da bo le-ta reprezentativen?

Spremenljivke

- Tudi: značilke, lastnosti, **atributi**, **razredi**, …
- Lahko jih delimo na:
  - neodvisne, odvisne
  - kvalitativne, kvantitativne
  - diskterne, zvezne
- Več kasneje – v poglavju o **merjenju**

(Per)centili

- Kaj je (**per)centil**? – poglejmo na primeru:

Recimo, da ste na testu motoričnih sposobnosti dosegli rezultat 35, največji možni rezultat pa je 50. Kaj vam to pove o vaši motorični sposobnosti? Kako motorično sposobni ste glede na ostale ljudi?

Bolj informativen podatek bi bil npr.: “kakšen odstotek ljudi je motorično manj sposobnih od mene?” → temu odstotku pravimo **(per)centil**.

Če je vaš rezultat npr. **65-i (per)centil**, to pomeni, da se je **65%** ljudi na testu motoričnih sposobnosti odrezalo **slabše** od vas. V vašem primeru je ***65-i (per)centil = 35***.

3 možne definicije (per)centila

1. **definicija:**

N-ti (per)centil je najmanjša vrednost, ki je *že večja od N%* vseh ostalih vrednosti.

2. **definicija:**

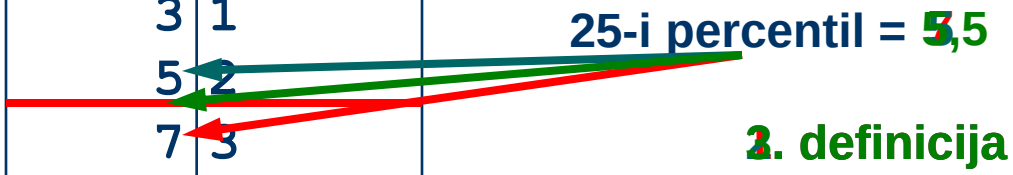
N-ti (per)centil je najmanjša vrednost, ki je *večja ali enaka N%* vseh ostalih vrednosti.

3. **definicija:**

“Interpolacija” med vrednostma iz 1. in 2. definicija (najbolj nedvoumna)

Definicije (per)centila – primer

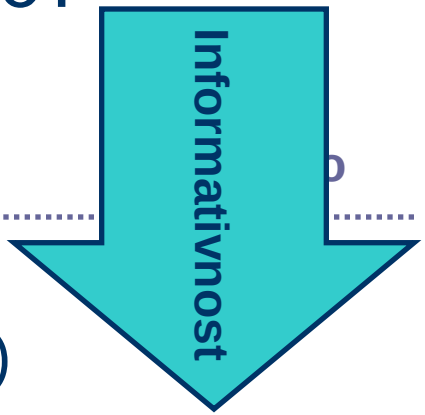
Rezultat	Razvrstitev
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8



Kako stvari izmerimo?

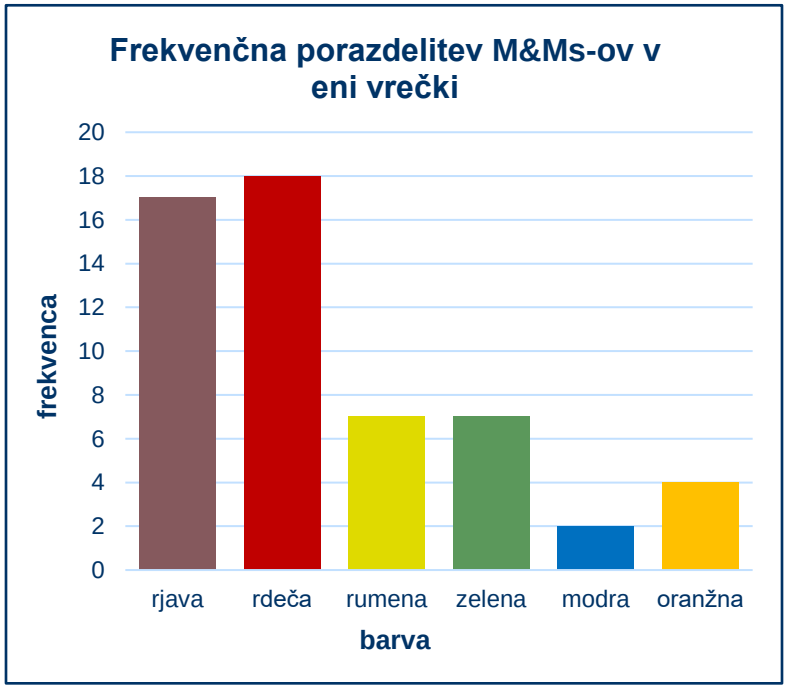
- V znanosti podatki tipično izhajajo iz meritev
- Na kakšne vse načine lahko merimo?
  - Nominalne (opisne) vrednosti
  - Ordinalne (urejene) vrednosti
  - Intervalne vrednosti
  - Vrednosti, ki ohranjajo razmerja (*ratio*)
- Pretvorbe med različnimi tipi vrednosti

  = osnova zbiranja podatkov / **napak**

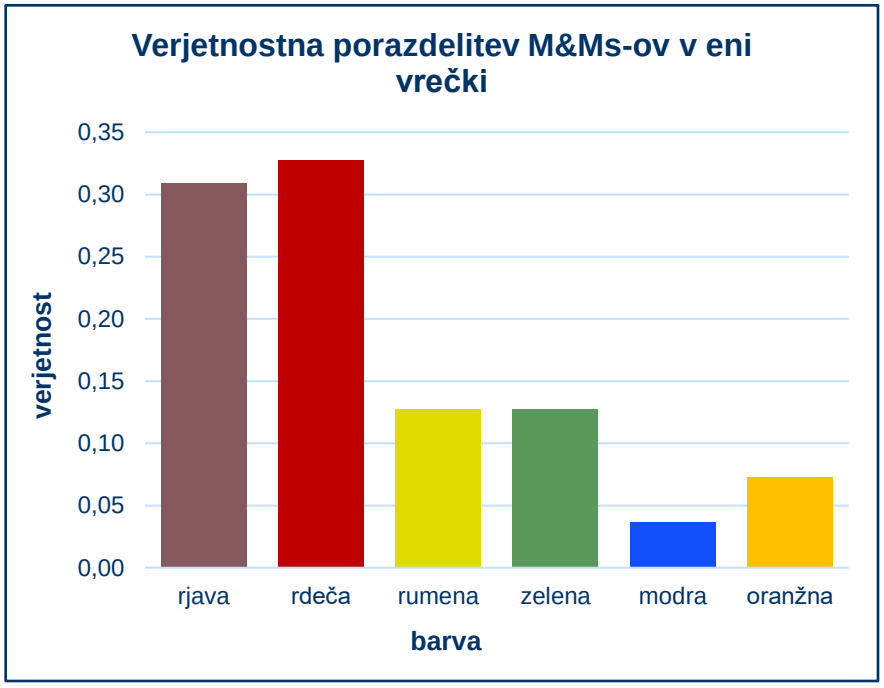


Porazdelitve diskretnih spremenljivk

Frekvenčna porazdelitev:

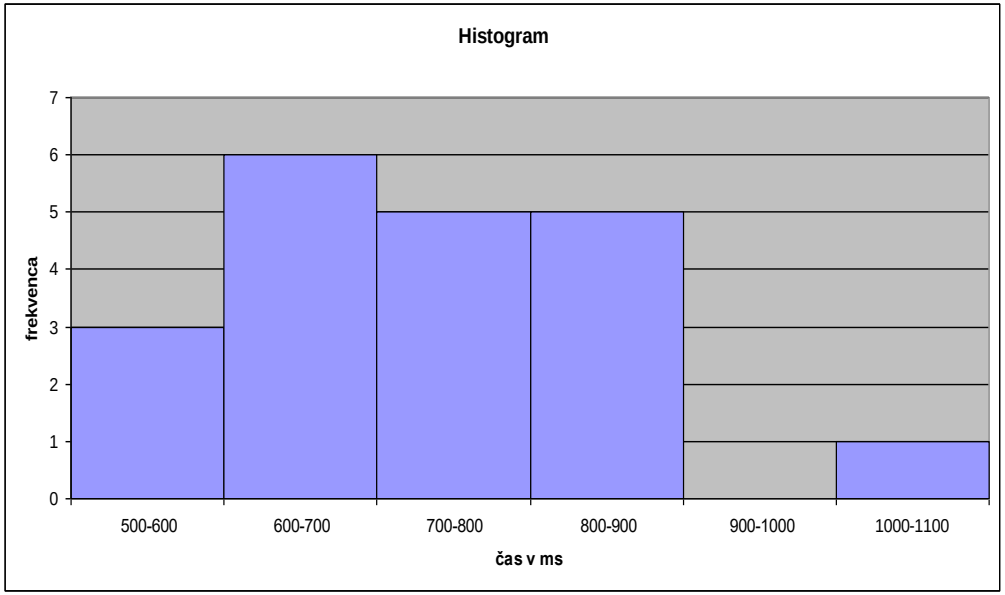


Verjetnostna porazdelitev:



Porazdelitve zveznih spremenljivk

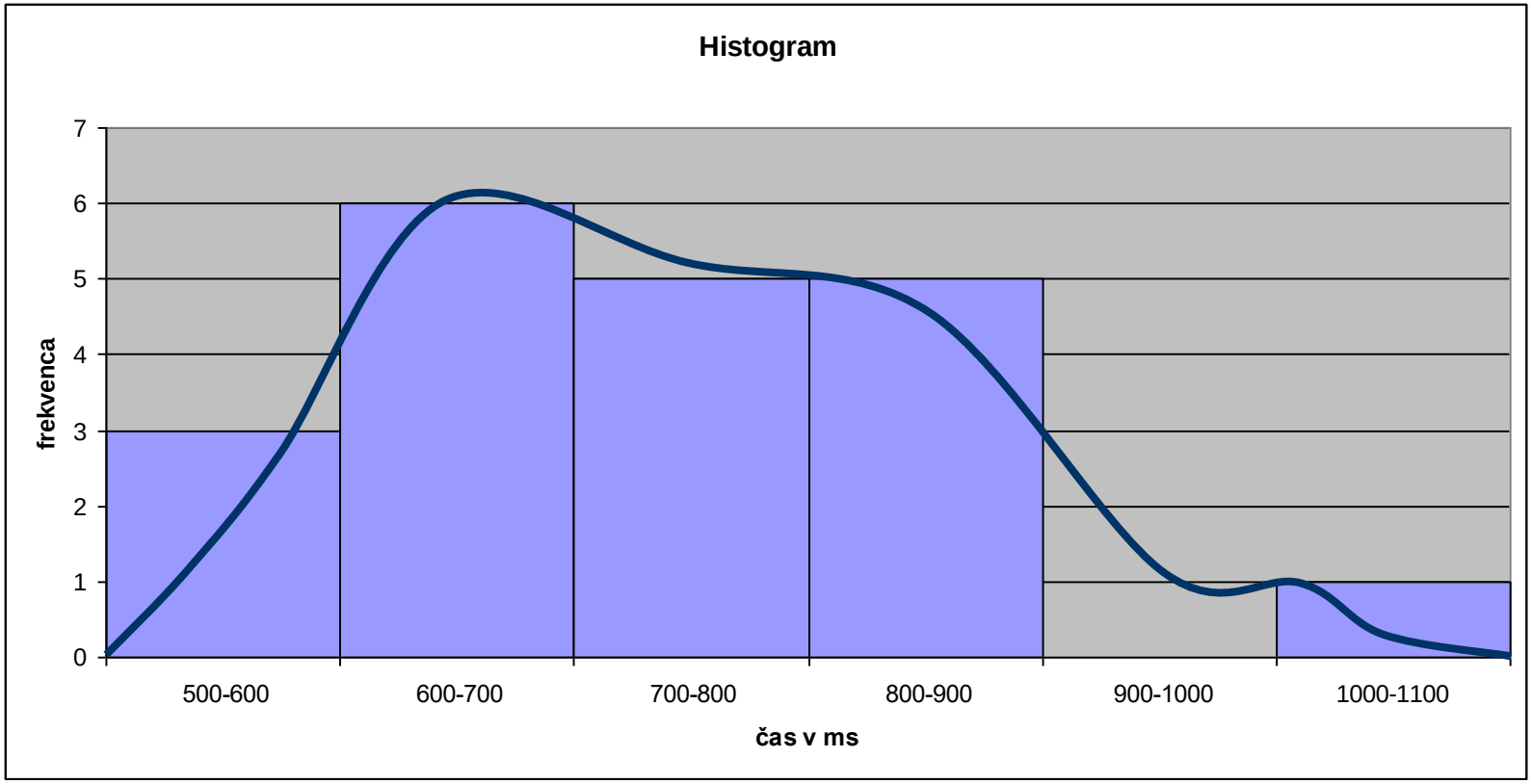
- Grupirana frekvenčna porazdelitev
  - grafično → histogram



Interval	Frekvenca
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Čas v *ms*
568
577
581
640
641
645
657
673
696
703
720
728
729
777
808
824
825
865
875
1007

Gostota verjetnosti



Linearne transformacije

- **Transformacija** = pretvorba
- **Linearna** = samo množenje s konstanto in/ali prištevanje konstante
  - če "originalne" in pretvorjene vrednosti predstavimo v koordinatnem sistemu, dobimo linearno funkcijo
- **Primeri:**
  - Pretvorba iz npr. inčev v centimetre (`x 2,54`)
  - Pretvorba iz ^{o}F v ^{o}C (`x 9/5 + 32`)