Statistics

Bor Bregant

27. maj 2024

Kazalo

0.1 Uvod

Statistika je veda, ki se ukvarja z zbiranjem, analiziranjem, interpretacijo, predstavitvijo in organizacijo podatkov. Izhaja iz *statisticum* (državni), saj je prvotno označevala analizo podatkov o državi. Njene uporabe presegajo matematiko in je temeljno orodje za raziskovanje na vseh področjih znanosti. Pomembno je, da jo uporabljamo odgovorno in etično ter kritično ocenjujemo kontekst in vir statističnih informacij, da se izognemo napačnim interpretacijam in zavajanju (zgled zavajanja).

Podatki in informacije

Podatki so niz vrednosti kvalitativnih spremenljivk.
Informacije so obdelani in interpretirani podatki.
Primer: temperatura po vsem svetu za zadnjih 100 let so podatki; analiza, ki ugotavlja, da globalna temperatura narašča, je informacija

Enote opazovanja in Spremenljivke

Enote opazovanja, za katere se zbirajo podatki, npr. oseba, gospodinjstvo, podjetje, izdelek, stavba, dogodek, država ...
Spremenljivka je katero koli značilno število ali količina, ki jo je mogoče izmeriti ali prešteti, na primer:

- Spol, starost, izobrazba, poklic, dohodek, narodnost ... (za osebo)
- Število članov, vrsta, lastništvo stanovanja, dohodek, internetna povezava ... (za gospodinjstvo)
- Število zaposlenih, lokacija, vrsta, sektor, prihodki ... (za podjetje)
- Dimenzije, teža, starost, barva, temperatura, cena ... (za izdelek)
- Dimenzije, lokacija, starost, lastništvo, materiali, cena ... (za stavbo)
- Površina, prebivalstvo, število podjetij, politična ureditev ... (za državo)

Tipi spremenljivk glede na vrednost in glede na mersko lestvico

Glede na vrednost

Kategorične (atributne) spremenljivke, npr. spol, izobrazba, barva, sektor, vrsta, regija
Številske spremenljivke:

- Zvezne (lahko imajo poljubne vrednosti), npr. točna starost, dohodek, cena, dimenzije, dolžina, širina, trajanje ...
- Diskretne (imajo le celoštevilske vrednosti), npr. letnica rojstva, velikost gospodinjstva, velikost podjetja, število udeležencev ...

Glede na mersko lestvico

- **Nominalne** spremenljivke: vrednosti se lahko razlikujejo le med seboj, razvrščanje ni možno, npr. spol, poklic, sektor, narodnost, regija ...
- **Ordinalne** spremenljivke: vrednosti so lahko razvrščene od najmanjše do največje, vendar razdalje med vrednostmi niso znane, npr. izobrazba, šolska ocena, stopnja strinjanja, stopnja zadovoljstva, tesnoba ...

- **Intervalne** spremenljivke: razlika med dvema vrednostma je smiselna, vendar ni dejanske ničelne vrednosti, je samo arbitrarna, npr. temperatura na lestvici Celzija, pH, koledarsko leto ...
- **Razmernostne** spremenljivke: imajo edinstveno in nearbitrarno ničelno vrednost, zato lahko izračunamo tudi razmerja, npr. temperatura po Kelvinovi lestvici, starost, dolžina, širina, višina, teža, velikost razreda, število udeležencev dogodka, dohodek ...

Kaj lahko izračunamo	**Nominalna**	**Ordinalna**	**Intervalna**	**Razmernostna**
Frekvenčna porazdelitev	✓	✓	✓	✓
Modus	✓	✓	✓	✓
Vrsti red vrednosti		✓	✓	✓
Mediana		✓	✓	✓
Povprečje			✓	✓
Razlika med vrednostmi			✓	✓
Seštevanje in odštevanje			✓	✓
Množenje in deljenje				✓

Python Example:

```
# Import necessary libraries
import pandas as pd

# Step 1: Read the CSV file
# Assume the CSV file is named 'data.csv' and located in the same directory as the s
df = pd.read_csv('data.csv')

# Step 2: Print the first few rows of the dataframe
print("First few rows of the dataframe:")
print(df.head())

# Step 3: Check the data types of each column
print("\nData types of each column:")
print(df.dtypes)

# Step 4: Convert a specific column to float (if necessary)
# Let's assume we have a column named 'income' which we want to convert to float
df['income'] = df['income'].astype(float)

# Verify the conversion
print("\nData types after conversion:")
print(df.dtypes)

# Step 5: Calculate the range of data in a numeric column
# Let's calculate the range for the 'income' column
income_range = df['income'].max() - df['income'].min()
print("\nRange of the 'income' column:")
print(income_range)

# Step 6: Label encoding for an ordinal categorical variable
# Let's assume we have an ordinal variable named 'education_level'
```

```
education_levels = {'High School': 1, 'Bachelor': 2, 'Master': 3, 'PhD': 4}
df['education_level'] = df['education_level'].map(education_levels)

# Verify the encoding
print("\nData after label encoding 'education_level':")
print(df.head())

# Step 7: One-hot encoding for a nominal categorical variable
# Let's assume we have a nominal variable named 'region'
df = pd.get_dummies(df, columns=['region'], prefix='region')

# Verify the one-hot encoding
print("\nData after one-hot encoding 'region':")
print(df.head())

# Additional Step: Descriptive statistics summary
print("\nDescriptive statistics of the dataframe:")
print(df.describe())

# Save the modified dataframe to a new CSV file
df.to_csv('modified_data.csv', index=False)

print("\nModified dataframe saved to 'modified_data.csv'.")
```

Populacija in vzorec

Populacija se nanaša na skupni niz opazovanj; pomembno jo je prostorsko in časovno opredeliti, npr.

- študenti Univerze na Primorskem v študijskem letu 2023/2024
- javni vrtci v Obalno-Kraški regiji na 1. 9. 2023
- gledališke predstave v Sloveniji v tednu od 19. do 25. 2. 2024
- knjige izdane v EU v januarju 2024

Vzorec se nanaša na niz podatkov, izbranih iz statistične populacije po določenem postopku, npr.

- sistematični vzorec 400 študentov UP
- naključni vzorec 200 knjig

Vrste statistične analize

Glede na namen:

- Opisna (deduktivna) statistika: analiza in opis zbranih podatkov brez težnje po posploševanju teh podatkov izven njihovega obsega
- Inferenčna (induktivna) statistika: sklepanje iz vzorca na populacijo

Glede na število sočasno analiziranih spremenljivk:

- Univariatna statistika: analiza ene spremenljivke
- Bivariatna statistika: analiza dveh spremenljivk, npr. hi-kvadrat, mere povezanosti, t-test, ANOVA, regresija, ...
- Multivariatna statistika: analiza več spremenljivk, npr. multipla regresija, analiza glavnih komponent, faktorska analiza, diskriminantna analiza ...

Koraki statistične analize

i Določitev vsebine in namena statistične študije, opredelitev objekta (enota in populacija) in vsebino opazovanja (spremenljivke)

ii Statistično opazovanje (celotne populacije ali vzorca)

iii Enostavna obdelava (urejanje, soritranje podatkov in izračun osnovnih karakteristik)

iv Analitična obravnava

Vaje

Vaja 1

Za naslednje spremenljivke definirajte nekaj možnih vrednosti in navedite, ali so zvezne ali diskretne, ter kakšna je raven merjenja:

- Število dnevnih poslov na ljubljanski borzi
- Temperatura v Kopru v stopinjah Celzija
- Življenjska doba osebnega računalnika
- Število dni letnega dopusta za zaposlene
- Dnevno prehojena razdalja:
  - a. kilometrih
  - b. korakih
- Leto neto dohodek učitelja
- Teža solate v gramih
- Število polic v knjižni omari
- Strinjanje s trditvijo na lestvici od 1 (Sploh se ne strinjam) do 5 (Povsem se strinjam)

Vaja 2

Za naslednje enote navedite nekaj primerov spremenljivk in določite njihovo mersko lestvico:

- Učenec
- Učitelj
- Razred
- Šola
- Učbenik

Vaja 3

Predstavljajte si, da proučujete pojav besede "trajnostni razvoj" v slovenskih srednješolskih učbenikih izdanih v obdobju od 2010 do 2020. Med njimi naključno izberete 250 učbenikov, v katerih preštejete, kolikokrat se pojavi beseda trajnostni razvoj.

- Kaj je enota analize in kaj je spremenljivka?
- Kaj je merska lestvica spremenljivke?
- Kakšen je vzorec in kako velik je?
- Kakšna je populacija?

Vaja 4

Recimo, da je imel Pokrajinski muzej v Kopru v letu 2023 natanko 20.000 obiskovalcev. Predstavljate si, da je vsak deseti obiskovalec muzeja prejel in izpolnil kratek vprašalnik:

- Kako velika je populacija?
- Kako velik je vzorec?
- Napiši vsaj štiri vprašanja, na podlagi katerih bi lahko definirali eno nominalno, eno ordinalno, eno intervalno in eno razmernostno spremenljivko.

Vaja 5

Razišči bazo podatkov raziskave PISA na spletni strani OECD

0.2 Opisna statistika in vizualizacija

Absolutna, relativna in grupirana frekvenčna porazdelitev

Grafična predsstavitev frelvenčnih porazdelitev za nominalne in ordinalne spremenljivke

Stolpčni grafikon ¿ tortni grafikon

Grafična predsstavitev frelvenčnih porazdelitev za intervalne in razmernostne spremenljivke

Histogram
Poligon
Ogiva (kumulativne frekvence)

Normalna porazdelitev

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

Slika
Asimetričnost (skewness)
V levo (k¡0), v desno (k¿0), če k med -1 in 1 še vedno normalna. Slika
Sploščenost (kurtosis)
Lepto (k¡0), mezo (k=0)m plati (k¿0), slika, če med -1,1 normalna

Rangiranje

Primer

Kvantili

Primer in imena

Mere centralne tendence

- **Modus** - vrednost, ki se najpogosteje pojavi v nizu vrednosti podatkov
- **Mediana** - vrednost, ki ločuje zgornjo polovico obsega razpona vrednosti od spodnje polovice
- **Aritmetična sredina** - povprečje niza vrednosti
- Druge mere (geometrijska, harmonična sredina, ...)

Primerjava med modusom, mediano in aritmetično za unimodalne asimetrične - slika

Mere variabilnosti (disperzije)

V kolikšni meri se vrednosti razlikujejo med seboj ter razlikujejo in odstopajo od povprečja. Delimo jih na:

- Absolutne mere (razpon, interkvartilni rang, absolutna deviacija aritmetične sredine/mediane, varianca in standardni odklon)

- Relativne mere (absolutne mere deljene s pripadajočo mero centralne tendence) se izračunajo samo za razmernostne spremenljivke; uporabljamo jih, ko želimo primerjati:
  - Dve porazdelitvi z zelo različno vrednostjo za nek mero centralne tendence;
  - Dve spremenljivki z različnima merskima enotama.

Theory:

Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images. Visualization makes it easier to detect patterns, trends, and outliers in groups of data.

Python Example:

```
import seaborn as sns
import pandas as pd

# Load an example dataset
data = sns.load_dataset("iris")

# Create a pairplot
sns.pairplot(data, hue="species")
plt.show()
```

0.3 Inferenčna statistika

Metode sklepanja iz vzorca na populacijo. Uporabljamo teorijo verjetnosti, da ocenimo, koliko lahko zaupamo rezultatom, pridobljenim na verjetnostnem vzorcu.

Vzorčenje

Je postopek izbire dela populacije, ki jo vključimo v raziskavo. Enote, ki so že izbrane pogosto ne vračamo v populacijo, če pa je velikost vzorca majhna, lahko to naredimo.

Metode vzorčenja

- **Verjetnostni vzorci**: Vsaka enota v populaciji ima znano neničelno verjetnost, da bo vključena v vzorec.
  - *Enostavno slučajno vzorčenje:* Vsaka enota ima enako in znano verjetnost izbire, ki ni enaka nič, notej so vsi možni vzorci enako verjetni. Primer: Z računalnikom naključno ustvarimo vzorec 100 dijakov, vpisanih na neko šolo v šolskem leti 2025/26 na podlagi seznama dijakov.
  - *Sistematično vzorčenje:* Iz vzorčnega okvirja vzamemo vsako k-to enoto. Vsaka enota ima enako verjetnost, da je izbrana v populacijo, toda vsi vzorci niso enako verjetni (npr. ne moremo hkrati izbrati četrte in pete enote, torej vzorčenje ni enostavno). Primer: Na podlagi seznama dijakov šole, ki je urejen po abecedi izberemo vsako deseto enoto. Naključno izberemo le prvo enoto npr. 2 in nadaljujemo z 12, 22, ...
  - *Stratificirano vzorčenje*: Popualcijo stratificiramo na podlagi vnaprej znanih informacij in nato izvedemo vzorčenje za vsak stratum posebej. Primer: Če ima šola 70% dijakov in 30% dijakinj, lahko vzamemo v vzorec enote proporcionalno glede na spol.
  - *Vzorčenje v skupinah*: Enote v populaciji so pogosto združene v skupinah, npr. učenci v razrede, razredi v šole, šole v države, itd. Tako lahko najprej izberemo vzorec skupin (npr. razredov) in naprej na temu vzorcu vzorčimo naprej. Primer: Na univerzi izberemo 5 od 15 programov in za vsakega od teh slučajno izberemo vzorec 100 študentov.
- **Neverjetnostni vzorci**: Verjetnosti izbir ne moremo izračunati.
  - *Priložnostni vzorci*: Primer: Državljanom pošljemo 1 milijon vprašalnikov. (slabo)
  - Ekspertna izbira
  - *Kvotno vzorčenje*: Primer: Med vzorčenjem nadziramo demografske značilnosti vzorca.

  Nauk: Velikost vzorca ni vse. Pomembnejša je njegova reprezentativnost. V idealni situaciji bi uporabljali verjetnostno vzorčenje. Ko to ni možno, je kvotno vzorčenje boljša izbira kot priložnostno.
  Natančnost in točnost s sliko. (precision, accuracy)

Velikost vzorca

Standardna napaka statistike je odvisna od velikosti vzorca.

Velikost vzorca za ocenjevanje aritmetične sredine navadno zahtevamo:

$$n > \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2,$$

kjer E označuje smiselno razliko med opazovanimi vrednostmi (navadno 1)
nekej nekej statistika

Intervali zaupanja

Parametre lahko ocenimo točkovno ali z intervalom. Intervali zaupanja kažejo na točnost ocene in podajo informacijo o njeni zanesljivosti.
S tveganjem α lahko rečemo, da interval(a, b) vsebuje parameter γ.

Slika normalne.

Širina intervala zaupanja je odvisna od:

- *Stopnje tveganja*: Višja je stopnja tveganja α, ožji je interval
- *Velikosti vzorca*: Večja je velikost vzorca n, ožji je interval.

Testiranje hipotez

Znanstvena metoda:

1. Opazovanje pojava
2. Postavljanje raziskovalnih vprašanj
3. Oblikovanje hipotez
4. Zbiranje podatkov
5. Sprejemanje hipotez in razvoj teorij in zakonitosti

Primer: Raziskovalno vprašanje: Ali obstajajo razlike med spoloma v matematični anksioznosti?
Hipoteza: Ženske imajo več matematične anksionosti kot moški ($H_0 : \mu_z - \mu_m > 0$).

Opomba: Hipoteza je vedno trditev, ki jo lahko poskušamo zavrniti. Hipoteza ni potrjena in tehnično ni pravilno reči, da je hipoteza sprejeta. Hipotezo lahko ali zavrnemo ali ne zavrnemo.

Ničelno hipotezo H_0 lahko neposredno preverimo in če je zavrnjena, je alternativna hipoteza pravilna (to je običajno naši cilj)
Alternativno hipotezo H_1 preverimo le posredno: če ničelne hipoteze ne moremo zavrniti, ničelne hipoteze ne sprejmemo, ampak sklenemo, da ni dovolj podatkov, da bi rekli, ali je razlika statistično pomembna

Primer: Ničelna hipoteza: Med ženskami in moškimi ni razlike v sposobnosti večopravilnosti. Alternativna hipoteza: Obstajajo razlike v sposobnosti večopravilnosti med ženskami in moškimi (dvostranski test). Alternativna hipoteza: Ženske so boljše pri večopravilnosti kot moški (enostransko testiranje)

Dve vrsti napake.

Stopnja značilnosti

Izberemo največjo verjetnost α stopnjo, do katere smo pripravljeni tvegati napako tipa I. Običajno se odločimo za 5% stopnjo značilnosti ($\alpha = 0.05$) , lahko pa je tudi 1% ($\alpha = 0.01$) ali nižja. Na podlagi izbranega α določimo kritično območje, kjer bo ničelna hipoteza zavrnjena. Pri tem upoštevamo, ali gre za dvostranski ali enostranski test

Slika two tests same probability level one tail, two tail.

Studentova t porazdelitev

Slika porazdelitve

Razpršenost je odvisna od tako imenovanih prostostnih stopenj (*angl. degrees of freedom*, *df*), ki so opredeljene kot velikost vzorca minus število parametrov populacije, ki jih je treba oceniti na podlagi vzorca.
Večji kot je vzorec, bližje je t porazdelitev normalni Z porazdelitvi

T test za en vzorec

Primerjava povprečja na vzorcu z določeno vrednostjo (npr. populacijskim parametrom ali nevtralno točko na lestvici)

$$t = \frac{mean - comparison}{standarderror}$$

Predpostavke:

- Slučajno vzorčenje, neodvisni vzorci
- Normalna porazdelitev podatkov (če je $N < 30$)

Primer: Ali se študenti pri preizkusu odrežejo bolje od naključja?
28 študentov je opravilo test s 100 vprašanji o besedilu, ki ga niso prebrali.
Vsako vprašanje je imelo 5 možnih odgovorov, zato bi pri povsem naključni izbiri pričakovali, da bo pravilno izbranih 20 postavk.
Vendar so v povprečju pravilno odgovorili na 46,57 vprašanj. S t-testom za en vzorec (angl. one-sample t-test) pokažemo, da so se študenti odrezali statistično značilno bolje od naključja ($M = 46.57, t(27) = 20.6, p < 0.001$).

Vaja 1

> Izbrali smo vzorec šestnajstih otrok in jih stehtali. Določi interval zaupanja za pravo vrednost aritmetične sredine s 5% tveganja.
> X: 35 37 29 26 31 32 28 40 27 33 33 34 31 30 29 38

Vaja 2

> Imamo podatke za vzorec učencev o tem, koliko ur tedensko porabijo za učenje doma. 78 jih je odgovorilo, da se pripravljajo na pouk 2 do 3 ure tedensko, 125 s jih uči 3 do 4 ure na teden, 103 se učijo vsak teden več kot 4 ure. Določi odstotek tistih učencev, ki tedensko posvečajo učenju najmanj časa in oceni ta odstotek v osnovni množici (z 1% tveganjem).

Vaja 3

> Učitelja matematike zanima, ali je povprečno število točk njegovih učencev na preizkusu znanja višje od nacionalnega povprečja, ki je 75. Izračunajte vrednost t-statistike pri stopnji tveganja 5%, določite kritično območje in sklepajte o rezultati za naslednje podatke:
> Za sodelovanje v raziskavi je bil izbran vzorec 25 učencev.
> Povprečna ocena iz matematičnega izpita v vzorcu je 80.
> Standardni odklon v vzorcu je 6.

Vaja 4

Želimo določiti minimalno velikost vzorca za pilotno študijo novega izobraževalnega programa. Standardni odklon učinka programa na bralne sposobnosti je 3 točke. Želite imeti 95% stopnjo zaupanja, da bo vaša ocena učinka natančna. Izračunajte minimalno velikost vzorca.

0.4 Bivariantna statistika

Vaja 1

...

0.5 Multivariantna statistika

Vaja 1

...