

Statistika

Bor Bregant

May 16, 2024

Contents

0.1	Uvod	4
0.2	Opisna statistika in vizualizacija	8
0.3	Inferenčna statistika	9

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 celoštevilске 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 script
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

Določitev vsebine in namena statistične študije, opredelitev objekta (enota in populacija) in vsebino opazovanja (spremenljivke) Statistično opazovanje (celotne populacije ali vzorca) Enostavna obdelava (urejanje, soritrnanje podatkov in izračun osnovnih karakteristik) Analitična obravnava

Vaje

0.2 Opisna statistika in vizualizacija

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