DSAI_Samenvatting

May 26, 2022

1 Datascience & AI

Samenvatting voor examen van AJ 2021-2022.

---

2 Module 1 Basisbegrippen, steekproefonderzoek

2.1 Basisbegrippen

Variabele = algemene eigenschap object, kan objecten onderscheiden. **Waarde** = Specifieke eigenschap, interpretatie van var.

Variabele	Waarde
Gender	Man
Hoogte	180 cm
Funny	Neen.

2.1.1 Meetniveaus

= variabele types. Bepalen beste analyse methode. (visualisatie, centrale tendens en spreiding, verband onderzoeken,…)

Kwalitatief = niet noodzakelijk numeriek. Beperkt aantal waardes.

Nominaal: categorieën zoals gender, ras, land, vorm,…
Ordinaal: Order, rank zoals militaire rank, onderwijsniveau,…

Kwantitatief = Numeriek met eenheid. Veel waardes die vaak uniek zijn.

Interval: Geen vast nulpunt => geen proporties. [1] (°C, °F)
Ratio: Absoluut nulpunt => wel proporties (bv afstand, energie, gewicht,…)

[1] 20 m is 1/3de (~33%) langer dan 15 meter (wel proportie) <-> 20°C is niet 1/3de warmer dan 15 °C (geen proportie)

Relaties tussen variabelen. variabelen hebben en verband als hun waardes systematisch veranderen.

	Pepsi	Coca Cola	Total
Like	56	24	80
Dislike	14	6	20
Total	70	30	100

Totalen zijn *Marginale totalen*

Onderzoek vaak naar **oorzakelijk verband** (frustratie lijd tot agressie, …).
Oorzaak: onafhankelijke variabele
Verband: Afhankelijke variabele

Een verband tussen 2 variabelen zijn niet noodzakelijk oorzakelijk verband!

2.2 Steekproef

Populatie: Volledige verzameling objecten/personen die je wilt onderzoeken
Steekproef: Deel van de populatie waarop metingen uitgevoerd worden.
In bepaalde gevallen is het resultaat van de steekproef toepasbaar op de volledige populatie.

Steekproefmethode: Bepalen populatie -> bepalen steekproefgrootte -> Kiezen van steekproefmethode (budget en tijd)

Hoe keuze maken voor steekproef?
aselecte steekproef: elk element van de populatie heef evenveel kans om gekozen te worden.
Niet aselecte steekproef: De elementen van een sample zijn niet random gekozen. Objecten die makkelijker verkregen worden zijn waarschijnlijker om deel te nemen aan de steekproef. (convenience sampling genoemd).

Stratified to variables: populatie verdeeld op basis van een kenmerk (bijvoorbeeld leeftijd,…). (ook kan vgm bij dit voorbeeld alles /10 gedaan worden (zie slides voorbeeld) en is dit ook stratified)

Gender	<=18	]18,25]	]25,40]	>40	Totaal
Vrouw	500	1500	1000	250	3250
man	400	1200	800	160	2560
Totaal	900	2700	1800	410	5810

2.2.1 Fouten

	Steekproeffout	niet steekproeffout
Accidental	Puur toeval	Onjuist antwoord aangeduid

	Steekproeffout	niet steekproeffout
Systematisch	Online onderzoek: mensen zonder internet uitgesloten. Straat onderzoek: enkel die op dat moment daar aan het wandelen is Vrijwilligers onderzoek: enkel geïnteresseerde mensen	Slecht of niet gecalibreerd meetmateriaal Waarde beïnvloed door het feit dat je het meet. Antwoorders liegen (bv aantal sigaretten per dag)

Algemene imports.

[ ]:
```
#imports
import numpy as np                                      # "Scientific computing"
import scipy.stats as stats                             # Statistical tests

import pandas as pd                                     # Data Frame
from pandas.api.types import CategoricalDtype

import matplotlib.pyplot as plt                         # Basic visualisation
from statsmodels.graphics.mosaicplot import mosaic      # Mosaic diagram
import seaborn as sns                                   # Advanced data␣
 ↪visualisation
import altair as alt                                    # Alternative visualisation␣
 ↪system
```

Python Module 1

[ ]:
```
#Import data van een csv file
ais = pd.read_csv('../data/ais.csv')
#indien geen , maar bijvoorbeeld ; gebruikt dan is het
#pd.read_csv(fileLink, delimiter=';')

##Eerste aantal lijnen tonen
ais.head()

#Aantal rijen en kolommen in een dataset printen
print(f"Aantal rijen: {len(ais)}")
#Aantal kolommen
print(f"Aantal kolommen: {len(ais.columns)}")
#Algemene info over dataset.
```

```
ais.info()
#lijntje * printn
print("*"*50)
#Aantal kolummen per type
print(ais.dtypes.value_counts())

#kolom als index instellen
ais.set_index(['id'])

#Voor een kolom categorie als meetvariabele instellen
ais.sex = ais.sex.astype("category")

#Kan ook variabelen als ordinaal aanduiden met een ordening. Bijvoorbeeld als␣
 ↪we voor sex zouden doen.
# Voorbeeld:
print(ais.sex.unique()) #uniek
sex_Type = CategoricalDtype(categories=['f','m'], ordered=True) #en orderen
ais.sex= ais.sex.astype(sex_Type)
#een kolom beschrijven
print(ais.ferr.describe())

#SELECTEREN DATA
#Toon de tweede rij
ais.iloc[[1]]

#Toon rij 4 tot en met 6
ais.iloc[4:7]
#Toon KOLOM 6 tem 8: (ferr, bmi, ssf)
ais.iloc[:,5:8]
#Toon 1 variabelen (pcBfat)
ais['pcBfat']

#Toon alles van specifieke query (sport=netball)
ais.query("(sport=='Netball')")
#Toon specifieke colom met specifieke query (colom wt van sport=netball)
ais.query("(sport=='Netball')").wt

#Toon allesmet een bmi>26
print("BMI ding")
print(ais[ais.bmi>26])

#Toon frequentie en dergelijke
bmiais = ais[ais.bmi>26]
sns.countplot(x=bmiais.sport, data=bmiais)

#Tel hoevaak een bepaalde categorie voorkomt
ais["Sport"].value_counts()
```

3 Module 2 Analyse van 1 variabele

3.1 Centrale tendens en spreiding

3.1.1 Maten van centrale tendens

Mean of Average De *arithmetic mean* is de som van alle waarden gedeeld door het aantal waarden.
$> \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Median Sorteer alle waarden en neem het middelste (gemiddelde bij een oneven).
Mode de mode is de waarde die het meest voorkomt in een dataset.

3.1.2 Maten van centrale spreiding

Range Absolute waarde van het verschil tussen het hoogste en laagste waarde.
Quartielen De quartielen van een gesorteerde set zijn 3 waarde die de set in 4 gelijke delen verdelen. Q_1, Q_2 , Q_3

Variantie: De variantie (S^2 of σ^2) is hetgemiddelde (mean) van het kwadraat van het verschil van de waardes van de dataset en het gemiddelde (arithmetic mean).
>S^2 = $\frac{1}{n-1} \sum_{i=i}^{n} (x_i - \overline{x})^2$

Standaard afwijking: De standaard afweiking (S of σ) is de wortel van de variantie

3.1.3 Samenvatting Centrale tendens en spreiding

Meetniveau	Center	Sprijdingsmaat
Kwalitatief	Mode	-
Kwantitatief	Avergae/mean Median	Variantie, standaard afwijking, range, interkwartielafstand

3.1.4 Samenvatting Symbolen

	Populatie	Steekproef
aantal elementen	N	n
Gemiddelde (mean)	µ	$\overline{x}$
variance	$\sigma^2 = \frac{\sum (x_i -)^2}{n-1}$	$S^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$
Standaard diviatie	σ	S

3.2 Data visualisatie

3.2.1 grafiek type overzicht

Meetniveau	Grafiek type
Kwalitatief	Staafdiagram
Kwantitatief	Boxplot Histogram Density plot

Taart diagrammen
vermijd gebruiken van taart diagrammen. Hoeken vergelijken is moeilijker dan lengtes, onbruikbaar voor veel categorieën

tips Assen labelen, duidelijke titel, eenheid, label die de grafiek verduidelijkt.

data distortion = zorgt voor fout interpreteren.