

DSAI_Samenvatting

May 26, 2022

1 Datascience & AI

Samenvatting voor examen van AJ 2021-2022. Door ydm#1001.

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,...)

¹20 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 leidt 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 heeft 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)

2.3 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
```

2.4 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 kolommen 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 ordenen
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 kolom met specifieke query (kolom wt van sport=netball)
ais.query("(sport=='Netball')").wt

#Toon alles met 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 hoe vaak 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.

$$> \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Median Sorteert alle waarden en neemt 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 waarden die de set in 4 gelijke delen verdelen. Q_1 , Q_2 , Q_3

Variantie: De variantie (S^2 of σ^2) is het gemiddelde (mean) van het kwadraat van het verschil van de waarden van de dataset en het gemiddelde (arithmetic mean).

$$> S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

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

3.1.3 Samenvatting Centrale tendens en spreiding

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

3.1.4 Samenvatting Symbolen

	Populatie	Steekproef
Aantal elementen	N	n
Gemiddelde (mean)	μ	\bar{x}
Variantie	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
Standaard deviatie	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.

3.3 Python Module 2

```
[ ]: #distributie van gevens voor sport (distribution)
sns.displot(data= ais["sport"])

#categorie plot voor sports
sns.catplot(data= ais, kind="count", x="sport")

#distribution met Kernel density estimate (soort van normaalverdeling achtige
↳ding te krijgen)
sns.displot(data=ais[ais.sex=="f"].ht, kde=True)

#dingen
rowers = ais[ais.sport == "Row"].ht
print(f"Mean: {rowers.mean()}")
print(f"Standard deviation: {rowers.std()}") # Pay attention: n-1 in the
↳denominator
print(f"Variance: {rowers.var()}") # Pay attention: n-1 in the
↳denominator
print(f"Skewness: {rowers.skew()}")
print(f"Kurtosis: {rowers.kurtosis()}")

# Median & co
print(f"Minimum: {rowers.min()}")
print(f"Median: {rowers.median()}")
print(f"Maximum: {rowers.max()}")
percentiles = [0.0, 0.25, 0.5, 0.75, 1.0]
```

```
print("Percentiles", percentiles, "\n", rows.quantile(percentiles))
print("Inter Quartile Range:", rows.quantile(.75) - rows.quantile(.25))
print(f"Range : {rows.max() - rows.min()}")
```

4 Module 3.1 De centrale limietstelling, betrouwbaarheidsintervallen

4.1 Kansverdeling van een steekproef

4.1.1 Kans

Kans is de relatieve frequentie van het voorkomen van een bepaald event (bij uitvoeren van groot aantal onafhankelijke experimenten)

- kansen zijn getallen aan een set toegewezen - Die sets zijn deel van een allesomvattende set, het *universum* Ω - De nummers (kansen) toegewezen aan een set voldoen aan 3 basis regels (axiom van kans) om overeen te komen met hoe kansen werken

1. Kansen zijn niet negatief $P(A) \geq 0$ voor elke A .
2. Het universum heeft een kans 1: $P(\Omega) = 1$.
3. Als A en B disjunct zijn ($A \cap B = \emptyset$) dan geldt $P(A \cup B) = P(A) + P(B)$ dit heet de somregel.

Eigenschappen

1. Complement regel: voor elke A geldt $P(\bar{A}) = 1 - P(A)$ als \bar{A} voorstelt dat A niet voorkomt.
2. Het onmogelijke event is kans nul: $P(\emptyset) = 0$
3. De algemene som regel: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Onafhankelijke events Een event is onafhankelijk als het voorkomen van dit event (of het weten dat dit voorkomt) de kans dat een ander event gebeurt niet beïnvloed. Wiskundig: $P(A \cap B) = p(A)p(B)$

4.1.2 Random variabele

Een random variable is een waarde toekennen aan verschillende gebeurtenissen. Bijvoorbeeld. 1 als je een J trekt uit een kaart spel, 2 bij een Q en 3 bij een K en tot slot 0 bij alle andere mogelijkheden

Kansverdeling functie (PDF) wiskundig. $f_x(x) = P(X = x)$

Expectation of a R.V. Verwachting van een random variabele is geschreven door x of $E(X)$ en is gegeven door $X = \sum_i x_i P(X = x_i) = \sum_i x_i f_x(x_i)$

TODO VOORBEELD KAARTSPEL SLIDES INVOREN

Variantie van een R.V. De variantie van een random variabele is bepaald door $\sigma^2 = \sum_i (x_i - x)^2 P(X = x_i) = \sum_i (x_i - x)^2 f_x(x_i)$.

Standaard afwijking: $\sigma_x = \sqrt{\sigma_x^2}$

Continue random variabele

5 Samenvatting Symbolen

	Populatie	Steekproef
Aantal elementen	N	n
Gemiddelde (mean)	μ	\bar{x}
Variantie	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
Standaard deviatie	σ	S

	Symbool	Formule
Expectation van een random value	μ_x of $E(X)$	$\sum_i x_i P(X = x_i) = \sum_i x_i f_x(x_i)$
Variantie van een random value	σ^2	$\sum_i (x_i - \mu)^2 P(X = x_i) = \sum_i (x_i - \mu)^2 f_x(x_i)$
Standaardafwijking van een R.V.	σ_x	$\sqrt{\sigma_x^2}$