samengevat

May 26, 2022

1 Datascience & AI

1.1 > Samenvatting voor examen van AJ 2021-2022.

1.2 ## Module 1 Basisbegrippen, steekproefonderzoek

1.3 Basisbegrippen

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

Variabele	Waarde
Gender	Man
Hoogte	180 cm
Funny	Neen.

1.3.1 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,…)

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

	Pepsi	Coca Cola	Total
Like	56	24	80

[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)

	Pepsi	Coca Cola	Total
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!

1.4 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

1.4.1 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)

[ ]:
```
#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
```