Data-Science 1

centrum spreiding

Karel de Grote Hogeschool

14/03/2023

Oefeningen huiswerk

- van welke kolommen in de enquête kan je relatieve frequenties berekenen?
- van welke kolommen in de enquête kan je cumulatieve frequenties berekenen?
- welke kolommen moet je in klassen opdelen om frequenties te kunnen berekenen?

Oefeningen huiswerk

- welk vak vinden de studenten het zwaarst? Aan de hand van welke frequenties kan je dit zien?
- welk vak vinden de studenten het minst boeiend? Aan de hand van welke frequenties kan je dit zien?

Oefeningen huiswerk

- wat is de lengte van de kleinste en de grootste student?
- in hoeveel klassen moet je de lengte opsplitsen volgens de methode van Scott?
- splits de kolom lengte op in zoveel klassen
- welke klasse komt het meest voor?
- welke klasse komt het minst voor?
- maak een plot van de frequenties van de klassen. Welke plot is hier aangewezen?

Oefeningen huiswerk

- bepaal alle mogelijke frequenties van besturingsysteem
- maak een plot van de absolute frequenties. Welk diagram is hier aangewezen?
- maak een plot van de percentielscores als dit mogelijk is

Oefeningen huiswerk

- bepaal alle mogelijke frequenties van informatica_belangrijk
- maak een plot van de absolute frequenties. Welk diagram is hier aangewezen?
- maak een plot van de percentielscores als dit mogelijk is
- hoeveel percent van de studenten vindt het extreem belangrijk om informatica te studeren?
- hoeveel percent van de studenten vinden het matig belangrijk of minder?

Oefeningen huiswerk

- bepaal de cumulatieve percentages voor de antwoorden op de vraag: “In hoeverre geloof je dat het belangrijk is om middelen gelijk en eerlijk te verdelen?”
- hoeveel percent van de studenten gaf een score van 3 of minder op deze vraag?
- hoeveel percent van de studenten gaf een score van 3 of meer op deze vraag?

Oefeningen huiswerk

- hoeveel studenten kiezen er voor iedere afstudeerrichting?
- wonen de meeste studenten in de stad of erbuiten?

Inhoud

- centrummaten
  - modus, mediaan, gemiddelde, andere gemiddelden
- spreidingsmaten
  - bereik, interkwartielafstand, standaardafwijking
- uitschieters

Centrummaten

Modus

- voorbeeld: zelfde als vorige week: laptops
- wat is de modus van cpuGeneration?

cpuGeneration	absolute frequentie
Sandy Bridge	63
Ivy Bridge	107
Haswell	166
Broadwell	218
Skylake	155
Kabylake	143

Modus

- vanaf welk meetniveau?
- gemakkelijk te bepalen adhv frequentietabel
- ongevoelig voor uitschieters
  - wat is dit?
  - waarom is de modus ongevoelig?
- soms moet je klassen maken: indeling kan veel invloed hebben!
- er kan meer dan 1 modus zijn

Mediaan

- hoe berekenen?
- welk meetniveau dus?

Mediaan

- ongevoelig voor extremen
- geen klassen nodig
- de waarden van de getallen speelt geen rol, enkel de volgorde
- je moet alle waarden sorteren: veel rekenwerk!
- als je mediaan incrementeel wil berekenen: hou frequentietabel bij
- geen exacte formule voor mediaan: moeilijk om wiskundige eigenschappen te vinden

Gemiddelde

- hoe berekenen?

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- welk meetniveau?

Gemiddelde

- vanuit een frequentietabel berekenen

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{m} f_i \cdot x_i$$

aantal keer sport/week	frequentie
0	18
1	20
2	22
3	10
4	3
5	1
6	1
totaal	75

Gemiddelde

- er is een formule => gemakkelijk om eigenschappen aan te tonen
- alle waarden spelen een rol
- ook uitschieters spelen een rol...
- bekomen waarde bestaat niet noodzakelijk

Andere gemiddelden

- rekenkundig gemiddelde
- gewogen gemiddelde
- meetkundig gemiddelde
- harmonisch gemiddelde
- voortschrijdend gemiddelde

Gewogen gemiddelde

$$\overline{x} = \frac{1}{\sum g_i} \cdot \sum_{i=1}^{m} g_i \cdot x_i$$

vak	sp	score
Computersystemen 1	6	18
OO programmeren 1	11	15
Datastr. en algo's	5	12
Data-analyse	3	10
Netwerkarchitectuur 1	3	18
Communicatie 1	7	13
User interfaces 1	4	17
Databanken 1	7	15
Software engineering 1	7	13
Boekhouden	3	12
Management accountig	4	12
totaal	60	??

Meetkundig gemiddelde

- koers van aandeel:

  +5%, +3%, -2%

- hoeveel gemiddeld gestegen of gedaald?

$$\overline{x} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

Harmonisch gemiddelde

- stel: auto rijdt bepaalde afstand
  - heen: 120 km/h
  - terug: 100 km/h
- hoeveel gemiddeld?

$$\frac{1}{\overline{x}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{x_i}$$

Voortschrijdend gemiddelde

- stel: data is een stream
- bereken gemiddelde van laatste n waarden
- wordt gebruikt bij forecasting

Spreidingsmaten

2

Voorbeeld

> x1 = [1, 2, 3, 5, 46, 87, 88, 89, 90]
> x2 = [42, 43, 44, 45, 46, 47, 47, 48, 49]

- wat is het gemiddelde en de mediaan?
- wat is het verschil tussen deze rijen getallen? == "spreiding"

Bereik

- = verschil tussen max en min waarde
- nadeel
  - uitschieters

Interkwartielafstand

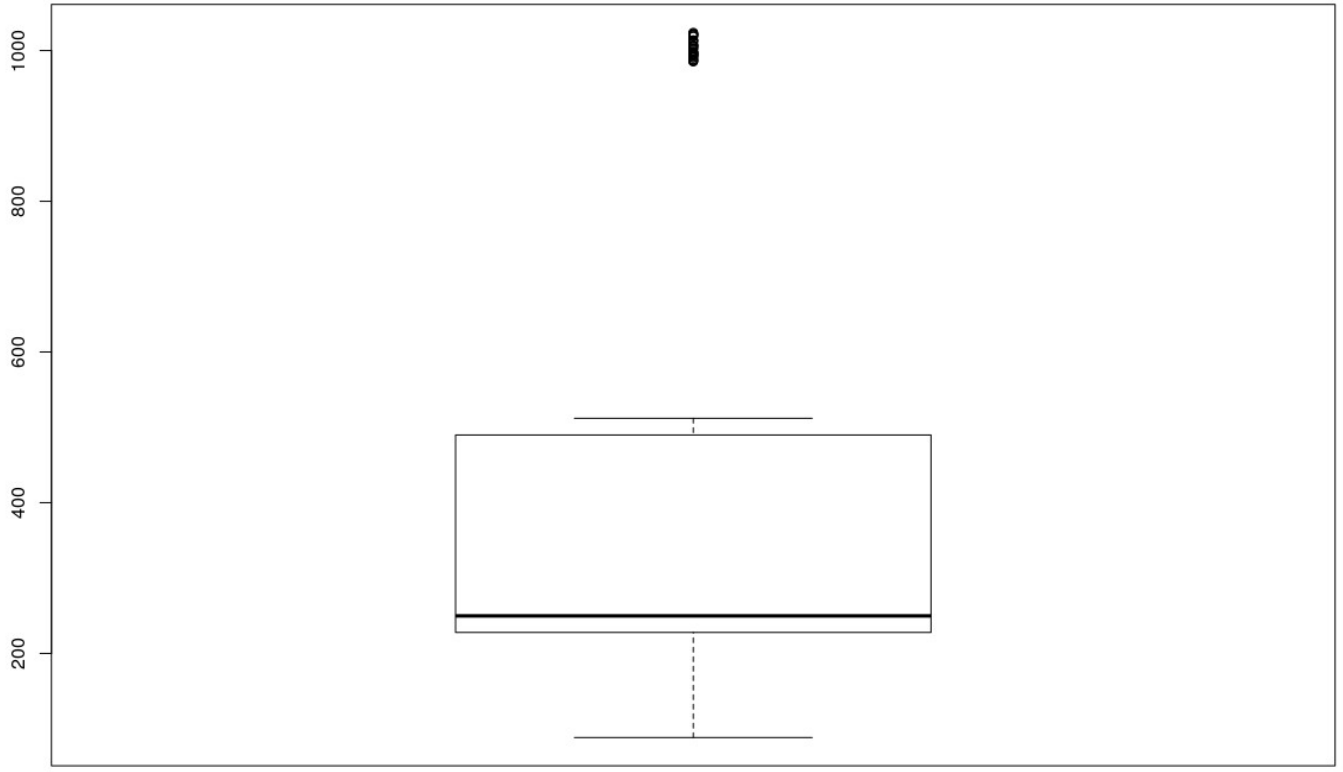
- kwartielen: variant op mediaan
  - wat is mediaan?
  - splits in 4 delen = kwartielen
  - splits in 10 delen = decielen
  - splits in 100 delen = percentielen

Interkwartielafstand

- IQR = Q3 - Q1
- dus...

  bereik wanneer je 25% grootste en kleinste waarden schrapt

Boxplot



Standaardafwijking

- spreiding = hoe dicht zitten de waarden ten opzichte van het centrum?
- neem centrum = gemiddelde
- kijk naar de verschillen tussen de waarden en het gemiddelde
- neem het gemiddelde van deze waarden
- (x-x.mean()).mean()
  - probleem?
  - oplossing?

Standaardafwijking

- mad = (x-x.mean()).abs().mean()
- var = ((x-x.mean() ** 2).mean()
- s = math.sqrt(var)

Bessel correctie

- normaal is formule dus:

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \mu \right)^2}$$

- maar in geval van steekproef:

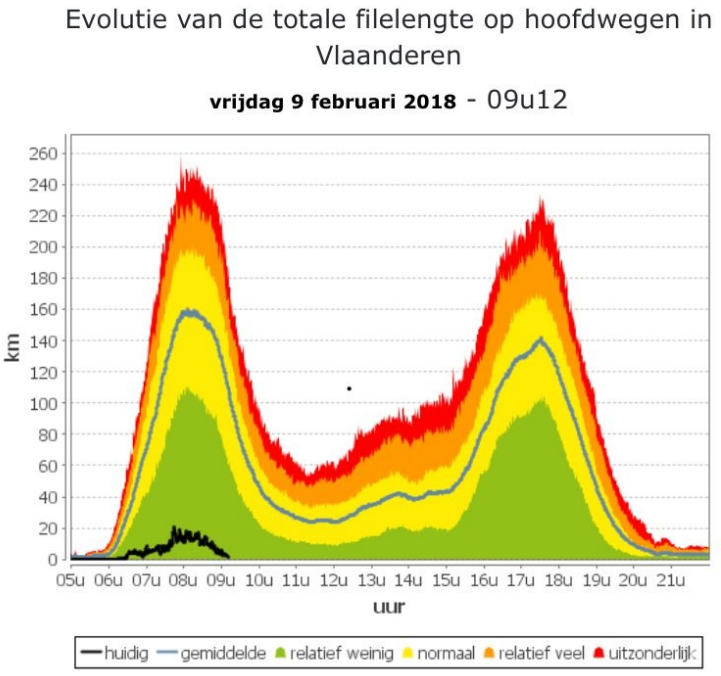
$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}$$

Uitschieters

Invloed van uitschieters

- uitschieter (outlier) = extreem hoge of lage waarde
- welke invloed heeft deze op
  - gemiddelde/standaardafwijking
  - mediaan/IQR
  - modus

Uitschieters vinden



- bepaal kwartielen Q1, Q2 en Q3
- bepaal IQR
- low = Q1 - 1.5 * IQR
- high = Q3 + 1.5 * IQR
- alles lager dan low of hoger dan high is een uitschieter
- "extreme uitschieter": vervang 1.5 door 3

Oefeningen

4

Oefeningen huiswerk

- Welke centrummaten kan je berekenen op de kolommen “gapmider”, “schrijfhand” en “informatica_belangrijk”?
- Welke spreidingsmaten kan je op deze kolommen berekenen?
- Welke spreidingsmaten combineer je best met welke centrummaten?

Oefeningen huiswerk

- Als je bij alle schoenmaten 5 optelt, wat gebeurt er dan met:
  - het gemiddelde
  - de mediaan
  - de standaardafwijking
  - de interkwartielafstand

Oefeningen huiswerk

- Als je alle schoenmaten deelt door 2, wat gebeurt er dan met:
  - het gemiddelde
  - de mediaan
  - de standaardafwijking
  - de interkwartielafstand

Oefeningen huiswerk

- Wat is de gemiddelde schoenmaat?
- Wat is de modus van bloedgroep? Wat betekent dit?
- Wat is de mediaan van respect? Wat betekent dit?
- Wat is de mediaan van informatica_belangrijk? Wat betekent dit?

Oefeningen huiswerk

- Maak een histogram van lengte. Welke vorm zie je hier?
- Wat is het gemiddelde? Kan je dit op de grafiek zien?
- Wat is de mediaan? Kan je dit op de grafiek zien?
- Wat is de standaardafwijking? Kan je dit op de grafiek zien?
- Wat is de interkwartielafstand? Kan je dit op de grafiek zien?
- Maak een boxplot van lengte. Welke centrummaat en spreidingsmaat zie je hier? Zijn er uitschieters?

Oefeningen huiswerk

- Bepaal de uitschieters voor lengte
- Hoeveel zijn er?
- Wat is de gemiddelde lengte wanneer je de uitschieters verwijdert?

Oefeningen huiswerk

- Maak een boxplot van schoenmaat en lengte naast elkaar. Wat zie je?
- Zet beide om naar z-scores en maak de plot weer. Zou er een verband zijn tussen de twee? Hint: zie volgende les

Oefeningen

- Zie Canvas
  - centrummaten
  - spreidingsmaten
  - marsbewoners