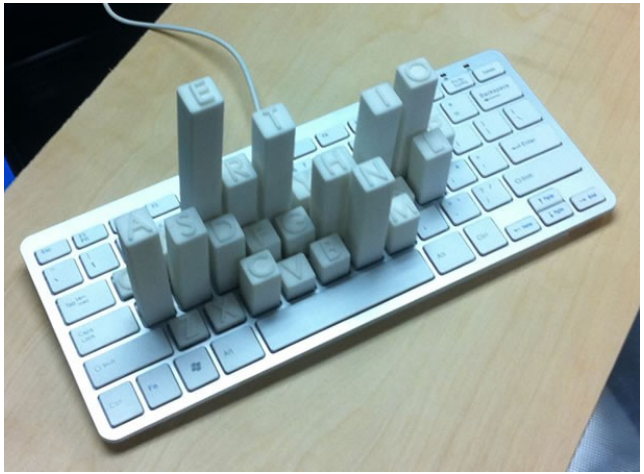
Data-Science 1

frequenties



Karel de Grote Hogeschool

07/03/2023

Klassikaal

- enquête
  - lees het bestand van de enquête in
  - maak de kolomnamen korter (zoek op)
  - zet alle datatypes juist
  - welke meetniveaus hebben de kolommen?

Inhoud

- voorbeeld
- absolute frequenties
- klassen
- relatieve frequenties
- cumulatieve frequenties
- cumulatieve percentages
- grafieken

Voorbeeld

De ruwe data

- bedrijf met 857 werknemers
- werknemers hebben een laptop
- we noteren per laptop
  - CPU generatie (Sandy Bridge, Ivy Bridge, Haswell, Broadwell, Skylake, Kabylake)
  - CPU type (i3, i5, i7)
  - RAM geheugen (in GB)
  - geformatteerde HD ruimte (in GB)
  - merk
- welk meetniveau? discreet of (quasi) continu?

De ruwe data

- resultaat (laptops.csv)

cpuGeneration	cpuType	RAM	diskspace	brand
Kabylake	i7	4	232,5	Toshiba
Kabylake	i5	2	992,5	Acer
Haswell	i7	16	495,6	Dell
Skylake	i7	4	217,2	Toshiba
Broadwell	i5	4	245,8	Acer
...	...	...	...	...

Absolute frequenties

2

Absolute frequenties

- tel hoeveel iedere waarde voorkomt
- voorbeeld: cpuType

cpuType	frequentie
i3	213
i5	556
i7	84

Absolute frequenties

- stel dat er NA-waarden zijn

cpuType	frequentie
i3	213
i5	556
i7	84
NA	4

Klassen

Probleem

- vanaf welk meetniveau kan je absolute frequenties bepalen?
- wat gebeurt er als de variabele (quasi) continu is?
  - voorbeeld: diskspace

    232.5 992.5 495.6 217.2 245.8 502.6 224.0 250.5 98.2
    484.4 482.9 246.2 485.8 221.5 484.5 501.7 ...
  - oplossing?

Opsplitsen in klassen

- groepeer waarden in "klassen"
  - dit zijn intervallen
  - voorbeeld: diskspace

    (0,100] (100,200] (200,300] (300,400] (400,500] (500,600] (600,700] (700,800] (800,900] (900,1000] (1000,1100]
  - iedere klasse heeft een klassebreedte en een klassemidden
- vervang iedere waarde door zijn klasse

  => wat is het resulterende meetniveau?

Hoeveel klassen?

- meer klassen: oorspronkelijk probleem komt weer
- minder klassen: resultaat is vager
- vuistregel: tussen 5 en 20 klassen

Hoeveel klassen?

- Sturges: $\lceil \log_2(n)+1 \rceil$
- Scott:
  - klassenbreedte = $\frac{3.5 \cdot \sigma}{\sqrt[3]{n}}$
  - aantal klassen = $\lceil (max(X)-min(X))/b \rceil$
- excel: $\lceil \sqrt{n} \rceil$

Relatieve frequenties

4

Relatieve frequenties

- absolute frequenties zijn moeilijk te interpreteren (je moet weten hoeveel elementen er in totaal zijn)
- dus: deel de frequenties door het aantal = relatieve frequenties
- in procent: doe *100

Relatieve frequenties

- voorbeeld: brand

brand	rel. freq.	in percenten
Acer	0,1517	15,2%
Apple	0,0758	7,6%
Asus	0,0758	7,6%
Dell	0,1879	18,8%
HP	0,2275	22,8%
Lenovo	0,0910	9,1%
Medion	0,0373	3,7%
Toshiba	0,1517	15,2%
NA	0,0012	0,1%

Cumulatieve frequenties

Cumulatieve frequenties

- telt hoeveel een bepaalde waarde of minder voorkomt
- voorbeeld: hoeveel laptops hebben een Haswell processor of minder?
- vanaf welk meetniveau?
- meestal worden de NA-waarden hier weggelaten

Cumulatieve frequenties

- voorbeeld: cpuGen

cpuGeneration	cum. freq.
Sandy Bridge	63
Ivy Bridge	170
Haswell	336
Broadwell	554
Skylake	709
Kabylake	852

Cumulatieve percentages

Cumulatieve percentages

- zelfde redenering als bij relatieve frequenties: percentages zijn duidelijker
- dus: deel door aantal en vermenigvuldig met 100

Cumulatieve percentages

- voorbeeld: cpuGen

cpuGeneration	cum. percentage
Sandy Bridge	7,4%
Ivy Bridge	20,0%
Haswell	39,4%
Broadwell	65,0%
Skylake	83,2%
Kabylake	100,0%

Percentielscores

- stel dat je resultaten van een test hebt:

0	1	2	3	4	5	6	7	8	9	10
1.1	10.3	11.4	33.6	55.4	65.1	84.2	98.1	100.0	100.0	100.0

- moeilijke of gemakkelijke test?
- hoeveel procent is gebuisd?
- oplossing: gebruik cumulatieve percentages als punten

  = punten ten opzichte van de groep
- nadeel: er is steeds 50% geslaagd, ongeacht kennis

Grafieken

Verschillende mogelijkheden

- standaard
  - taartdiagram
  - staafdiagram
  - histogram
- andere
  - spider plots
  - word clouds

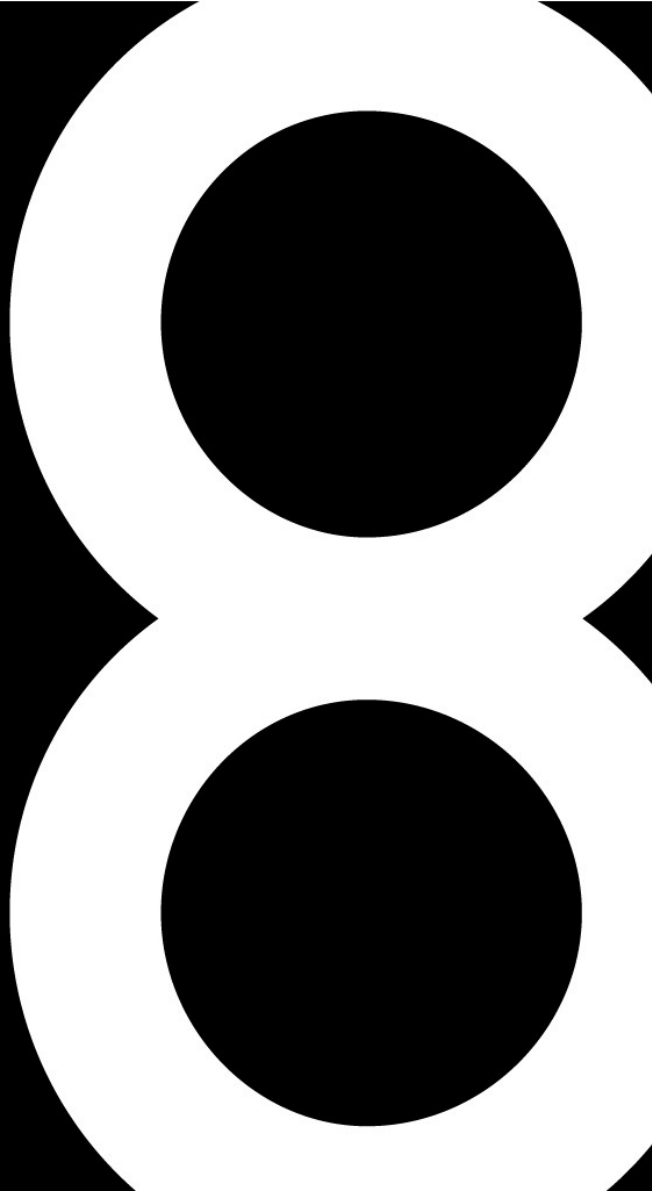
Grafieken in het algemeen

- altijd assen benoemen!
- titel geven
- eenheden vermelden in de assen!

Grafieken

- zie Jupyter notebook

Oefeningen



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 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?

Oefeningen

- zie Canvas
  - frequentieverdelingen
  - zonneopbrengst
  - webserver
  - beeldverwerking