Data Advanced

Data Representatie

- IT

Lector:
Heidi Tans

1. Inleiding

De dataset “slaagcijfers 2016-2017.xlsx” bevat een aantal gegevens van 435 studenten:

- Naam student
- Geslacht student: M – V
- Vooropleiding Algemeen: ASO – TSO – BSO
- Vooropleiding IT: veel – matig – geen
- Bis-student: ja - neen
- Score op 5 OLOD’s (afgerond op geheel)
- Studiepunten per OLOD
- Aantal uren besteed OLOD1 (lessenperiode + examenperiode)
- Aantal uren besteedd OLOD2 (lessenperiode + examenperiode)
- Gemiddeld aantal uren besteed aan studies (alle OLOD’s) op weekbasis
- Geslaagd: ja – neen
- Procent
- Aantal OLOD’s tweede zit

2. Gegevens verzamelen

Gegevens zijn het vertrekpunt van statistisch onderzoek.

Een aantal vragen dringen zich op:

Wat zijn gegevens?

Waar halen we gegevens?

Hoe 'betrouwbaar' zijn gegevens?

2.1. Wat zijn gegevens?

Gegevens zijn feiten en cijfers die verzameld, geanalyseerd en samengevat worden voor presentatie en interpretatie.

Er zijn verschillende soorten van gegevens. We onderscheiden

kwalitatieve gegevens (ordinaal – nominaal)

kwantitatieve gegevens (discreet – continu)

Afhankelijk van het type gegeven worden er andere analyses uitgevoerd.

Kwalitatieve gegevens

Kwalitatieve gegevens zijn labels of namen die gebruikt worden om eigenschappen van elementen aan te geven. Deze gegevens zijn niet-numeriek van aard en hebben een beperkt aantal uitkomstencategorieën (De categorieën kunnen wel met een getal aangegeven worden, maar dit is enkel een code die geen verdere betekenis heeft).

Deze kwalitatieve gegevens kunnen verder onderverdeeld worden in

nominale gegevens (niet – geordende categorische systemen)

ordinale gegevens (geordende categorische systemen)

Voorbeeld (slaagcijfer 2016-2017.xlsx)

Geslacht

Vooropleiding IT

Kwantitatieve gegevens

Kwantitatieve gegevens geven een hoeveelheid of grootte aan. Zulke gegevens worden verkregen door tellen, meten, … Kwantitatieve gegevens zijn altijd numeriek.

Deze kwantitatieve gegevens kunnen verder onderverdeeld worden in discrete gegevens en continue gegevens.

discrete gegevens

Men spreekt van discrete gegevens als de observaties van die aard zijn dat zij worden uitgedrukt door getallen die niet willekeurig dicht bij elkaar kunnen liggen. In de praktijk heeft men meestal te maken met gehele getallen (+ beperkt aantal verschillende uitkomsten).

continue gegevens

Als de mogelijke numerieke waarden in een zeker bereik willekeurig dicht bij elkaar kunnen liggen, hebben we continue gegevens. In principe kunnen de gegevens elke numerieke waarde aannemen.

Voor de statistische verwerking van continue gegevens zal men deze data dikwijls in groepjes samenvatten.

Voorbeeld (slaagcijfer 2016-2017.xlsx)

Studiepunten per OLOD

Gemiddeld aantal uren gestudeerd op weekbasis

Aantal OLOD's tweede zit

Afhankelijk van het soort gegevens (kwantitatief / kwalitatief) worden andere statistische analyses toegepast.

2.2. Waar halen we deze gegevens (gegevensbronnen)?

We kunnen gegevens verkrijgen bij organisaties die gespecialiseerd zijn in het verzamelen en beheren van gegevens. Maar soms zijn de gegevens die nodig zijn voor een bepaalde toepassing al aanwezig (vb databases met gegevens over de werknemers van een bedrijf, gegevens over koopgedrag van klanten, gegevens met betrekking tot studenten, …).

Voorbeeld

http://statbel.fgov.be/nl/statistieken/opendata/

http://archive.ics.uci.edu/ml/datasets.html

Wanneer geen gegevens beschikbaar zijn via instanties of bestaande bronnen, kunnen ze ook door statistisch onderzoek bekomen worden.

Bij dit statistisch onderzoek is o.a. een onderscheid te maken tussen: experimenteren & waarnemen

Experimenteren

In experimenteel onderzoek worden de variabelen van belang vastgesteld. Daarna wordt bestudeerd hoe bepaalde factoren de variabelen in het onderzoek beïnvloeden.

Voorbeeld

Invloed van Study Buddy op resultaat van student

Variabele van belang = score OLOD

Factor die variabele beïnvloedt = study buddy

Om gegevens te verkrijgen over de invloed van een study buddy op de score op een OLOD worden een aantal proefpersonen geselecteerd. De gegevens omtrent de socre van deze proefpersonen worden genoteerd voor en na het gebruik maken van een study buddy (bvb eerste zit en tweede zit).

Waarnemen

Bij waarnemend onderzoek wordt niet geprobeerd om variabelen te manipuleren.

De enquête is de meest gebruikte soort van waarnemend onderzoek.

Bij een enquête worden een aantal onderzoeksvragen gesteld en voorgelegd aan de proefpersonen.

Een opmerking bij waarnemend gegevens verzamelen is dat deze manier vaak veel tijd en geld kost.

Voorbeeld

Bevraging EVA – OLOD

2.3. Hoe betrouwbaar zijn deze gegevens (meetfouten)?

Statistici dienen zich altijd bewust te zijn van mogelijke foute gegevens in statistisch onderzoek. Het gebruik van foute gegevens kan in bepaalde gevallen erger zijn dan helemaal geen gegevens te hebben en geen conclusies te trekken. Speciale procedures kunnen gebruikt worden om de interne consistentie van de gegevens te controleren.

Blindelings gebruiken van gegevens die voorhanden zijn, kan gevaarlijk zijn.

Waar het fout kan gaan…

SLECHT VOORBEELD VAN STEEKPROEF

https://www.demorgen.be/buitenland/een-op-de-vijf-vlaamse-moslims-heeft-begrip-voor-is-en-haar-manier-van-actievoeren-b8338104/

http://www.demorgen.be/opinie/de-islam-enquete-is-een-voorbeeld-van-hoe-het-niet-moet-ba5df0e6/Ravgg/

Foutieve gegevens

22 jaar werkervaring van een 20 – jarige medewerkster van het bedrijf.

Bekijk observatie 78 in de gegeven dataset.

SLECHTE ENQUETE

Stel dat we een groep proefpersonen een enquête zouden sturen met daarin de vraag "Vult U graag enquêtes in?" en we kijken naar de teruggestuurde formulieren, dan zouden we waarschijnlijk de conclusie kunnen trekken dat de overgrote meerderheid graag enquêtes invult!!
Zo past de steekproef zichzelf aan. 't Is eigenlijk net zo dom als een steekproef per e-mail houden en de vraag "Heeft U een computer?" stellen. De conclusie zal ongetwijfeld zijn dat 100% van de mensen een computer heeft.

Dit zijn natuurlijk wel heel voor de hand liggende voorbeelden, maar soms is het fout zijn van een steekproef slechter te zien. Zo wilden twee leerlingen op de middelbare school onderzoeken hoeveel er gerookt werd onder scholieren. Ze gingen aan het begin van de pauze bij de buitendeur staan en vroegen de eerste 50 leerlingen die naar buiten kwamen: Rook je?" Helaas komen natuurlijk in de pauze de rokers het eerst naar buiten.....

FOUTE CONCLUSIES

http://peilingpraktijken.nl/weblog/2015/01/542/ ➔ 3 jongeren / 3 vakanties ➔ 100% kans om slechte vakantie te hebben.

Aantal geslaagden over de jaren heen (maar totaal aantal wordt weggelaten):

We vergelijken het slaagcijfer van de opleiding Slavistiek en de opleiding IT.
In beide opleidingen zijn er evenveel geslaagden in het eerste jaar: slaagcijfer is gelijkaardig maar…. In de opleiding slavistiek zitten maar 10% van het aantal studenten in de opleiding IT!!!

3. Gegevens voorstellen

De meeste statistische informatie in kranten, tijdschriften, rapporten en andere publicaties bestaat uit gegevens die zo zijn samengevat en gepresenteerd dat de lezer ze eenvoudig kan begrijpen. Dergelijke samenvattingen van gegevens, zowel in tabellen, grafieken als getallen, vallen onder wat men noemt beschrijvende statistiek.

In deze paragraaf wordt dieper ingegaan op het presenteren van gegevens in tabellen en de grafische voorstelling ervan.

3.1. Frequentietabel kwalitatieve gegevens

Dataset: slaagcijfers 20162017_verkort

n: totaal aantal gegevens

Definitie

De absolute frequentie f_i van waarneming x_i is het aantal keer dat deze waarneming voorkomt in de gegevens. De som van de absolute frequentie is gelijk aan het totaal aantal observaties: $\sum_{i=1}^{k} f_i = n$

Vooropleiding Algemeen	f_i
ASO	1
TSO	15
BSO	4
	20

In bovenstaande dataset komen 15 studenten uit TSO. Dit is zeer veel (in het achterhoofd onthoud je natuurlijk dat dit 15 studenten van de 20 zijn).

Moesten we niet 20 maar 200 studenten geobserveerd hebben, dan is dit weinig.

Vandaar dat het begrip relatieve frequentie in het leven geroepen is.

Definitie

De relatieve frequentie φ_i wordt als volgt berekend: $\varphi_i = \frac{f_i}{n}$ met $\sum_{i=1}^{k} \varphi_i = 1$.

Voorbeeld (slaagcijfers 2016-2017_verkort.xlsx)

- **Vooropleiding Algemeen**
- **Tweede zit**

Vooropleiding Algemeen	f_i	φ_i
ASO	1	0.05
TSO	15	0.75
BSO	4	0.2
	20	**1**

Tweede zit	f_i	φ_i
zwaar	5	0.25
matig	4	0.2
licht	5	0.25
geen	6	0.3
	20	**1**

3.2. Frequentietabel kwantitatieve discrete gegevens

De verschillende gegevens rangschikken we in **stijgende** volgorde zodat we een gerangschikte frequentietabel bekomen.

Definitie

Veronderstel dat $x_1, \ldots, x_k$ de k verschillende waarnemingen zijn (gerangschikt van klein naar groot); $f_1, \ldots, f_k$ de corresponderende frequenties en n het totaal aantal gegevens.

Cumulatieve absolute frequentie: $$cf_m = \sum_{j=1}^{m} f_j$$

Cumulatieve relatieve frequentie: $$c\varphi_m = \sum_{j=1}^{m} \varphi_j = \sum_{j=1}^{m} \frac{f_j}{n}$$

Voorbeeld (slaagcijfer 2016-2017_verkort.xlsx): score OLOD1

Score OLOD1	f_i	φ_i	cf_i	$c\varphi_i$
2	2	0.1	2	0.1
3	0	0	2	0.1
4	0	0	2	0.1
5	5	0.25	7	0.35
6	0	0	7	0.35
7	4	0.2	11	0.55
8	0	0	11	0.55
9	1	0.05	12	0.6
10	1	0.05	13	0.65
11	1	0.05	14	0.7
12	0	0	14	0.7
13	0	0	14	0.7
14	1	0.05	15	0.75
15	2	0.1	17	0.85
16	2	0.1	19	0.95
17	1	0.05	20	1
	20	**1**		

3.3. Frequentietabel kwantitatieve continue gegevens

Wanneer we een groot aantal continue gegevens hebben, heeft het begrip frequentietabel, zoals gehanteerd in vorige secties weinig zin (ook bij zeer veel verschillende discrete gegevens is dit zo). Immers alle of haast alle gegevens zullen verschillend zijn zodat alle frequenties f_i bijna gelijk zijn aan 1.

Maw. de frequentietabel zal totaal geen vereenvoudigde weergave zijn.

Voorbeeld (slaagcijfers 2016-2017_verkort.xlsx : Score OLOD1 voor afronding)

Zonder klassen

Score OLOD1 (voor afronding)	f_i	φ_i	cf_i	$c\varphi_i$
1.89	1	0.05	1	0.05
2.06	1	0.05	2	0.1
	…	…	…	…
	…	…	…	…
16.86	1	0.05	19	0.95
16.14	1	0.05	20	1
	20	**1**		

Met klassen

Continue waarnemingen kunnen wel overzichtelijk voorgesteld worden door gebruik te maken van **frequentietabellen met klassenindeling**.

Klassen zijn zelf geconstrueerde intervallen zodat elke waarneming tot één en slechts één klasse behoort. Aangezien er in de praktijk geen strikte afspraken bestaan voor de keuze van deze klassen, volgen we hier één bepaalde methode, nl. deze waarbij het aantal klassen niet minder is dan 5 en niet meer is dan 15, waarbij de klassen allen even breed zijn en waarbij de intervallen links gesloten en rechts open zijn.

Werkwijze voor het opstellen van een frequentietabel met klassenindeling:

- Zoek het grootste en kleinste waarnemingsgetal (in ons voorbeeld is dit 1.89 en 16.86)
- Bereken het verschil tussen deze extreme waarden (16.86 – 1.89 = 14.97).
- Deel dit verschil door 5 en door 15 en kies een klassenbreedte b tussen deze uitkomsten ($0.998 \leq$ klassenbreedte ≤ 2.994 ; kies b = 2). ➔ 8 klasses

Score OLOD1 (voor afronding)	f_i	φ_i	cf_i	$c\varphi_i$
[1.89 ; 3.89 [	2	0.1	2	0.1
[3.89 ; 5.89 [	5	0.25	7	0.35
	…	…	…	…
	…	…	…	…
[13.89 ; 15.89 [	2	0.1	17	0.85
[15.89 ; 17.89 [	3	0.15	20	1
	20	**1**		

Definities:

- **Klassengrenzen**: zijn de kleinste en grootste grens van een klasse, in die zin dat de onderste grens in die klasse wel en de bovenste grens niet kan bereikt worden.
  Zo bevat de klasse [15.89; 17.89[alle getallen die groter of gelijk zijn aan 15.89 en strikt kleiner dan 17.89

- **Klassenbreedte**: is het verschil tussen de grootste en kleinste klassengrens van een klasse.

- **Klassenmidden**: is de helft van de som van de grootste en kleinste klassengrens van een klasse. Zo is het klassenmidden van de klasse [15.89 ; 17.89[gelijk aan 16.89

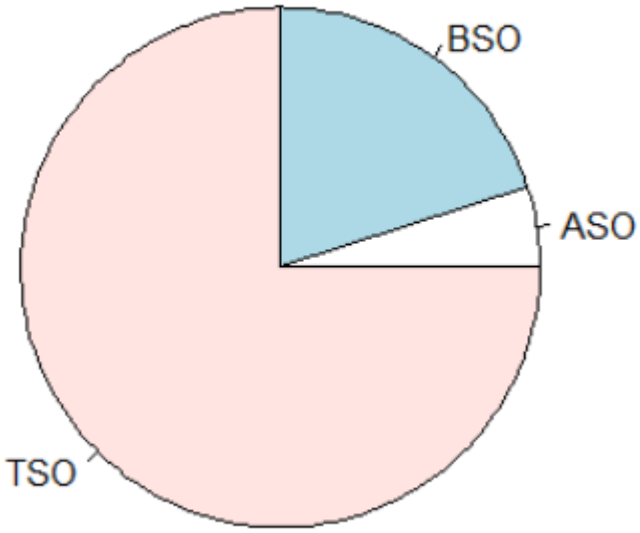
- **Klassenfrequentie**: de klassenfrequentie van de i – de klasse is het aantal waarnemingen dat tot deze klasse behoort.

3.4. Grafische voorstelling

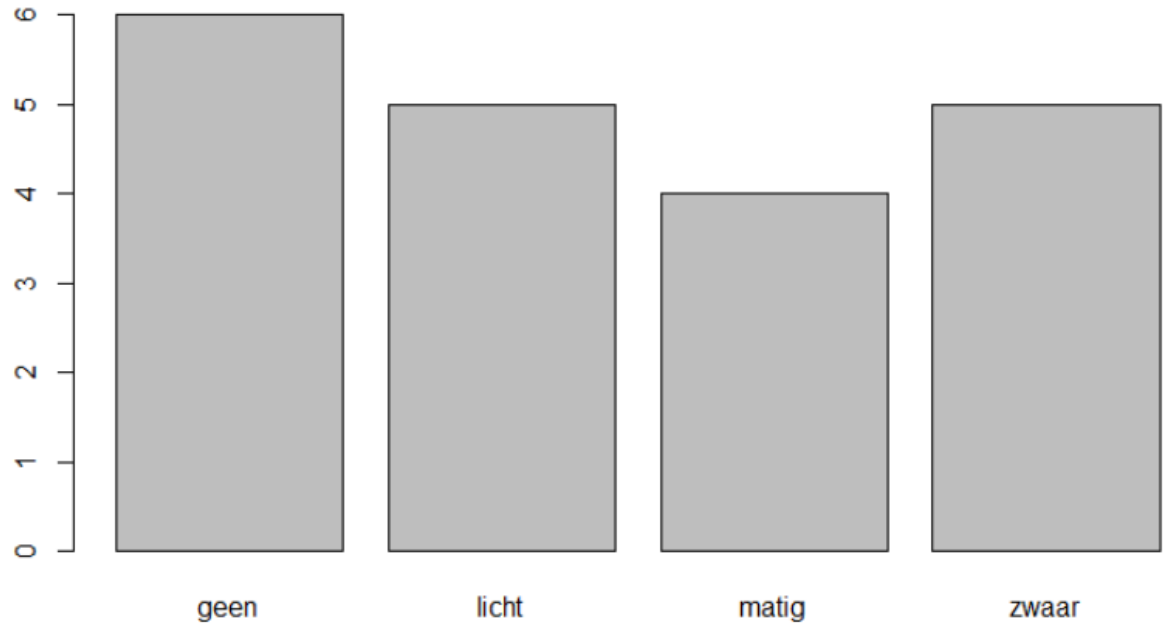
Grafieken zijn een efficiënte manier om gegevens voor te stellen. Veelgebruikte types zijn het

- cirkeldiagram
- staafdiagram (kwalitatief en kwantitatief discreet)
- histogram (continu)

Voorbeeld (slaagcijfer 20162017_verkort): Cirkeldiagram (vooropleiding)

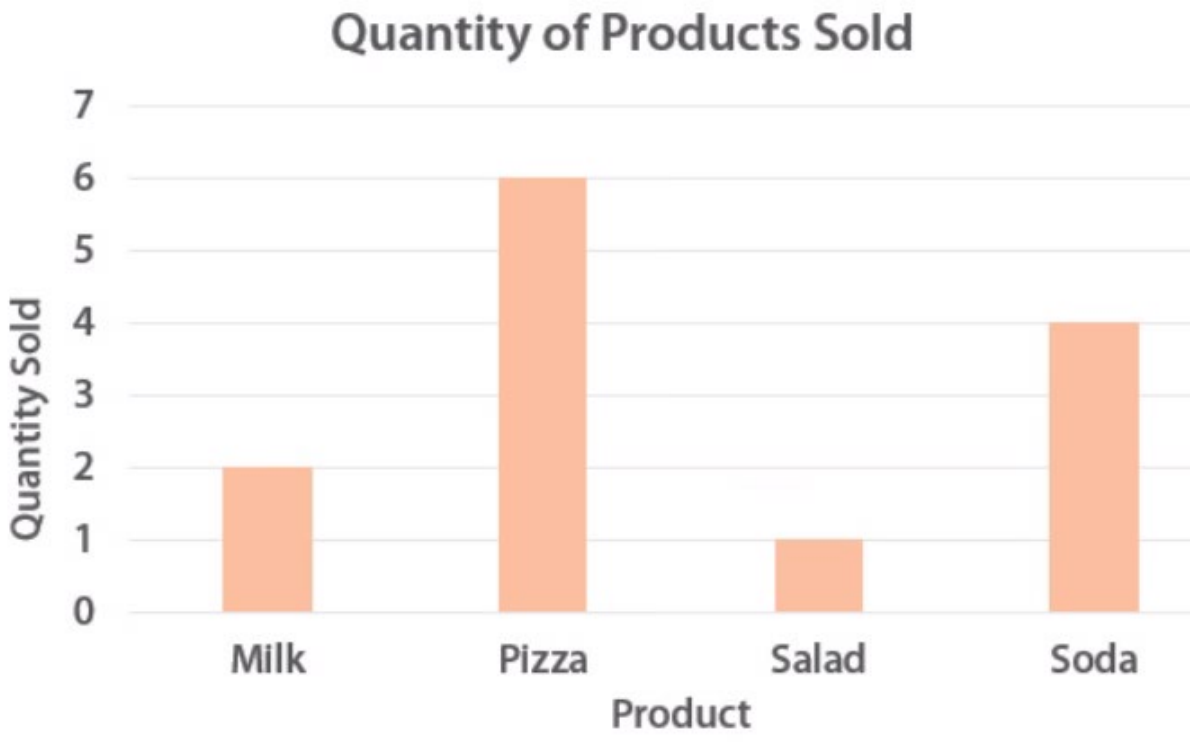


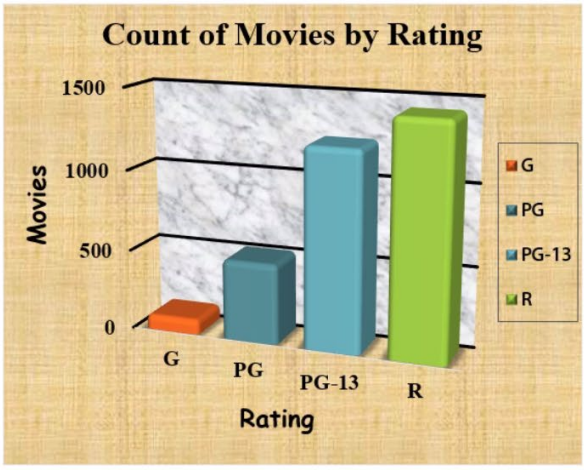
Voorbeeld (slaagcijfer 20162017_verkort): Staafdiagram (tweede zit)

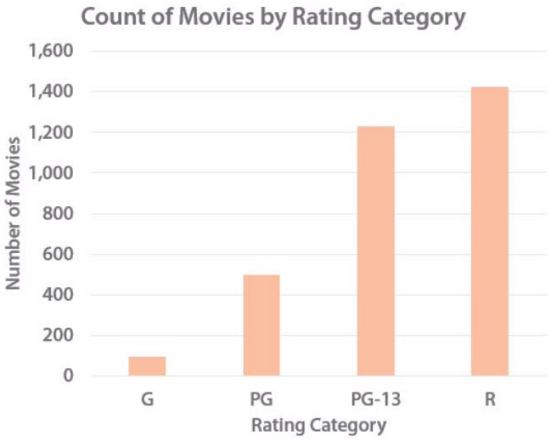


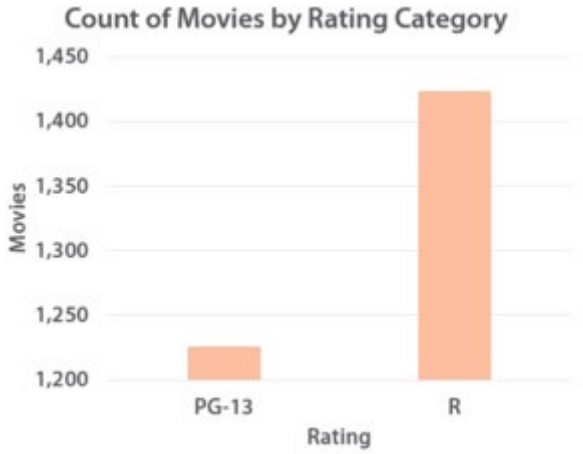
http://peilingpraktijken.nl/weblog/2015/01/cirkeldiagram-of-staafdiagram/

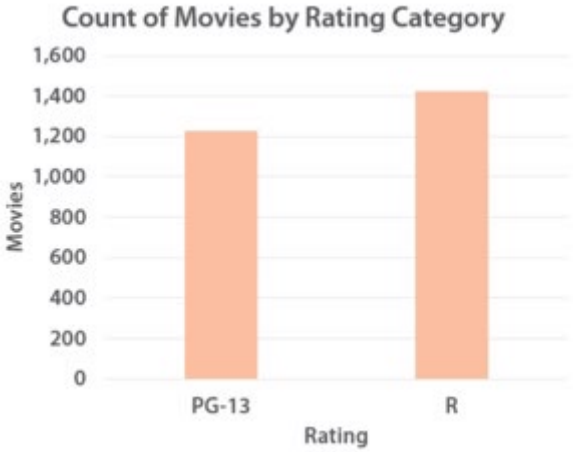
ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1




Count of Movies by Rating
1500
1000
500
0
Movies
G
PG
PG-13
R
Rating
G
PG
PG-13
R


Count of Movies by Rating Category
1,600
1,400
1,200
1,000
800
600
400
200
0
Number of Movies
G
PG
PG-13
R
Rating Category

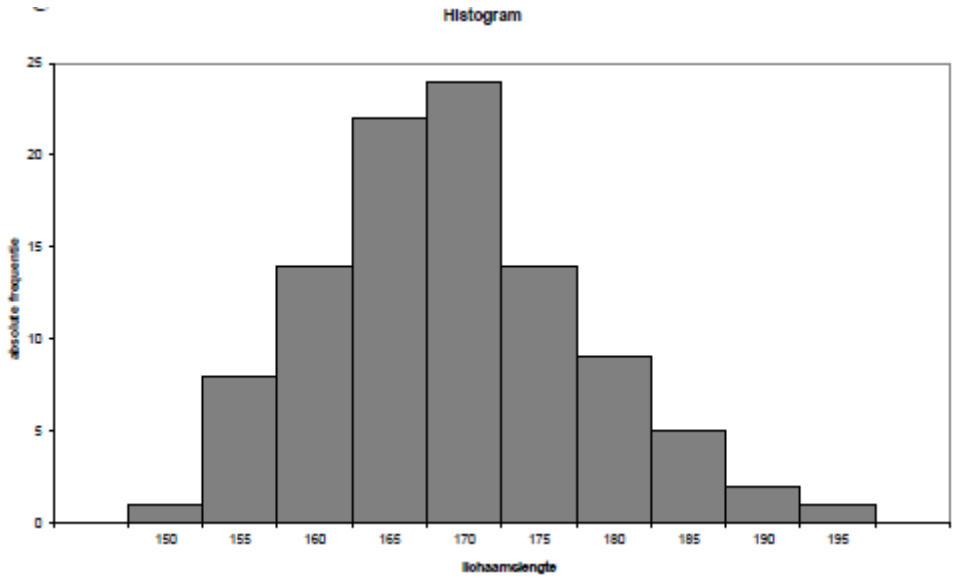

Count of Movies by Rating Category
1,450
1,400
1,350
1,300
1,250
1,200
Movies
PG-13
R
Rating


Count of Movies by Rating Category
1,600
1,400
1,200
1,000
800
600
400
200
0
Movies
PG-13
R
Rating

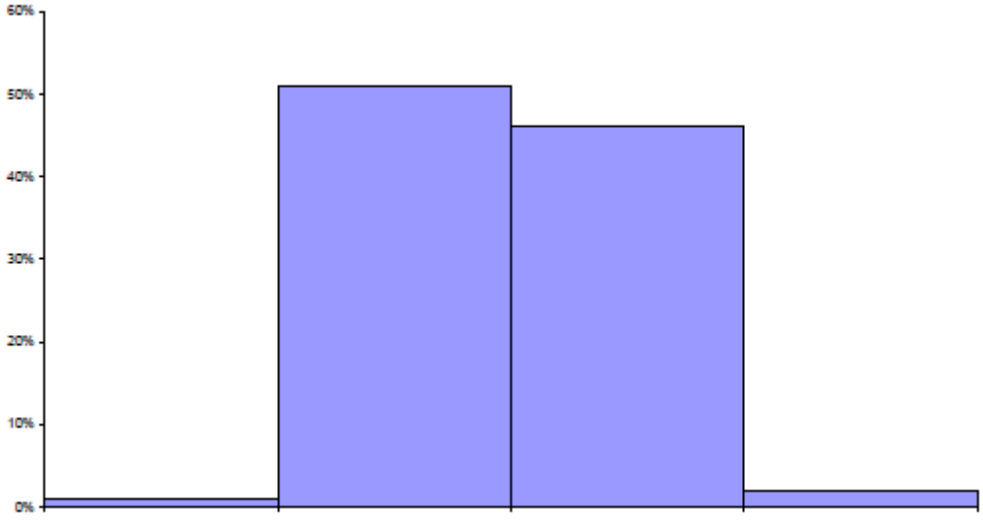
Opmerking:

Wanneer de klassenbreedte bij het maken van een frequentietabel anders gekozen wordt, dan ziet het histogram er uiteraard anders uit.

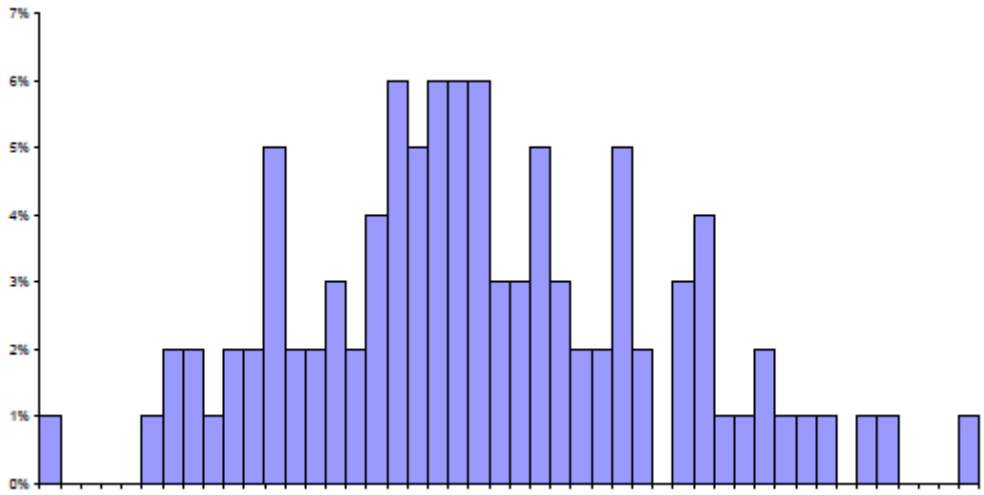
Histogram van de lichaamslengte van 100 2de jaars studenten TIN



Als het aantal intervallen te klein is, dan gaat het algemene patroon van de verdeling verloren.

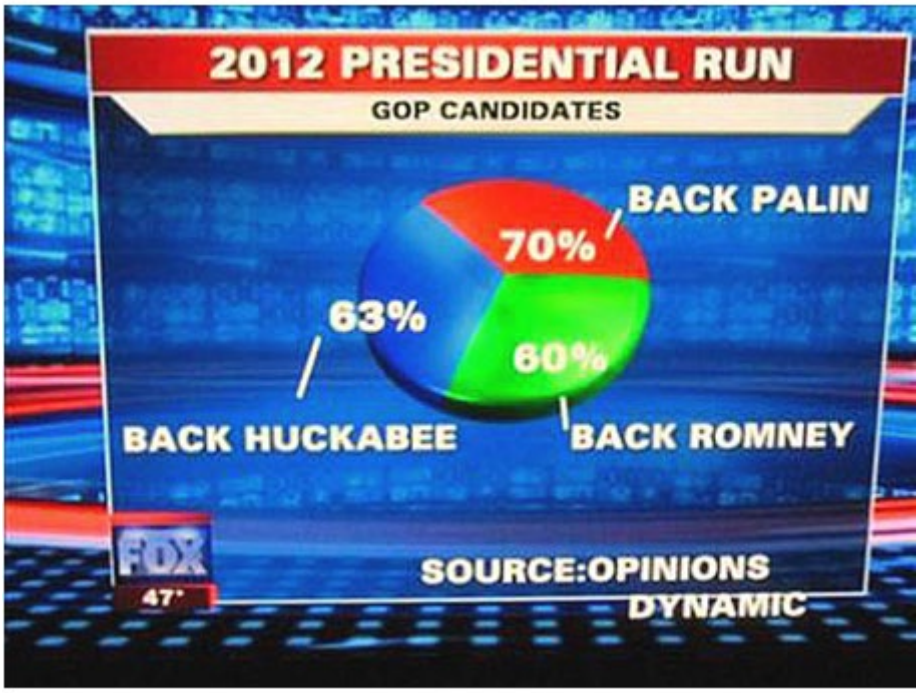


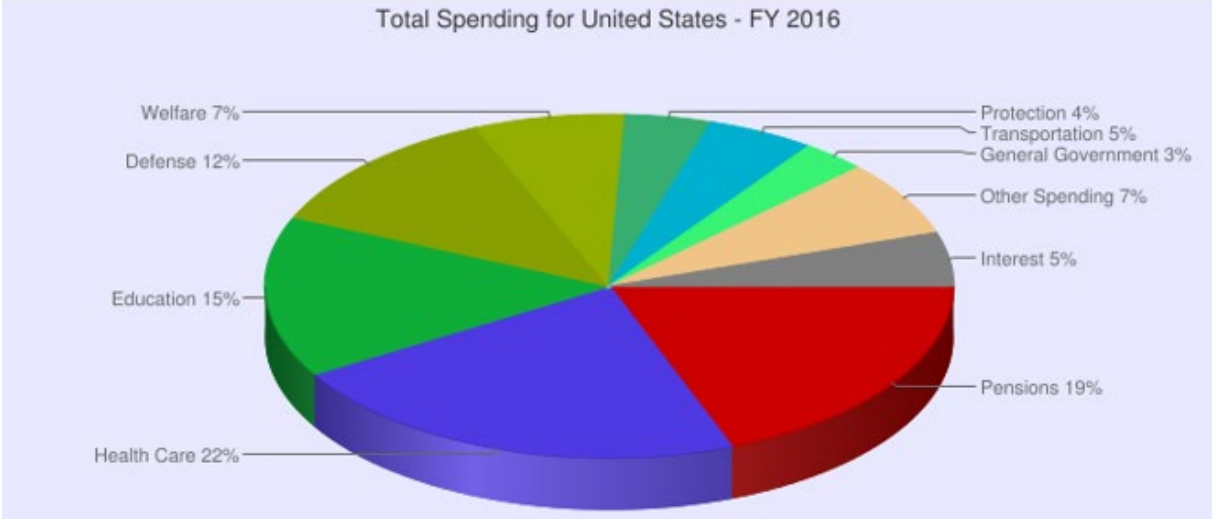
Als het aantal intervallen te groot is, dan komen alle (irrelevante) details (veroorzaakt door het toeval dat in elke steekproef aanwezig is) op de voorgrond. Het histogram wordt minder overzichtelijk.

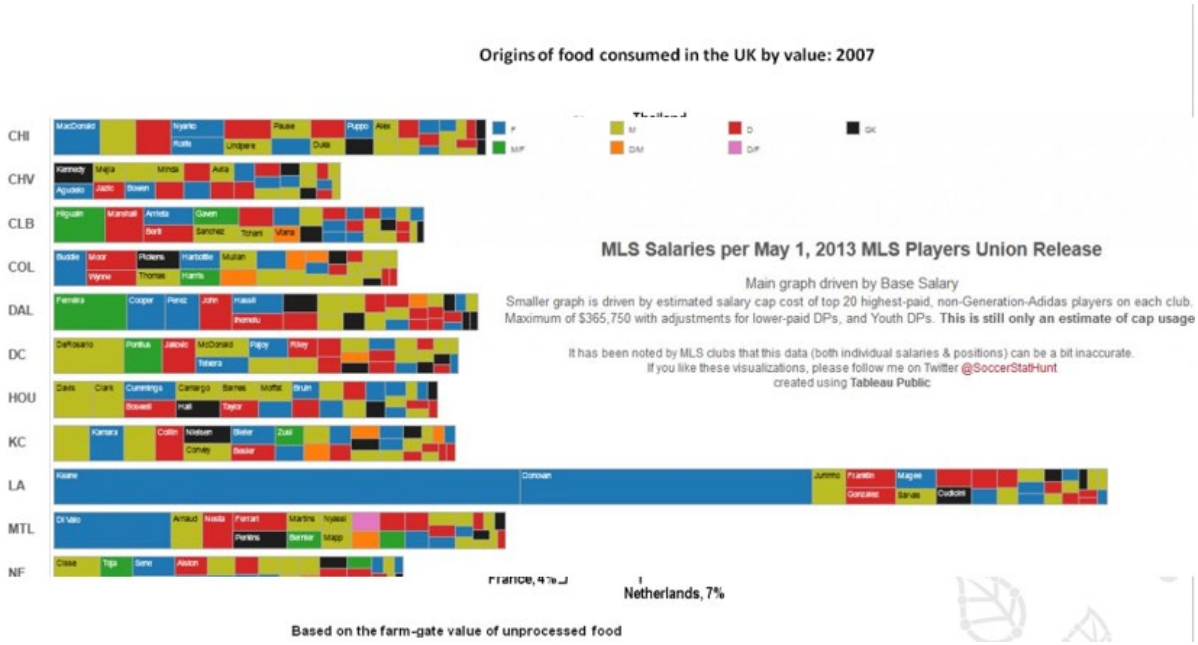


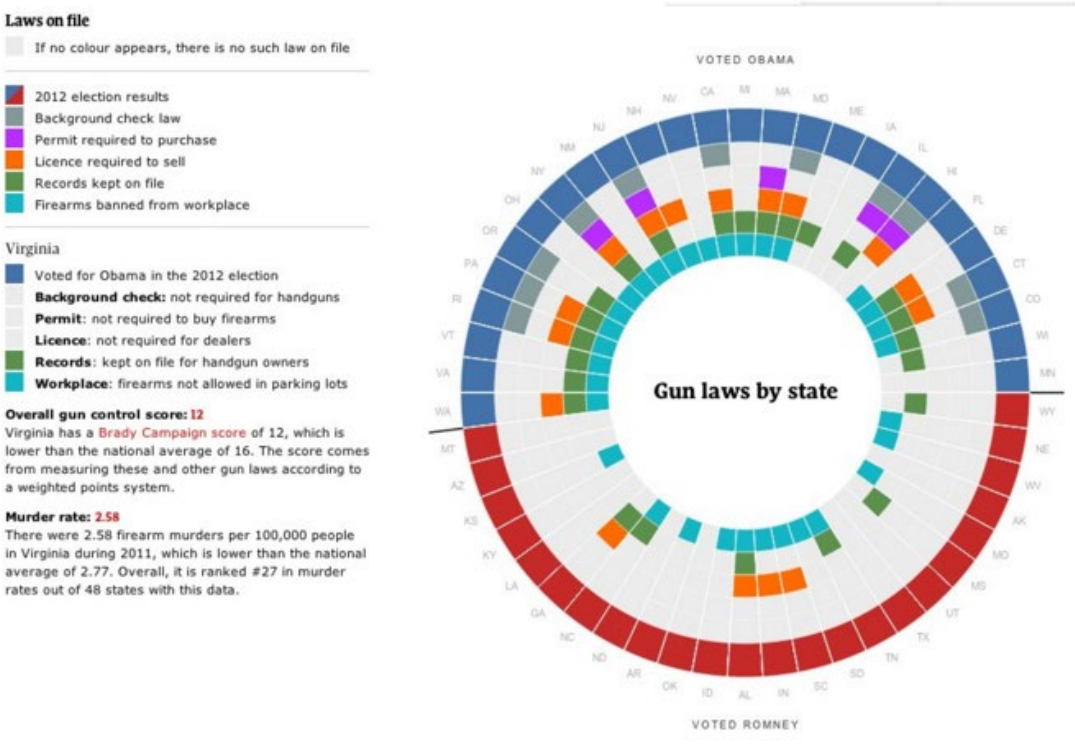
Het is aan te raden de invloed van verschillende klassenbreedten op het uitzicht van het histogram na te gaan.

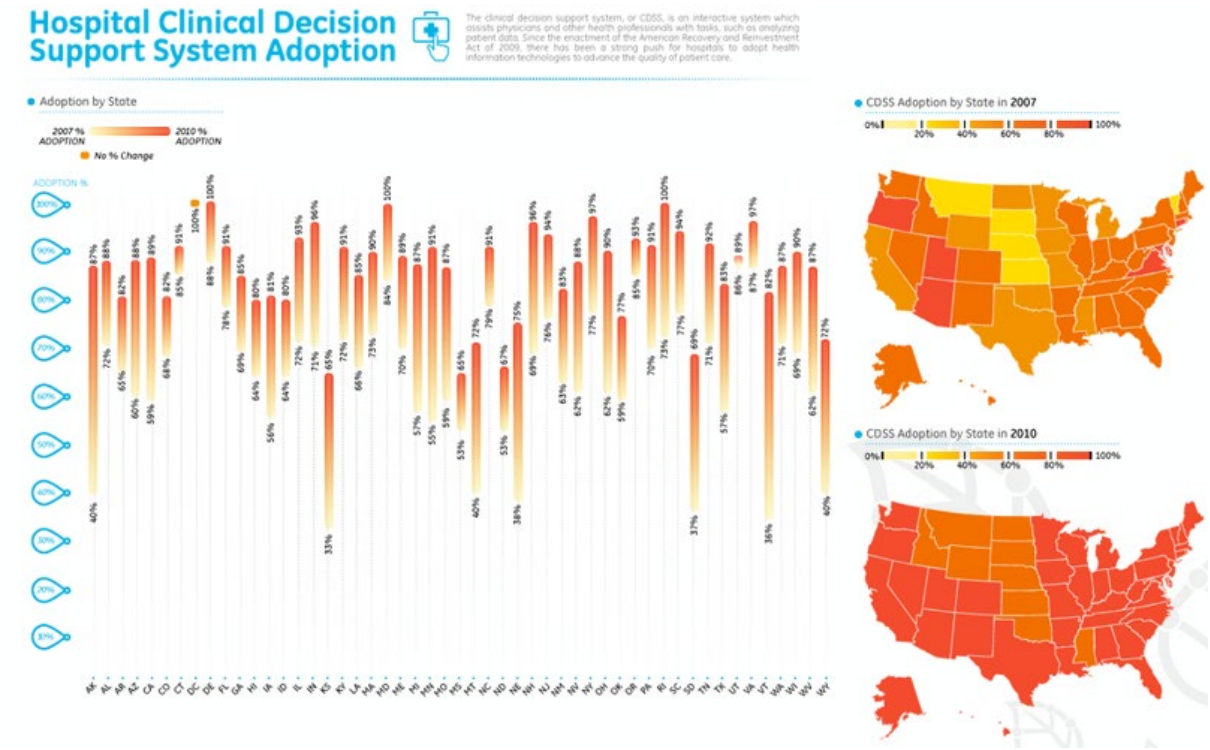
Waar het fout kan gaan...

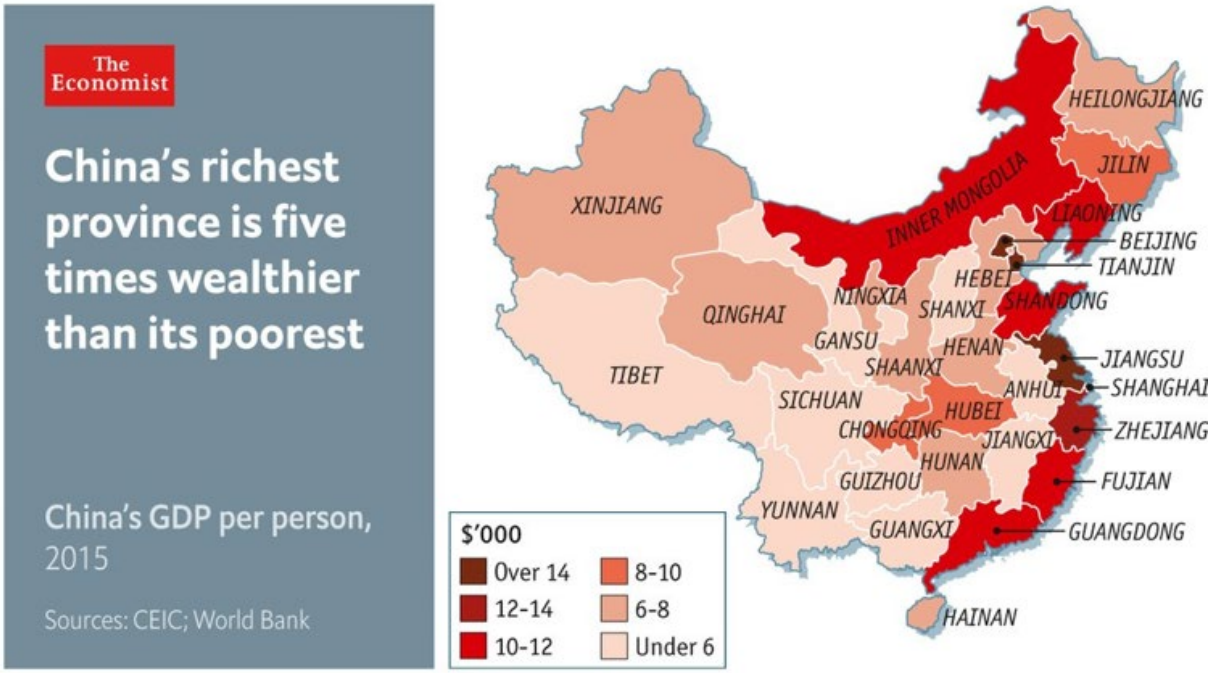




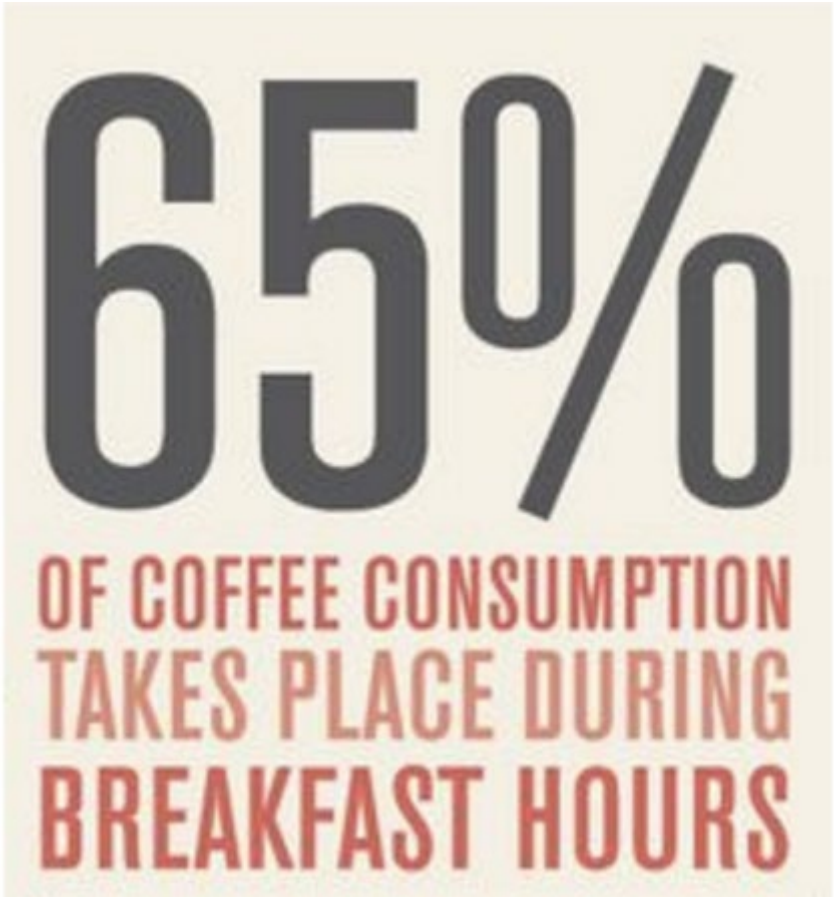

Origins of food consumed in the UK by value: 2007
CHI
CHV
CLB
COL
DAL
DC
HOU
KC
LA
MTL
NE
MLS Salaries per May 1, 2013 MLS Players Union Release
Main graph driven by Base Salary
Smaller graph is driven by estimated salary cap cost of top 20 highest-paid, non-Generation-Adidas players on each club.
Maximum of $365,750 with adjustments for lower-paid DPs, and Youth DPs. This is still only an estimate of cap usage
It has been noted by MLS clubs that this data (both individual salaries & positions) can be a bit inaccurate.
If you like these visualizations, please follow me on Twitter @SoccerStatHunt
created using Tableau Public
Netherlands, 7%
Based on the farm-gate value of unprocessed food


Laws on file
If no colour appears, there is no such law on file
2012 election results
Background check law
Permit required to purchase
Licence required to sell
Records kept on file
Firearms banned from workplace
Virginia
Voted for Obama in the 2012 election
Background check: not required for handguns
Permit: not required to buy firearms
Licence: not required for dealers
Records: kept on file for handgun owners
Workplace: firearms not allowed in parking lots
Overall gun control score: 12
Virginia has a Brady Campaign score of 12, which is lower than the national average of 16. The score comes from measuring these and other gun laws according to a weighted points system.
Murder rate: 2.58
There were 2.58 firearm murders per 100,000 people in Virginia during 2011, which is lower than the national average of 2.77. Overall, it is ranked #27 in murder rates out of 48 states with this data.
VOTED OBAMA
Gun laws by state
VOTED ROMNEY


Hospital Clinical Decision Support System Adoption
The clinical decision support system, or CDSS, is an interactive system which assists physicians and other health professionals with tasks, such as analyzing patient data. Since the enactment of the American Recovery and Reinvestment Act of 2009, there has been a strong push for hospitals to adopt health information technologies to advance the quality of patient care.
Adoption by State
2007 % ADOPTION
2010 % ADOPTION
No % Change
CDSS Adoption by State in 2007
CDSS Adoption by State in 2010
0%
20%
40%
60%
80%
100%


The Economist
China's richest province is five times wealthier than its poorest
China's GDP per person, 2015
Sources: CEIC; World Bank
$'000
Over 14
12-14
10-12
8-10
6-8
Under 6
HEILONGJIANG
JILIN
LIAONING
INNER MONGOLIA
XINJIANG
BEIJING
TIANJIN
HEBEI
NINGXIA
SHANXI
SHANDONG
QINGHAI
GANSU
HENAN
JIANGSU
SHAANXI
TIBET
ANHUI
SHANGHAI
SICHUAN
HUBEI
CHONGQING
JIANGXI
ZHEJIANG
HUNAN
GUIZHOU
FUJIAN
YUNNAN
GUANGXI
GUANGDONG
HAINAN

Large view (Source: *The Economist*)



4. Gegevens samenvatten

Om **kwantitatieve gegevens** verder te analyseren, associeert men er zogenaamde centrum- en spreidingsmaten aan, die de locatie en spreiding van de gegevens weergeven.

4.1. Kengetallen voor locatie

Definitie: rekenkundig gemiddelde

De karakteristiek plaats of locatie van een gegevensset vatten we samen door één enkel getal: het **rekenkundig gemiddelde**. Dit gemiddelde geeft een soort "midden" aan van de gegevensset.

Het gemiddelde van n observaties $x_1, \cdots, x_n$ is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Opmerking:

Indien de gegevens gegeven zijn in een frequentietabel (n gegevens waarvan k verschillend)

Dan wordt het gemiddelde $\bar{x}$ als volgt berekend: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$

Indien de gegevens gegeven zijn in een frequentietabel met klassenindeling, dan gebruiken we voor de berekening van $\bar{x}$ niet x_i maar de klassenmiddens m_i met de bijbehorende klassenfrequenties. Dit geeft een benadering voor $\bar{x}$: $\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} m_i f_i$

Een groot **nadeel** aan het gemiddelde als spreidingsmaat is dat het zeer gevoelig is voor extremen. Dit wordt duidelijk in volgend voorbeeld:

A: 5 10 5 3 7 $\quad \bar{x}_A = \frac{5+10+5+3+7}{5} = 6$

B: 5 10 1000 3 7 $\quad \bar{x}_B = \frac{5+10+1000+3+7}{5} = 205$

Een mogelijke oplossing is de meest extreme uitkomsten te verwijderen voordat men het gemiddelde bepaalt. Men spreekt dan van een **trimmed mean**.

B: 5 10 3 7 $\quad \bar{x}_B = \frac{5+10+3+7}{4} = 6.25$

Als men over een voldoend aantal waarnemingen beschikt, kan men de hoogste en laagste 5% van de waarnemingen weglaten.

Er zijn situaties waarin de formule van het rekenkundige gemiddelde niet zomaar mag toegepast worden omdat de uitkomsten $x_1, \dots, x_n$ een verschillend gewicht hebben. Bijvoorbeeld bij het berekenen van je eindpercentage op je rapport, krijgt niet elk OLOD hetzelfde gewicht. Dit gewicht is afhankelijk van de studiepunten.

Definitie: gewogen rekenkundig gemiddelde

Het gewogen (rekenkundig) gemiddelde van een reeks numerieke gegevens $x_1, \dots, x_n$ met gewichten $w_1, \dots, w_n$ is de som van alle waarnemingen vermenigvuldigd met het juiste gewicht, gedeeld door de som van de gewichten: $\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots w_n x_n}{w_1 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Spreidingsmaten voor locatie die niet gevoelig zijn aan uitschieters / extreme waarden, zijn oa de mediaan en de modus.

Definitie

De mediaan van een rij van n gegevens (**gerangschikt van klein naar groot**) is

De middelste waarde als n oneven is

Het rekenkundig gemiddelde van de middelste twee gegevens als n even is

Voorbeeld

5 gegevens van klein naar groot:	32	42	46	46	54	➔ mediaan = 46		
8 gegevens van klein naar groot:	2	3	4	7	8	10	10	15
					➔ mediaan = (7+8)/2 = 7.5			

Opmerking

Bij gegroepeerde gegevens (frequentietabel met klassenindeling) nemen we als benadering voor de mediaan het klassenmidden van de klasse waarin het middelste (of de middelste twee) getal(len) zich bevind(en). Vallen de middelste twee getallen toevallig in naburige klassen dan neemt men het rekenkundig gemiddelde van de twee klassenmiddens van deze klassen

De zogenaamde kwartielen vormen een uitbreiding van de mediaan. De mediaan verdeelt een geordende rij gegevens in 2 gelijke delen. Wanneer we onze geordende rij gegevens in 4 willen delen, kan dit met behulp van kwartielen:

Definitie: kwartielen

In de statistiek is een kwartiel één van de drie waarden die een geordende set data in vier gelijke delen opdeelt. Elk deel is dus een kwart van de dataset. Men spreekt van eerste, tweede en derde kwartiel en noteert deze als Q1, Q2 en Q3

Definitie: modus

De modus is de observatie die het vaakst voorkomt.

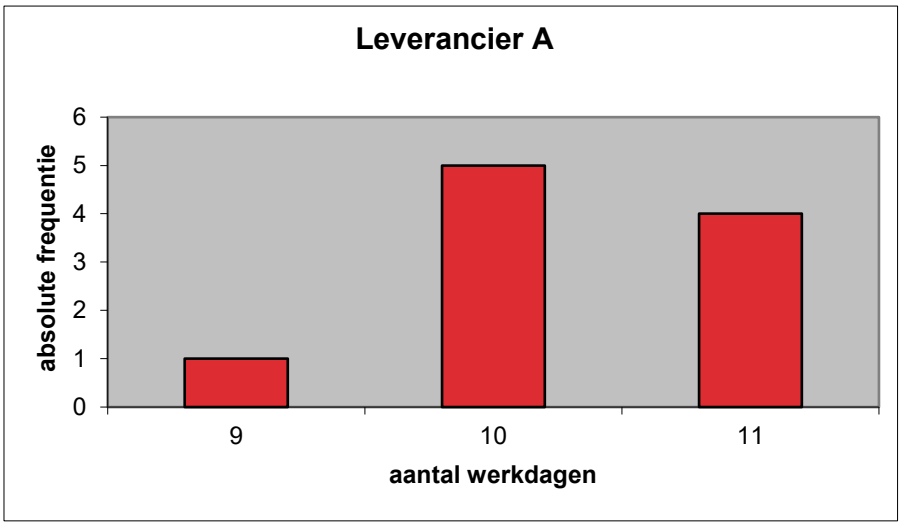
4.2. Kengetallen voor spreiding

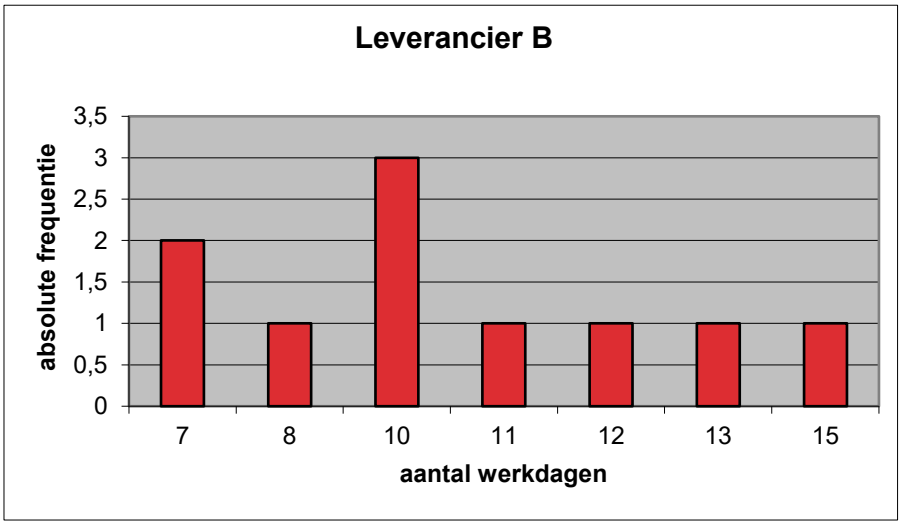
Voorbeeld

De afdeling inkoop van een grote fabrikant plaatst regelmatig bestellingen bij twee verschillende leveranciers. Volgende tabel geeft het aantal dagen weer dat nodig was om 10 bestellingen te leveren.

A:	11	10	9	10	11	11	10	11	10	10
B:	8	10	13	7	10	11	10	7	15	12

Wanneer we het gemiddelde aantal werkdagen berekenen zowel voor leverancier A als B, bekomen we 10.3. Toch kunnen we niet stellen dat beide leveranciers even betrouwbaar zijn wat betreft het op tijd leveren. Aan volgende staafdiagrammen zien we dat de spreiding rond het gemiddelde van deze twee leveranciers verschillend is (de spreiding bij leverancier B is groter dan bij A).





Als we willen weten hoe sterk de individuele gegevens $x_1, \cdots, x_n$ afwijken van hun gemiddelde maken we gebruik van de variantie.

Definitie

- De **variantie** s^2 van de gegevens $x_1, \cdots, x_n$ met gemiddelde $\bar{x}$ is gelijk aan

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

- Als $x_1, \cdots, x_k$ verschillende gegevens zijn met respectievelijke frequenties $f_1, \cdots, f_k$ (bij n gegevens), dan is de variantie s^2 gelijk aan $$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2$$

- Als de gegevens gegeven zijn in een frequentietabel met klassenindeling, dan gebruiken we de klassenmiddens m_i ipv. de gegevens x_i . Een benadering voor de variantie is dan

$$s^2 \approx \frac{1}{n-1}\sum_{i=1}^{k} f_i(m_i - \bar{x})^2$$

Definitie

De standaardafwijking s van de gegevens $x_1, \cdots, x_n$ is de positieve vierkantswortel uit de variantie

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

Analoog voor de andere definities van de variantie.

Definitie: spreidingsbreedte R (Range)

De **spreidingsbreedte** van n gegevens $x_1, \cdots, x_n$: R = grootste waarde – kleinste waarde

Opmerking

Net zoals het gemiddelde is de spreidingsbreedte uiterst gevoelig voor uitschieters.

Definitie

De interkwartielafstand van n gegevens $x_1, \cdots, x_n$ is gedefinieerd als

$$IKA = Q_3 - Q_1$$

De interkwartielafstand geeft de spreidingsbreedte weer van de middelste 50% van de gegevens en is dus niet gevoelig aan uitschieters.

4.3. Visuele voorstelling van locatie en spreiding

De boxplot is een eenvoudige grafische samenvatting van enkele belangrijke kengetallen van een gegevensset. Met een boxplot kan in één oogopslag informatie over locatie en spreiding van de verschillende gegevensverzamelingen vergeleken worden.

De boxplot wordt getekend mbv. de 5 volgende getallen (**vijf – getallenrésumé**)

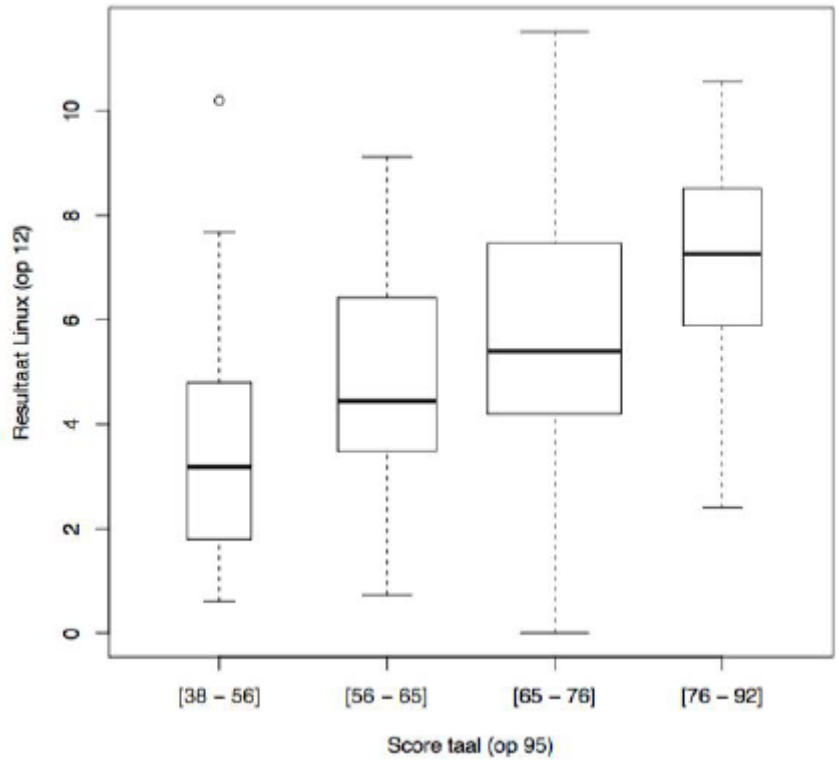
- het kleinste gegeven
- het eerste kwartiel Q_1
- de mediaan of tweede kwartiel Q_2
- het derde kwartiel Q_3
- het grootste gegeven

Tussen elk van de opeenvolgende getallen ligt telkens 25% van de gegevens.

Schematisch gaan we als volgt te werk:

Teken tegenover een schaal een rechthoek die begint bij het eerste kwartiel Q_1 en eindigt bij het derde kwartiel Q_3.
Teken een verticale lijn in de rechthoek op de plaats van de mediaan.
Teken links en rechts een horizontale lijn die uit de rechthoek komt en links gaat tot de kleinste en rechts gaat tot de grootste waarde.
Uitschieters kan je ook aflezen van de boxplot.

Voorbeeld



Een boxplot geeft ons heel wat informatie over gegevens:

- centrum: mediaan (eventueel gemiddelde)
- spreiding: de lengte van de box is de IKA
- Scheefheid: de positie van de mediaan ten opzichte van het eerste en het derde kwartiel geeft reeds aan of er asymmetrie is of niet. Bij symmetrische verdelingen (zie ook volgend deel) ligt de mediaan in het midden van de box, bij rechtsscheve verdelingen in de onderste helft en bij linkksscheve verdelingen in de bovenste helft van de box. Bovendien kunnen we naar de lengtes van de 'takken' kijken. Rechtse scheefheid geeft aanleiding tot boxplots waarbij de bovenste tak langer is dan de onderste. Bij linkse scheefheid is dit natuurlijk andersom
- Zwaarte van de staarten: Indien veel mogelijke uitschieters worden aangeduid, wijst dit vaak op een zwaarstaartige verdeling. De grenzen van de takken zijn immers op een zodanige manier bepaald dat, indien alle gegevens uit een normale verdeling komen, we mogen verwachten dat ongeveer 1% van de gegevens toch als mogelijke uitschieter wordt aangeduid. Vinden we er beduidend meer, dan wijst dit ofwel op de aanwezigheid van echte uitschieters, ofwel op het niet-normaal zijn van de gegevens (door scheefheid of zwaardere staarten).

Boxplots zijn ook zeer interessant om verschillende verzamelingen van gegevens met elkaar te vergelijken. Men kan onmiddellijk meerdere karakteristieken tegelijk bekijken.