

Statistiek

Hoofdstuk 1. Statistiek, gegevens en een kritische houding

1.1. De wetenschap statistiek

Statistiek de wetenschap van gegevens

- verzamelen
- evalueren (classificeren, samenvatten, organiseren, analyseren)
- interpreteren

1.2. 2 Soorten statistische toepassingen

Steekproef trekken gegevens selecteren uit een grotere verzameling waarvan we de kenmerken willen schatten

- Beschrijvende statistiek **beschrijven** van verzamelde gegevens
- Verklarende statistiek trekt **conclusies** over de gehele groep op basis van een deel (steekproef) van deze groep

1.3. 6 Basiselementen van de statistiek

Experimentele eenheden	de bestudeerde objecten
Populatie	de verzameling experimentele eenheden
Variabele	kenmerk of eigenschap van een individuele eenheid uit de populatie – Metten: getallen toekennen aan variabelen – Census van de populatie: resultaat als we een variabele meten voor iedere eenheid van een populatie
Steekproef	deelverzameling van de populatie
Statistische gevolgtrekking/conclusie	veralgemening vanuit de steekproef naar de populatie
Betrouwbaarheidsmaat	uitspraak over de onzekerheid van de statistische gevolgtrekking

H1: Statistiek, gegevens en statistisch denken 7

Basiselementen van de statistiek

Elementen van de **beschrijvende** statistiek

- Populatie en steekproef
- Onderzochte variabele(n)
- Numerieke synthesesmiddelen (grafieken, tabellen, maatstaven)
- Opsporen van gegevenspatronen

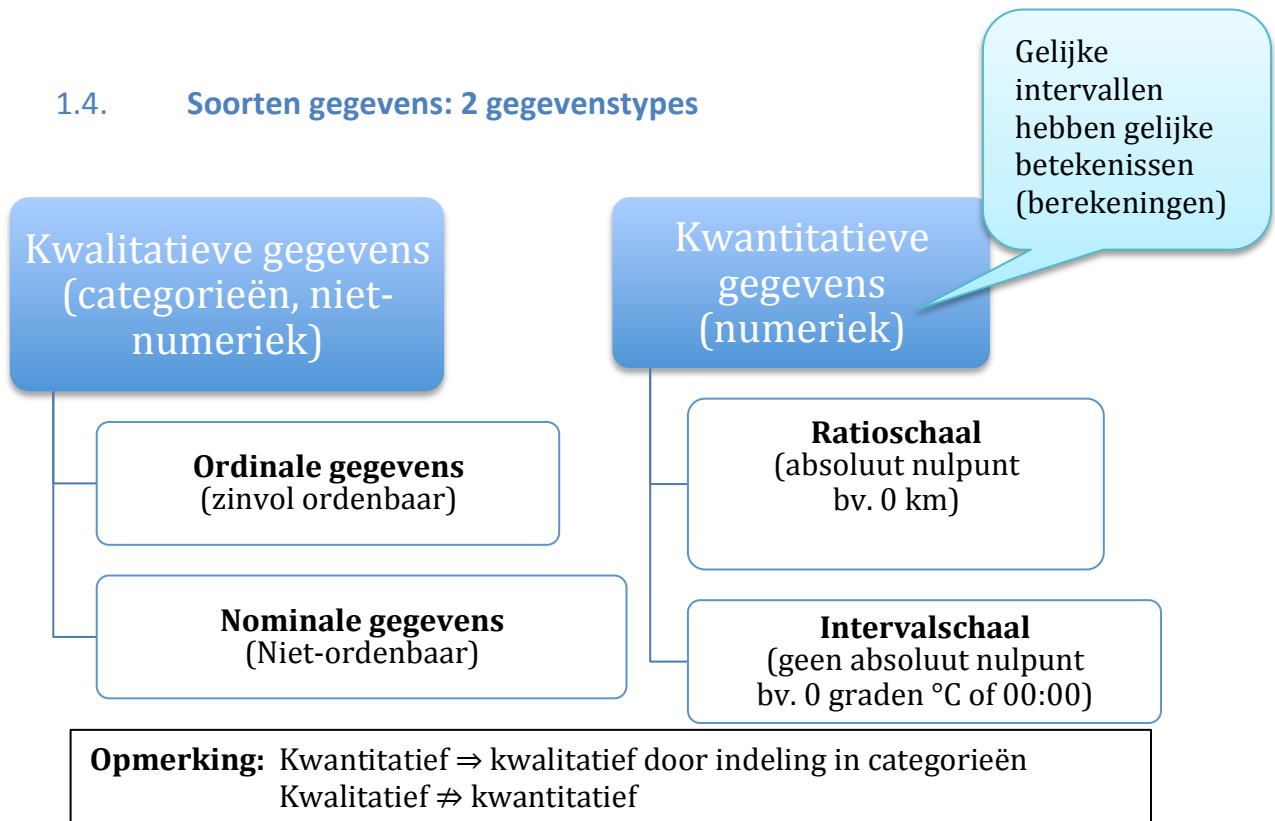
H1: Statistiek, gegevens en statistisch denken 8

Basiselementen van de statistiek

Elementen van de **verklarende** statistiek

- Populatie
- Onderzochte variabele(n)
- Steekproef uit de populatie
- Conclusie over de populatie, op basis van de steekproef
- Betrouwbaarheidsmaat voor de gevolgtrekking

1.4. Soorten gegevens: 2 gegevenstypes



1.5. Gegevens verzamelen

Gegevensbronnen: 4 methodes

1. **Publicaties**
= boek, tijdschrift, krant, website
→ primaire bron: maker is betrokkene, ooggetuige,...
→ secundaire bron: maker is niet direct betrokken, vaak andere tijd
2. **(Ontworpen/gecontroleerd) experiment**
= effecten van een behandeling onderzoeken
→ experimentele groep/ behandelingsgroep: krijgt behandeling
→ controlegroep: krijgt geen behandeling of krijgt placebo
3. **Enquête (survey)**
= gegevens via vragenlijsten, gesteld aan een steekproef van personen
4. **Waarnemend onderzoek/ observatie**
= onderzoeken van experimentele eenheden in hun natuurlijke omgeving, zonder interactie of behandeling!

Steekproeven trekken

- Nodig voor verklarende statistiek
- Steekproeven moeten *representatief* zijn voor gehele populatie!

$\Rightarrow$ representativiteit via **aselecte steekproef**
= Elke deelverzameling van dezelfde omvang uit de populatie heeft dezelfde kans om gekozen te worden

1.6. Blijf kritisch: fouten in gegevens

⇒ fouten doordat steekproef niet aselekt (niet representatief) is!

3 Soorten vertekeningen ("bias")	Vertekening door (zelf)selectie	Vertekening door non-respons	Meetfouten
Wat	een deelverzameling van de populatie is uitgesloten van het onderzoek	onderzoekers zijn niet in staat om gegevens te verkrijgen over alle eenheden in de steekproef	slordigheden in gegevens
Oorzaak	wie op een enquête reageert heeft vaak al uitgesproken meningen over het onderwerp	oorzaak zoeken, hoe verwerken we dit statisch?	- dubbelzinnige of suggestieve vragen - enquêteur - apparatuur - verkeerde omzetting van gegevens

Hoofdstuk 2. Het beschrijven van gegevensverzamelingen

2.1. Kwalitatieve gegevens beschrijven

→ niet-numeriek, ingedeeld in categorieën of klassen

– **Frequentie per klasse of klassefrequentie:** # gegevens in een klasse

– **Relatieve klassefrequentie:**

$$\frac{\text{Klassefrequentie}}{\text{totaal \# waarnemingen}}$$

– **Klassepercentage:**

relatieve frequentie x 100

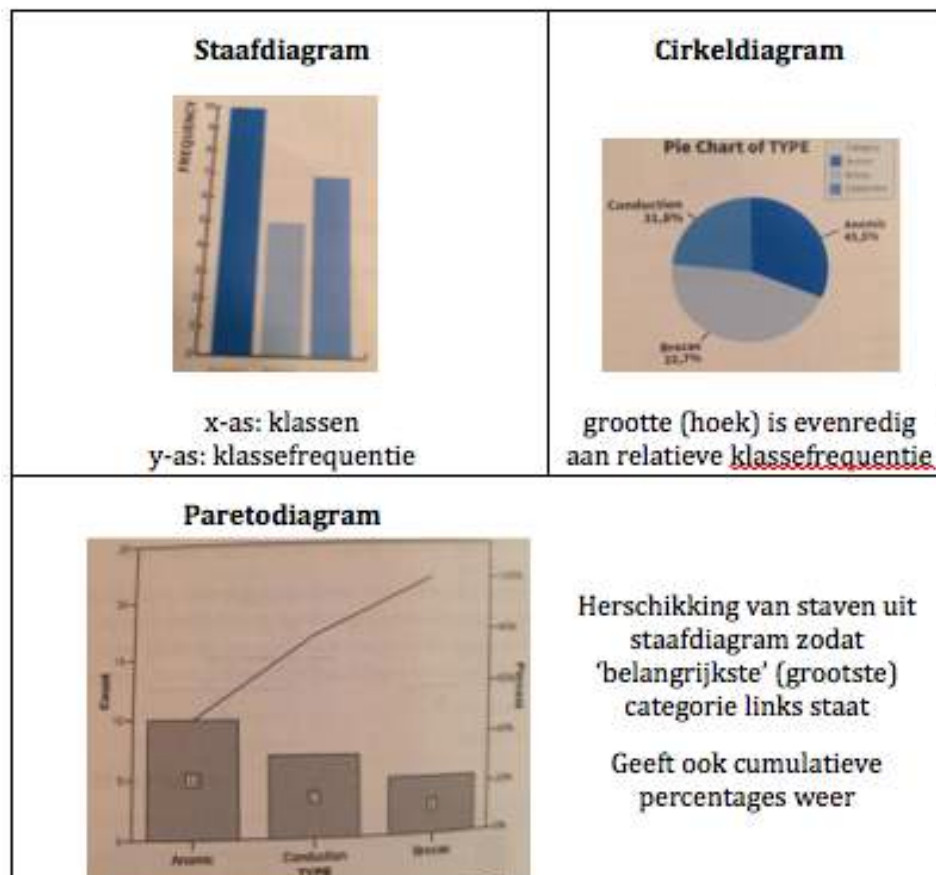
– **Cumulatieve percentage van a**

steekprofelementen met een waarde $\leq a$

DEGREE		Cumulative klassepercentage			
		Klassefrequentie	Klassepercentage	Valid Percent	Cumulative Percent
		Frequency	Percent	Valid Percent	Percent
Valid	Bachelors	6	24.0	24.0	24.0
	Law	3	12.0	12.0	36.0
	Masters	9	36.0	36.0	72.0
	None	4	16.0	16.0	88.0
	PhD	3	12.0	12.0	100.0
	Total	25	100.0	100.0	

→ Waarden in 1^e kolom 1st geordend van klein naar groot, anders interpretatie-probleem

Weergave



2.2. Kwantitatieve gegevens beschrijven met grafische methoden

Weergave

Staaftdiagram (kwal)
≠ histogram (kwant)

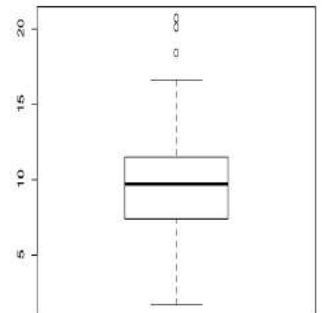
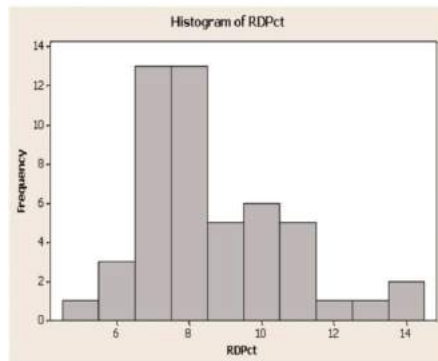
Frequentietabel

Histogram

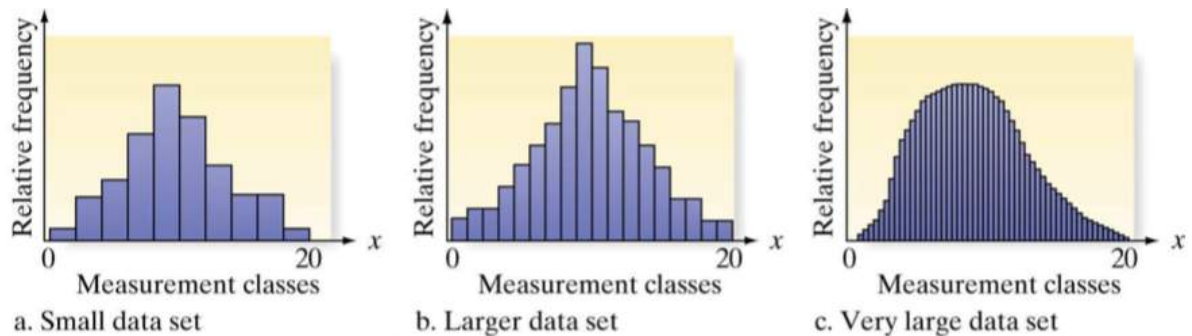
Boxplot

TABLE 2.3 Class Intervals, Frequencies, and Relative Frequencies for the 50 R&D Measurements

Class	Class Interval	Class Frequency	Class Relative Frequency
1	4.5–5.5	1	1/50 = .02
2	5.5–6.5	3	3/50 = .06
3	6.5–7.5	13	13/50 = .26
4	7.5–8.5	13	13/50 = .26
5	8.5–9.5	5	5/50 = .10
6	9.5–10.5	6	6/50 = .12
7	10.5–11.5	5	5/50 = .10
8	11.5–12.5	1	1/50 = .02
9	12.5–13.5	1	1/50 = .02
10	13.5–14.5	2	2/50 = .04
Totals		50	1.00



Effect van het aantal waarnemingen op de klassenbreedte

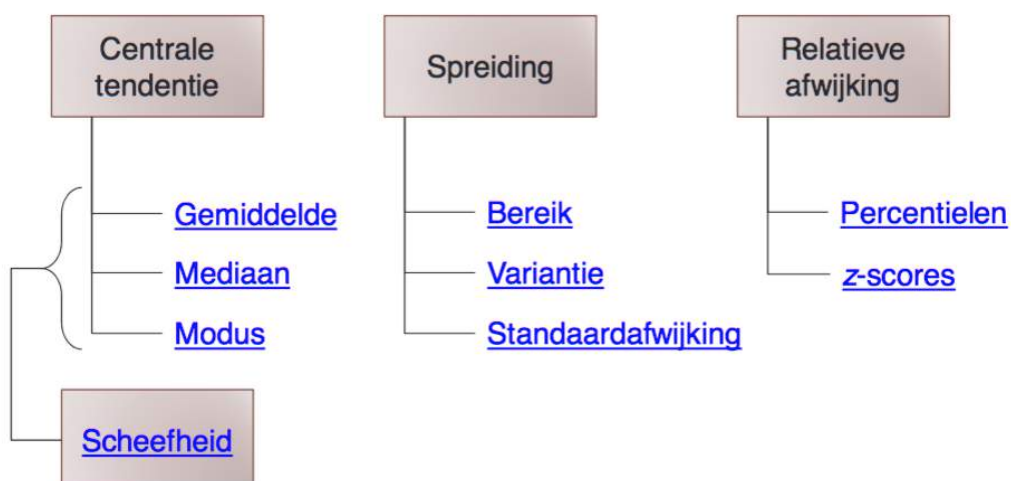


Bepalen aantal klasse

Aantal waarden in gegevensverzameling	Aantal klassen (vuistregel)
< 25	5 - 6
25-50	7 - 14
> 50	15 -20

→ $\sqrt{n}$: regel vh #klassen

2.3. Kwantitatieve gegevens: eigenschappen en maatstaven



CENTRALE TENDENTIE

1) Het (rekenkundig) gemiddelde

Is de som van alle waarden in de gegevens-verzameling gedeeld door het totaal aantal waarden.

- Beïnvloed door alle uitkomsten, o.a. door uitschieters of extreme waarden.
- Symbolen: $\bar{x}$ (steekproef) en μ (populatie)
- $\bar{x}$ is een schatting van μ (nauwkeuriger met stijgend aantal waarden).
- Formule:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

2) De mediaan

= middelste waarde na het rangschikken op grootte van de waarden.

- Bij oneven n is de mediaan gelijk aan de middelste waarde.
- Bij even n is de mediaan de halve som van de twee middelste waarden.

Bepaling voor ruwe data:

1. Bepaal $k = \frac{n+1}{2}$. Dit geeft het rangnummer van de mediaan.
2. Is k een geheel getal, dan is de k -de waarde de mediaan.
Is k geen geheel getal, dan is de mediaan de halve som van de $[k]$ -de en de $[k]$ -de waarde.

- De mediaan is minder gevoelig voor extreem grote of kleine waarden dan het gemiddelde.
- Minstens 50% der waarden zijn $\leq$, en minstens 50% der waarnemingen zijn $\geq$ de mediaan.

$[k] \rightarrow$ 'k nr beneden afronden tot op dichtstbijzijnde geheel getal'.
 $\lceil k \rceil \rightarrow$ 'k nr boven afronden tot op dichtstbijzijnde geheel getal'.

3) De modus

= de vaakst voorkomende waarde in de gegevensverzameling.

- Zinvol voor zowel kwalitatieve al kwantitatieve gegevens.
- Gegevensverzameling kan meer dan één modus hebben (multimodaal).
- Bij in klassen gegroepeerde gegevens spreekt men van een **modale klasse** – een klasse met hoogste frequentie.

VOORBEELD 1

Gegevens: (#9)

1 3 5 6 8 8 9 11 12

Gemiddelde:

$$\bar{x} = \frac{63}{9} = 7$$

Mediaan: 8

$$k = \frac{9+1}{2} = 5^e$$

Modus: 8

VOORBEELD 2

Gegevens: (#9)

1 3 5 6 8 9 9 11 12

Gemiddelde:

$$\bar{x} = \frac{64}{9} = 7,1...$$

Mediaan: 9

$$k = \frac{9+1}{2} = 5^e$$

Modus: 9

VOORBEELD 3

Gegevens: (#10)

1 3 5 6 8 9 9 11 12
100

Gemiddelde: 16,4

$$\bar{x} = \frac{164}{10} = 16,4$$

Mediaan: 8,5

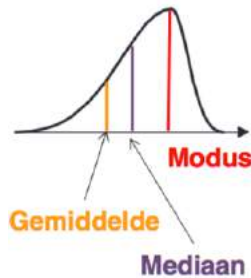
$$k = \frac{10+1}{2} = 5,5^e$$

$\rightarrow$ halve som van 5^e en 6^e waarde

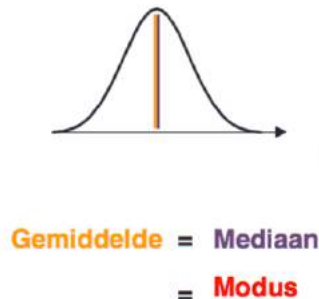
Modus: 9

SCHEEFHEID ("skewness")

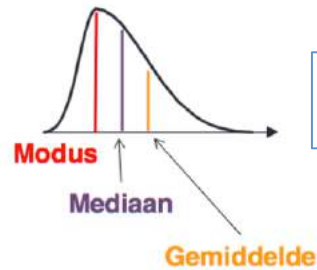
Linksscheef



Symmetrisch



Rechtsscheef



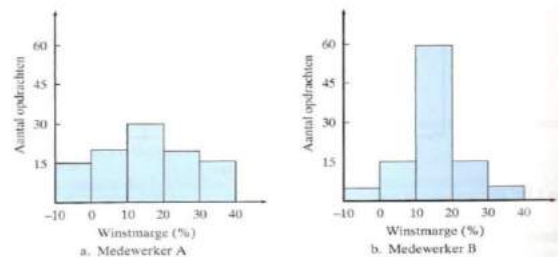
Links: $\bar{x} < Me$
Rechts: $\bar{x} > Me$

SPREIDING

1) Bereik

= het verschil tussen de grootste waarde en de kleinste waarde.

Ongevoelig voor variatie bij grote gegevensverzamelingen:



Spreading: max - min

2) Variantie

= de som van de gekwadrateerde verschillen met het gemiddelde, gedeeld door $(n - 1)$.

- Uitgedrukt in 'eenheden-kwadraat' (bemoeilijkt interpretatie).
- Waarom kwadrateren? Laten we het kwadraat weg, dan hebben we altijd 0 als uitkomst (eig. deviatiescores)
- Formule (definitie):

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

= Afwijking van score x_i (voor persoon i) tov gem. $\bar{x}$

⇒ BELANGRIJK!!

Eigenschap deviatiescores

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = 0$$

Bewijs

$$\begin{aligned} &(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) \\ &= (x_1 + x_2 + \dots + x_n) - (\bar{x} + \bar{x} + \dots + \bar{x}) \\ &= (n \cdot \bar{x}) - (n \cdot \bar{x}) \\ &= 0 \end{aligned}$$

Som vd verschillen
vh gem is altijd 0
→ kwadrateren



Rekenformule steekproefvariantie

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

↪ Hiermee vermijd je berekenen vd deviatiescores (zonder bewijs)

Steekproefvariantie: waarom delen door $n - 1$ en niet n ?

Men wil de populatievariantie (σ^2) schatten door de steekproefvariantie (s^2).

- Als het populatiegemiddelde (μ) onbekend is, dient men dit eerst te schatten (door $\bar{x}$).
- Echter, de steekproefwaarden x_i zullen zich in het algemeen 'scharen' rond $\bar{x}$ en niet rond μ . Deviaties $|x_i - \bar{x}|$ zijn in het algemeen kleiner dan deviaties $|x_i - \mu|$.
- Hierdoor zal de teller van s^2 vaak kleiner uitvallen dan de teller van σ^2 . Om een 'goede' schatter te bekomen dienen we dit te compenseren via een noemer die kleiner is dan n .

3) Standaardafwijking

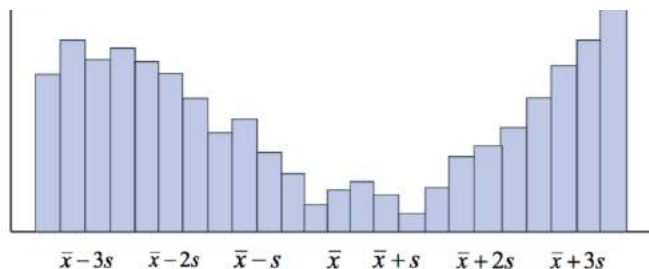
= de positieve vierkantswortel van de steekproefvariantie.

- Uitgedrukt in dezelfde eenheden als de gegevenswaarden.
- Formule:

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

Betekenis van de standaardafwijking: verdelingsregels

➤ Regel van Tsjebysev: verdeling niet heuvelvormig



Geen apriori info over het aant. waarden.

Minstens 3/4 van de waarden

Minstens 8/9 van de waarden

Het percentage van de waarnemingen die binnen k standaardafwijkingen van het gemiddelde vallen is ten minste

$$100 \left(1 - \frac{1}{k^2}\right) \text{ procent}$$

Als $k=2$, hebben we minimaal 75%.

Als $k=3$, minimaal 88,9%

Als $k=1$, minimaal 0%.

De regel van Tsjebysev maakt geen veronderstellingen over de vorm van de gegevensverdeling. Voor $k=1$ is ze waardeloos. Hoe meer gekend is over de vorm, des te nauwkeuriger de benedengrenzen van de percentages (zie: empirische regel)

Conceptvraag

Tien studenten hebben de volgende scores op het examen
Wis I-A behaald

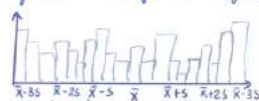
6, 7, 8, 9, 9, 11, 11, 12, 13, 14 → $n=10$

Iemand vindt dat $s = 1,26$.
 $\Rightarrow s = 1,26$ is niet mogelijk
interval moet groter zijn dan
 $\neq 5\%$ ($\frac{3}{4}$) maar niet het geval

Leg uit, zonder de standaardafwijking na te rekenen, waarom dit resultaat niet correct kan zijn.

$$\bar{x} = \frac{6+7+8+9+9+11+11+12+13+14}{10} = \frac{100}{10} = 10$$

Volgens stelling van Tsjebysev:



$$\bar{x} - 2s \Rightarrow 10 - 2 \cdot 1,26 = 7,48$$

$$\bar{x} + 2s \Rightarrow 10 + 2 \cdot 1,26 = 12,52$$

zie (a)

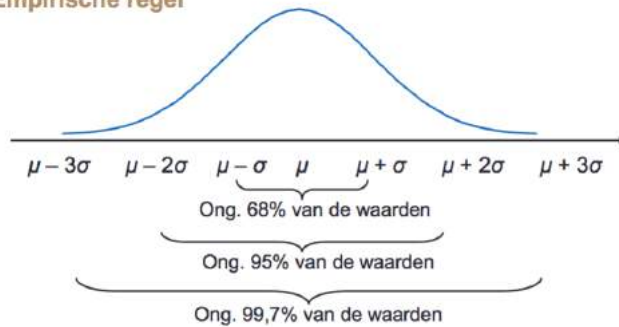
minstens $\frac{3}{4}$ waarden
minstens $\frac{8}{9}$ waarden

- Empirische regel (68 – 95 - 99,7 regel): verdeling heuvelvormig

Gauss-curve

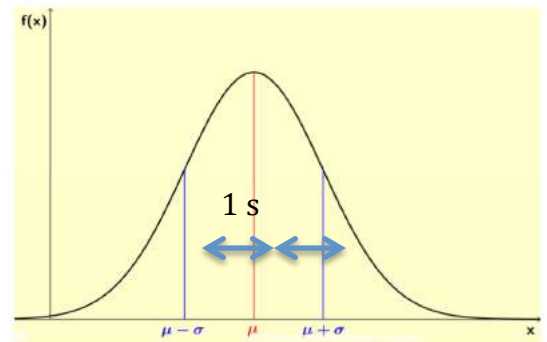
Betekenis van de standaardafwijking

Empirische regel

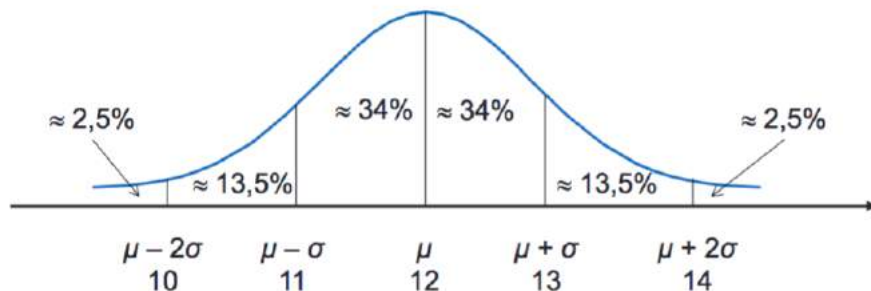


De normale verdeling

(als voorbeeld van een symmetrische en heuvelvormige verdeling)



Je koopt een nieuwe smartphone. Specificaties fabrikant: gemiddelde spreekijd is 12 uur, standaardafwijking is 1 uur. De verdeling van de spreekijden is symmetrisch en heuvelvormig. Jouw toestel haalt net geen 10 uur spreekijd. Heeft de fabrikant onjuiste specificaties meegedeeld?



Samengevat: maten van spreiding

Maatstaf	Waarnemingen afkomstig uit...	
	de hele populatie	een steekproef
Rekenkundig gemiddelde	$\mu = \frac{\sum x_i}{N}$	$\bar{x} = \frac{\sum x_i}{n}$
Variantie	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$
Standaardafwijking	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

N is de populatiegrootte en n de steekproefgrootte

MATEN VAN RELATIEVE AFWIJING

1) Percentielen

- **Percentielen** zijn grenswaarden die de *gerangschikte* meetwaarden in 100 groepen van gelijke grootte verdelen. Er zijn dus 99 percentielen.
- Voor elke $p = 1, 2, \dots, 99$, is het p -de percentiel een **grenswaarde** waarbij minstens $p\%$ van de meetwaarden kleiner en minstens $(100 - p)\%$ groter is dan deze **grenswaarde**.
 - De mediaan is het 50^{ste} percentiel.
 - Het 25^{ste} percentiel heet ook **eerste kwartiel** (Q_1); het 75^{ste} percentiel heet het **derde kwartiel** (Q_3).

2) Z-scores

De z-score van een meetwaarde is de afstand tussen die waarde en het gemiddelde, uitgedrukt in aantal standaardafwijkingen.

- Hiermee kan je verschillende gegevensverzamelingen met elkaar vergelijken.

- Formules:

$$z = \frac{x - \mu}{\sigma} \quad (\text{populatie})$$

$$z = \frac{x - \bar{x}}{s} \quad (\text{steekproef})$$

Bv. z-score = 1
Dan zit je 1
standaardafw. v/h
gemiddelde

Uitschieters berekenen: boxplot en z-score

Uitschieter (of uitbijter)

= een meetwaarde die ongewoon groot of klein is tov de rest van de gegevensverzameling

- Oorzaken
 - Onjuiste meetwaarde
 - Meetwaarde afkomstig uit een andere populatie
 - Juiste, maar zeldzame gebeurtenis

Vuistregels

• Boxplots

- Meetwaarden tussen binnenste en buitenste omheiningen zijn 'mogelijke' uitschieters.
- Meetwaarden voorbij de buitenste omheiningen zijn 'vrijwel zeker' uitschieters.

KA = kwartielafstand
= $Q_3 - Q_1$

Uitschieters:
 $\geq 1,5 \text{ KA} + Q_3$
 $\leq Q_1 - 1,5 \text{ KA}$

• Z-scores

Uitschieters: > 3 of < -3

- Meetwaarden met z-scores groter dan 3 in absolute waarde worden als uitschieters beschouwd.
- Indien de gegevensverdeling zeer scheef is, kunnen waarden met z-scores groter dan 1 in absolute waarde ook als uitschieters beschouwd worden.

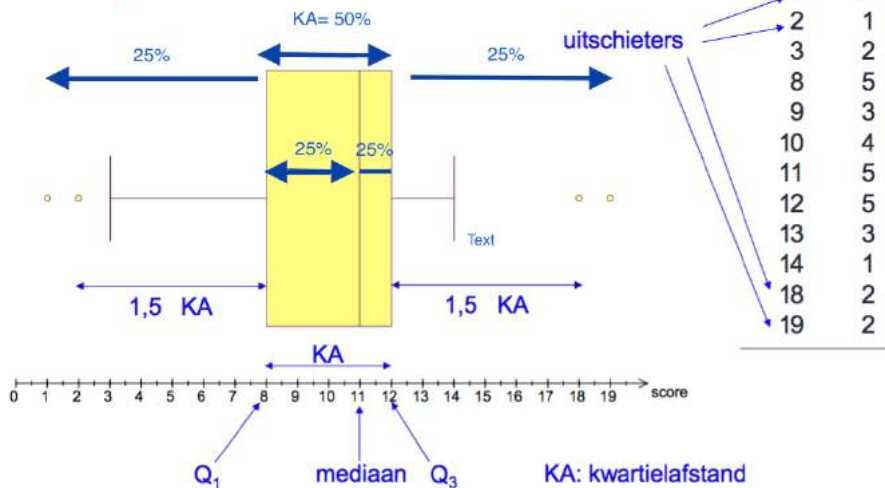
Let op:

Als snorharen berekend zijn mag je niet meteen waarden van formule pakken, maar kijken welke de laagste/hogste waarde is die net geen uitschieter is en die als snorharen nemen

- Hoe detecteren?

➤ Boxplot

Boxplot



Q1: 1^e kwartiel (25%)

Mediaan (50%)

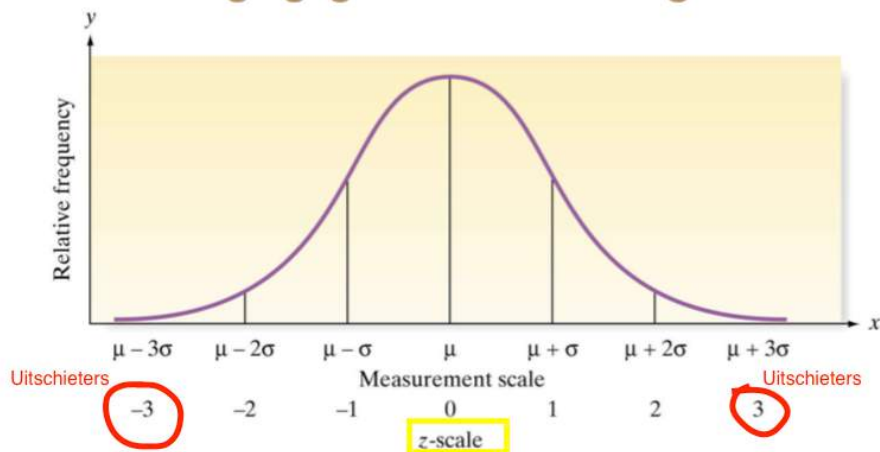
Q3: 3^e kwartiel (75%)

KA: kwartielafstand
= $Q3 - Q1$ (hierin ligt 50% vd waarden)

Bereik: grootste – kleinste waarde

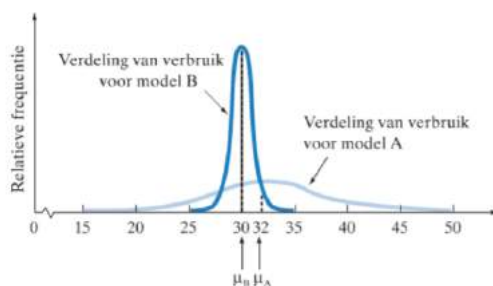
➤ Z-score

Empirische regel voor symmetrische en heuvelvormige gegevensverzamelingen



2.4. De waarheid verdraaien met statistiek

- Schaalmanipulatie (breedte, hoogte)
- Betekenis van het woord 'gemiddelde' in de omgangstaal?
- Discussie van centrale tendentie zonder rekening te houden met spreiding



Welk automodel koop je als je enkel het gemiddeld brandstofverbruik kent?

En als je ook de standaardafwijking kent?

Waarom?

Figuur 2.37: Aantal mijlen dat je kunt afleggen met twee automodellen

Hoofdstuk 3. Kansrekening

3.1. Gebeurtenissen, uitkomstenruimten en kans

- **Experiment**
Handeling of een proces van waarnemen dat tot een niet-zekere uitkomst leidt
- **Enkelvoudige gebeurtenis (uitkomst)**
De meest fundamentele uitkomst van een experiment
- **Uitkomstenruimte S ("sample space")**
De verzameling van alle uitkomsten van een experiment

Voorbeeld 1 Experiment: dobbelsteen werpen Uitkomst: bv. 6 Uitkomstenruimte: {1, 2, 3, 4, 5, 6} → Elke uitkomst heeft 1/6 kans	Voorbeeld 2 Experiment: 2 muntstukken opgooien Uitkomst: KM (kruis-munt) Uitkomstenruimte: {KM, MK, MM, KK} → Elke uitkomst heeft 1/4 kans
---	---

- **Gebeurtenis**
Een verzameling van uitkomsten
- **Kans op een gebeurtenis**
Is de som van de kansen van alle uitkomsten behorende tot de gebeurtenis

Voorbeeld Experiment: 1 dobbelsteen gooien Gebeurtenis: $A = \{\text{meer dan 4 ogen}\} = \{5, 6\}$ Kans op A = $P(A) = (\text{kans op 5}) + (\text{kans op 6})$
--

Hoe de kans op een uitkomst van een experiment bepalen?

Kansregels voor uitkomsten

- Kans op uitkomst is een getal tussen 0 en 1 (grenzen inbegrepen)
- Som van de kansen van alle mogelijke uitkomsten moet 1 zijn

Kansdefinities

➤ Kansdefinitie van Laplace

In een experiment met *gelijkwaardige (even waarschijnlijke) uitkomsten* geldt:

$$P(\text{uitkomst}) = \frac{1}{\text{aantal uitkomsten}}$$

➤ Wet van de grote aantallen

- Bij het herhaald uitvoeren van een experiment zal de relatieve frequentie van een uitkomst x een limietwaarde aannemen
- Deze limietwaarde kan als kans op de uitkomst x opgevat worden

$$P(x) = \lim_{n \rightarrow +\infty} \frac{n_x}{n}$$

n_x = aantal keer dat uitkomst x optreedt

n = aantal herhalingen van het experiment

→ Betekenis: Als men een experiment zich oneindig keer herhaalt dan zal de kans op een bepaalde uitkomst een bepaalde waarde benaderen

Hoe de kans op een gebeurtenis bepalen?

1. Definieer het experiment
2. Maak een lijst van alle uitkomsten
3. Ken kansen toe aan de uitkomsten
4. Bepaal welke uitkomsten tot de gebeurtenis behoren
5. Tel de kansen op van de uitkomsten uit de vorige stap

Voorbeeld

Experiment: twee onvervalste dobbelstenen werpen.

Gevraagd: $P(\text{8 ogen})$ en $P(\text{minimaal 9 ogen})$

Oplossing: Alle uitkomsten (36 in het totaal) zijn gelijkwaardig ondersteld en hebben dus elk eenzelfde kans van $1/36$.

Dobbelsteen 1	Dobbelsteen 2					
	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

$$P(\text{8 ogen}) = \frac{1}{36} + \frac{1}{36} + \dots + \frac{1}{36} = \frac{5}{36}$$

$$P(\text{minstens 9 ogen}) = \frac{1}{36} + \frac{1}{36} + \dots + \frac{1}{36} = \frac{10}{36}$$

Wat als het aantal uitkomsten te groot is om op te sommen?

➤ Combinatieregels

Aantal verschillende steekproeven van grootte n te bepalen uit een populatie van omvang N

$$\binom{N}{n} = \frac{N!}{n! (N - n)!}$$

$$n! = n \cdot (n - 1) \dots 3 \cdot 2 \cdot 1$$

Per afspraak: $0! = 1$

→ Aantal combinaties van n elementen uit een totaal van N elementen

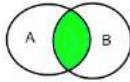
3.2. – 3.3. – 3.4. Soorten gebeurtenissen

▪ Samengestelde gebeurtenissen

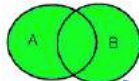
= een combinatie van twee of meerdere gebeurtenissen

2 soorten:

- Doorsnede $\cap$ ("en")
 $A \cap B$



- Unie $\cup$ ("of")
 $A \cup B$



Somregel voor kansen

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

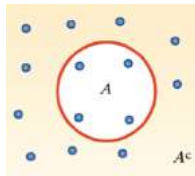
Waarbij $P(A \cap B) = P(A) \cdot P(B)$

→ Uitkomsten in de doorsnede niet dubbel meetellen!

→ Venndiagram

▪ Complementaire gebeurtenissen

= complement van een gebeurtenis A is de gebeurtenis dat A niet voorkomt



Complementregel

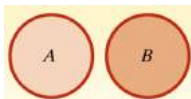
$$P(A) + P(A^c) = 1$$

OF

$$P(A^c) = 1 - P(A)$$

▪ Disjuncte gebeurtenissen

= gebeurtenissen die geen gemeenschappelijke uitkomsten bevatten (of waarbij een bepaalde uitkomst niet tegelijk kan optreden)



Somregel voor disjuncte gebeurtenissen

$$P(A \cup B) = P(A) + P(B)$$

Want A en B zijn disjunct

$$\Leftrightarrow P(A \cap B) = 0$$

Doorsnede en unie van gebeurtenissen

Onderstaande kruistabel geeft een steekproef weer van n werknemers volgens leeftijd en inkomen. Je selecteert lukraak een werknemer uit de steekproef. Elke cel geeft de kans dat je de werknemer in de betreffende leeftijd-inkomencategorie aantreft.

Gebeurtenis A: werknemer is ouder dan 50j

Gebeurtenis B: werknemer heeft inkomen tussen 2500 en 5000 euro

Doorsnede.
Als 'EN' te interpreteren.

Leeftijd	Maandinkomen (in euro)		
	< 2500	2500 – 5000	> 5000
< 30j	5%	12%	10%
30j - 50j	14%	22%	16%
> 50j	8%	10%	3%

$$P(A) = 0,21 \quad P(B) = 0,44 \quad P(A \cap B) = 0,10$$

$$P(A \cup B) = 0,08 + 0,10 + 0,03 + 0,12 + 0,22 = 0,55$$

Dit resultaat is ook te verkrijgen met de SOMREGEL (zie volgende dia)

Unie. Als (niet exclusieve) 'OF' te interpreteren.

Somregel: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Voorbeeld

Experiment: lukraak een kaart trekken uit een spel kaarten.

Gevraagd: de kans dat de getrokken kaart een aas is of een rode kleur heeft.

Type	Kleur		Totaal
	Rood	Zwart	
Aas	2	2	4
Geen aas	24	24	48
Totaal	26	26	52

$$\begin{aligned}
 P(\text{Rood} \cup \text{Aas}) &= \frac{24 + 2 + 2}{52} \\
 &= \frac{26}{52} + \frac{4}{52} - \frac{2}{52} \\
 &= P(\text{Rood}) + P(\text{Aas}) - P(\text{Rood} \cap \text{Aas})
 \end{aligned}$$

3.5. – 3.6. – 3.7. Voorwaardelijkheid, afhankelijkheid en kansbomen

■ Voorwaardelijke kans $P(A|B)$

De kans dat gebeurtenis A optreedt als je weet dat gebeurtenis B is opgetreden
→ kans op A gegeven B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

met $P(B) \neq 0$

■ Productregel

De kans op de doorsnede van 2 gebeurtenissen kan worden berekend met de productregel die gebruik maakt van voorwaardelijke kansen.

$$\begin{aligned} \text{Uit } P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ \text{En } P(B|A) &= \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} \end{aligned}$$

$$\begin{aligned} \text{volgen: } P(A \cap B) &= P(B|A) \cdot P(A) \\ &= P(A|B) \cdot P(B) \end{aligned}$$

■ Onafhankelijke gebeurtenissen

Twee gebeurtenissen zijn onderling onafhankelijk als de kans dat de ene gebeurtenis voorkomt niet afhangt van het optreden van de andere gebeurtenis

LET OP: A en B zijn onafhankelijk

$\Leftrightarrow$ A en B zijn disjunct

$$\rightarrow A \cap B = \emptyset$$

$$\rightarrow P(A \cap B) = 0$$

$$\rightarrow P(A) \cdot P(B) = 0$$

→ A of B moet 0 zijn en lege verzameling zijn ($A=\emptyset$ of $B=\emptyset$)

A en B zijn onafhankelijk

$$\Leftrightarrow \begin{aligned} P(A|B) &= P(A) \\ P(B|A) &= P(B) \end{aligned}$$

$\Rightarrow$ Productregel voor onafh. geb.

A en B zijn onafhankelijk

$$\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

Hier is een tegenvoorbeeld:

Stel, experiment = 2 muntstukken opgooien.

Uitkomstenruimte $S = \{KK, KM, MK, MM\}$

Neem A : {K bij muntstuk 1}, dus $A = \{KK, KM\}$.

Neem B : {K bij muntstuk 2}, dus $B = \{KK, MK\}$.

Duidelijk zijn A en B niet disjunct, want $A \cap B = \{KK\}$.

Maar, A en B zijn onafhankelijk, want

$P(A \cap B) = P(A) \cdot P(B)$, omdat $P(A \cap B) = 1/4$, $P(A) = 2/4$ en $P(B) = 2/4$.

■ Regel van Bayes

Regel waarmee je, als de voorwaardelijke kansen $P(B|A)$ gegeven zijn, de kans op het 'omgekeerde' $P(A|B)$ kunt berekenen

De regel van Bayes

(algemene vorm)

Gegeven k paarsgewijs disjuncte gebeurtenissen $B_1, B_2, \dots, B_k$, waarbij $P(B_1) + P(B_2) + \dots + P(B_k) = 1$, dan

$$\begin{aligned} P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\ &= \frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \dots + P(B_k)P(A|B_k)} \end{aligned}$$

Uit de productregel volgt

$$\begin{aligned} P(B|A) &= \frac{P(B \cap A)}{P(A)} \\ &= \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} \end{aligned}$$

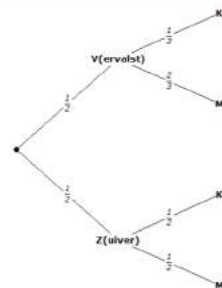
$\Rightarrow$ Regel van Bayes

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Kansboom

→ handig voor het berekenen van de kans op doorsneden van gebeurtenissen

Je beschikt over 50 zuivere en 50 vervalste muntstukken met kans op munt gelijk aan 2/3.
Als je willekeurig een muntstuk opgooit, hoe groot is de kans dat het munt is?



$$P(M) = P(M \text{ en } V) + P(M \text{ en } Z) \\ = P(M | V)P(V) + P(M | Z)P(Z) \\ = 2/3 \cdot 1/2 + 1/2 \cdot 1/2 = 7/12$$

Voorwaardelijke kans

Voorbeeld

		Marktrente		Totaal aant. obs.
		daalt	stijgt (B)	
Prijs aandeel	daalt (A)	10	85	95
	stijgt	90	15	105
	Totaal	100	100	200

$$P(A) = 95/200 = 0,475$$

$$P(A|B) = P(A \text{ doorsnede } B) / P(B) = (85/200) / (100/200) = 0,85$$

Onafhankelijkheid

Stel $A = \{100 \text{ keer na elkaar 'kruis' werpen}\}$,
 $B = \{101\text{-ste worp geeft 'kruis'}\}$.

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \\ = \frac{2^{100}}{2^{101}} = \frac{1}{2} \\ P(B) = \frac{1}{2}$$

Je gooit een onvervalst muntstuk 100 keer op en noteert telkens de uitkomst. Stel, op een heel uitzonderlijke dag heb je 100 keer na elkaar 'kruis' gegooit. Wat is dan de kans dat je bij de 101-ste worp opnieuw 'kruis' bekomt?

A en B heten 'onafhankelijk'

Afhankelijke gebeurtenissen

Angst voor examen wiskunde vs aantal uren wiskunde in laatste jaar SO

Aantal	Graad van angst			Totaal
	geen/weinig	matig	veel/erg veel	
minder dan 3			4	4
3 - 4	9	21	91	121
5 of meer	16	20	31	67
Totaal	25	41	126	192

A: {3 - 4 uren wis}, B: {veel/erg veel angst}

$$P(B) = \frac{126}{192} \approx 0,66$$

A en B zijn niet onafhankelijk

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{91/192}{121/192} = \frac{91}{121} \approx 0,75$$

De regel van Bayes: voorbeeld

- Stel $B = \{\text{besmet}\}$ en $A = \{\text{test positief}\}$. Er is gevraagd: $P(B|A)$.
- Gegevens: $P(B) = 0,01$ en $P(B^c) = 1 - P(B) = 0,99$.
- Test is 95% accuraat wil zeggen: $P(A|B) = 0,95$ en $P(A|B^c) = 0,05$.

Je wordt getest op een zeldzame ziekte. • Test is voor 95% accuraat. • Kans op besmetting door ziekte: 1%. Stel dat je positief test op deze ziekte. Wat is dan de kans dat je deze ziekte ook hebt?

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \\ = \frac{(0,95) \cdot (0,01)}{(0,95) \cdot (0,01) + (0,05) \cdot (0,99)} \approx 0,161$$

Conceptvraag

Enveloppe I	Enveloppe II
€10	€20
€20	€20

Je hebt twee gesloten enveloppen. Eén envelop bevat een biljet van 10€ en een biljet van 20€. De andere envelop bevat twee biljetten van 20€. Je kiest lukraak een envelop en neemt er willekeurig een biljet uit. Als dat biljet 20€ waard is, wat is dan de kans dat je de envelop beet hebt met nog een biljet van 20€ erin?

- A. 1/4
B. 1/2
C. 2/3
D. 1
E. Geen van bovenstaande.

ENVELOP	BILJET	TIJD
I	10	25%
I	20	25%
II	20	25%
II	20	25%

2/3 van de tijd heb je envelop II in handen

A = geselecteerde envelop = II
B = test van blindelings €20 getrokken
→ $P(A) = 1/2$
→ $P(A \text{ complement}) = 1 - 1/2 = 1/2$
Gekend: $P(B|A) = 1$
 $P(B|A \text{ complement}) = 1/2$
Bayes: $(1 \cdot 1/2) / (1 \cdot 1/2 + 1/2 \cdot 1/2)$
 $= 2/3 = 66\% \text{ ongeveer}$

Hoofdstuk 4. Stochastische variabelen met een discrete kansverdeling

4.1. Twee soorten stochastische variabelen

Stochastische variabelen

= Variabele die numerieke waarden aanneemt bij toevallige uitkomsten van experiment

→ bij elke uitkomst wordt één en slechts één waarde toegekend

→ 2 soorten

- **Discrete stochastische variabelen**
→ Kunnen slechts een *eindig of aftelbaar* aantal waarden aannemen
- **Continue stochastische variabelen**
→ Neemt een *oneindig en niet aftelbaar* aantal waarden aan, te vergelijken met een interval of halfrechte op de reële getallen

LET OP: Stochastische variabele neemt oneindig aantal waarden aan
 $\neq$ variabele = continu

Bv casinospel: reeks spelen achter elkaar

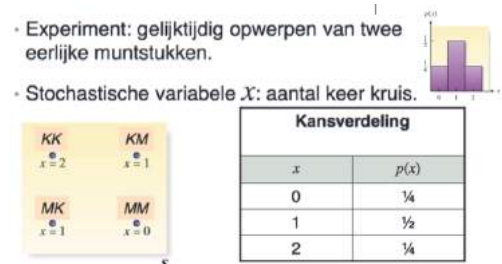
Reeks: $\{1, 2, 3, \dots, n\}$ & $F = 1^e$ keer winnen, of $n+1$ als er geen winst is

⇒ Oneindig # uitkomsten maar wel discreet (je kan ze op rij zetten)

4.2. Discrete kansverdelingen

Discrete kansverdeling

= grafiek, tabel of formule die de kansen vastlegt die horen bij de mogelijke waarden die de stochastische variabele kan aannemen



Eigenschappen

- $p(x) \geq 0$ voor alle waarden van x
- $\sum_x p(x) = 1$

Kans kan niet negatief zijn

Som van alle kansen is 1

4.3. Synthesematen van een stochastische variabele

- **Verwachtingswaarde**
= gewogen gemiddelde van de mogelijke waarden van de variabele
- **Variantie**
= gewogen gemiddelde van de gekwadrateerde afwijkingen tov de verwachtingswaarde
- **Standaardafwijking**
= positieve vierkantswortel van de variantie

$$\mu = E(x) = \sum x \cdot p(x)$$

$$\sigma^2 = E[(x - \mu)^2] = \sum (x - \mu)^2 p(x)$$

$$\sigma = \sqrt{\sum (x - \mu)^2 p(x)}$$

!!!

	Regel van Tsjebysev	Empirische regel
Geldt voor:	elke kansverdeling	heuvelvormige en symmetrische kansverdelingen
$P(\mu - \sigma < x < \mu + \sigma)$	≥ 0	$\approx 0,68$
$P(\mu - 2\sigma < x < \mu + 2\sigma)$	$\geq 3/4$	$\approx 0,95$
$P(\mu - 3\sigma < x < \mu + 3\sigma)$	$\geq 8/9$	$\approx 0,997$

Synthesematen: rekenschema

- Experiment: gelijktijdig opwerpen van twee eerlijke muntstukken.
- Stochastische variabele X : aantal keer kruis.

x	Versch. uitkomsten $p(x)$	Gemiddelde μ $x p(x)$	Afwijking $x - \mu$	$(x - \mu)^2$	Variantie $(x - \mu)^2 p(x)$
0	$1/4$	$0 \cdot 1/4 = 0$	$0 - 1 = -1$	$(-1)^2 = 1$	$1 \cdot 1/4 = 1/4$
1	$1/2$	$1/2 \cdot 1 = 1/2$	$1 - 1 = 0$	$0^2 = 0$	$0 \cdot 1/2 = 0$
2	$1/4$	$2 \cdot 1/4 = 1/2$	$2 - 1 = 1$	$1^2 = 1$	$1 \cdot 1/4 = 1/4$
Totaal		SOM = 1	SOM = 0		SOM = $1/2$

Conceptvraag

Gegeven de kansverdeling

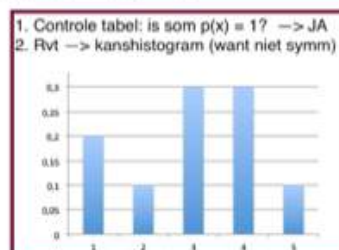
x	1	2	3	4	5
$p(x)$	0,2	0,1	0,3	0,3	0,1

met $\mu = 1 \cdot (0,2) + 2 \cdot (0,1) + 3 \cdot (0,3) + 4 \cdot (0,3) + 5 \cdot (0,1) = 3$.

Beantwoord zonder σ uit te rekenen:

Juist of fout?

$\sigma < 1$



Rekenen uit ongerijmde $\sigma < 1$

$$P[\mu - 2\sigma < x < \mu + 2\sigma]$$

$1 < \quad < 5$ (want $\mu = 3$)

$$\Leftrightarrow P[1 < x < 5]$$

$$= P[x = 2] + P[x = 3] + P[x = 4]$$

$$= 0,1 + 0,3 + 0,3 = 0,7$$

DUS

$$P[\mu - 2\sigma < x < \mu + 2\sigma] < 0,7$$

→ Kan niet volgens RvT
(moet $\geq 0,75$ zijn)

DUS $\sigma < 1$ is fout

4.4. De binomiale verdeling

Binomiale stochastische variabele

= Aantal 'successen' in een binomiaal experiment

$$p(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

$$= \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x}$$

waarbij n = aantal trials

x = aantal keer 'succes' bij n trials

p = vaste kans op succes per trial

Kenmerken

- Rij van n identieke deelexperimenten ('trials')
- Elke trial heeft 2 uitkomsten: S ('succes') en M ('mislukking')
- De kans op S (en dus ook op M) is dezelfde bij elke trial
- De trials zijn onafhankelijk van elkaar

Let op!

Wil niet zeggen dat kans op S = kans op M!

Eigenschappen

- Verwachtingswaarde

$$\mu = \sum_{x=1}^n x \cdot p(x) = n \cdot p$$

- Variantie

$$\sigma^2 = \sum_{x=1}^n (x - n \cdot p)^2 \cdot p(x)$$

$$= n \cdot p \cdot (1 - p)$$

- Standaardafwijking

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}$$

Verklaring bij formule

$$p(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

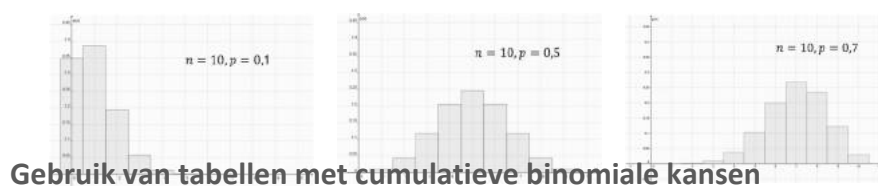
Welke trials = successen?

mogelijkheden om k trials te kiezen uit n

Kans op k successen en (n-k) mislukkingen

Toepassing van productregel voor onafh. geb.

Binomiale kanshistogrammen



Gebruik van tabellen met cumulatieve binomiale kansen

n	k	0	1	2	3	4	5	6	7	8	9	10
25	0	0.778	0.277	0.072	0.004	0.000						
	1	0.974	0.642	0.271	0.027	0.000						
	2	0.998	0.873	0.537	0.098	0.000						
	3	1.000	0.966	0.764	0.224	0.033						
	4	1.000	0.993	0.902	0.421	0.098						
	5	1.000	0.999	0.967	0.617	0.191						
	6	1.000	1.000	0.991	0.780	0.341	0.074	0.001	0.000	0.000	0.000	0.000
	7	1.000	1.000	0.998	0.891	0.512	0.154	0.022	0.001	0.000	0.000	0.000
	8	1.000	1.000	1.000	0.953	0.677	0.354	0.054	0.004	0.000	0.000	0.000
	9	1.000	1.000	1.000	0.983	0.811	0.425	0.115	0.013	0.000	0.000	0.000
	10	1.000	1.000	1.000	0.994	0.902	0.586	0.212	0.034	0.002	0.000	0.000
	11	1.000	1.000	1.000	0.998	0.956	0.732	0.345	0.078	0.006	0.000	0.000
	12	1.000	1.000	1.000	1.000	0.983	0.846	0.500	0.154	0.017	0.000	0.000
	13	1.000	1.000	1.000	1.000	0.994	0.922	0.655	0.268	0.044	0.002	0.000
	14	1.000	1.000	1.000	1.000	0.998	0.966	0.788	0.414	0.098	0.006	0.000
	15	1.000	1.000	1.000	1.000	1.000	0.987	0.885	0.575	0.189	0.017	0.000
	16	1.000	1.000	1.000	1.000	1.000	0.996	0.946	0.726	0.321	0.047	0.000
	17	1.000	1.000	1.000	1.000	1.000	0.999	0.978	0.846	0.488	0.109	0.002
	18	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.926	0.659	0.220	0.009
	19	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.971	0.807	0.383	0.033
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.910	0.579	0.098
	21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.967	0.766	0.236
	22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.902	0.463
	23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.973	0.729
	24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.928	0.723
p =		0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90

Zij x binomiaal verdeeld met n = 25 en p = 0,40. Dan,

$$P(x \leq 9) \approx 0.425$$

Aantal keer 'rood' in een serie van 15 roulettespelen. Je hebt een klas van $N = 10$ studenten, waarvan 6 meisjes en de rest jongens. Je pikt er achtereenvolgens aselekt twee studenten uit. Stel x = aantal meisjes.

1. Trial = 1 roulettespel ; $n = 15$

2. 'succes' = rood
'mislukking' = zwart/groen

Roulette: 37 vakjes
1 groen
18 zwart
18 rood

3. $P(\text{rood}) = 18/37$

4. Speluitkomsten zijn onafhankelijk
→ binomiale stochastische variable!

Voorwaarden = vervuld

1. Trial: student aselekt kiezen ; $n = 2$

2. 'Succes' = meisje

Voorwaarden 3 & 4 niet voldaan

3. Trial 1: $P(\text{meisje}) = 6/10 = 0,6$
Trial 2: $P(\text{meisje}) = 5/9$ (als trial 1 meisje)
OF $= 6/9$ (als trial 1 jongen)

4. Uitkomstmogelijkheden van trial 2 worden beïnvloed door uitkomst van trial 1
→ geen binomiale

Examen met 25 J/F-vragen.
Stel x = aantal correct beantwoorde vragen.

1. Trial: vraag beantwoorden ; $n = 25$

2. 'Succes' = correct beantwoorden

3. $P(\text{succes}) = 1/2$ (VW: bij gokken!!! → veronderstelling)

4. Antwoorden van de vragen enkel onafhankelijk bij gokken (anders waarschijnlijk niet onafhankelijk)
→ binomiale (VW: bij gokken)

1. Trial: student aselekt kiezen ; $n = 2$

VUISTREGEL 'GROOT GENOEG'
 $n < 5\%$ van N

2. 'Succes' = meisje

Voorwaarden = vervuld

3. Trial 1: $P(\text{meisje}) = 60/100 = 0,6$
Trial 2: $P(\text{meisje}) = 59/99$ (als trial 1 meisje) $\approx 0,6$
OF $= 69/99$ (als trial 1 jongen)

4. Uitkomstmogelijkheden van trial 2 quasi niet beïnvloed door uitkomst van trial 1, omdat de groep voldoende omvangrijk is en aselekt werd gekozen
→ binomiale

- Rol 5 keer achter elkaar een dobbelsteen en noteer het aantal keer dat je minstens vijf ogen hebt.
- Wat is de kans dat dit aantal gelijk is aan 4?
- Oplossing:

Trial: dobbelsteen werpen ; $n = 5$

'Succes' = 5 of 6 ogen
 $P(\text{succes}) = 2/6$ of $1/3$
 $P(4) = (5 \text{ uit } 4) \cdot (1/3)^4 \cdot (2/3)^1$
 $= 2/(3)^5 \approx 0,008$

Oefening

Bij de fabricage van geheugenchips is 5% van de chips defect. Door middel van controlemetingen worden deze defecte chips verzameld.

Bereken de kans dat bij 25 metingen het aantal defecte chips

- gelijk is aan nul.
- ten hoogste 2 is.
- gelijk is aan 2.
- ten minste 2 is.

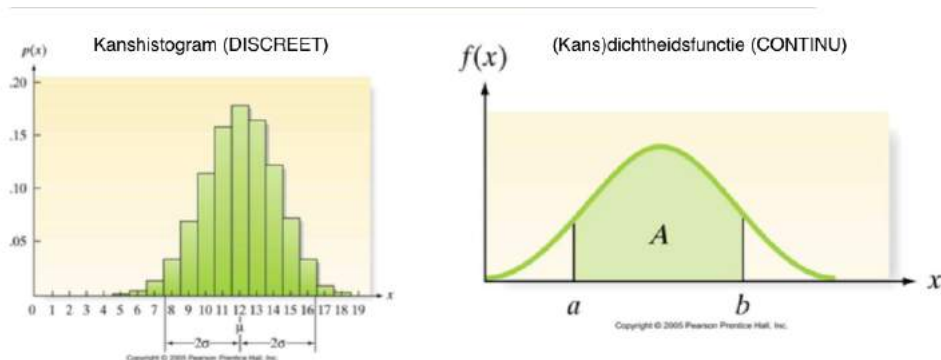
Gegeven:
 $n = 25$
 $p = 0,05$

Oplossing:
Trial: chip onderzoeken (S =defect en M =werkend)
 $p(x) = (25 \text{ uit } x) \cdot p^x \cdot (1-p)^{n-x}$ → te moeilijk
TABEL:
A: $P(x=0) = P(x \leq 0) \approx 0,277$
B: $P(x \leq 2) \approx 0,873$
C: $P(x=2) = P(x \leq 2) - P(x \leq 1) \approx 0,873 - 0,642 = 0,231$
 $P(x=k) = P(x \leq k) - P(x \leq k-1)$
→ somregel disjuncte gebeurtenissen
D: $P(x \geq 2) = 1 - P(x \leq 1) \approx 1 - 0,642 = 0,358$

Hoofdstuk 5. Stochastische variabelen met een continue kansverdeling

5.1. Continue kansverdelingen

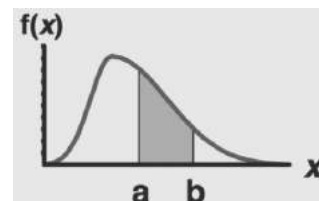
- Wordt bepaald door een functie $f(x)$ die de rol overneemt van het kanshistogram bij discrete stochastische variabelen
- De functie $f(x)$ is de **(kans)dichtheidsfunctie** van de continue stochastische variabele



Eigenschappen

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \quad \text{Tot. opp.}$$

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$



Opmerkingen

$$P(x = a) = P(a \leq x \leq a) = \int_a^a f(x)dx = 0$$

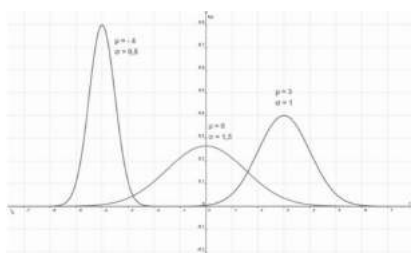
$$P(a < x < b) = P(a \leq x \leq b) = P(a \leq x < b) = P(a < x \leq b)$$

5.2. Normale verdeling (Gauss-curve) $x \sim N(\mu, \sigma)$

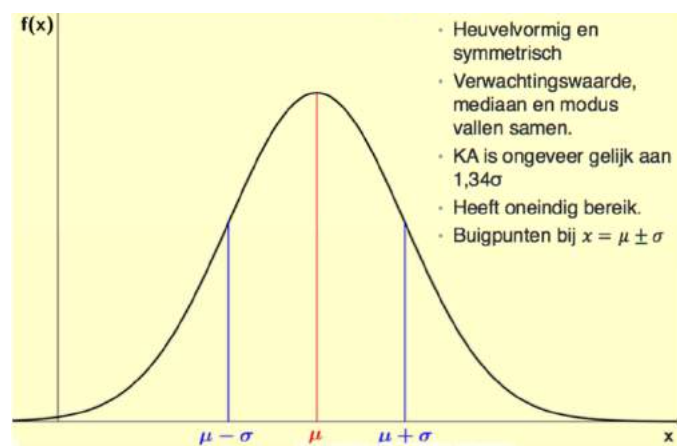
Belang

- Goede beschrijving van heel wat stochastische variabelen
- Vaak gebruikt als benadering van discrete kansverdelingen, zoals de binomiale
- Vormt de basis van de verklarende statistiek

Kenmerken



μ = locatieparameter
 σ = dispersieparameter



Kansdichtheidsfunctie

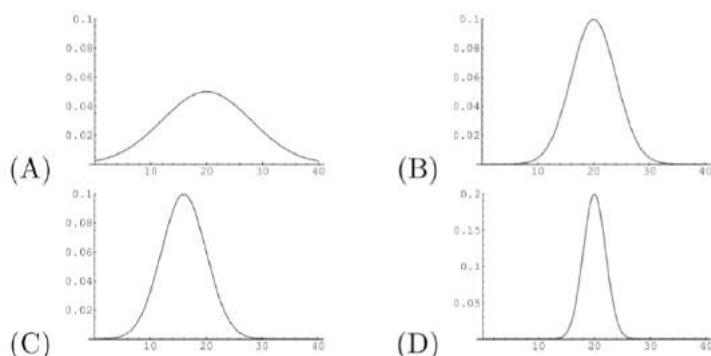
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (x \in]-\infty, +\infty[)$$

Met μ = verwachtingswaarde
En σ = standaardafwijking

Invloed van μ en σ

Conceptvraag

Welke van de volgende grafieken kan de kansdichtheidsfunctie zijn van een normaal verdeelde veranderlijke met $\mu = 20$ en $\sigma = 4$?



→ B = juist

- Buigpunten op $\mu + \sigma$ en $\mu - \sigma$
 $\mu + \sigma = 20 + 4 = 24$
 $\mu - \sigma = 20 - 4 = 16$
- Rond μ gecentreerd ($\mu=20$)
- Empirische regel nagaan $x \sim N(\mu, \sigma)$
 $= P[\mu - 2\sigma < x < \mu + 2\sigma]$
 $\Rightarrow 95\%$ van waarden
 $12 < x < 28$

Standaardnormale verdeling $x \sim N(0, 1)$

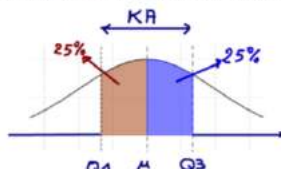
- Normale verdeling met $\mu = 0$ en $\sigma = 1$
- Notatie: z
- Dichtheidsfunctie $f(z) \rightarrow$
- Eigenschap**
 x normaal verdeeld met μ en σ
 $\Leftrightarrow z = \frac{x-\mu}{\sigma}$ standaardnormaal verdeeld

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}z^2}$$

- Standaardnormale tabel
- Bewijs eigenschap 1,34 KA**

Oefening 5.5. Toon aan dat voor elke normale verdeling geldt KA = (1,35) σ

$$KA = Q3 - Q1$$



$$\begin{aligned} P(\mu < x < Q3) &= 0,25 \\ \Leftrightarrow P(0 < z < \frac{Q3-\mu}{\sigma}) &= 0,25 \\ \Leftrightarrow P(0 < z < z_0) &= 0,25 \text{ met } z_0 = \frac{Q3-\mu}{\sigma} \\ \Leftrightarrow z_0 &\approx 0,67 \quad (*) \\ \Rightarrow Q3 &\approx \mu + 0,67 \cdot \sigma \\ \text{analoog: } Q1 &\approx \mu - 0,67 \cdot \sigma \end{aligned} \quad \Rightarrow KA = Q3 - Q1 \approx \mu + 0,67 \cdot \sigma - (\mu - 0,67 \cdot \sigma)$$

$$KA \approx 1,34 \sigma$$

(*) Opm: de z-score ligt iets hoger,
vandaan: $KA \approx 1,35 \sigma$

Standaardnormale tabel

z	00	01	02	03	04	05	06	07	08
0	0000	0040	0080	0120	0160	0199	0239	0279	0319
1	0358	0398	0438	0478	0517	0557	0596	0636	0675
2	0714	0753	0793	0832	0871	0910	0948	0987	1026
3	1065	1104	1143	1181	1219	1258	1296	1335	1374
4	1413	1451	1489	1527	1565	1603	1641	1679	1717
5	1755	1793	1831	1868	1905	1943	1980	2017	2054
6	2091	2128	2165	2202	2239	2276	2312	2349	2386
7	2423	2459	2495	2532	2568	2604	2641	2677	2713
8	2749	2785	2821	2857	2893	2929	2965	3001	3037
9	3073	3108	3144	3179	3215	3251	3286	3321	3357
10	3392	3428	3463	3498	3534	3569	3605	3641	3676
11	3712	3747	3783	3819	3854	3890	3925	3961	3996
12	4032	4067	4103	4138	4174	4209	4245	4281	4316
13	4352	4387	4423	4458	4494	4529	4565	4601	4636
14	4672	4708	4743	4779	4814	4850	4886	4921	4957

5.3. Beschrijvende methoden om te bepalen of een verdeling normaal is

Beschrijvende methoden

- Histogram bij benadering heuvelachtig en symmetrisch?
- Gaat de empirische regel op?
- Via normaliteitsplot

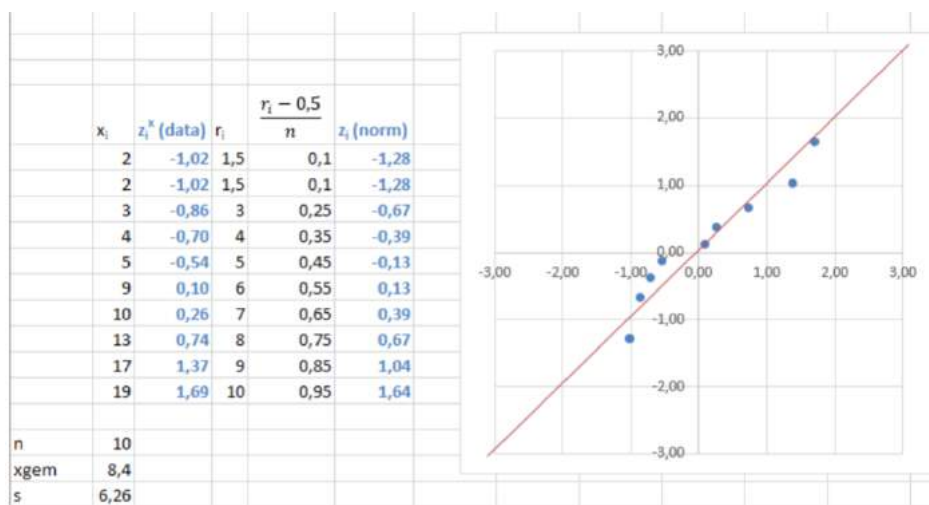
Normaliteitsplot

= Als de gegevens normaal verdeeld zijn, dan liggen de punten van een normaliteitsplot op een rechte

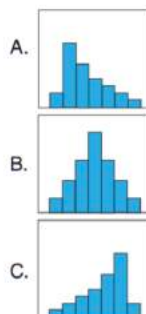
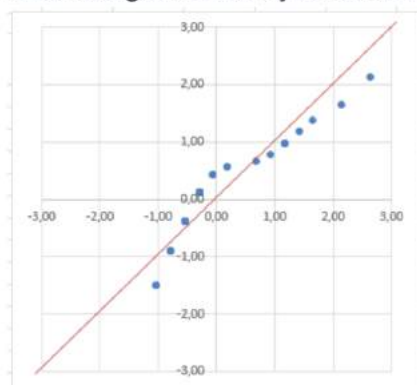
1. Rangschik en label de gegevens in oplopende volgorde als $x_1, x_2, x_3, \dots, x_n$.
2. Bepaal het rangnummer r_i van elke x_i
3. Gebruik de standaardnormale tabel om de Z_i -waarde te vinden zodat

$$P(Z < z_i) = \frac{r_i - 0,5}{n} \quad \text{voor elke } i = 1, 2, \dots, n$$
Merk op: $P(Z < z_i) \approx P(x < x_i)$
4. Plot de koppels $(x_1, z_1), (x_2, z_2), \dots, (x_n, z_n)$ of $(z_1, x_1), (z_2, x_2), \dots, (z_n, x_n)$

Grafisch



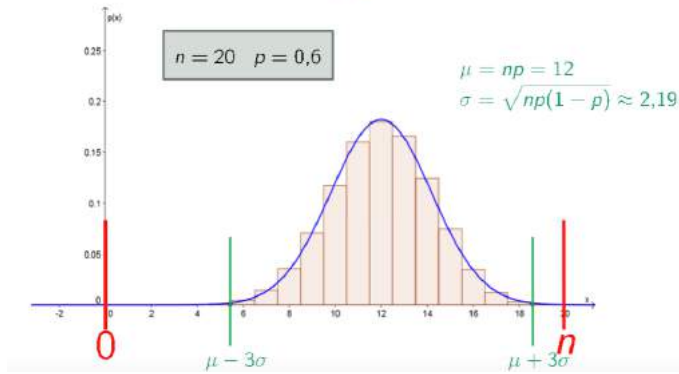
Welk histogram hoort bij de normaliteitsplot?



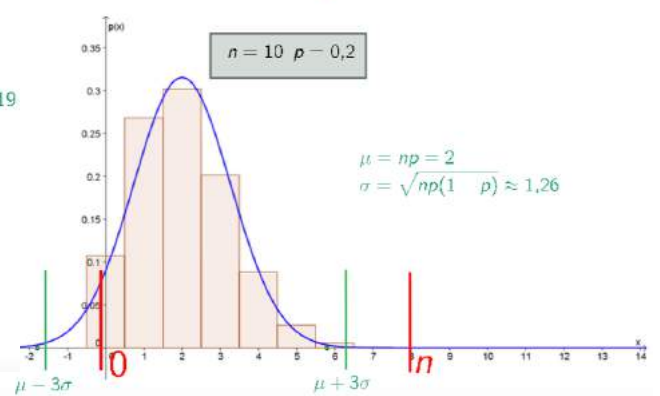
Gegevens op rechte: symmetrisch (B)
 Gegevens onder rechte: linksscheef (C)
 Gegevens boven rechte: rechtsscheef (A)

Binomiale verdeling met een normale verdeling benaderen

Goede benadering

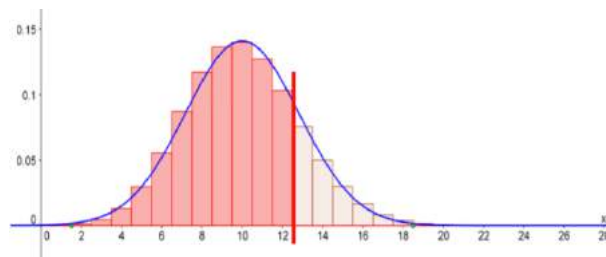


Slechte benadering



Continuïteitscorrectie

- Belangrijk wanneer je een discrete kansverdeling benadert door een continue
- $P(x \leq k) = \text{som van de staafhoogtes tem } k$
 $\approx \text{oppervlakte onder de normale curve tot } k + 0,5$



Hoofdstuk 6. **Verdelingen van steekproefgrootheden**

6.1. **De verdeling van een steekproefgrootheid**