FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

Statistiek I

Samenvatting

Prof. dr. Carette

Opleiding: bachelor of science in de Handelswetenschappen
Academiejaar 2016–2017

UNIVERSITEIT GENT

Inhoudsopgave

Hoofdstuk 1: Statistiek, gegevens en statistisch denken

De wetenschapsstatistiek

Statistiek is de wetenschap van gegevens. Zij omvat het verzamelen, classificeren, samenvatten, organiseren,…

Statistische toepassingen in bedrijven

	Beschrijvende statistiek	**Verklarende statistiek**
Omschrijving	- Voor verbanden te ontdekken	- Voor schatten, nemen van beslissingen en het voorspellen
Elementen	-De populatie (of steekproef) -Een of meer variabele -Tabellen, grafieken,… -Vermelding van de patronen die in de samenvatting naar voren komen	-De populatie -Een of meer variabele -De steekproef -De conclusie over de populatie -Een betrouwbaarheidsmaat voor de conclusie

Basiselementen van de statistiek

- Experimentele eenheid
- Populatie
- Variabele → census
- Steekproef
- Statistische conclusie
- Betrouwbaarheidsmaat

Soorten gegevens

Kwalitatief	**Nominaal**	*Opsommen waarden, benoemen*	*Man/vrouw Abonnee, geen abonnee*
	Ordinaal	*Ordenen waarden*	*Eens/oneens Goud/zilver/brons*
Kwantitatief	**Intervalvariabele**	*Verschillen interpreteerbaar, gelijke intervallen*	*13:00,14:00,… IQ, Temperatuur*
	Ratiovariabele	*Verhoudingen, met absoluut nulpunt*	*Leeftijd, inkomen, punten examen*

Kwantitatief is op een natuurlijk voorkomende numerieke schaal

Hoofdstuk 2: Beschrijven van gegevensverzamelingen

Gegevens beschrijven

Beschrijvende maten:

- **Frequentie**: hoeveel meetwaarden er voorkomen in een bepaalde klasse
- **Relatieve frequentie**: $\frac{Frequentie}{totaal\ aantal\ waarden}$
- **Percentage**: $relatieve\ frequentie\ x\ 100$

Soorten grafische voorstellingen

- **Kwalitatieve gegevens**
  - **Cirkeldiagram**
  - **Staafdiagram**
- **Kwantitatieve gegevens**
  - **Frequentietabel**
  - **Histogram**

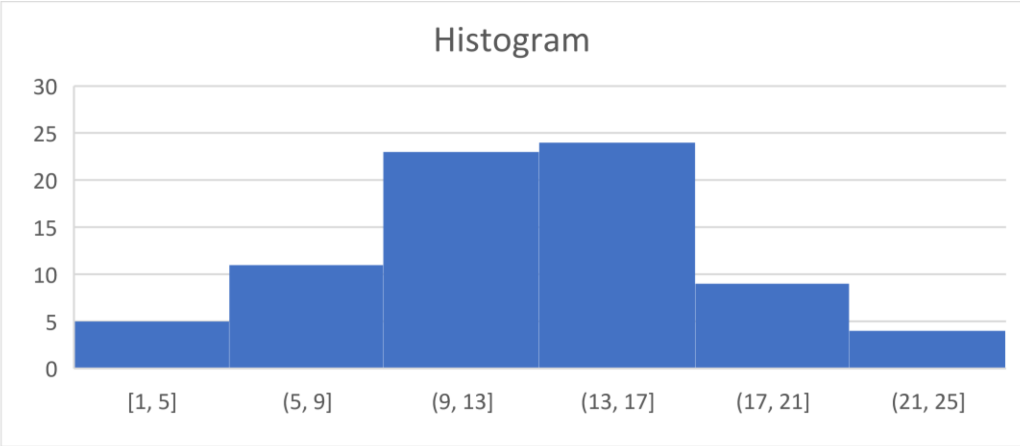
Beschrijven met grafische methode

- Frequentietabel

  →frequentie van de waarden die voorkomen in een bepaalde klasse, bijvoorbeeld voor een histogram

Klasse	Frequentie
1-5	5
6-9	11
9-13	21
13-17	23
17-21	8
21-25	4

- Histogram



Sommatienotatie

$$\boxed{\sum_{i=1}^{n} Xi}$$

Beschrijving van het centrum

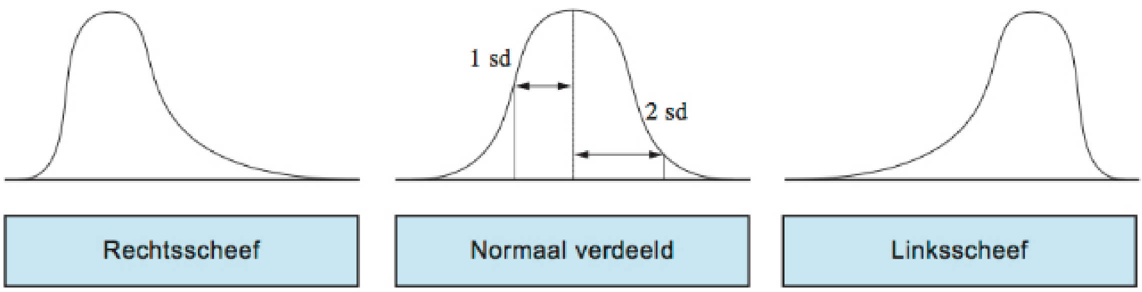
- Gemiddelde

→Het gemiddelde van een verzameling kwantitatieve gegevens is de som van de meetresultaten gedeeld door het aantal gegevens.

$$\boxed{\begin{array}{c}\textbf{\textit{Steekproef gemiddelde}} \\ \bar{x} = \dfrac{\sum_{i=1}^{n} Xi}{n}\end{array}}$$

- Mediaan

→Het middelste getal indien de getallen in opklimmende of dalende waarden genoteerd zijn



- Modus

  →*Is de vaakst voorkomende waarde in een gegevensverzameling*

Beschrijving van de spreiding

- Bereik

  →*De grootste meetwaarde min de kleinste meetwaarde*

- Steekproefvariantie

Steekproefvariantie:

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

Hoe groter de variantie hoe meer de afzonderlijke waarden verschillen, en hoe meer ze van het gemiddelde scharen

→ Men deelt door n-1 ipv n omdat ze meer rond $\bar{x}$ gaan scharen dan rond μ. Dit wordt gecompenseerd door -1 te doen in de noemer

Regel van Tsjebysjev:

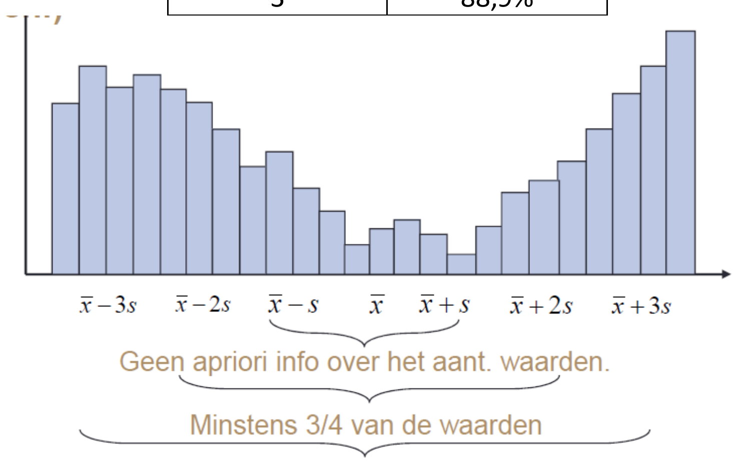
→Geldt enkel voor gegevens die niet heuvelvormig verdeeld zijn

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

$$100.(1 - \frac{1}{k^2})$$

Van de gegevens.

K	%
1	0%
2	75%
3	88,9%



- Deviatiescore

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_i - \bar{x}) = 0$$

$$Bewijs:$$

$$\begin{aligned}&(x_i - \bar{x}) + (x_i - \bar{x}) + \cdots + (x_i - \bar{x}) \\ &= (x_1 + x_2 + \cdots + x_n) - (\bar{x} + \bar{x} + \cdots + \bar{x}) \\ &= (n\,.\,\bar{x}) - (n\,.\,\bar{x}) \\ &= 0\end{aligned}$$

- Standaardafwijking/ standaarddeviatie

→*positieve wortel van de steekproefvariantie*

$$s = \sqrt{s^2}$$

Symbolen:

S^2= variantie van de steekproef

S=standaardafwijking van de steekproef

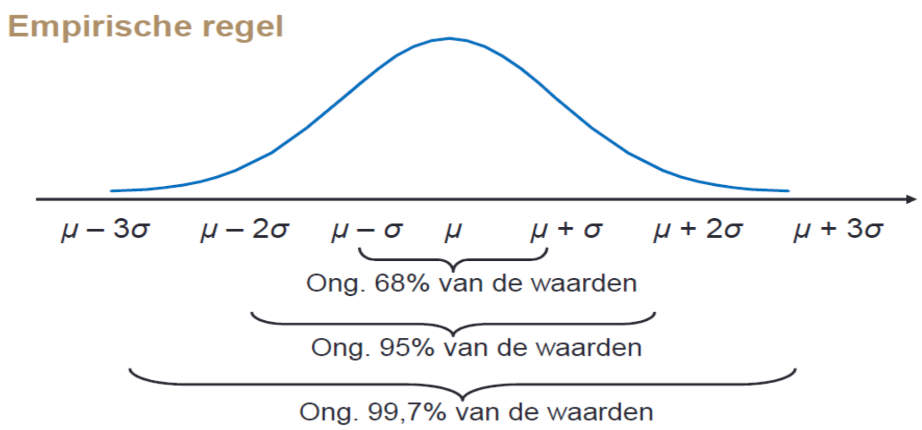
σ^2= variantie van de populatie

σ=standaardafwijking van de populatie

Betekenis van de standaardafwijking

Interval	Procentueel aantal waarden binnen het interval	Z-waarden
$(\mu - \sigma, \mu + \sigma)$	68%	-1 en 1
$(\mu - 2\sigma, \mu + 2\sigma)$	95%	-2 en 2
$(\mu - 3\sigma, \mu + 3\sigma)$	99,7%	-3 en 3

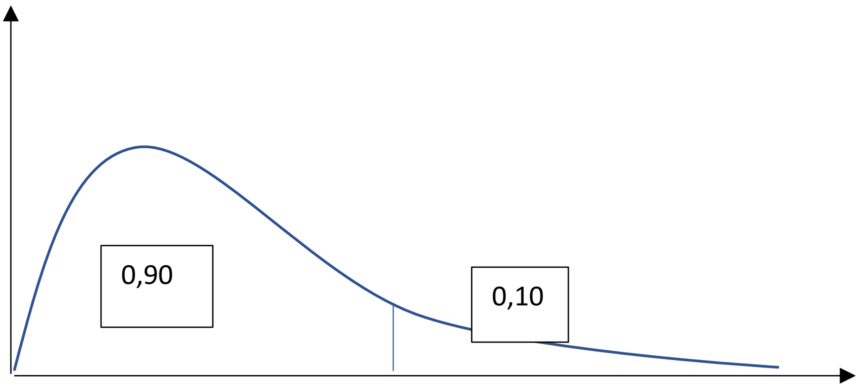
→*Indien de gegevens heuvelvormig verdeeld zijn.*



Hoeveel wijkt een waarneming af van de overige waarnemingen?

- Percentiel

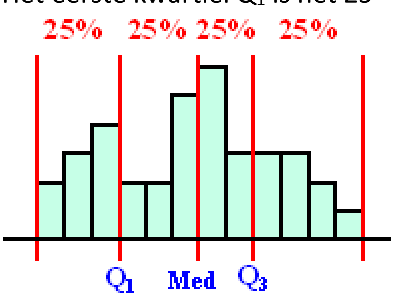
  →p% valt onder het P^e percentiel, en (100-p)% erboven valt



Maten van spreiding

- Kwartielen

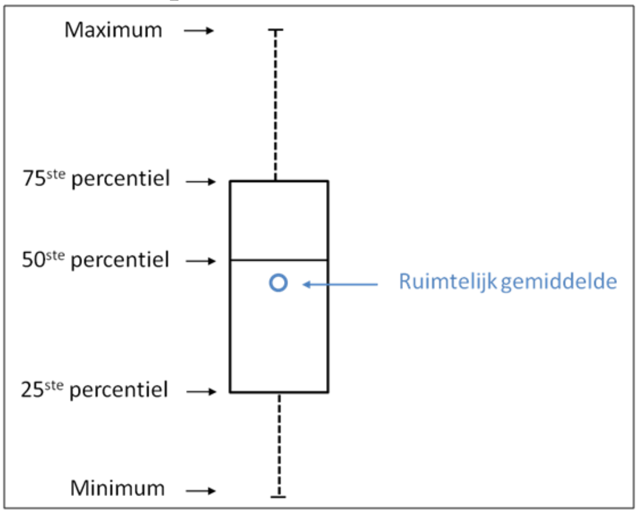
Het eerste kwartiel Q_1 is het 25^{ste} percentiel van de gegevensverzameling



- Interkwartielafstand

$$\boxed{KA = Q_3 - Q_1}$$

➢ Box-plots



Maten van relatieve afwijking

➢ Z-score

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

$$Z = \frac{X - \mu}{\sigma}$$
$$Populatie$$

$$Z = \frac{x - \bar{x}}{s}$$
$$Steekproef$$

Samenvattende tabel

	Waarnemingen afkomstig uit…	
Maatstaf	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	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
Mate van relatieve afwijking z-score	$z = \frac{X - \mu}{\sigma}$	$z = \frac{X - \bar{X}}{s}$

Uitschieters opsporen

- Uitschieters

  Wat zijn uitschieters?

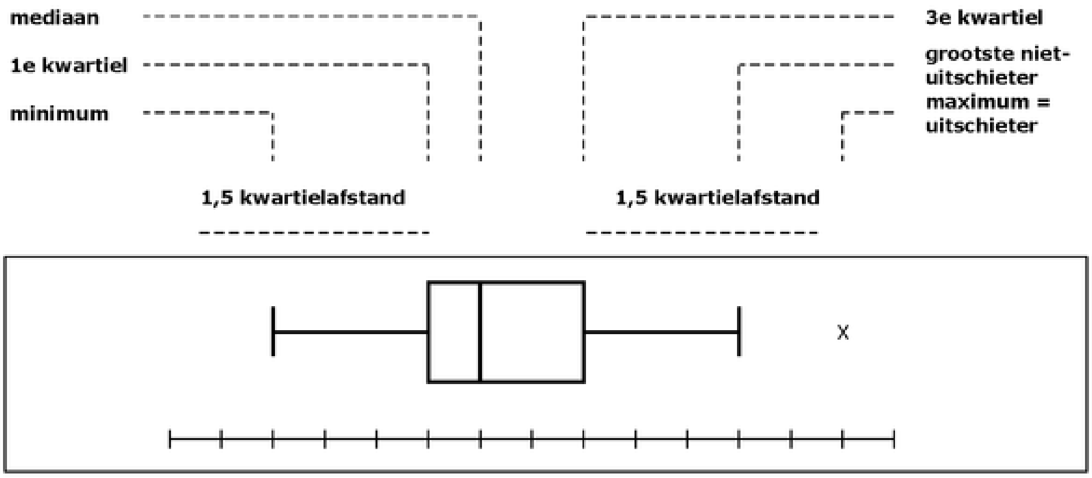
  →Onjuiste meetwaarden

  →zeldzame gebeurtenissen

- Boxplots

→ Tussen de binnenste en de buitenste omheining kunnen er uitschieters voorkomen

→ buiten de buitenste omheining zijn het bijna zeker altijd uitschieters



- Z-scores

Misleidende grafieken

- Schaalmanipulatie

Hoofdstuk 3: Kansrekening

Gebeurtenissen, uitkomstenruimten en kansen

Experiment: Een handeling of proces waarbij de uitkomst niet zekerheid kan worden voorspeld.
Een enkelvoudige gebeurtenis: De meest fundamentele uitkomst van een experiment.
Uitkomstenruimte: Is de verzameling van alle uitkomsten van een experiment. *(= S)*
Gebeurtenis: Een verzameling uitkomsten. Gebruik hoofdletter.

Kansregels:

- De kansen van alle uitkomsten moeten liggen tussen *0 & 1.*
- De kansen van alle uitkomsten in een uitkomstenruimte moeten opgeteld gelijk zijn aan *1.*

Stappen voor het berekenen van de kans:

- *Definieer* het experiment
- *Lijst* van alle *mogelijke uitkomsten* opstellen
- Men kent *kansen toe* aan de *uitkomsten*
- Bepaal *welke uitkomsten* die met de desbetreffende gebeurtenis *overeenkomen*
- *Tel de kansen op*

$\boldsymbol{Combinatieregel}$:

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

Met
$N = \quad elementen\ in\ populatie$
$n = hoeveel\ men\ er\ selecteert$

$\boldsymbol{Kansdefinitie\ van\ Laplace}$:

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

Wet van grote aantallen:

$$\boxed{P(X) = \lim_{n \to +\infty} \left(\frac{n_x}{n}\right)}$$

$$Waarbij: \begin{cases} n_X = het\ aantal\ keer\ dat\ de\ uitkomst\ x\ optreedt \\ n = aantal\ herhalingen\ van\ het\ experiment \end{cases}$$

Verenigingen en doorsneden:

- Vereniging
  →van twee gebeurtenissen A en B is de gebeurtenis als de beide gevallen optreden

$$A \cup B = als\ A\ of\ B\ of\ beide$$

- Doorsnede
  →van twee gebeurtenissen is als de gebeurtenis tegelijk optreden. De uitkomsten behoren zowel tot de verzameling A als B

$$A \cap B = A\ en\ B$$

Complementaire gebeurtenissen

→het complement van een gebeurtenis A is de kans dat de gebeurtenis A niet optreedt.

Het optellen van complementaire gebeurtenissen

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

De somregel en disjuncte gebeurtenissen

$$\boldsymbol{Somregel\ v\acute{o}or\ kansen}$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Disjuncte gebeurtenissen
  $Als\ A \cap B\ niet\ bestaat, dus\ als\ er\ geen\ gemeenschappelijke\ uitkomsten\ heeft$

  Kans op de vereniging van twee disjuncte gebeurtenissen:

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

Voorwaardelijke kans

$$\boldsymbol{Voorwaardelijke\ kans}$$
$$P(A|\ B) = \frac{P(A \cap B)}{P(B)} \quad met\ P(B) \neq 0$$

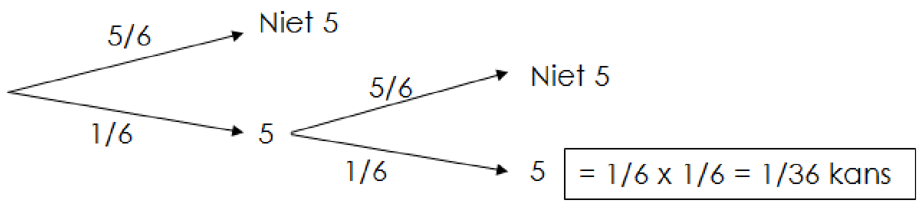
Productregel en onafhankelijke gebeurtenissen

- Productregel

$$P(A \cap B) = P(A)P(B/A)$$

$$P(A \cap B) = P(B)P(A/B)$$

- Kansboom



- Onafhankelijke gebeurtenissen

$$\begin{aligned}&Ze\ zijn\ onafhankelijk\ indien\\ &P(A/B) = P(A)\ of\\ &P(B/A) = P(B)\end{aligned}$$

De kans op de doorsnede van twee onafhankelijke gebeurtenissen:

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

$$\textbf{Regel van Bayes}:$$

$$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^C)}$$

$hieruit\ volgt$:

$$\boxed{P(B/A) = \frac{P(A/B)P(B)}{P(A/B)P(B) + P(A/B^C)P(B^C)}}$$

Hoofdstuk 4: Discrete kansverdelingen

<table>
<tr><th colspan="2">Stochastische variabelen
→Variabele die numerieke waarden aanneemt bij de toevallige uitkomsten van een experiment
→Hangt af van het toeval
→Bij elke uitkomst wordt één en slechts één waarde aangenomen</th></tr>
<tr><td>Discrete stochastische variabelen</td><td>Continue stochastische variabelen.</td></tr>
<tr><td colspan="2"><u>Definitie</u></td></tr>
<tr><td>→ Kunnen slecht een <u>eindig aantal waarden</u> aannemen. Bv: aantal ogen van een dobbelsteen, # verkopen per week
→Telbaar</td><td>→ Kunnen een <u>oneindig aantal waarden</u> aannemen</td></tr>
<tr><td colspan="2"><u>Eigenschappen van kansverdelingen</u></td></tr>
<tr><td>1. $P(x) \geq 0 \; voor \; alle \; waarden \; van \; x$
2. $\sum_x p(x) = 1$</td><td></td></tr>
</table>

Discrete stochastische variabelen

<table>
<tr><th colspan="2"><u>Synthesematen van een discrete stochastische variabele</u></th></tr>
<tr><td>Verwachtingswaarde</td><td>→Is het gewogen gemiddelde van de mogelijke waarden van de variabele
$$\mu = E(x) = \sum x\, p(x)$$</td></tr>
<tr><td>Variantie</td><td>→Is het gewogen gemiddelde van de gekwadrateerde afwijkingen t.o.v. de verwachtingswaarde
$$\sigma^2 = E[(x-\mu)^2] = \sum_x (x-\mu)^2\, p(x)$$</td></tr>
<tr><td>Standaardafwijking</td><td>→Is de positieve vierkantswortel van de variantie
$$\sigma = \sqrt{\sum_x (x-\mu)^2 \; p(x)}$$</td></tr>
</table>

- Synthesematen van een stochastische variabele

	Regel van Tsjebysjev	**Empirische regel**
Geldt voor:	*Elke kansverdeling*	*Heuvelvormige en symmetrische kansverdelingen*
$P(\mu - \sigma < x < \mu + \sigma)$	≥ 0	≈ 0,68
$P(\mu - 2\sigma < x < \mu + 2\sigma)$	≥3/4 =75%	≈ 0,95
$P(\mu - 3\sigma < x < \mu + 3\sigma)$	≥8/9 =88.9%	≈ 0,997

Discrete kansverdelingen: Binomiale verdeling

→Variabele die numerieke waarden aanneemt bij de toevallige uitkomsten van een experiment
→Bij elke uitkomst wordt één en slechts één waarde aangenomen

➢ Kenmerken

Binomiaal experiment:

- Maar 2 uitkomsten: succes of geen succes
- Rij van *n* identieke deelexperimenten (=Trial)
- Elke trial heeft twee 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
  → Het aantal successen van een binomiaal experiment noemt men een **binomiale stochastische variabele**

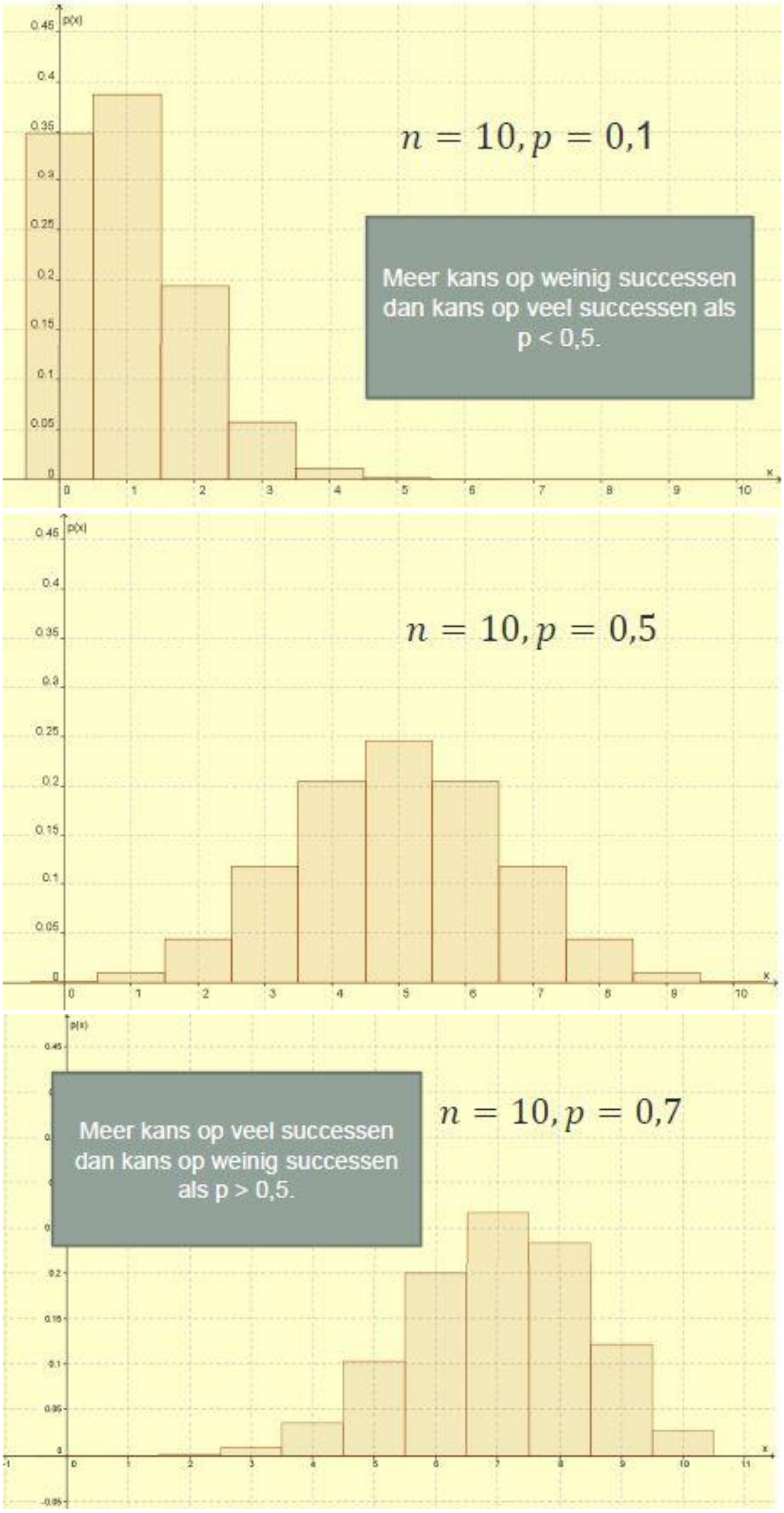
Opmerking: een experiment dat niet binomiaal is is bijvoorbeeld aselect iemand kiezen uit een groep mensen zonder terugzetting. Dit is omdat het niet onafhankelijk is en omdat de kans elke keer verandert. Dit is altijd behalve als n<5% van N

$$\boldsymbol{Binomiale\ verdeling}$$
$$p(k) = \binom{n}{k}(p^k)(1-p)^{n-k}$$

Met:
$k = kans\ op\ succes$
$n - k = kans\ op\ mislukking$
$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$

$Verwachtingswaarde: \mu = np$
$Variantie: \sigma^2 = n.p.q$
$q = (1-p)$
$Standaardafwijking: \sigma = \sqrt{npq}$

- Binomiaal kanshistogram:



- Eigenschappen

Eigenschappen van een binomiale verdeling	
Verwachtingswaarde	$\mu = \sum_{x=1}^{n} x\,p(x) = np$
Variantie	$\sigma^2 = \sum_{x=1}^{n} (x-np)^2\,p(x) = n\,p(1-p)$
Standaardafwijking	$\sigma = \sqrt{n\,p(1-p)}$

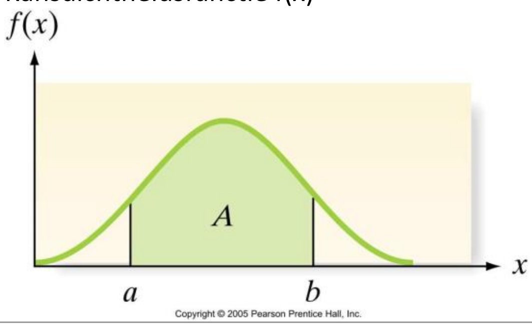
- Gebruik van de tabel Cumulatieve kansen

p =			0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99
n = 5	k =	0	0,951	0,774	0,590	0,328	0,168	0,078	0,031	0,010	0,002	0,000	0,000	0,000	0,000
		1	0,999	0,977	0,919	0,737	0,528	0,337	0,188	0,087	0,031	0,007	0,000	0,000	0,000
		2	1,000	0,999	0,991	0,942	0,837	0,683	0,500	0,317	0,163	0,058	0,009	0,001	0,000
		3	1,000	1,000	1,000	0,993	0,969	0,913	0,813	0,663	0,472	0,263	0,081	0,023	0,001
		4	1,000	1,000	1,000	1,000	0,998	0,990	0,969	0,922	0,832	0,672	0,410	0,226	0,049
n = 6	k =	0	0,941	0,735	0,531	0,262	0,118	0,047	0,016	0,004	0,001	0,000	0,000	0,000	0,000
		1	0,999	0,967	0,886	0,655	0,420	0,233	0,109	0,041	0,011	0,002	0,000	0,000	0,000
		2	1,000	0,998	0,984	0,901	0,744	0,544	0,344	0,179	0,070	0,017	0,001	0,000	0,000
		3	1,000	1,000	0,999	0,983	0,930	0,821	0,656	0,456	0,256	0,099	0,016	0,002	0,000
		4	1,000	1,000	1,000	0,998	0,989	0,959	0,891	0,767	0,580	0,345	0,114	0,033	0,001
		5	1,000	1,000	1,000	1,000	0,999	0,996	0,984	0,953	0,882	0,738	0,469	0,265	0,059
n = 7	k =	0	0,932	0,698	0,478	0,210	0,082	0,028	0,008	0,002	0,000	0,000	0,000	0,000	0,000
		1	0,998	0,956	0,850	0,577	0,329	0,159	0,063	0,019	0,004	0,000	0,000	0,000	0,000
		2	1,000	0,996	0,974	0,852	0,647	0,420	0,227	0,096	0,029	0,005	0,000	0,000	0,000

Bijvoorbeeld als x binomiaal verdeeld is met n=5 en p=0,40
→dan is $p(x \leq 4) \approx 0{,}990$

<u>Hoofdstuk 5: continue kansverdeling</u>

<u>Soorten kansverdeling</u>

<table>
<tr><th></th><th>Discrete kansverdeling</th><th>Continue kansverdeling</th></tr>
<tr><td></td><td></td><td>Kansdichtheidsfunctie f(x)

</td></tr>
<tr><td>Eigenschappen</td><td></td><td>$$\int_{-\infty}^{+\infty} f(x)dx = 1$$
$$P(a \leq x \leq b) = \int_a^b f(x)dx$$
$$\begin{aligned} P(a \leq x \leq b) &= P(a < x < b) \\ &= P(a < x \leq b) \\ &= P(a \leq x < b) \end{aligned}$$</td></tr>
</table>

- De normale verdeling

$$Kansdichtheidsfunctie\ f(x)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})} \qquad met\ x\ \epsilon]-\infty.+\infty[$$

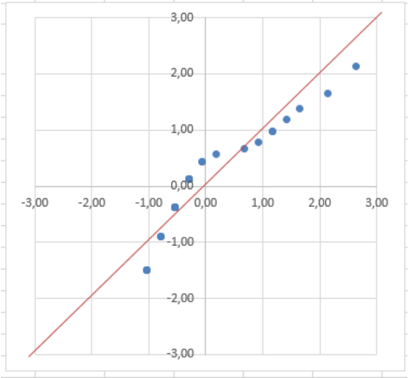
- Standaardnormale verdeling

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad met\ x\ \epsilon]-\infty.+\infty[$$

$$\mu = 0$$
$$\sigma = 1$$

$z = \frac{x-\mu}{\sigma}$ is het aantal standaardafwijkingen tussen $x\ en\ \mu$

Zijn gegevens afkomstig uit een normale verdeling?

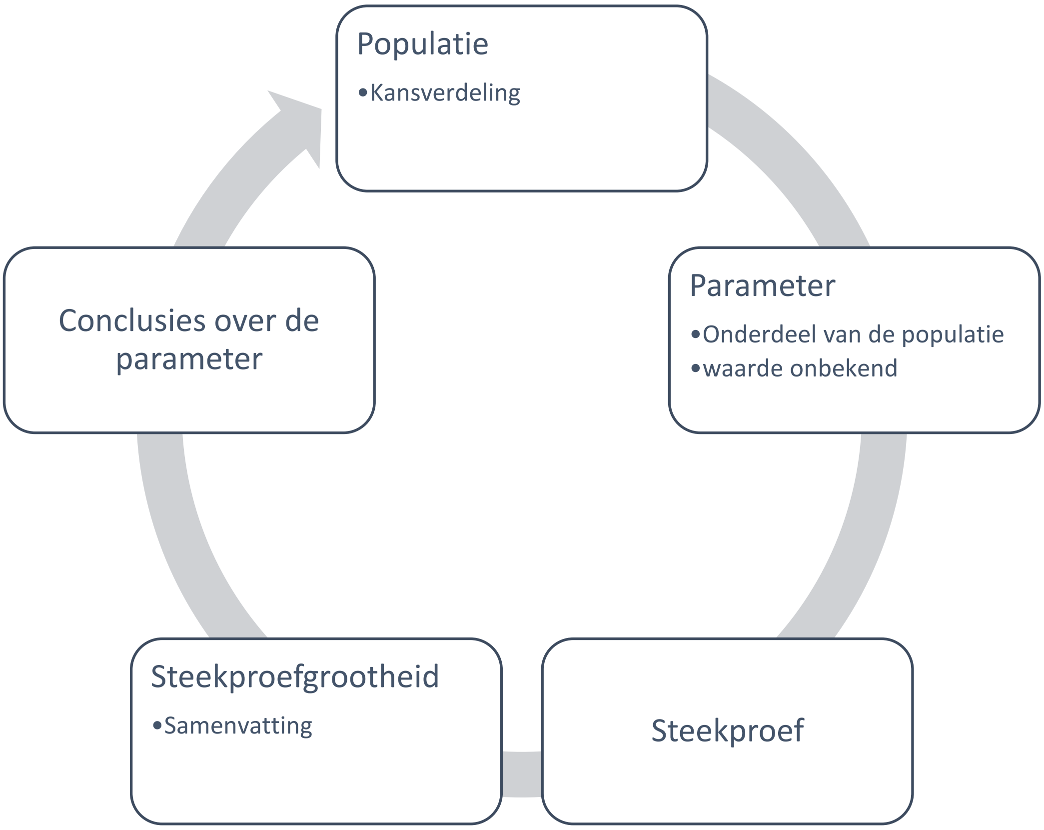
- Maak een histogram en kijk of deze symmetrisch is
- Intervallen berekenen:
  - $\overline{x} \pm s \leftrightarrow 68\%\ van\ de\ gegevens$
  - $\overline{x} \pm 2s \leftrightarrow 95\%\ van\ de\ gegevens$
  - $\overline{x} \pm 3s \leftrightarrow 100\%\ van\ de\ gegevens$
- Normaliteitsplot:
  - Kijken of de punten ongeveer allemaal op een rechte lijn liggen
  - 

Het benaderen van een binomiale verdeling door een normale verdeling

1. Stel de waarde van n en p vast
2. Bereken het interval $\mu \pm 3\sigma = np \pm 3\sqrt{npq}$
   a. Met $q = (1 - p)$
   b. Moet $\in [0, n]$
3. Druk de binomiale kans uit in de vorm $P(x \leq a)\ of\ P(x \leq b) - P(x \leq a)$
4. De continuïteitscorrectie is $(a + 0{,}5)$ dus
$$z = \frac{(a + 0{,}5) - \mu}{\sigma}$$
5. Schets de benaderde nominale verdeling en maak gebruik van de tabel om de bijhorende kansen te berekenen.

Hoofdstuk 6: Verdelingen van steekproefgrootheden

Inleiding



De verdeling van de steekproefgrootheid

De **steekproefgrootheid** heeft een kansverdeling: *de uitkomst van een steekproef is namelijk niet zeker en hangt van het toeval af.*

De **steekproefverdeling** van een steekproefgrootheid die voor een steekproef van *n* metingen wordt berekend, is de kansverdeling van deze steekproefgrootheid

De centrale limietstelling

Wordt een steekproef genomen uit een populatie met een normale verdeling, dan zal de kansverdeling van $\overline{x}$ een normale verdeling zijn.

Eigenschappen:

- $\mu_{\overline{x}} = E(\overline{x}) = \mu$

- $Standaardafwijking: \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

$$\frac{De\ standaardafwijking\ van\ de\ populatie\ waaruit\ de\ steekproef\ genomen\ wordt}{Vierkantswortel\ uit\ de\ grootte\ van\ de\ steekproef}$$

De centrale limietstelling

Indien een aselecte steekproef met *n* waarnemingen die uit een willekeurige populatie worden genomen met verwachtingswaarde µ en standaardafwijking σ, als *n* groot genoeg is zal de kansverdeling van $\overline{x}$ een normale verdeling hebben . Hoe groter de steekproef, des te beter de benadering van de kansverdeling van $\overline{x}$

$$\mu_{\overline{x}} = \mu$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Indien n tov van N meer dan 5% bedraagt, moet men voor $\sigma_{\overline{x}}$ nog eens vermenigvuldigen met $\sqrt{\frac{(N-n)}{(N-1)}}$