Samenvatting

-OIMB-

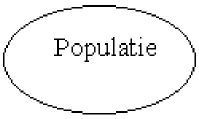
Inhoudsopgave

NOVIO MAGUM
ESV
1989
PECUNIA NON OLET

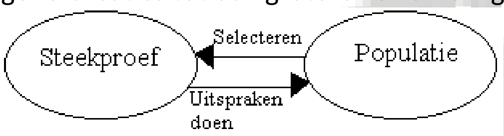
Bovemij Crowe EY Grant Thornton KPMG pwc

College 1: (Hoofdstuk 1 en 2)

Beschrijvende Statistiek: Gebruikt numerieke en grafische methoden om patronen in een gegevensverzameling te ontdekken, om informatie die uit de gegevensverzameling kan worden gewonnen samen te vatten, en om deze informatie op een overzichtelijke manier te presenteren.



Verklarende Statistiek (Inductieve of Toetsende Statistiek): Gebruikt steekproefgegevens voor het maken van schattingen, het nemen van beslissingen en het doen van voorspellingen of voor andere generalisaties tot een grotere verzameling gegevens.



Populatie: Een verzameling eenheden (meestal personen, objecten, transacties of gebeurtenissen) die we willen bestuderen.
Variabele: Kenmerk of eigenschap van een individuele eenheid van een populatie.
Meten: Het proces waarbij we getallen toekennen aan variabelen van individuele eenheden van een populatie.
Steekproef: Een deelverzameling van de eenheden van een populatie.
Statistische gevolgtrekking: Schatting, voorspelling of een andere generalisatie met betrekking tot de populatie die gebaseerd is op informatie uit een steekproef.
Betrouwbaarheidsmaat: Uitspraak (meestal kwantitatief) over de mate van onzekerheid die gepaard gaat met een statistische gevolgtrekking.

4 elementen beschrijvende statistiek:
- Relevante populatie.
- Een of meer variabelen.
- Tabellen, grafieken of numerieke hulpmiddelen om een samenvatting te geven.
- Conclusies met betrekking tot de gegevens, gebaseerd op de patronen die naar voren zijn gekomen.

5 elementen verklarende statistiek:
- Relevante populatie.
- Een of meer variabelen.
- Steekproef van populatie-eenheden.
- Gevolgtrekking over de populatie, gebaseerd op informatie in de steekproef.
- Een betrouwbaarheidsmaat voor de gevolgtrekking.

Proces: Een reeks acties of operaties die input omzet in output. Een proces produceert of genereert output in de loop van de tijd.
Blackbox: Een proces waarvan de operaties of acties onbekend of niet gespecificeerd zijn.
Kwantitatieve gegevens: Meetwaarden die worden geregistreerd op een van nature voorkomende numerieke schaal.
Kwalitatieve gegevens: Meetwaarden die niet op een natuurlijk voorkomende numerieke schaal kunnen worden gemeten; ze kunnen alleen worden geclassificeerd in een categorie uit een groep categorieën.

Gegevens verkrijgen uit:

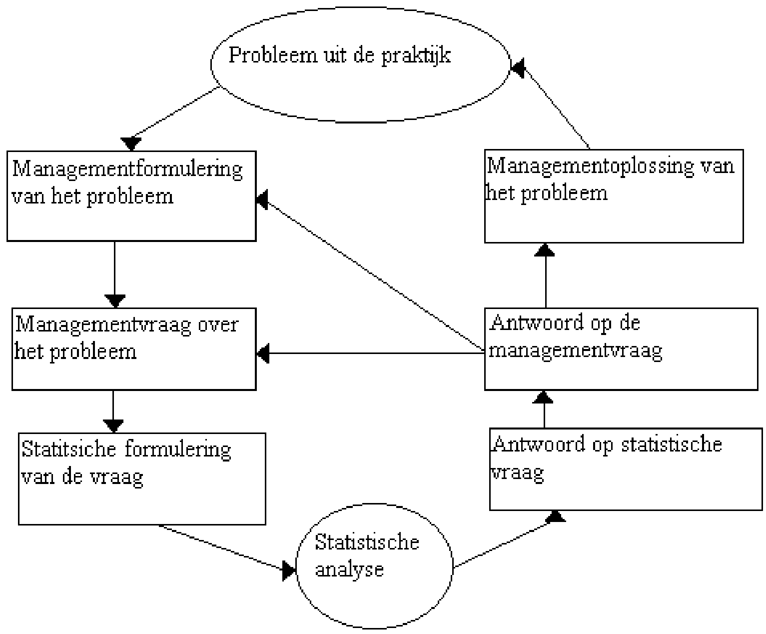
- Gepubliceerde bron.
- Experiment.
- Enquête.
- Observatie.

Onderzoek:

- Onderzoeksvraag
- Onderzoekseenheid
- Variabele (meetniveau)
- Datamatrix
- Relaties tussen variabelen
- Steekproef *
- Statistische schatting *
- Betrouwbaarheidsmeting (significantie) *

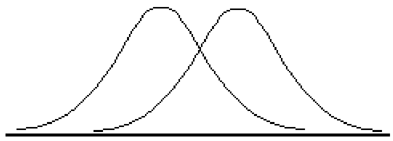
(* Inductieve/Toetsende/Verklarende Statistiek)

Statistiek: Verzamelen, classificeren, samenvatten, organiseren, analyseren, interpreteren van numerieke gegevens.
Statistisch denken: Omvat het toepassen van rationeel denken en van de wetenschap statistiek om kritisch gegevens en gevolgtrekkingen te kunnen beoordelen. Een essentieel punt bij dit gedachteproces is het feit dat er variatie bestaat in populaties en procesgegevens.

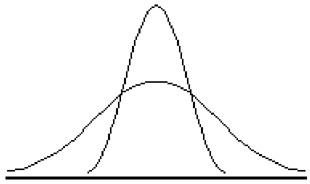


Centrale tendentie:
Gemiddelde
Mediaan
Modus

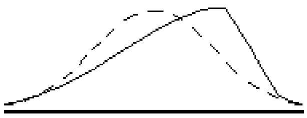
Bovemij Crowe EY Grant Thornton KPMG pwc



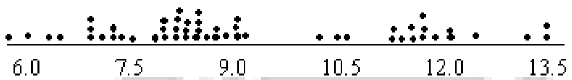
Spreiding/Variatie:
Bereik
Kwartielafstand
Variantie
Standaard Deviatie



Vorm:
Scheefheid



Frequentie per klasse: Het aantal waarnemingen in de gegevensverzameling dat tot een bepaalde klasse behoort.
Relatieve frequentie per klasse: De frequentie per klasse gedeeld door het aantal waarnemingen in de gegevensverzameling.
Pareto-analyse: Het indelen van items in categorieën en het bepalen welke categorieën de meeste waarnemingen bevatten.
Puntendiagram: De waarde van elke meting in de gegevensverzameling wordt door een punt op de horizontale schaal aangegeven.



Stam-en-bladdiagram: In dit diagram is de stam het deel van de meetwaarde (percentage) links van de decimaalkomma, terwijl het resterende deel, rechts van de decimaalkomma, het blad is.

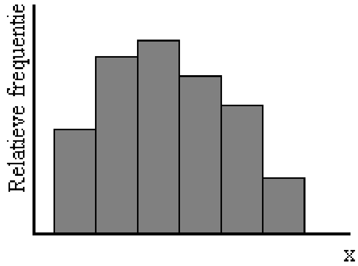
Stem	Leaves
5	269
6	055568999
7	1122457789
8	001222458
9	02455679
10	1556
11	137
12	
13	255

Leaf digit unit = 0,1

Maximum = 13,5
Minimum = 5,2
Mediaan = 8,05

Bovemij Crowe EY GrantThornton KPMG pwc

Histogram: De horizontale as die de meetwaarde (percentage) aangeeft is verdeeld in intervallen van gelijke grootte.



Sommatienotatie: Deze sommatiemethode sommeert de metingen van de variabele die rechts van het sommatieteken staat, beginnend met de 1e meting en eindigend met de ne meting.

$$\sum_{i=1}^{n} i \qquad \sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

Gemiddelde: De som van de metingen gedeeld door het aantal metingen in de gegevensverzameling.

Steekproefgemiddelde ($\bar{x}$):

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

Populatiegemiddelde (μ)

Mediaan (m): Het middelste getal wanneer de meetwaarden in opklimmende (of dalende) orde worden gerangschikt.

- n = oneven, mediaan = middelste getal.
- n = even, mediaan = gemiddelde van de twee middelste getallen.

Modus: De meting die het vaakst voorkomt in de gegevensverzameling.
Bereik: De grootste meetwaarde minus de kleinste meetwaarde (maximum – minimum = bereik).
Steekproefvariantie (Var): (in het geval van een steekproef met n meetwaarden) De som van de gekwadrateerde afwijkingen van het gemiddelde, gedeeld door (n – 1).

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{(n-1)} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2}{(n-1)}$$

Standaarddeviatie van de steekproef (s/σ): Positieve wortel uit de steekproefvariantie.

Bovemij Crowe EY GrantThornton KPMG pwc

$s / \sigma = \sqrt{Var} = \sqrt{s^2} = \sqrt{\sigma^2}$

Empirische regel (spreiding):

- Ongeveer 68 % van de meetwaarden zal binnen een standaarddeviatie van het gemiddelde vallen, dat wil zeggen binnen het interval ($\bar{x}$ - s, $\bar{x}$ + s) voor steekproeven en ($\mu - \sigma$, $\mu + \sigma$) voor populaties.
- Ongeveer 95 % van de meetwaarden zal binnen twee standaarddeviaties vallen, dat wil zeggen binnen het interval ($\bar{x}$ - 2s, $\bar{x}$ + 2s) voor steekproeven en ($\mu - 2\sigma$, $\mu + 2\sigma$) voor populaties.
- Ongeveer 99,7 % van de meetwaarden (vrijwel alle meetwaarden) zal binnen drie standaarddeviaties vallen, dat wil zeggen binnen het interval ($\bar{x}$ - 3s, $\bar{x}$ + 3s) voor steekproeven en ($\mu - 3\sigma$, $\mu + 3\sigma$) voor populaties.

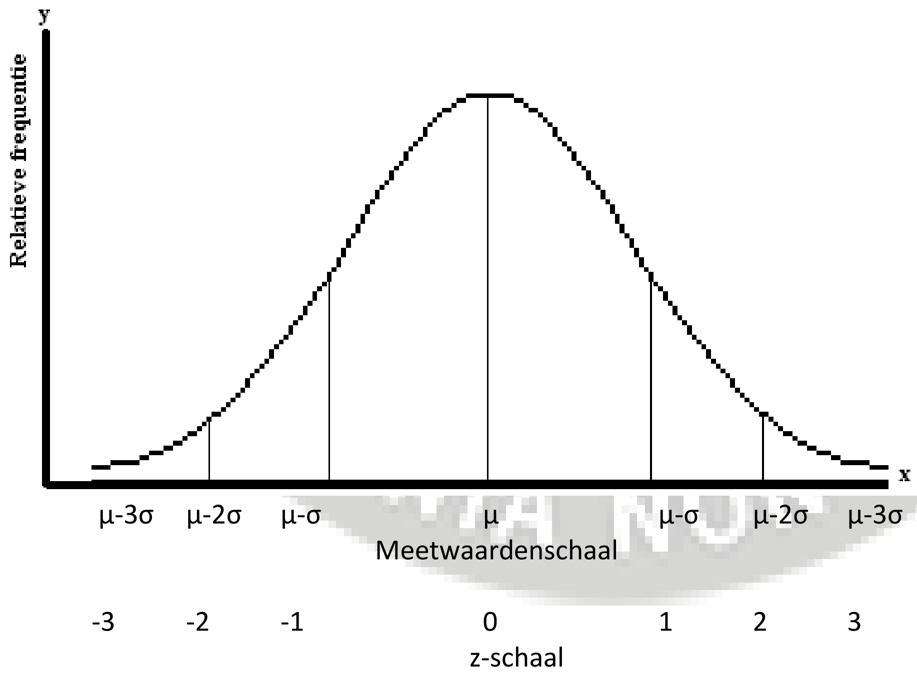
Percentiel: Voor elke verzameling van n meetwaarden (gesorteerd in klimmende of dalende volgorde), is het p^e percentiel een getal zodanig dat p % van de meetwaarden onder het p^e percentiel valt, en $(100 - p^e)$ eronder valt.

z-score: Vertegenwoordigt de afstand tussen een gegeven meetwaarde x en het gemiddelde, uitgedrukt in standaarddeviaties.

z-score (steekproef) = $\dfrac{x - \bar{x}}{s}$

z-score (populatie) = $\dfrac{x - \mu}{\sigma}$

- Ongeveer 68 % van de meetwaarden zal een z-score tussen –1 en 1 hebben.
- Ongeveer 95 % van de meetwaarden zal een z-score tussen –2 en 2 hebben.
- Ongeveer 99,7 % van de meetwaarden (vrijwel alle meetwaarden) zal een z-score tussen –3 en 3 hebben.

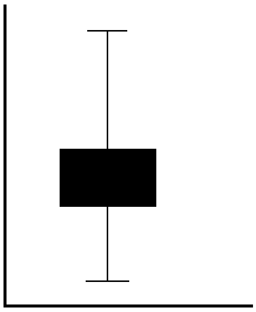


Uitbijter: Een waarneming (of meetwaarde) die uitzonderlijk groot of klein is ten opzichte van de overige waarden in een gegevensverzameling.

Oorzaken:

- De meetwaarde is incorrect waargenomen, geregistreerd of in de computer ingevoerd.
- De meetwaarde is afkomstig uit een andere populatie.
- De meetwaarde is correct, maar vertegenwoordigt een (toevallige) zeldzame gebeurtenis.

Box-plot: Weergave van de gegevensverzameling gebaseerd op kwartielen van een gegevensverzameling.



Eerste kwartiel = 25e percentiel van de gegevensverzameling.
Tweede kwartiel = 50e percentiel van de gegevensverzameling (mediaan).
Derde kwartiel = 75e percentiel van de gegevensverzameling.

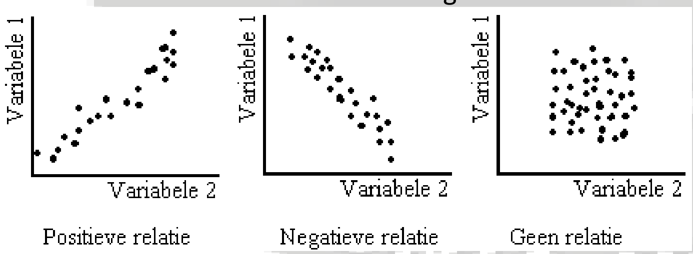
Kwartielafstand (KA) = Afstand tussen het eerste en het derde kwartiel:
KA = Q3 – Q1

Inner fence = Q1 – 1,5 (KA)
Outer fence = Q3 + 1,5 (KA)

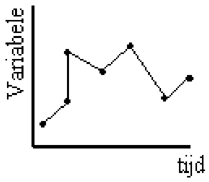
Vuistregels voor het vinden van uitbijters:

- **Box-plot:** Waarnemingen tussen de inner en de outer fences worden geacht mogelijke uitbijters te zijn. Waarnemingen die buiten de outer fence vallen, zijn bijna zeker uitbijters.
- **z-scores:** Waarnemingen met z-scores hoger dan 3 in absolute waarde worden als uitbijters beschouwd. (Voor sommige scheve verdelingen kunnen z-scores groter dan 1 in absolute waarde uitbijters zijn.)

Spreidingsdiagram (puntenwolk): Tweedimensionale grafiek die een bivariate relatie (relatie tussen twee kwantitatieve variabelen), met de waarden van de ene variabele langs de verticale as, en de waarden van de andere variabele langs de horizontale as.



Tijdreeks: Gegevens die in de loop van de tijd worden geproduceerd en geregistreerd.



Waarheid verdraaien:

- Afstand van X-as uitrekken op inpersen.
- Verticale as bij een ander punt dan 0 laten beginnen of onderbreking in verticale as.
- Breedte staven van de staafdiagram veranderen (dikker of dunner).
- Bijschrift veranderen.

➢ Verdeling (weglaten).

College 2: (Hoofdstuk 3, 4 en 5)
Experiment: Een handeling of een proces van waarnemen dat tot één enkele uitkomst leidt die niet met zekerheid kan worden voorspeld.
Uitkomst: De meest fundamentele uitkomst van een experiment.
Uitkomstenruimte (S): Verzameling van alle uitkomsten van dat experiment.
(wanneer je een munt 2 maal opgooit: S = [KK, MM, KM, MK])
Kans van een uitkomst: Een getal tussen 0 en 1 (inclusief deze waarden), dat de waarschijnlijkheid meet dat dit de uitkomst is als het experiment wordt uitgevoerd. (Relatieve frequentie)
Kansregels voor uitkomsten:

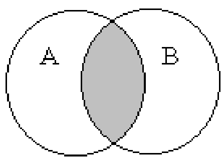
- ➢ Alle kansen van uitkomsten moeten liggen tussen 0 en 1 (inclusief deze twee waarden).
- ➢ De kansen van alle uitkomsten in een uitkomstenruimte moeten bij elkaar opgeteld gelijk zijn aan 1.

Gebeurtenis: Een specifieke verzameling uitkomsten.
Kansen op een gebeurtenis (P(A)): De kans op een gebeurtenis A wordt berekent door de kansen van de uitkomsten in de uitkomstenruimte voor A bij elkaar op te tellen.
Stappen voor het berekenen van de kans op een gebeurtenis:

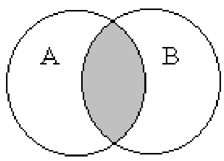
1. Definieer het experiment (beschrijf het waarnemingsproces en soort waarneming).
2. Maak een lijst van de uitkomsten.
3. Ken kansen toe aan de uitkomsten.
4. Bepaal welke verzameling uitkomsten tot de desbetreffende gebeurtenis hoort.
5. Tel de kansen van de uitkomsten bij elkaar op, om de kans op de gebeurtenis te verkrijgen.

Samengestelde gebeurtenissen:

- ➢ **Vereniging van twee gebeurtenissen:** Gebeurtenis die voorkomt als de eerste of de tweede of allebei de gebeurtenissen voorkomen als het experiment één keer wordt uitgevoerd. Bij gebeurtenis A of B of beide: **A U B**.
- ➢ **Doorsnede van twee gebeurtenissen:** Gebeurtenis die voorkomt als zowel de eerste als de tweede gebeurtenis voorkomt als het gesprek één keer wordt uitgevoerd. Bij gebeurtenis A en B: **A ∩ B**.

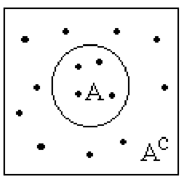


Het witte+grijze gebied is A U B.



Het grijze gebied is A ∩ B.

Complement van een gebeurtenis: De gebeurtenis dat de gebeurtenis niet voorkomt.
Complement van A geven we aan met A^C, de gebeurtenis die bestaat uit alle uitkomsten die niet in gebeurtenis A zitten.

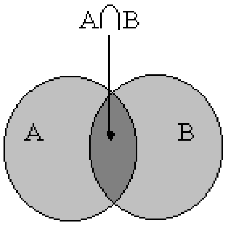


Optellen van kansen van complementaire gebeurtenissen:

Bovemij Crowe EY GrantThornton KPMG pwc

Som van de kansen van een complementaire gebeurtenis is gelijk aan 1, dat wil zeggen: P(A) + (P A^C) = 1 (Waarbij P staat voor kans)

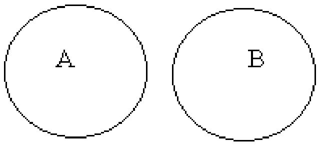
Somregel voor gebeurtenissen:
P (A U B) = P(A) + P(B) - A ∩ B



Totale gekleurde gebied is A ⋃ B.

Disjunct van een gebeurtenis: Als A ∩ B geen uitkomsten bevat, dat wil zeggen dat A en B geen gemeenschappelijke uitkomsten hebben.
P (A ∩ B) = 0

Kans van twee disjunctieve gebeurtenissen: Als twee gebeurtenissen A en B disjunctief zijn is de kans van de vereniging van A en B gelijk aan de som van de kansen van A en B; ofwel
P (A U B) = P(A) + P(B).
(deze formule is onjuist als de gebeurtenissen niet disjunctief zijn)



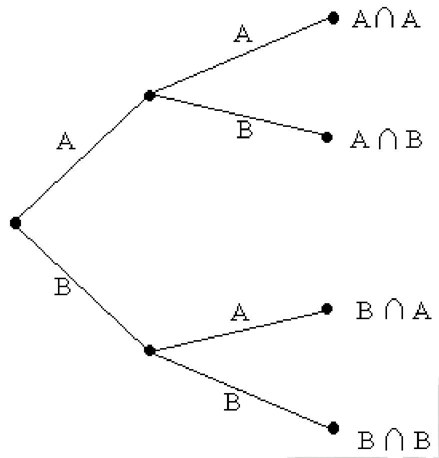
Onvoorwaardelijke kans: Geen speciale voorwaarden aangenomen, buiten die welke het experiment definiëren.
Voorwaardelijke kans: Een kans die rekening houdt met 'extra kennis'.
Formule voor voorwaardelijke kans: (dat gebeurtenis A voorkomt als gegeven is dat gebeurtenis B voorkomt)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$ (We veronderstellen dat P (B) ≠ 0)

Productregel voor kansen: Formule voor de kans op een doorsnede van gebeurtenissen A en B.
P (A ∩ B) = P (A) * P (A | B) ,of P (A ∩ B) = P (B) * P (A | B)

Kansboom:

Bovemij Crowe EY GrantThornton KPMG pwc



Onafhankelijke gebeurtenis als:

- Gebeurtenis die voorkomt, beïnvloed niet de kans op een andere gebeurtenis.
- Impliceert geen causaliteit.
- P (A | B) = P (A)
- P (A | B) = P (B)

Disjuctieve gebeurtenis is een afhankelijke gebeurtenis omdat P (A | B) ≠ P (A)

Kans van de doorsnede van twee onafhankelijke gebeurtenissen: Als de gebeurtenissen A en B onafhankelijk zijn, is de kans van de doorsnede van A en B gelijk aan het product van de kansen op A en B; ofwel
$P(A \cap B) = P(A) * P(B)$

Aselecte steekproef (willekeurige steekproef): Manier waarop n elementen uit de populatie worden geselecteerd, dat elke verzameling van n elementen in de populatie dezelfde kans heeft om geselecteerd te worden.
Random-getallengenerator: Hulpmiddel dat door onderzoekers wordt gebruikt om met willekeurige getallen een aselecte steekproef te genereren.

Random variabele: Een variabele die numerieke waarden aanneemt die horen bij de willekeurige uitkomsten van een experiment, waarbij één (en slechts één) waarde aan elke uitkomst wordt toegekend.

- **Discrete random variabelen:** random variabelen die een telbaar aantal waarden kunnen aannemen.
- **Continue random variabelen:** random variabelen die waarden kunnen aannemen die corresponderen met elk van de punten van één of meer intervallen.

Gemiddelde (verwachtingswaarde) van een discrete random variabele: (maat voor centrale tendentie)
$\mu = E(x) = \sum x p(x)$

Variantie van een discrete random variabele: (maat voor spreiding, verwachtingswaarde van de gekwadrateerde afstand van het gemiddelde)
$\sigma^2 \quad = \quad E[(x-\mu)^2] \quad = \quad \sum (x-\mu)^2 p(x)$

Standaarddeviatie van een discrete random variabele: Is gelijk aan de positieve wortel van de variantie.

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

	Emperische regel
	Geldt voor kansverdelingen die heuvelvormig en symmetrisch zijn
P $(\mu - \sigma < x < \mu - \sigma)$	$\approx$ 0,68
P $(\mu - 2\sigma < x < \mu - 2\sigma)$	$\approx$ 0,95
P $(\mu - 3\sigma < x < \mu - 3\sigma)$	$\approx$ 1,00

Dichotome resultaten: Resultaten met twee mogelijke alternatieven. (ja-nee of goedgekeurd-afgekeurd)

Binomiale random variabele: Variabele waarvan één van de twee mogelijke uitkomsten gerealiseerd wordt.

- Het experiment bestaat uit n identieke proeven.
- Er zijn slechts twee mogelijke uitkomsten van elke proef (A of B).
- De kans op de ene uitkomst (A) blijft hetzelfde van proef tot proef. De kans op de ene variabele (A) wordt aangeduid met 'p', en de kans op de andere variabele (B) wordt aangeduid met 'q'.
  $q = 1 - p$
- De proeven zijn onafhankelijk.
- De binomiale random variabele x is het aantal keren A in n proeven.

Gemiddelde, variantie en standaarddeviatie voor een binomiale rendom verdeling:

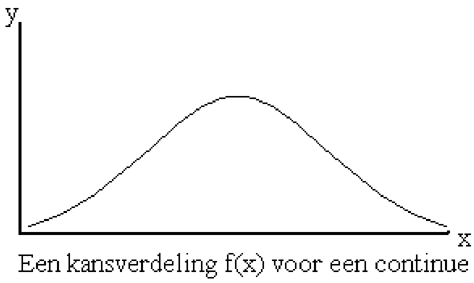
Gemiddelde: $\mu = n * p$

Variantie: $\sigma^2 = n * p * q$

Standaarddeviatie: $\sigma = \sqrt{(n * p * q)}$

Cumulatieve binomiale kansen: P $(x \leq k)$ → k = waarden, x = random variabele.

Continue random variabele: Gladde kromme, een functie van x, die wordt aangeduid met het symbool f(x) en wordt **kansdichtheidsfunctie (kdf)**, **frequentiefunctie** of **kansverdeling** genoemd.

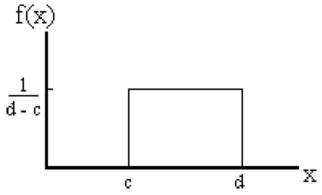


Een kansverdeling f(x) voor een continue random variabele

Een continue kansverdeling:

- Uniforme verdeling
- Normale verdeling
- Exponentiële verdeling
- Andere verdelingen

Uniforme kansverdeling: Continue random variabele die over het hele bereik van waarden even waarschijnlijk zijn bezitten een dergelijke kansverdeling. Een uniforme kansverdeling heeft een rechthoekige vorm. Ook wel 'randomness distribution' genoemd omdat een manier om een uniforme random variabele te genereren , bestaat uit het uitvoeren van een experiment waarin een punt willekeurig wordt gekozen tussen de punten c en d op de horizontale as.



Stel random variabele x heeft alleen waarde binnen het interval '$c \leq x \leq d$' ,

hoogte van het interval is gelijk aan → 1 / (d – c)

Totale oppervlakte gebied onder f(x)→

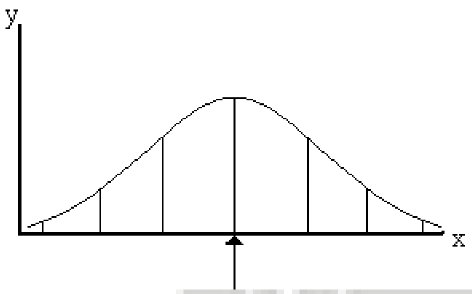
totale opp. rechthoek = basis * hoogte = $(d-c) * \left(\frac{1}{d-c}\right) = 1$

Kansverdeling, gemiddelde en standaarddeviatie van een uniforme random variabele x:

$f(x) = \frac{1}{d-c}$ $\quad c \leq x \leq d$

$\mu = \frac{c+d}{2}$ $\quad \sigma = \frac{d-c}{\sqrt{12}}$

Normale verdeling (bell curve): Normaal verdeelde variabele die een klokvormige kansverdeling heeft.

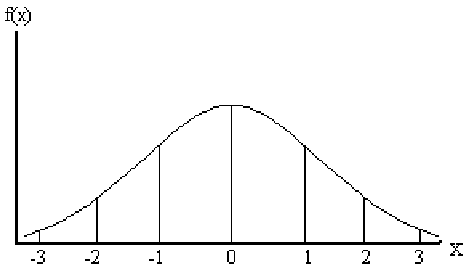


μ (gemiddelde/mediaan/modus)

Kansverdeling voor een normaal verdeelde variabele x:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-(1/2)[(x-\mu)/\sigma]^2}$$

μ = gemiddelde van de normaal verdeelde variabele
σ = standaarddeviatie
π = 3,14159… (constante)
e = 2,71828… (constante)



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

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

Standaard normale verdeling: Normale verdeling met $\mu = 0$ en $\sigma = 1$. Een random variabele met een standaardnormale verdeling, aangegeven met het symbool 'z', wordt een 'standaard normaal verdeelde variabele' genoemd.
Als x een normaal verdeelde variabele is met een gemiddelde 'μ' en standaarddeviatie 'σ', dan heeft de random variabele z, gedefinieerd door de formulel/:

$$z = \frac{x - \mu}{\sigma}$$

een standaardnormale verdeling. De waarde 'z' geeft het aantal standaarddeviaties tussen x en μ.

Stappen voor het bepalen van een kans die bij een normaal verdeelde variabele hoort:

1. Schets de normale verdeling en geef het gemiddelde van de random variabele x aan. Arceer het gebied dat overeenkomt met de kans die je wilt bepalen.
2. Reken de grenzen van het gearceerde gebied van x-waarden om naar waarden van de standaardnormaal verdeelde variabele z met de formule:
   $$z = \frac{x - \mu}{\sigma}$$
   Geef de berekende z-waarden aan onder de corresponderende x-waarden in je schets.
3. Gebruik de tabel (IV, appendix B) om de oppervlakte te vinden die met de z-waarden corresponderen.

Hoe men kan bepalen of gegevens afkomstig zijn uit een bij benadering normale verdeling:

1. Construeer een histogram of stam-en-bladdiagram voor de gegevens en bekijk de vorm van de grafiek. Als de gegevens een ongeveer normale verdeling hebben, zal de vorm van het histogram of van het stam-en-bladdiagram op de normale kromme lijken.
2. Bereken de intervallen $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$, en bereken het percentage meetwaarden dat binnen elk van deze intervallen valt. Als de gegevens ongeveer normaal verdeeld zijn, zullen deze percentages ongeveer gelijk zijn aan respectievelijk 68%, 95% en 100%.
3. Bepaal de kwartielafstand (KA) en de standaarddeviatie s, voor de steekproef, en de bereken vervolgens de verhouding KA/s. Als de gegevens ongeveer normaal verdeeld zijn is KA/s ≈ 13.
4. Construeer een normaliteitsplot voor de gegevens. Als de gegevens ongeveer normaal verdeeld zijn, zullen de punten (ongeveer) op een rechte lijn liggen.

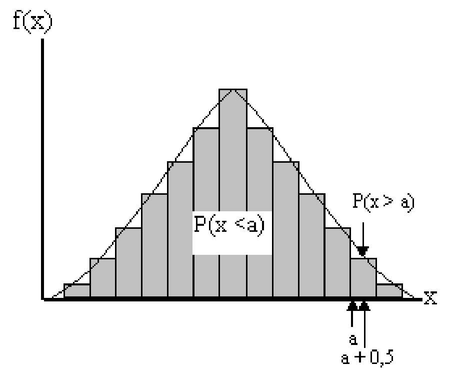
Normaliteitsplot: Spreidingsdiagram met de gesorteerde gegevenswaarden op de ene as, en de overeenkomende verwachte z-waarden van een standaardnormale verdeling op de andere as.

Continuïteitscorrectie: Dat wil zeggen dat we een correctie toepassen op de discrete verdeling, zodat deze door de continue verdeling kan worden benaderd.

Het gebruik van een normale verdeling om binomiale kansen te benaderen:

1. Bereken, nadat je n en p hebt bepaald voor een binomiale verdeling het interval
   $\mu \pm 3\sigma = np \pm 3\sqrt{npq}$
   Als het interval binnen het bereik van 0 en n ligt, zal de normale verdeling een goede benadering geven voor de kansen van de meeste binomiale gebeurtenissen.
2. Druk de binomiale kans die moet worden benaderd uit in de vorm $P(x \leq a)$ of $P(x \leq b) - P(x \leq a)$.
3. Voor elke relevante waarde a is de continuïteitscorrectie gelijk aan (a + 0,5), en de corresponderende standaardnormale z-waarde is

   $$z = \frac{(a + 0{,}5) - \mu}{\sigma}$$



4. Schets de benaderende normale verdeling en arceer het gebied dat correspondeert met de kans van de relevante gebeurtenis. Controleer of de rechthoeken die je in het gearceerde gebied hebt opgenomen, corresponderen met de kans op de gebeurtenis die je wilt benaderen. Gebruik de tabel (IV, Appendix B) en de z-waarden die je in stap 3 hebt berekend, om de oppervlakte van het gearceerde gebied te vinden. Dit is de benaderde kans van de binomiale gebeurtenis.

$$P(x \leq a) = P\left[z \leq \frac{(a + 0{,}5) - \mu}{\sigma}\right]$$ → Normale benadering van binomiale kans.

College 3: (Hoofdstuk 6 en 7)

Parameter (θ (= théta)): Numerieke beschrijvende maat van een populatie.
Steekproefgrootheid: Numerieke beschrijvende maar van een steekproef, wordt berekend uit waarnemingen in de steekproef.
Steekproefverdeling: (van een steekproefgrootheid uit een steekproef van n metingen is) de kansverdeling van een steekproefgrootheid.

Steekproefverdeling van $\bar{x}$ →

- Werkelijk gemiddelde van de steekproefverdeling ($\mu_{\bar{x}}$)

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

  0

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

- Werkelijke standaarddeviatie (standaardfout) van de steekproefverdeling ($\sigma_{\bar{x}}$)

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

  $\sigma_{\bar{x}} > \sigma$

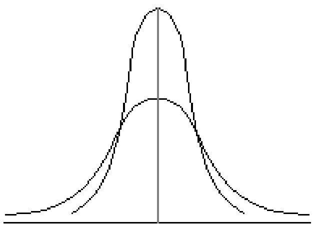
- z-score van de steekproefverdeling

  $$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

  Met de z-score kan de kans afgeleid worden, gebruik makend van de z-tabel (appendix B, tabel IV).

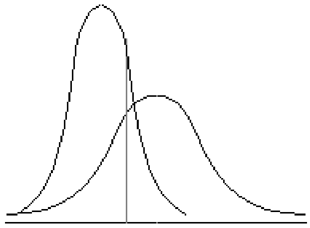
Puntschatter: (van een populatieparameter) een regel/formule die ons verteld hoe we de steekproefgegevens moeten gebruiken om één enkel getal te berekenen dat gebruikt kan worden als een schatting van de populatieparameter. (per definitie onbetrouwbaar)
Schattingsfout: Het verschil tussen de schatting en de werkelijke waarde van de populatieparameter.

Zuivere schatter: Een steekproefgrootheid waarvan de steekproefverdeling een gemiddelde heeft dat gelijk is aan de populatieparameter die deze grootheid moet schatten.

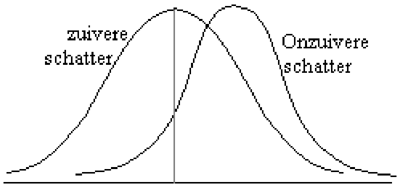


θ

Onzuivere schatter: Een steekproefgrootheid waarvan de steekproefverdeling een gemiddelde heeft dat niet gelijk is aan de populatieparameter die deze grootheid moet schatten.



θ

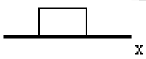
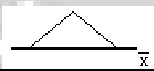
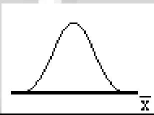
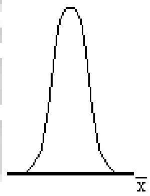
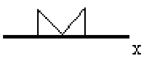
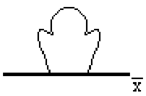
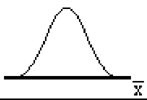
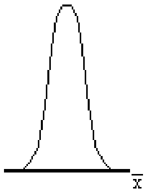
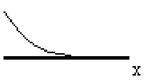
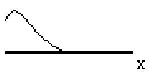
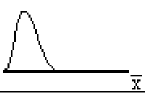
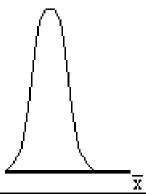


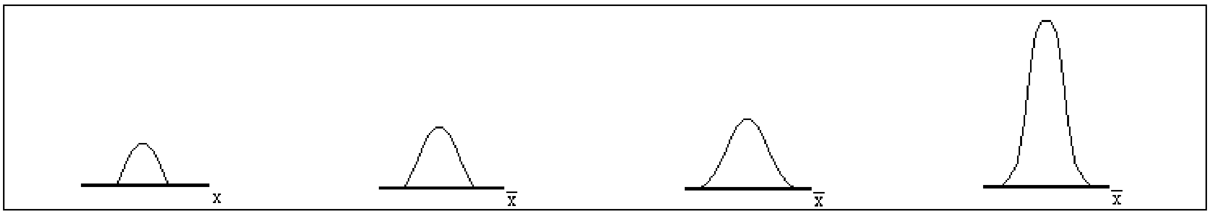
θ

- ➢ Hoe groter de steekproefomvang (n), des te efficiënter (dichter bij de werkelijkheid) is de steekproefverdeling.
- ➢ Hoe kleiner de standaarddeviatie, des te efficiënter is de steekproefverdeling.
- ➢ De standaarddeviatie van de steekproefverdeling neemt af als de steekproefomvang toeneemt.

Als een aselecte steekproef van n waarnemingen uit een populatie met een normale verdeling wordt genomen, zal de steekproefverdeling van $\bar{x}$ een normale verdeling zijn.

Als n groot genoeg is, zal de steekproefverdeling van $\bar{x}$ uit een aselecte steekproef van n waarneming uit een willekeurige populatie met gemiddelde μ en standaarddeviatie σ een normale verdeling benaderen met gemiddelde en standaardeviatie. Hoe groter de steekproef is, des te beter zal de normale benadering van de steekproefverdeling van $\bar{x}$ zijn. <u>(Centrale limietstelling)</u>

Orginele populatie	Steekproefverdeling van $\bar{x}$ voor n = 2	Steekproefverdeling van $\bar{x}$ voor n = 5	Steekproefverdeling van $\bar{x}$ voor n = 30
			
			
			



Intervalschatter (betrouwbaarheidsinterval): Formule die met behulp van steekproefgegevens een interval dat de populatieparameter schat berekend.
Betrouwbaarheidscoëfficiënt: De kans dat een aselect gekozen betrouwbaarheidsinterval de populatieparameter bevat.
Betrouwbaarheid: Betrouwbaarheidscoëfficiënt uitgedrukt als percentage.

α (coëfficiënt van onbetrouwbaarheid) = 1 - betrouwbaarheid

Betrouwbaarheid 100 (1 – α)	**α**	**α / 2**	**z α/2**
90 %	0,10	0,05	1,645
95 %	0,05	0,025	1,96
99 %	0,01	0,005	2,575

Bij σ = onbekend kan de standaarddeviatie van de steekproef gebruikt worden, mits n groot is.

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

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Interpretatie van een betrouwbaarheidsinterval voor een populatiegemiddelde: Bij een 100(1 – α)% betrouwbaarheidsinterval voor μ vormen, kunnen we zeggen dat we er voor 100(1 – α)% zeker van zijn dat μ tussen de ondergrens en de bovengrens van het betrouwbaarheidsinterval ligt. (Deze uitspraak geeft het vertrouwen in het schattingsproces weer, en niet zozeer het vertrouwen in het specifieke interval dat uit de steekproefgegevens is berekend)

Betrouwbaarheidsinterval (z-verdeling) = $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

T-verdeling: Lijkt op z-verdeling, het verschil zit in het feit dat de t-verdeling variabeler is dan de z-verdeling, de t-verdeling heeft dan ook 2 variabelen ($\bar{x}$ en s) en de z-verdeling slechts 1 ($\bar{x}$).
Vrijheidsgraden (n – 1): Afhankelijkheid van de variatie in de steekproefverdeling van t van de steekproefomvang.

Bij een kleinere n wordt in plaats van de z-verdeling gebruik gemaakt van de t-verdeling, waarbij n = n – 1.

Betrouwbaarheidsinterval (t-verdeling) = $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$
(aangenomen wordt dat de aselecte steekproef bij benadering normaal verdeeld is)

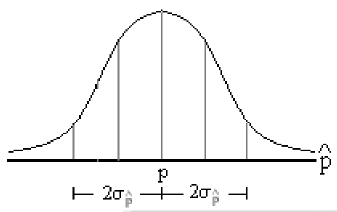
Dakjes (^) = geven aan dat de symbolen daaronder schattingen zijn.

Bovemij Crowe EY GrantThornton KPMG pwc

$\hat{p}$: Populatiefractie.

Steekproefverdeling van $\hat{p}$:

- Het gemiddelde van de steekproefverdeling van $\hat{p}$ is p; dat betekent dat $\hat{p}$ een zuivere schatter van p is.
- De standaarddeviatie van de steekproefverdeling van $\hat{p}$ is $\sqrt{pq / n}$; dat betekent dat $\sigma_{\hat{p}} = \sqrt{pq / n}$; waarbij $q = 1 - p$.
- Bij grote steekproeven is de steekproefverdeling van $\hat{p}$ bij benadering normaal. Voorwaarde is wel dat de z-waarden voor $\hat{p} \pm 3\sigma_{\hat{p}}$ tussen 0 en 1 in liggen.



Betrouwbaarheidsinterval voor een grote steekproef van p:

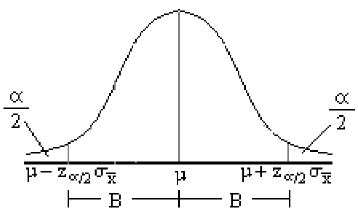
$\hat{p} \pm z_{\alpha/2}\, \sigma_{\hat{p}} \quad = \quad \hat{p} \pm z_{\alpha/2} \sqrt{pq / n} \quad \approx \quad \hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}\hat{q} / n}$

met $\hat{p} = x / n$ en $\hat{q} = 1 - \hat{p}$.

Als n groot genoeg is, kan $\hat{p}$ de waarde van p benaderen in de formule voor $\sigma_{\hat{p}}$.

Gecorigeerd (1 – α)% betrouwbaarheidsinterval voor een populatiefractie p:
$\tilde{p} \pm z_{\alpha/2} \sqrt{\tilde{p}(1 - \tilde{p}) / (n + 4)}$
Waarbij $\tilde{p} = (x + 2) / (n + 4)$ de gecorrigeerde steekproeffractie van de waarnemingen, waarbij x het aantal successen in de steekproef is, en n de steekproefomvang.

Begrenzing (B): Gebied waarbinnen de schatting van μ 100(1 – α) betrouwbaar is. De begrenzing B is gelijk aan de halve breedte van het betrouwbaarheidsinterval.



(bij een populatiefractie vervang je simpelweg μ door p en $\bar{x}$ door $\hat{p}$)

Bepaling van de steekproefomvang voor een 100(1 – α)% betrouwbaarheidsinterval voor μ:
Om μ te kunnen schatten binnen een begrenzing B met een 100(1 – α)% betrouwbaarheid, wordt de vereiste steekproefomvang als volgt bepaald:

$$B = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\underline{(z_{\alpha/2})^2\, \sigma^2}$

n = B²

(σ is vaak onbekend, deze kun je schatten door de standaarddeviatie (s) van de steekproef te gebruiken→ σ = s
of door een schatting te maken met behulp van het bereik → σ ≈ R / 4)

Bepaling van de steekproefomvang voor een 100(1 – α)% betrouwbaarheidsinterval voor p:
Om een binomiale kans p te kunnen schatten binnen een begrenzing B met een 100(1 – α)% betrouwbaarheid, wordt de vereiste steekproefomvang als volgt bepaald:

$B = z\,\alpha/2\ \sqrt{(pq / n)}$

$$n = \frac{(z\,\alpha/2)^2\, pq^2}{B^2}$$

(pq is vaak onbekend, deze kan geschat worden door de steekproeffractie van successen $\hat{p}$ te gbruiken uit een eerdere steekproef)

Vuistregel voor de eindige-populatiecorrectiefactor:
Gebruik de eindige-populatiecorrectiefactor als n / N = 0,05.

Enkelvoudige steekproef uit een eindige populatie van grootte N:

- **Schatting van het populatiegemiddelde**
  Geschatte standaardfout/standaarddeviatie:
  $$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$
  Benaderd 95% betrouwbaarheidsinterval: $\bar{x} \pm 2\hat{\sigma}_{\bar{x}}$

- **Schatting van de populatiefractie**
  Geschatte standaardfout/standaarddeviatie:
  $$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ \sqrt{\frac{N-n}{N}}$$
  Benaderd 95% betrouwbaarheidsinterval: $\hat{p} \pm 2\hat{\sigma}_{\hat{p}}$

- (De betrouwbaarheidsintervallen zijn benaderd, omdat gebruik wordt gemaakt van 2 om de waarde z0,025 = 1,96 te benaderen.

<u>**Steekproefonderzoek:**</u>

- **Steekproefonderzoek:** Gebruikte term voor het nemen van steekproeven van populaties.

- **Steekproefonderzoeksontwerpen:** Ontwerpen van hoe een steekproef zal worden uitgevoerd.

  Gestratificeerde steekproef: Als de steekproefeenheden die uit de populatie genomen zullen worden, fysiek gescheiden kunnen worden in twee of meer groepen steekproefeenheden (**= strata**), waarbij de variatie in respons binnen een stratum kleiner is dan de variatie binnen de gehele populatie.
  + geeft schatters met kleinere standaardfout (dan de enkelvoudige aselecte steekproef)
  + representativiteit hoger (dan de enkelvoudige aselecte steekproef)
  + kosten waarschijnlijk lager (dan de enkelvoudige aselecte steekproef)

Systematische steekproef: Systematisch proefpersonen selecteren voor een steekproef.
+ makkelijker (dan de enkelvoudige aselecte steekproef)
- systematische steekproefvertekening, er mogen geen cycli bestaan.

Gerandomiseerde responssteekproef: Steekproef waarbij de proefpersoon random vragen toegewezen krijgt.
+ eerlijk antwoord op gevoelige vragen; vragen waar men neigt naar een 'vals' antwoord.

- **Non-respons:** Steekproefeenheden in een steekproef geven geen steekproefwaarnemingen, kan leiden tot sterk vertekende resultaten. Komt voor bij bijvoorbeeld enquêtes via post of telefoon.

- **Fouten in steekproefonderzoek:**

  Steekproefhandeling: *Fout bij het nemen van de steekproef, bijvoorbeeld een deel van de populatie weglaten.*

  Non-interviews: *Informatie wordt slechts voor een deel van de steekproef verkregen, dit veroorzaakt problemen door de eventuele verschillen tussen de geïnterviewden en de niet-geïnterviewden.*

  Geschiktheid van de respondent: *Bij uitval van een respondent is de vervangende respondent niet altijd even goed.*

  Het begrijpen van de concepten: *Sommige respondenten begrijpen niet goed wat er met de vragen bedoeld wordt.*

  Gebrek aan kennis: *Respondenten beschikken niet altijd over de gevraagde informatie of proberen deze niet te verkrijgen.*

  Verbergen van de waarheid: *Respondenten antwoorden niet altijd naar waarheid, bijvoorbeeld uit angst of wantrouwen.*

  Suggestieve vragen: *De formulering van de vraag leidt de respondent een bepaalde kant op bij het beantwoorden.*

  Verwerkingsfouten: *Fouten bij het coderen, invoeren, etc.*

  Conceptuele problemen: *Wat wordt verlangd, kan verschillen van wat het onderzoek in feite doet (bijvoorbeeld door het halen van een deadline)*

  Enquêteurfouten: *Enquêteurs die fouten maken bij het oplezen van de vraag of de antwoorden in eigen woorden weergeven kunnen daarmee tot vertekeningen leiden.*

College 4: (Hoofdstuk 8)

Elementen van het toetsen van een hypothese:

- **Nulhypothese (Ho):** Hypothese die aangenomen zal worden, tenzij de gegevens overtuigend bewijs leveren waaruit blijkt dat de hypothese onjuist is. Ho : µ = ...

Bovemij Crowe EY GrantThornton KPMG pwc

- **Alternatieve (onderzoeks) hypothese (Ha):** Hypothese die de nulhypothese tegen spreekt en alleen wordt aangenomen als de gegevens overtuigend bewijs leveren van de waarheid van de hypothese. Ha : μ = …
- **Toetsingsgrootheid:** Een steekproefgrootheid die wordt gebruikt om te beslissen of de nulhypothese moet worden verworpen.
- **Verwerpingsgebied:** De waarden van de toetsingsgrootheid waarvoor de nulhypothese zal worden verworpen.
- **Aannames:** Duidelijke uitspraken over de aannames die zijn gemaakt over de populatie(s) waaruit de steekproef wordt genomen.
- **Experiment en berekening van de toetsingsgrootheid:** Uitvoeren van een steekproefexperiment en bepalen van de waarde van de toetsingsgrootheid.
- **Conclusie:**
  Als de waarde van de toetsingsgrootheid in het verwerpingsgebied valt, wordt de nulhypothese verworpen en is de conclusie dat de alternatieve hypothese juist is. Dit proces van hypothese toetsen is leidt voor 100α% van de gevallen onterecht tot de conclusie dat Ho juist is. (**Type I fout**)
  Als de waarde van de toetsingsgrootheid niet in het verwerpingsgebied valt, wordt de nulhypothese niet verworpen. Er wordt niet geconcludeerd dat de nulhypothese juist is, omdat de kans β dat de nulhypothese onterecht behouden wordt onbekend is. (**Type II fout**)

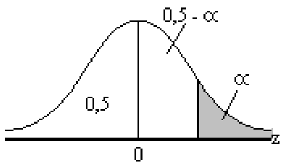
	Werkelijke stand van zaken	
Conclusie	**Ho is juist**	**Ha is juist**
Accepteer Ho (veronderstel Ho is juist)	Juiste conclusie	Type II fout (kans β)
Verwerp Ho (veronderstel Ha is juist)	Type I fout (kans α)	Juiste conclusie

Kiezen van een nulhypothese en een alternatieve hypthese:

- **Alternatieve hypothese**
  Rechts-eenzijdig Ha: μ > …
  Links-eenzijdig Ha: μ < …
  Tweezijdig Ha: $\mu \neq$ …
- **Nulhypothese**
  Rechts-eenzijdig Ho: μ = …
  Links-eenzijdig Ho: μ = …
  Tweezijdig Ho: μ = …

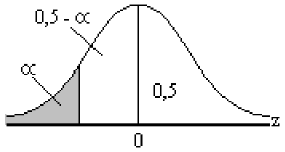
Verwerpingsgebied:

- Rechts-eenzijdig Ha: μ > …

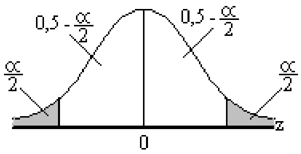


Bovemij Crowe EY GrantThornton KPMG pwc

- Links-eenzijdig Ha: μ < ...



- Tweezijdig Ha: μ ≠ ...



	Alternatieve hypothesen		
	Links-eenzijdig	**Rechts-eenzijdig**	**Tweezijdig**
$\alpha = 0{,}10$	z < -1,28	z > 1,28	z < -1,645 of z > 1,645
$\alpha = 0{,}05$	z < -1,645	z > -1,645	z < -1,96 of z > 1,96
$\alpha = 0{,}01$	z < -2,33	z > -2,33	z < -2,575 of z > 2,575

Toetsingsgrootheid: Meet het aantal standaarddeviaties tussen de waargenomen waarde van $\bar{x}$ en de waarde μ uit de nulhypothese.

Toetsingsgrootheid = $\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$

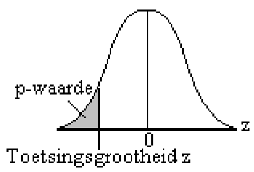
Toetsen van een hypothese over μ met een grote steekproef:

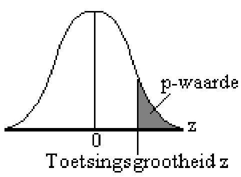
	Eenzijdige toets	**Tweezijdige toets**
Hypothesen	Ho: $\mu = \mu_0$ Ha: $\mu < \mu_0$ of Ha: $\mu < \mu_0$	Ho: $\mu = \mu_0$ Ha: $\mu \neq \mu_0$
Toetsingsgrootheid (z)	$\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$	$\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$
Verwerpingsgebied	$z < -z\alpha$ of $z > z\alpha$	$\lvert z\rvert > z\alpha/2$
waarbij za / za/2 zo gekozen zijn dat:	$P(z > z\alpha) = \alpha$	$P(z > z\alpha/2) = \alpha / 2$

Waargenomen significantieniveau (p-waarde): Bij een specifieke statistische toets de kans (aangenomen dat Ho juist is) op het waarnemen van een waarde van de toetsingsgrootheid die in ten minste even sterke mate in strijd is met de nulhypothese en in overeenstemming met de alternatieve hypothese, als de werkelijke waarde die uit de steekproefgegevens is berekend.

Berekening p-waarde:

- Bepalen van de waarde van de toetsingsgrootheid z voor het resultaat van het steekproefexperiment.
- Eenzijdige toets→



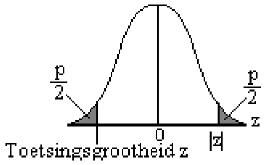


Bovemij Crowe EY GrantThornton KPMG pwc

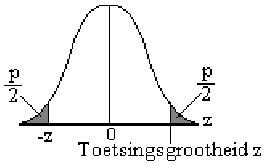
(Ha: $\mu < \mu_o$)

(Ha: $\mu < \mu_o$)

Tweezijdige toets→



(Toetsingsgrootheid z negatief)



(Toetsingsgrootheid z positief)

Beslissing over nulhypothese op basis van p-waarde:

- Kies maximumwaarde α die je nog acceptabel vindt.
- Als het waargenomen significantieniveau (p-niveau) lager is dan de gekozen waarde van α, verwerp dan de nulhypothese.
  Als het waargenomen significantieniveau (p-niveau) hoger is (gelijk aan) dan de gekozen waarde van α, verwerp de nulhypothese dan niet.

Toetsen van een hypothese over μ voor een kleine steekproef:

	Eenzijdige toets	**Tweezijdige toets**
Hypothesen	Ho: $\mu = \mu_o$ Ha: $\mu < \mu_o$ of Ha: $\mu < \mu_o$	Ho: $\mu = \mu_o$ Ha: $\mu \neq \mu_o$
Toetsingsgrootheid (t)	$\frac{\bar{x} - \mu_o}{s / \sqrt{n}}$	$\frac{\bar{x} - \mu_o}{s / \sqrt{n}}$
Verwerpingsgebied (waarbij ta en ta/2 gebaseerd zijn op (n – 1) vrijheidsgraden)	$t < -t\alpha$ of $t > t\alpha$	$\lvert t \rvert > t\alpha/2$

Toetsen van een hypothese over μ voor een kleine steekproef:

	Eenzijdige toets	**Tweezijdige toets**
Hypothesen	Ho: $p = p_o$ Ha: $p < p_o$ of Ha: $p < p_o$	Ho: $p = p_o$ Ha: $p \neq p_o$
Toetsingsgrootheid (z) (waarbij, volgens Ho, $\sigma_{\hat{p}} = \sqrt{(p_o q_o / n)}$ en $q_o = 1 - p_o$)	$\frac{\hat{p} - p_o}{\sigma_{\hat{p}}}$	$\frac{\hat{p} - p_o}{\sigma_{\hat{p}}}$
Verwerpingsgebied	$z < -z\alpha$ of $z > z\alpha$	$\lvert z \rvert > z\alpha/2$

(aanname dat het experiment binomiaal is en de steekproef groot genoeg, zodat het interval $p_o \pm 3\sigma_{\hat{p}}$ niet 0 of 1 omvat)

Berekenen van β voor een toets over μ bij een grote steekproef:

- Bereken de waarde(n) van $\bar{x}$ die overeenkomt/overeenkomen met de grens/grenzen van het verwerpingsgebied.

  Rechts-eenzijdige toets: $\bar{x}_O = \mu_o + z\alpha\sigma_{\bar{x}} \approx \mu_o + z\alpha\,(s / \sqrt{n})$
  Links-eenzijdige toets: $\bar{x}_O = \mu_o - z\alpha\sigma_{\bar{x}} \approx \mu_o - z\alpha\,(s / \sqrt{n})$

Bovemij Crowe EY GrantThornton KPMG pwc

Tweezijdige toets: $\bar{x}_{O,L} = \mu_o + z_{\alpha/2}\,\sigma_{\bar{x}} \approx \mu_o + z_{\alpha/2}\,(s / \sqrt{n})$
$\bar{x}_{O,R} = \mu_o - z_{\alpha/2}\,\sigma_{\bar{x}} \approx \mu_o - z_{\alpha/2}\,(s / \sqrt{n})$

- Specificeer de waarde van μ_a in de alternatieve hypothese waarvoor de waarde van β wordt berekent. Reken dan de grenswaarde(n) van $\bar{x}_O$ om naar z, waarbij je de alternatieve verdeling met gemiddelde μ_a gebruikt.

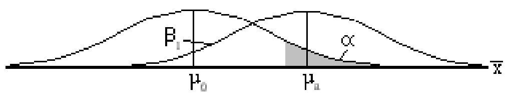
  $$z = \frac{\bar{x}_O - \mu_a}{\sigma_{\bar{x}}}$$

  Maak een schets van de alternatieve verdeling (gecentreerd door μ_a), en arceer het acceptatiegebied (niet-verwerpingsgebied).

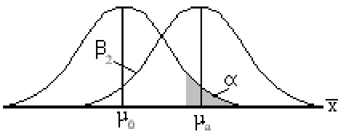
Onderscheidingsvermogen van een toets: De kans dat de toets terecht leidt tot het verwerpen van de nulhypothese voor een bepaalde waarde van μ in de alternatieve hypothese. Het onderscheidingsvermogen van een toets is gelijk aan $1 - \beta$ voor het desbetreffende alternatief.

Eigenschappen van β en van het onderscheidingsvermogen:

- Voor een vaste n en α neemt de waarde van β af en neemt het onderscheidingsvermogen toe als de afstand tussen de gespecificeerde nulwaarde μ_o en de gespecificeerde alternatieve waarde μ_a toeneemt.
- Voor een vaste n en een gegeven waarden van μ_o en μ_a neemt de waarde van β toe en neemt het onderscheidingsvermogen af als de waarde van α kleiner wordt.
- Voor een vaste α en gegeven waarde van μ_o en μ_a neemt de waarde van β af en neemt het onderscheidingsvermogen toe als de steekproefgrootte n toeneemt.



Kleine n



Grote n, vast α (β2<β1)

Chi-kwadraat (χ^2) verdeling: Steekproefverdeling van s^2 als de gegevens uit een normaal verdeelde populatie afkomstig zijn.

Betrouwbaarheidsinterval σ^2:
(Door middel van chi-kwadraatverdeling met (n – 1) vrijheidsgraden)

$$\frac{(n-1)\,s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)\,s^2}{\chi^2_{(1-\alpha/2)}}$$

Toetsen van een hypothese over σ^2:

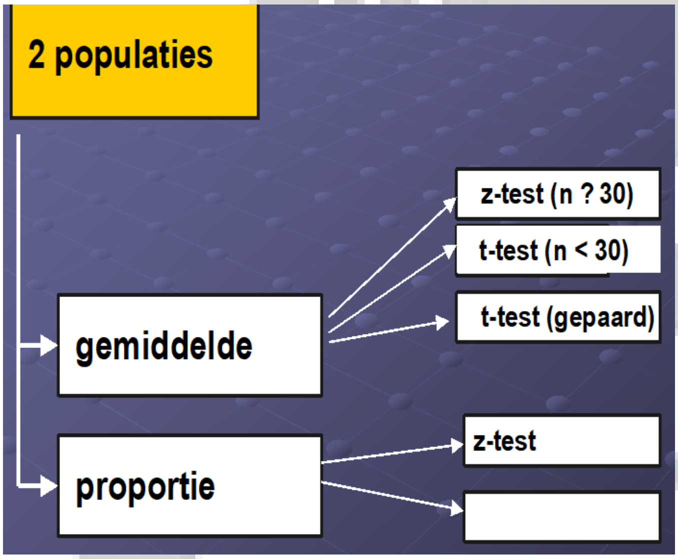
	Eenzijdige toets	Tweezijdige toets
Hypothesen	Ho: $\sigma^2 = \sigma^2_o$ Ha: $\sigma^2 < \sigma^2_o$ of Ha: $\sigma^2 < \sigma^2_o$	Ho: $\sigma^2 = \sigma^2_o$ Ha: $\sigma^2 \neq \sigma^2_o$
Toetsingsgrootheid (χ^2)	$\frac{(n-1)\,s^2}{\sigma^2}$	$\frac{(n-1)\,s^2}{\sigma^2}$
Verwerpingsgebied	$\chi^2 < \chi^2_{(1-\alpha)}$ of $\chi^2 > \chi^2_{\alpha}$	$\lvert \chi^2 \rvert > \chi^2_{\alpha/2}$

(Aanname is dat de populatie waaruit de steekproef is genomen bij benadering normaal verdeeld is)
(Kleine tot matige afwijkingen van normaliteit maken de chi-kwadraattoets ongeldig)

College 5: (Hoofdstuk 9)

Eigenschappen van de steekproefverdeling van ($\bar{x}_1 - \bar{x}_2$):

- Het gemiddelde van de steekproef verdeling ($\bar{x}_1 - \bar{x}_2$) is ($\mu_1 - \mu_2$).
- Als twee steekproeven onafhankelijk zijn, dan is de standaarddeviatie van de steekproefverdeling
  $$\sigma_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
  waarbij σ_1^2 en σ_2^2 de varianties zijn van de twee populaties waaruit de steekproeven worden genomen, en n_1 respectievelijk n_2 de steekproefomvang is. $\sigma_{(\bar{x}_1-\bar{x}_2)}$ wordt ook wel de standaardfout van de steekproefgrootheid ($\bar{x}_1 - \bar{x}_2$) genoemd.
- De steekproefverdeling van ($\bar{x}_1 - \bar{x}_2$) is door de centrale limietstelling bij benadering normaal voor grote steekproeven.



Betrouwbaarheidsinterval voor ($\mu_1 - \mu_2$) bij grote steekproeven:

$(\bar{x}_1 - \bar{x}_2) \pm z\alpha/2\, \sigma_{(\bar{x}_1-\bar{x}_2)} = \qquad (\bar{x}_1 - \bar{x}_2) \pm z\alpha/2 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(aannames: aselecte steekproef waarbij n_1 en n_2 zijn groot genoeg (> 30), zodat de verdeling normaal is en s1 en s2 een goede afspiegeling zijn van σ_1^2 en σ_2^2.

Toetsen van een hypothese voor ($\mu_1 - \mu_2$) bij grote steekproeven:

- <u>Hypothesen:</u>
  (Eenzijdige toets) — (Tweezijdige toets)
  Ho: ($\mu_1 - \mu_2$) = Do — Ho: ($\mu_1 - \mu_2$) = Do
  Ha: ($\mu_1 - \mu_2$) < Do of Ha: ($\mu_1 - \mu_2$) < Do — Ha: ($\mu_1 - \mu_2$) ≠ Do
  (Do = het te toetsen verschil tussen de gemiddelden)

- <u>Toetsingsgrootheid:</u>
  $$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sigma_{(\bar{x}_1-\bar{x}_2)}} \quad \text{waarbij} \quad \sigma_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- <u>Verwerpingsgebied:</u>
  $z < -z\alpha$ of $z > z\alpha$ $\qquad\qquad$ $|z| > z\alpha/2$

(aannames: zelfde als bij het betrouwbaarheidsinterval bij grote steekproeven)

Gecombineerde steekproefschatter van μ^2 (sp^2):

$$sp^2 = \frac{(n1-1)\,s1^2 + (n2-1)\,s2^2}{(n1-1)+(n2-1)} = \frac{(n1-1)\,s1^2 + (n2-1)\,s2^2}{n1+n2-2}$$

of

$$sp^2 = \frac{\Sigma(x1-\bar{x}1)^2 + \Sigma(x2-\bar{x}2)^2}{n1+n2-2}$$

Betrouwbaarheidsinterval voor ($\mu1$ - $\mu2$) bij kleine steekproeven (onafhankelijke steekproeven):

$$(\bar{x}1 - \bar{x}2) \pm t\alpha/2 \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$\text{waarbij } sp^2 = \frac{(n1-1)\,s1^2 + (n2-1)\,s2^2}{n1+n2-2}$$

en $t\alpha/2$ is gebaseerd op $(n1 + n2 - 2)$ vrijheidsgraden.

(aannames: Beide populaties waaruit de steekproeven worden genomen hebben relatieve frequentieverdelingen die bij benadering normaal zijn, de populatievarianties zijn gelijk en de steekproeven zijn aselect en onafhankelijk uit de populatie gekozen.)

Toetsen van een hypothese voor ($\mu1$ - $\mu2$) bij kleine steekproeven (onafhankelijke steekproeven):

- <u>Hypothesen:</u>
  (Eenzijdige toets) $\qquad\qquad$ (Tweezijdige toets)
  Ho: $(\mu1 - \mu2) = Do$ $\qquad\qquad$ Ho: $(\mu1 - \mu2) = Do$
  Ha: $(\mu1 - \mu2) < Do$ of Ha: $(\mu1 - \mu2) < Do$ $\qquad\qquad$ Ha: $(\mu1 - \mu2) \neq Do$
  (Do = het te toetsen verschil tussen de gemiddelden (de onderzoeker vermoed 0))

- <u>Toetsingsgrootheid:</u>
  $$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

- <u>Verwerpingsgebied:</u>
  $t < -t\alpha$ of $t > t\alpha$ $\qquad\qquad$ $|t| > t\alpha/2$
  waarbij $t\alpha$ en $t\alpha/2$ gebaseerd zijn op $(n1 + n2 - 2)$ vrijheidsgraden.

(aannames: zelfde als bij het betrouwbaarheidsinterval bij kleine steekproeven)

Betrouwbaarheidsinterval voor μD = ($\mu1$ - $\mu2$) bij gepaarde waarnemingen:
Grote steekproef

<u>σD</u> $\qquad\qquad$ <u>sD</u>

Bovemij Crowe EY GrantThornton KPMG pwc

$\bar{x}_D \pm z_{\alpha/2} \sqrt{n_D} \approx \qquad \bar{x}_D \pm z_{\alpha/2} \sqrt{n_D}$

(Aanname: steekproefverschillen worden aselect gekozen uit de populatie van verschillen.)

Kleine steekproef

$\bar{x}_D \pm t_{\alpha/2} \dfrac{s_D}{\sqrt{n_D}}$ waarbij $t_{\alpha/2}$ gebaseerd is op $(n_D - 1)$ vrijheidsgraden.

(Aannames: de relatieve frequentieverdeling van de populatie van verschillen is normaal en de steekproefverschillen worden aselect gekozen uit de populatie van verschillen.)

Toetsen voor $\mu_D = (\mu_1 - \mu_2)$ bij gepaarde waarnemingen:

- Hypothesen:

(Eenzijdige toets)	(Tweezijdige toets)
Ho: $\mu_D = D_o$	Ho: $\mu_D = D_o$
Ha: $\mu_D < D_o$ of Ha: $\mu_D < D_o$	Ha: $\mu_D \neq D_o$

(Do = het te toetsen verschil tussen de gemiddelden)

Grote steekproef

- Toetsingsgrootheid:

$$z = \frac{\bar{x}_D - D_o}{\sigma_D/\sqrt{n_D}} \approx \frac{\bar{x}_D - D_o}{s_D/\sqrt{n_D}}$$

- Verwerpingsgebied:

$z < -z_\alpha$ of $z > z_\alpha$ $\qquad$ $|z| > z_{\alpha/2}$

(Aanname: de verschillen worden aselect gekozen uit de populatie verschillen.)

Kleine steekproef

- Toetsingsgrootheid:

$$t = \frac{\bar{x}_D - D_o}{s_D/\sqrt{n_D}}$$

- Verwerpingsgebied:

$t < -t_\alpha$ of $t > t_\alpha$ $\qquad$ $|t| > t_{\alpha/2}$

waarbij t_α en $t_{\alpha/2}$ gebaseerd zijn op $(n_1 + n_2 – 2)$ vrijheidsgraden.

(Aannames: de relatieve frequentieverdeling van de populatie van verschillen is normaal en de steekproefverschillen worden aselect gekozen uit de populatie van verschillen.)

Eigenschappen van de steekproefverdeling van $(\hat{p}_1 - \hat{p}_2)$:

- Het gemiddelde van de steekproefverdeling van $(\hat{p}_1 - \hat{p}_2)$ is $(p_1 - p_2)$, dat wil zeggen:
  $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
  $(\hat{p}_1 - \hat{p}_2)$ is dus een zuivere schatter van $(p_1 - p_2)$
- De standaarddeviatie van de steekproefverdeling van $(\hat{p}_1 - \hat{p}_2)$ is:
  $$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$
- Als de steekproefomvang n_1 respectievelijk n_2 groot is, is de steekproefverdeling van $(\hat{p}_1 - \hat{p}_2)$ bij benadering normaal.

Bovemij Crowe EY GrantThornton KPMG pwc

100(1 - α)% Betrouwbaarheidsinterval voor ($\hat{p}_1 - \hat{p}_2$) bij grote steekproeven:

$$(\hat{p}_1 - \hat{p}_2) \pm z\alpha/2\, \sigma_{(\hat{p}_1 - \hat{p}_2)} = (\hat{p}_1 - \hat{p}_2) \pm z\alpha/2 \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

$$\approx (\hat{p}_1 - \hat{p}_2) \pm z\alpha/2 \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

(Aannames: Onafhankelijke aselecte steekproeven, die groot genoeg zijn zodat de normale verdeling een goede benadering vormt voor de steekproefverdeling van $\hat{p}_1$ en $\hat{p}_2$.)

Toetsen van een hypothese voor (p1 - p2) bij grote steekproeven:

- Hypothesen:

  (Eenzijdige toets)
  Ho: (p1 - p2) = 0
  Ha: (p1 - p2) < 0 of Ha: (p1 - p2) < 0

  (Tweezijdige toets)
  Ho: (p1 - p2) = 0
  Ha: (p1 - p2) ≠ 0

- Toetsingsgrootheid:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$$

- Verwerpingsgebied:

  z < -zα of z > zα

  |z| > zα/2

$$\left(\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \approx \sqrt{\frac{\hat{p}\hat{q}_1}{n_1} + \frac{\hat{p}\hat{q}_2}{n_2}} \quad \text{waarbij } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}\right)$$

(Aanname: zelfde als voor het betrouwbaarheidsinterval voor ($\hat{p}_1 - \hat{p}_2$) bij een grote steekproef)
(mag alleen gebruikt worden indien: Po ± 3 σ^p (Po ± 3√(poqo / n)) en σ ; 1)

Bepaling van de steekproefomvang voor het schatten van µ1 – µ2:

Om µ1 – µ2 binnen een gegeven begrenzing met kans (1 - α) te schatten wordt de volgende formule met de gewenste betrouwbaarheid voor even grote steekproeven gebruikt:

$$n_1 = n_2 = \frac{(z_{a/2})^2(\sigma_1^2 + \sigma_2^2)}{B^2}$$

(σ_1^2 en σ_2^2 moeten gesubstitueerd worden door s_1^2 en s_2^2, deze schattingen kunnen gebaseerd zijn op eerdere steekproeven of op de meest conservatieve schatting s = R / 4 (R = bereik).)

Bepaling van de steekproefomvang voor het schatten van p1 - p2:

Om p1 - p2 binnen een gegeven begrenzing B met kans (1 - α) te schatten wordt de volgende formule met de gewenste betrouwbaarheid voor even grote steekproeven gebruikt:

$$n_1 = n_2 = \frac{(z_{a/2})^2(p_1q_1^2 + p_2q_2^2)}{B^2}$$

(p1 en p2 moeten gesubstitueerd worden, deze schattingen kunnen gebaseerd zijn op eerdere steekproeven of op de meest conservatieve schatting p1 = p2= 0,5)

F-toets voor gelijke populatievarianties:

- Hypotheses:

  (Eenzijdige toets)
  Ho: $\sigma_1^2 = \sigma_2^2$
  Ha: $\sigma_1^2 < \sigma_2^2$ of Ha: $\sigma_1^2 < \sigma_2^2$

  (Tweezijdige toets)
  Ho: $\sigma_1^2 = \sigma_2^2$
  Ha: $\sigma_1^2 \neq \sigma_2^2$

Bovemij Crowe EY GrantThornton KPMG pwc

- Toetsingsgrootheid:

$$F = \frac{s2^2}{s1^2} \quad \text{(of)} \quad F = \frac{s1^2}{s2^2}$$

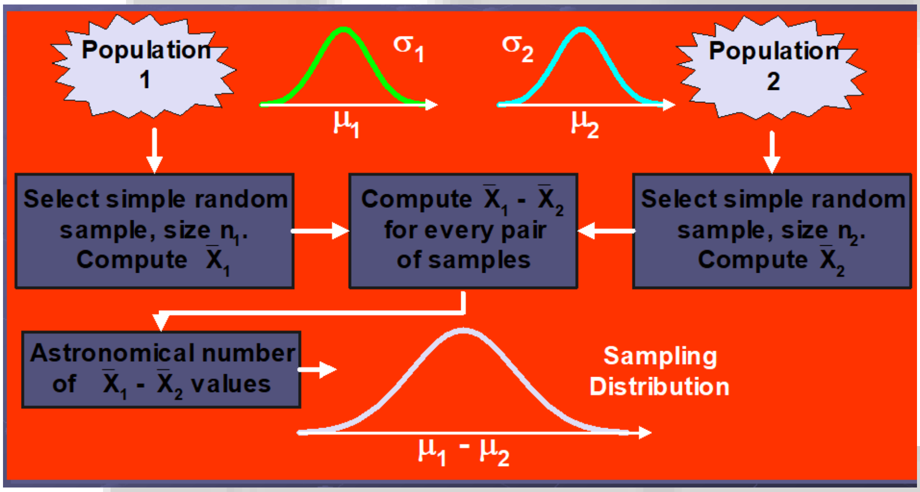
$$F = \frac{\text{Grootste steekproefvariantie}}{\text{Kleinste steekproefvariantie}} = \frac{s1^2}{s2^2} \text{ als } s1^2 > s2^2 \quad \text{(of)} \quad = \frac{s2^2}{s1^2} \text{ als } s2^2 > s1^2$$

- Verwerpingsgebied:

  $F > F\alpha$ $F > F\alpha/2$

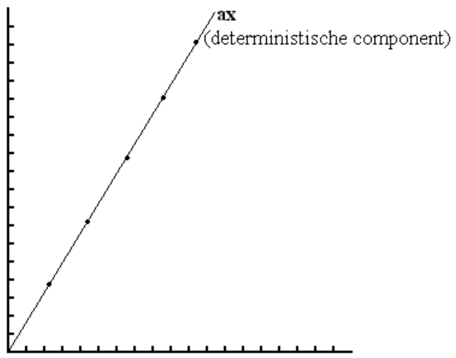
  ($F\alpha$ en $F\alpha/2$ gebaseerd zijn op v1 vrijheidsgraden in de teller en v2 vrijheidsgraden in de noemer.)

(Aannames: Steekproeven zijn aselecte, onafhankelijk en de populaties waaruit de steekproeven worden genomen zijn normaal verdeeld.)



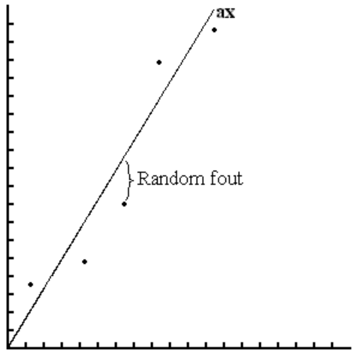
College 6: (Hoofdstuk 10 en 15.6)

Deterministisch model: Een model dat een exacte relatie tussen variabelen veronderstelt.



Random fout: Variantie in het model door random verschijnselen.

Bovemij Crowe EY Grant Thornton KPMG pwc



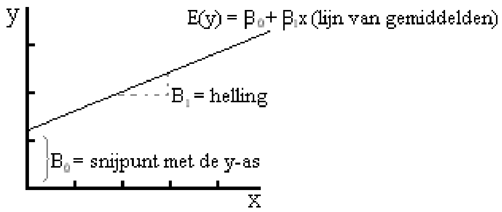
Algemene vorm van een kansmodel:
y = deterministische component + random fout

Waarbij:
y = relevante variabele
gem. waarde random fout = 0
gem. waarde y, E(y) = deterministische component van het model

Lineair model: Model waarbij in een grafiek het deterministische deel van het model een rechte lijn vormt.

Een eerste orde (lineair) model:
$y = \beta 0 + \beta 1x + \varepsilon$

waarbij:
y = afhankelijke variabele of responsvariabele (variabele waarvoor een model wordt gevormd)
x = onafhankelijke variabele of voorspellende variabele (variabele die wordt gebruikt als voorspeller van y)
$E(y) = \beta 0 + \beta 1x$ = deterministische component (lijn van gemiddelden)
ε (epsilon) = random-foutcomponent
β0 (bèta nul) = snijpunt van de lijn met de y-as
β1 (bèta één) = helling van de lijn, de toename/afname in de deterministische component van y voor elke toename/afname van een eenheid x



Ontwikkelen van een model:

- Ontwikkel een hypothese over de deterministische component van het model, die een verband legt tussen het gemiddelde E(y) en de onafhankelijke variabele x.
- Gebruik de steekproefgegevens om de onbekende parameters is het model te schatten.
- Specificeer de kansverdeling van de random-foutcomponent, en schat de standaarddeviatie van deze verdeling.
- Geef de statistische beoordeling van de bruikbaarheid van het model.
- Indien de conclusie luidt dat het model bruikbaar is, gebruik het dan voor het doen van voorspellingen, het maken van schattingen en andere doeleinden.

Kleinste-kwadratenlijn:

$y = \beta_0 + \beta_1 x$

- Som van de fouten (SE, sum of errors) is gelijk aan nul.
- Som van kwadraten van de fouten (SSE, sum of squared errors) is kleiner dan voor enig ander lineair model.

Formules voor de kleinste-kwadratenschatting:

- Helling: $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$
- Snijpunt met y-as: $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Waarbij:

$$SSxy = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$SSxx = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

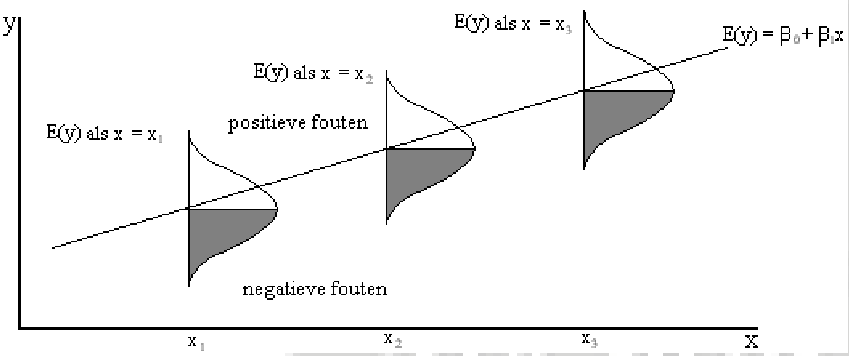
SSxy = som van de kruisproducten x * y

SSxx = som van de kwadraten van de x-waarden (SSyy = som van de kwadraten van de y-waarden)

n = steekproefomvang

Modelaannames:

1. Het gemiddelde van de kansverdeling is nul.
2. De variantie van de kansverdeling van ε is constant voor alle waarden van de onafhankelijke variabele x.
3. De kansverdeling van ε is normaal.
4. De waarden van ε die bij twee waargenomen waarden van y horen, zijn onafhankelijk.



Schatting van σ^2 voor een eerste orde (lineair) model:

$$s^2 = \frac{SSE}{\text{Aantal vrijheidsgraden voor fout}} = \frac{SSE}{n-2}$$

waarbij:

$$SSE = \sum(y_i - \hat{y}) = SSyy - \hat{\beta}_1\, SSxy$$

$$SSyy = \sum(y_i - \hat{y}) = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Om de standaarddeviatie σ van ε (**geschatte standaardfout van het regressiemodel**) te schatten, berekenen we,

Bovemij Crowe EY GrantThornton KPMG pwc

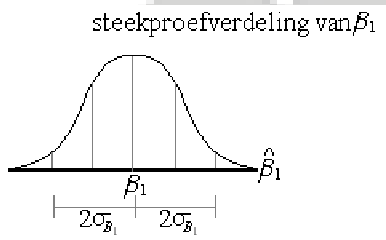
$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$

(niet afronden, neem minimaal 6 significante cijfers voor SSyy, $\hat{\beta}_1$ en SSxy)

Interpretatie van s, de geschatte standaarddeviatie van ε:
Er (wordt verwacht dat) de meeste (ong. 95%) van de waargenomen y-waarden binnen 2s van de waarde $\hat{y}$ van de kleinste-kwadratenvoorspelling zullen liggen.

Steekproefverdeling van β1:
Als de vier aannames over ε aannemen, zal de steekproefverdeling van de kleinste-kwadratenschatter β1 van de helling normaal zijn, met gemiddelde β1 en standaarddeviatie $\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$.

Om $\sigma_{\hat{\beta}_1}$ te schatten wordt gebruik gemaakt van $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$ (**de geschatte standaardfout van de kleinste-kwadratenhelling β1**)



Bruikbaarheidstoets voor een model: enkelvoudige lineaire regressie:

- Hypothesen:

	Eenzijdig:	Tweezijdig
	H0 : β1 = 0	H0 : β1 = 0
	Ha : β1 < 0	Ha : β1 ≠ 0
	Ha : β1 > 0	

- Toetsingsgrootheid:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\frac{s}{\sqrt{SS_{xx}}}}$$

- Verwerpingsgebied:

	Eenzijdig	Tweezijdig
	t < -tα	|t| < tα/2
	t > -tα	

(Aannames:vier aannames betreffende ε, pag. 28)

Interpretatie van p-waarden voor β-coëfficiënt bij regressie:
Omdat vrijwel ieder statistisch softwarepakket een tweezijdige p-waarde voor elk van beide β-parameters in het regressiemodel geven, is het noodzakelijk om de uitdraai bij een eenzijdige toets als volgt aan te passen:

Rechts-eenzijdige toets (Ha : β1 > 0):	p-waarde =	p/2	als t > 0
		1 - p/2	als t < 0
Links-eenzijdige toets (Ha : β1 > 0):	p-waarde =	p/2	als t < 0
		1 – p/2	als t > 0

Bovemij Crowe EY Grant Thornton KPMG pwc

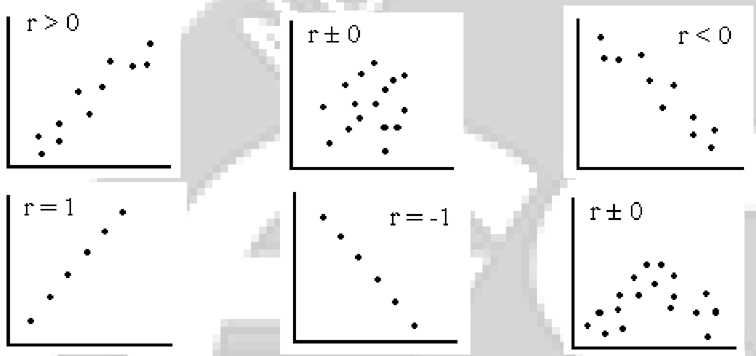
(waarbij p de p-waarde is die op de uitdraai vermeld staan en t de waarde van de toetsingsgrootheid is)

Een 100(1 – α)% betrouwbaarheidsinterval voor de enkelvoudige lineaire regressie helling β1:
$\hat{\beta}_1 \pm t_{\alpha/2}\, s_{\hat{\beta}_1}$

waarbij de geschatte standaardfout $\hat{\beta}_1$ wordt berekend uit $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$ en $t_{\alpha/2}$ gebaseerd is op (n – 2) vrijheidsgraden.
(Aannames: de vier aannames betreffende ε, pag. 28)

Pearson productmoment correlatiecoëfficiënt (r): Een maat voor de sterkte van het lineaire verband tussen twee variabelen x en y.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\, SS_{yy}}}$$



Determinatiecoëfficiënt (r^2): representeert de fractie van de totale steekproefspreiding rond $\bar{y}$, die verklaard wordt door het lineaire verband tussen y en x.

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Praktische interpretatie van de determinatiecoëfficiënt (r^2): Ongeveer 100(r^2)% van de steekproefvariantie in y (gemeten als de totale som van de kwadraten van de afwijkingen van de steekproefwaarden van y en van hun gemiddelde $\bar{y}$) kan verklaard worden door het gebruik van x voor het voorspellen van y in het lineaire model.

Steekproeffouten voor de schatter van het gemiddelde van y en voor de voorspeller van een individuele nieuwe waarde van y:

1. De standaarddeviatie van de steekproefverdeling van de schatter $\hat{y}$ van het gemiddelde van y voor een specifieke waarde van x, xp, is
   $$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$
   waarbij: σ de standaarddeviatie is van de random fout ε. We noemen $\sigma_{\hat{y}}$ de standaardfout van $\hat{y}$.

2. De standaarddeviatie van de voorspellingsfout voor de voorspeller $\hat{y}$ van een individuele nieuwe y-waarde voor een specifieke waarde van x is
   $$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$
   waarbij: : σ de standaarddeviatie is van de random fout ε. We noemen $\sigma_{(y-\hat{y})}$ de

standaardfout van $\hat{y}$.
$(y - \hat{y})$ = voorspellingsfout

Een 100(1 – α)% betrouwbaarheidsinterval voor de gemiddelde waarde van y voor x = xp:
$\hat{y} \pm t_{\alpha/2}$ (geschatte standaardfout van $\hat{y}$)

$$\hat{y} \pm t_{\alpha/2}\, s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$
waarbij: t α/2 gebaseerd is op (n – 2) vrijheidsgraden.

Een 100(1 – α)% betrouwbaarheidsinterval' voor een individuele waarde van y voor x = xp:
$\hat{y} \pm t_{\alpha/2}$ (geschatte standaardfout van de voorspelling)

$$\hat{y} \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$
waarbij: t α/2 gebaseerd is op (n – 2) vrijheidsgraden.

Waarschuwing:
De kleinste kwadratenmethode kan een slechte weergave zijn van het werkelijke model voor waarden van x die buiten het bereik van de waarden van x in je steekproef vallen.

Rangcorrelatiecoëfficiënt van Spearman (rs): De maat voor de correlatie tussen rangnummers.

$$r_s = \frac{SS_{uv}}{\sqrt{SS_{uu}\, SS_{vv}}}$$

waarbij:

$$SS_{uv} = \sum(u_i - \bar{u})(v_i - \bar{v}) = \sum u_i v_i - \frac{(\sum u_i)(\sum v_i)}{n}$$

$$SS_{uu} = \sum(u_i - \bar{u})^2 = \sum u_i^2 - \frac{(\sum u_i)^2}{n}$$

$$SS_{vv} = \sum(v_i - \bar{v})^2 = \sum v_i^2 - \frac{(\sum v_i)^2}{n}$$

ui = rangnummer van de i^de waarneming in de steekproef 1
vi = rangnummer van de i^de waarneming in de steekproef 2
n = aantal paren waarnemingen (aantal waarnemingen in elke steekproef)

Vereenvoudigde formule voor rs':

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

waarbij: di = ui - vi (verschil in rangnummers van de i^de waarneming voor steekproef 1 en 2)

De niet-parametrische toets voor rangcorrelatie van Spearman:
- Hypothesen:

Eenzijdig:	Tweezijdig:
H0 : ρ = 0	H0 : ρ = 0
Ha : ρ < 0	Ha : ρ ≠ 0
Ha : ρ > 0	

Bovemij Crowe EY GrantThornton KPMG pwc

- Toetsingsgrootheid:
  rs (steekproefrangcorrelatie)

- Verwerpingsgebied:

  rs > rs,α |rs| > rs,α/2
  rs < -rs.α

ρ (rho) = populatiecorrelatiecoëfficiënt
Waarbij: rs,α of rs,α/2 de waarde in tabel XVII is die correspondeert met het oppervlak α in de in de rechterstaart van de verdeling en n paren waarnemingen.
(Aannames: De steekproef van experimentele eenheden waarvoor de twee variabelen worden gemeten aselect gekozen wordt en de kansverdelingen van de twee variabelen zijn continu)
Gelijke meetwaarden: Ken aan gelijke meetwaarden een rangnummer toe dat het gemiddelde is van de rangnummers die ze gekregen zouden hebben als ze ongelijk, maar wel opeenvolgend geweest waren. Bijvoorbeeld meetwaarde 3 en 4, ken dan aan elk het rangnummer ((3 + 4)/2) 3,5 toe. Het aantal gelijke rangnummers moet klein zijn ten opzichte van het aantal waarnemingen.

College 7: (Hoofdstuk 11)

Het algemene multipele-regressiemodel:
$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta kxk + \varepsilon$

waarbij:
y = afhankelijke variabele
x1, x2, …, x3 = onafhankelijke variabelen
$E(y) = \beta 0 + \beta 1 + \beta 2 + \ldots + \beta k$ = deterministisch deel van het model
βi bepaalde de bijdrage van de onafhankelijke variabele xi

Het analyseren van een multiple-regressiemodel:
1. Bepaal de deterministische component.
2. Gebruik de steekproefgegevens om de onbekende parameters (β0, β1, β2…, βk) te schatten.
3. Specificeer de kansverdeling van de random-foutterm, maak een schatting van de standaarddeviatie σ van deze verdeling.
4. Controleer of aan de aannames van ε zijn voldaan, breng zo nodig wijzigingen aan in het model.
5. Geef een statistische beoordeling van de bruikbaarheid van het model.
6. Indien het model bruikbaar is, gebruik het dan voor voorspellingen en schattingen.

Een eerste-ordemodel voor vijf kwantitatieve onafhankelijke variabele:
$E(y) = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \beta 4x4 + \beta 5x5$
Waarbij x1, x2, …, x3 kwantitatieve variabelen zijn die geen functie zijn van andere onafhankelijke variabelen.
Waarschuwing:
De interpretatie van de β-parameters in een multiple-regressiemodel hangt af van de termen die in het model worden gespecificeerd. De interpretatie is alleen geldig voor een eerste-orde lineair model!

Aannames voor de random fout ε:
1. Voor elke reeks waarden van x1, x2, …, x3 heeft de random fout ε een normale kansverdeling met een gemiddelde gelijk aan 0 en een variantie gelijk aan σ^2.
2. De random fouten zijn onderling onafhankelijk.

Bovemij Crowe EY GrantThornton KPMG pwc

Schatter van σ^2 voor een multiple-regressiemodel met k onafhankelijke variabelen:

$$s^2 = \frac{SSE}{n - \text{aantal geschatte } \beta\text{-parameters}} = \frac{SSE}{n - (k + 1)}$$

Toets van een individuele parametercoëfficiënt in het multipele-regressiemodel:

- Hypothesen:

Eenzijdig:	Tweezijdig
$H0 : \beta i = 0$	$H0 : \beta i = 0$
$Ha : \beta i < 0$	$Ha : \beta i \neq 0$
$Ha : \beta i > 0$	

- Toetsingsgrootheid:

$$t = \frac{\hat{\beta}_i}{s_{\hat{\beta}i}}$$

- Verwerpingsgebied:

$t > t\alpha$	$\lvert t \rvert > t\alpha/2$
$t < -t\alpha$	

Waarbij $t\alpha$ en $t\alpha/2$ gebaseerd zijn op $n – (k + 1)$ vrijheidsgraden en n = aantal waarnemingen, k + 1 = aantal β-parameters in het model.

Een 100(1 – α)% betrouwbaarheidsinterval voor een β-parameter:

$$\hat{\beta}_i \pm t_{\alpha/2}\, s_{\hat{\beta}i}$$

Waarbij $t\alpha/2$ $n – (k + 1)$ vrijheidsgraden heeft en n = aantal waarnemingen, k + 1 = aantal β-parameters in het model.

Waarschuwing:

Als je $H0 : \beta i = 0$ niet kunt verwerpen, zijn er verschillende conclusies mogelijk:

- Er bestaat geen verband tussen y en x.
- Er bestaat een verband tussen y en x, maar er is een type II fout opgetreden.
- Er bestaat een verband tussen y en x, maar dit verband is complexer dan een lineair verband. Het enige dat je over de β-parameter kunt zeggen is dat er voldoende (als je $H0 : \beta i = 0$ verwerpt) of onvoldoende (als je $H0 : \beta i = 0$ niet verwerpt) bewijs is voor een lineaire relatie tussen y en x.

Multipele-determinatiecoëfficiënt (R^2):

$$R^2 = 1 - \frac{SSE}{SSyy} = \frac{SSyy - SSE}{SSyy} = \frac{\text{Verklaarde spreiding}}{\text{Totale spreiding}}$$

Waarschuwing:

Gebruik in het multipele-regressieanalyse de waarde R^2 als maat voor de bruikbaarheid van het model voor het voorspellen van y alleen als de steekproef aanzienlijk meer meetwaarden bevat dat het aantal β-parameters in het model.

Gecorrigeerde multipele-regressiecoëfficiënt:

$$R_a^2 = 1 - \left[\frac{(n-1)}{n-(k+1)}\right]\left(\frac{SSE}{SS_{yy}}\right)$$

$$= 1 - \left[\frac{(n-1)}{n-(k+1)}\right](1 - R^2)$$

Bovemij Crowe EY Grant Thornton KPMG pwc

Het toetsen van de algemene bruikbaarheid van het model: de variantieanalyse F-toets:

- Hypothesen:
  $H0 : \beta 1 = \beta 2 = … = \beta k = 0$
  Ha : Ten minste één $\beta i \neq 0$

- Toetsingsgrootheid:

$$F = \frac{(SSyy - SSE) / k}{SSE / [n - (k + 1)]} = \frac{R^2 / k}{(1 - R^2) / [n - (k+1)]}$$

$$= \frac{\text{Kwadratisch gemiddelde (model)}}{\text{Kwadratisch gemiddelde (fout)}}$$

- Verwerpingsgebied:
  $F > F\alpha$, met k vrijheidsgraden in de teller en $[n - (k + 1)]$ vrijheidsgraden in de noemer.

Waarschuwing:
Een verwerping van de nulhypothese in de globale F-toets leidt tot de conclusie dat het model statistisch bruikbaar is. Het wil niet zeggen dat het model het beste is, maar is meer een maatstaf waaraan het model dient te voldoen wil het in aanmerking komen.

Aanbeveling voor het toetsen van de bruikbaarheid van een multipele-regressiemodel:

1. Eerst de bruikbaarheid van het model toetsen met behulp van de F-toets.
   $H0 : \beta 1 = \beta 2 = … = \beta k = 0$
   Als het model bruikbaar is (dus de nulhypothese verworpen) kan verder gegaan worden, anders moeten er aanpassingen gemaakt worden of een nieuw model.
2. Voer t-toetsen uit op die β-parameters waarin je het meest bent geïnteresseerd. Het is verstandig om het aantal β's dat wordt getest te beperken en het uitvoeren van t-toetsen leidt tot een verhoogde kans op een fout van de eerste soort.

College 8: (Hoofdstuk 16)

Eigenschappen van het multinomiale experiment:

1. Het experiment bestaat uit n identieke deelexperimenten.
2. Er zijn k mogelijke uitkomsten van elk deelexperiment. Deze uitkomsten worden klassen of cellen genoemd.
3. De kansen van de k uitkomsten, die worden aangegeven met p1, p2, …, pk blijven hetzelfde van deelexperiment tot deelexperiment, waarbij $p1 + p2 + … + pk = 1$
4. De deelexperimenten zijn onderling onafhankelijk.
5. De relevante variabelen zijn de tellingen, n1, n2, …, nk van het aantal waarnemingen dat in elk van de k uitkomsten (klassen) valt. (ni = telling voor cel i)

Het toetsen van hypothese over multinomiale kansen: enkelvoudige tabel:

- Hypothesen:
  H0: p1 = p1,0,p2 = p2,0,…, pk = pk,0 (p1,0, p2,0,…, pk,0 zijn de waarden van de multinomiale kansen)
  Ha: ten minste één van de multinomiale kansen is niet gelijk aan de hypothetische waarde.

- Toetsingsgrootheid:

$$\chi^2 = \sum \frac{[ni - E(ni)]^2}{E(ni)}$$

Bovemij Crowe EY GrantThornton KPMG pwc

waarbij $E(n_i) = n\, p_{i,0}$ de verwachte celfrequentie is, dat wil zeggen het verwachte aantal uitkomsten van type i, als we aannemen dat H0 juist is. De totale steekproefomvang is n.

- Verwerpingsgebied:
  $\chi^2 > \chi_\alpha^2$
  waarbij χ_α^2 $(k - 1)$ vrijheidsgraden heeft.

(Aannames: Er is een multinomiale experiment uitgevoerd, de steekproefomvang n is groot genoeg, zodat voor elke cel de verwachte frequentie $E(n_i)$ gelijk is aan 5 of meer.)

Algemene vorm voor een kruistabelanalyse: een toets voor onafhankelijkheid:

- Hypothesen:
  H0: De twee classificeringen zijn onafhankelijk.
  Ha: De twee classificeringen zijn afhankelijk.

- Toetsingsgrootheid:
  $$\chi^2 = \sum \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}$$
  waarbij $\hat{E}(n_{ij}) = \frac{r_i c_j}{n}$
  r_i = totaal voor rij i.
  c_j = totaal voor kollom j.

- Verwerpingsgebied:
  $\chi^2 > \chi_\alpha^2$
  waarbij χ_α^2 $(r - 1)(c - 1)$ vrijheidsgraden heeft.

(Aannames: De n waargenomen tellingen zijn een aselecte steekproef uit de populatie die we onderzoeken. We kunnen dit beschouwen als een experiment met r * c uitkomsten.
De steekproefomvang n is groot genoeg, zodat voor elke cel de verwachte telling $\hat{E}(n_{ij})$ gelijk is aan 5 of meer.)

SPSS commando's:

- **Frequentietabel**
  Analyse → discriptive statistics → frequencies (statistics en/of charts invoeren)

- **Hercoderen**
  Transform → recode into different variables → 'naam geven' → change → old & new values → value labels invoeren

- **Kruistabel**
  Analyse → discriptive statistics → crosstabs (statistics en cells invoeren)

- **Selecteren van case**
  Data → select cases

- **Data gegevens over case**
  Analyse → discriptive statistics → explore

- **Twee gemiddelden vergelijken**
  Analyse → compare meanse → one sample T test

Bovemij Crowe EY GrantThornton KPMG pwc

- **Twee groepen vergelijken**
  Analyse → compare means → independent sample T test

- **Twee variabelen met elkaar vergelijken**
  Analyse → compare means → paired sample T test

- **Bivariate samenhang**
  Analyse → correlate → bivariate

- **Lineaire regressie analyse**
  Analyse → regression → lineair

Kennen en kunnen lijstje

Kennen

- Beschrijvende statistiek
- Verklarende statistiek
- Populatie
- Variabele
- Steekproef
- Statistische gevolgtrekking
- Betrouwbaarheids maat
- Proces
- Blackbox
- Kwantitatieve gegevens
- Kwalitatieve gegevens
- Statistiek
- Statisch denken
- Frequentie per klasse
- Relatieve frequentie per klasse
- Pareto-analyse
- Puntendiagram
- Stam-en-bladdiagram
- Histogram
- Sommatienotatie
- Gemiddelde
- Steekproefgemiddel de
- Populatiegemiddeld e
- Mediaan
- Modus
- Bereik
- Steekproefvariantie
- Standaarddeviatie
- Percentiel
- Uitbijter
- Boxplot
- Spreidingsdiagram
- Tijdreeks
- Experiment
- Uitkomstenruimte
- Kans van uitkomst
- Kans op een gebeurtenis (P)
- Complement van een gebeurtenis
- Onvoorwaardelijke kans
- Voorwaardelijke kans
- Aselecte steekproef
- Random variabele
- Dichotome resultaten
- Binomiale random variabele
- Continue random variabele
- Uniforme kansverdeling
- Normaliteitsplot
- Continuïteitscorrect ie
- Parameter
- Steekproefgroothei d
- Steekproefverdeling
- Puntschatter
- Schattingsfout

Bovemij Crowe EY GrantThornton KPMG pwc

- Zuivere schatter
- Onzuivere schatter
- Intervalschatter
- Betrouwbaarheidscoëfficiënt
- Betrouwbaarheid
- Gestratificeerde steekproef
- Systematische steekproef
- Gerandomiseerde responssteekproef
- Non-respons
- Chi-kwadraat verdeling
- Normaliteitsplot
- Continuïteitscorrectie
- Deterministisch model
- Random fout

Kunnen

- Elementen van beschrijvende statistiek herkennen
- Elementen van verklarende statistiek herkennen
- Scheefheid bepalen
- De empirische regel kennen en toepassen
- Z-score berekenen en opzoeken
- Oorzaken van een uitbijter herleiden
- Uitbijters vinden
- Kansregels voor uitkomsten
- Stappenplan voor berekenen van kans op gebeurtenis kennen
- Noteren van samengestelde gebeurtenissen
- Optellen van kansen van complementaire gebeurtenissen
- Somregel voor gebeurtenissen
- Formule voor voorwaardelijke kans kennen
- Een kansboom begrijpen
- Normaalverdeling kennen
- Stappenplan bepalen kans normaal verdeelde variabele
- Binomiale kansen benaderen met een normale verdeling
- Steekproefomvang bepalen
- Fouten in steekproefonderzoek
- De elementen van een hypothese toetsen kennen
- Verschil tussen Type I en type II fouten
- P-waarde berekenen
- Hypothese toetsen
- Eigenschappen van de steekproefverdeling
- F-toets
- Algemene vorm van een kansmodel
- Eerste orde model
- Multipele-regressiemodel
- Multipele-determinatiecoëfficiënt
- Eigenschappen van multinomiale experiment benoemen

Disclaimer

ESV Nijmegen spant zich in om de inhoud van dit materiaal zo vaak mogelijk te actualiseren en/of aan te vullen. Ondanks deze zorg en aandacht is het mogelijk dat inhoud onvolledig en/of onjuist is. Het aangeboden materiaal is het hulpmiddel bij het studeren naast de literatuur die opgegeven is door de docent. De aangeboden materialen worden aangeboden zonder enige vorm van garantie of aanspraak op juistheid.

Alle rechten van intellectuele eigendom betreffende deze materialen liggen bij ESV Nijmegen. Kopiëren, verspreiden en elk ander gebruik van deze materialen is niet toegestaan zonder schriftelijke toestemming van ESV Nijmegen, behoudens en slechts voor zover anders bepaald in regelingen van dwingend recht, tenzij bij specifieke materialen anders aangegeven is.

Tips en opmerkingen over de samenvatting kunnen naar secretaris@esvnijmegen.nl gestuurd worden.