

STATISTIEK EXAM 2023-2024

Prof M. Hubert

Liam Luys

KU Leuven

Veel succes met het examen!

Ik heb geprobeerd een samenvatting te maken van de cursustekst.

Geen garantie dat alles er in staat en/of juist is, doe zeker de boeken ook open.

Foutje gevonden? Laat het zeker weten, dan pas ik het aan!

1. Doel van de Statistiek

1.1 Inleiding

Doel Statistiek

- Antwoorden formuleren op onderzoeksvragen over populatie, gebaseerd op waarnemingen van een steekproef uit die populatie

Statistische Analyse

1. Opzetten Relevante Experimenten
 - Verzamelen van gegevens
2. Exploratieve Analyse van Gegevens
 - Beschrijvende Statistiek
3. Beslissen adhv wiskundige modellen
 - Statistische Inferentie

1.2 Verzamelen van Gegevens

Populatie

- Wie of wat je wil onderzoeken
- Verzameling alle onderzoekseenheden

Steekproef

- Hoeveel onderzoekseenheden je onderzoekt
- Deelverzameling van de populatie
- Steekproefgrootte n
- Moet representatief zijn voor de populatie, hoe groter n , hoe meer info

Variabele X

- Welke kenmerken je onderzoekt
- Kan verschillende waardes/uitkomsten hebben
- Uitkomstverzameling φ

Toevalsvariabele X

- Uitkomst van X ligt niet op voorhand vast, hangt af van toeval
- Notatie altijd met hoofdletter!

Resultaten Steekproef grootte n

- Bij 1 toevalsvariabele X
 - $x_1, x_2, \dots, x_n$
- Bij 2 toevalsvariabelen X en Y
 - $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Bij p toevalsvariabelen $X_1, \dots, X_p$
 - $x_{ij}, \quad i = \{1, \dots, n\}, \quad j = \{1, \dots, p\}$

Gegevensmatrix

- Rijen: Onderzoekseenheden
- Kolom: Toevalsvariabelen

1.3 Classificatie van Gegevens

Gegevens opdelen in types met elk hun onderzoeksstrategieën

1.3.1 Kwalitatieve en kwantitatieve Gegevens

Kwalitatieve (categorische) variabelen

- Nominaal
 - o Niveaus worden enkel gebruikt om onderzoekseenheden te klasseren
 - o Niet interpreteren als meer of minder
 - o Code voor elke categorie die kenmerk voorstelt, geen wiskundige betekenis
Voorbeeld: Geslacht, Provincies, Kleur van object...
- Ordinaal
 - o Elementen geordend volgens bepaalde ordeningsklasse
 - o Klein - Groot, Veel - Weinig, Sterk - Zwak, Eerste - Laatste, ...
 - o Voorgesteld door cijfers om te vergelijken, niet rekenen
Voorbeeld: Inkomstenniveaus, gevaarcodes, opleidingsniveau

$$\varphi = \{m_1, m_2, \dots, m_k\}$$

Kwantitatieve (metrische) variabelen

- Verschil tussen meetwaarden in hoeveelheden
- Vaste meeteenheid
- Rekenen mogelijk, verschillen en hoeveelheden berekenbaar (numerieke betekenis)
Voorbeeld: Leeftijd, tijdsduur, ...

$$\varphi = \{m_1, m_2, \dots, m_k\} \subset \mathbb{R}$$

1.3.2 Discrete en Continue Gegevens

Op basis van uitkomstenverzameling φ onderscheid maken

Discreet

- Beperkt aantal mogelijke uitkomsten
- Tussen twee opeenvolgende waardes ligt geen derde waarde

Continu

- Onbeperkt (bijna) aantal mogelijke uitkomsten
- Voor elke twee waardes in φ is het mogelijk er een derde te vinden die ertussen ligt

Meetniveau	Uitkomstenverzameling	
	Discreet	Continu
Kwalitatief (Nominaal/Ordinaal)	eindig: $\mathcal{S} = \{m_1, m_2, \dots, m_k\}$	
Kwantitatief	(on)eindig: $\mathcal{S} = \{m_1, m_2, \dots, m_k\} \subset \mathbb{R}$	oneindig: $\mathcal{S} \subset \mathbb{R}$

Belangrijke Opmerkingen

- Nauwkeurigheid van meetgegevens is beperkt tot nauwkeurigheid meetinstrument
 - o Hierdoor zullen continue variabele discreet worden voorgesteld in steekproefgegevens
- Als discrete toevalsvariabele zeer veel verschillende mogelijke uitkomsten heeft wordt deze vaak beschouwd als continue toevalsvariabele

- Kwalitatieve Gegevens zijn steeds discreet

1.4 Van Steekproef naar Populatie

Het doel is om antwoorden op vragen over de populatie te beantwoorden adhv gegevens over een steekproef

Niet altijd makkelijk

- Keuze van steekproef
- Grootte van steekproef
- Resultaten van steekproef

Wet van grote aantallen

- Hoe groter de steekproef, hoe beter de overgang naar populatie
- Zeer belangrijk in statistiek

2. Beschrijvende Statistiek

2.1 Absolute en Relatieve Frequentie

Hoe zijn uitkomsten in (steekproef)groep verdeeld?

2.1.1 Staafdiagram voor een kwalitatieve variabele

Kwalitatieve Variabele

- Gegevens: $x_1, \dots, x_n$
- $\varphi = \{m_1, m_2, \dots, m_k\}$

Absolute Frequentie n_j

- Aantal uitkomsten x_i die gelijk zijn aan m_j

Relatieve Frequentie / Steekproefproportie

- $f_{j,n} = n_j/n$
- Som van alle relatieve frequenties = 1

$$\sum_{j=1}^k f_{j,n} = \frac{n_1}{n} + \frac{n_2}{n} + \dots + \frac{n_k}{n} = 1$$

- Kan je voorstellen in frequentietabel

Grafische Voorstelling (Discreet)

- Staafdiagram
- Taartdiagram
- Donutdiagram

2.1.2 Histogram voor een kwantitatieve Variabele

Kwantitatieve Variabele

- $\varphi = \{m_1, m_2, \dots, m_k\} \subset \mathbb{R}$
 - o φ eindig
 - Frequentietabel, Staafdiagram, ...
 - o φ oneindig
 - Discretisatie / Groepering van gegevens

Discretisatie

- φ opdelen in k aangrenzende deelintervallen/klassen.
 - o Klassen van gelijke breedte met klassenbreedte Δ
- Een nieuwe discrete uitkomstverzameling ontwikkelen met $\varphi = \{m_1, m_2, \dots, m_k\}$, met m_j de klassenmiddens

Grafische Voorstelling (Continu)

- Histogram
 - o Verschillende klassenbreedtes geven vaak een ander beeld van de gegevens!

Frequentiedichtheden

- Om een goed beeld weer te geven van de gegevens in een histogram
 - Verticale herschaling door hoogte van balk gelijk te nemen als relatieve frequentie gedeeld door klassenbreedte
- $$h_j = h(m_j) = \frac{f_{j,n}}{\Delta_j} = \frac{n_j}{\Delta_j n}$$

Histogrammen gebaseerd op frequentiedichtheden stellen relatieve frequenties voor dmv oppervlaktes, niet hoogtes!

2.2 Vormen en Verdelingen

Patroon Histogram = Weerspiegeling verdeling van beschikbare steekproefgetallen

2.2.1 Symmetrische Verdelingen

Symmetrisch Histogram

- Aan linker- en rechterzijde hetzelfde gedrag
- Horizontale spiegeling om het midden

2.2.2 Normale Verdeling

Normale / Gaussische Verdeling

- Type van symmetrische verdeling
- Neemt langs beide zijden af
- Hoe groter steekproef, hoe meer histogram klokcurve benadert

Niet altijd mogelijk om direct te zien of gegevens normaal verdeeld zijn (zeker niet bij kleinere steekproefgrootte)

2.2.3 Andere Verdelingen

Scheve Verdeling

- Rechtsscheef: Rechts zwakt trager af dan links
- Linksscheef: Links zwakt trager af dan rechts

Verdeling met lange/zware staarten

- Staarten vallen trager naar nul dan bij klokvorm

Verdeling met korte/lichte staarten

- Staarten vallen sneller naar nul dan bij klokvorm

Bimodale Verdeling

- Twee / Meerdere toppen voorkomen in histogram

Uitschieters

- Eén getal of weinig getallen niet bij de rest van de gegevensgroep
 - o Klein of groot ten opzichte van de rest
- Belangrijk om te achterhalen waarom (Fout? Toeval? ...)
- Nadenken wat ermee te doen
 - o Verwijderen? Vaak niet nuttig wegens infoverlies

2.3 Cumulatieve Frequenties en Kwantielen

2.3.1 Cumulatieve Verdelingsfunctie

Voor metrische (kwantitatieve) gegevens kan cumulatieve verdelingsfunctie vervanging bieden voor frequentietabellen

Empirische Cumulatieve Verdelingsfunctie

- Relatief aantal steekproefuitkomsten x_i die kleiner zijn dan x
- $\hat{F}_n(x) = \frac{1}{n}$ (met n aantal waarden x_i waarvoor geldt: $x_i \leq x$)
- $\hat{F}_n(x_{(i)}) = \frac{i}{n} \rightarrow$ indien geen samenvallende waarden zijn

2.3.2 Kwantielfunctie

Gegeven een percentage p , welk getal x situeert zich in de overeenkomstige positie in de steekproef?

Empirische Kwantielfunctie $\hat{Q}_n(p)$

- Inverse van verdelingsfunctie $\hat{F}_n(x)$
- Geeft x waarvoor $\hat{F}_n(x) \geq p$ met $0 \leq p \leq 1$
- $\hat{Q}_n(p)$ is het kleinste getal x waarvoor $\hat{F}_n(x) \geq p$
- $\hat{Q}_n(p) = x_{(i)}$ indien $\frac{i-1}{n} < p \leq \frac{i}{n}$

Kwantielen

- Kwantielfuncties met
 - o $p = 0.25$ (Q_1)
 - o $p = 0.50$ (Q_2) (mediaan)
 - o $p = 0.75$ (Q_3)
- Bij symmetrie: $Q_2 - Q_1 \approx Q_3 - Q_2$
- Bij asymmetrie:
 - o $Q_2 - Q_1 \gg Q_3 - Q_2 \rightarrow$ Rechtsscheef
 - o $Q_2 - Q_1 \ll Q_3 - Q_2 \rightarrow$ Linksscheef

2.4 Centrumkenmerken

Hoe redeneren over gegevens?

- Grafische voorstellingen
 - o Handig om visueel te bekijken (diagrammen, histogrammen, ...)
- Numerieke waarden
 - o Informatie reduceren naar getallen

2.4.1 Steekproefgemiddelde

Steekproefgemiddelde $\bar{x}$

- Meest bekende manier om centrum van verdeling aan te duiden
- Zeer gevoelig voor uitschieters!

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

$$- \bar{x} = \frac{1}{n} \sum_{j=1}^k m_j n_j = \sum_{j=1}^k m_j f_{j,n} \rightarrow \text{Bij Discrete gegevens}$$

$\varphi = \{m_1, m_2, \dots, m_k\}$ $k = \# \text{mogelijke uitkomsten}$ $n_j = \text{absolute frequentie}$

2.4.2 Mediaan

Steekproefmediaan

- Geeft de middelste waarde van de geordende steekproef weer
- Veel minder gevoelig voor uitschieters!

$$- \text{med}(x_1, \dots, x_n) = \begin{cases} x_{((n+1)/2)} & \rightarrow n \text{ onveven} \\ (x_{(n/2)} + x_{((n/2)+1)})/2 & \rightarrow n \text{ even} \end{cases}$$

Bij symmetrische verdeling: Gemiddelde en mediaan ongeveer gelijk!

2.4.3 Getrimd Gemiddelde

Getrimd Gemiddelde

- Vast percentage van de kleinste en grootste gegevens wordt weggelaten, van de overige gegevens wordt het gemiddelde berekend
- Voordelen
 - o Meer reguliere observaties worden behouden dan bij berekening van mediaan
- Nadelen
 - o Indien trimpercentage te groot --> veel reguliere observaties weggelaten
 - o Indien trimpercentage te klein --> te weinig uitschieters verwijderd

2.4.4 Modus

Modus

- Meest voorkomend element in oplossingsverzameling
- Hoogste frequentie

2.5 Spreidingskenmerken

Hoe liggen gegevens verspreid rond de centrummaten?

2.5.1 Bereik

Bereik

- Afstand tussen grootste en kleinste waarde
- $R = x_{(n)} - x_{(1)}$
- Nadelen
 - o Weinig informatief, Afhankelijk steekproefgrootte, erg gevoelig aan uitschieters

2.5.2 Standaardafwijking en Variantie

Gemiddelde afstand tot gemiddelde

- Als een goede spreidingsmaat?

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

- Wordt niet gebruikt, wel aangepast naar andere grootte

Standaarddeviatie / Standaardafwijking

- Maat voor de spreiding van gegevens
- Goede wiskundige eigenschappen
- Zeer gevoelig voor uitschieters

$$- s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \xrightarrow{\text{Indien discrete gegevens}} s = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (m_j - \bar{x})^2 n_j} = \sqrt{\frac{n}{n-1} \sum_{j=1}^n (m_j - \bar{x})^2 f_{j,n}}$$

Variantie

- Kwadraat van standaarddeviatie

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

Interpretatie Standaardafwijking

- Hoe excentrisch observatie of steekproefwaarde is ten opzichte van centrumkenmerk (gemiddelde)
- Bij Normale Verdeling
 - o Ongeveer 2/3 van gegevens binnen één standaardafwijking van steekproefgemiddelde
 - o Ongeveer 95% binnen twee standaardafwijkingen van steekproefgemiddelde
 - o Die regels kunnen ook geformuleerde adhv de z-score

Z-Score

- $z_i = \frac{x_i - \bar{x}}{s}$ met s de standaardafwijking
- Bij normale verdeling:
 - o 2/3 Gegevens: $-1 < z < 1$
 - o 95% gegevens: $-2 < z < 2$

2.5.3 Interkwartielafstand

Interkwartielafstand IQR

- Lengte van gebied rond de mediaan met ongeveer de helft van de gegevens
- Veel minder invloed van uitschieters
- $IQR = \hat{Q}_n(0.75) - \hat{Q}_n(0.25)$

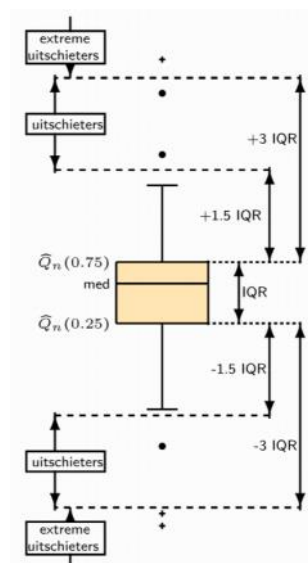
- Normale verdeling: $\frac{IQR}{s} \sim 1.34$

2.5.4 Median Absolute Deviation

Median Absolute Deviation MAD

- Mediaan van afstanden van elke observatie tot de mediaan
- Zeer ongevoelig aan uitschieters
- $MAD = med_{i=1, \dots, n} |x_i - med(x_1, \dots, x_n)|$
- Normale verdeling: $\frac{MAD}{s} \sim 0.67 \Rightarrow \frac{MAD}{0.67} \sim s$ en $\frac{IQR}{MAD} \sim 2$

2.6 Boxplot



Inhoud Boxplot

- Centrum
 - o Mediaan zijn als centrummaat aangeduid
- Spreiding
 - o Lengte van de doos is de Interkwartielafstand IQR
- Scheefheid
 - o Symmetrie --> Mediaan in midden van de doos
 - o Rechtsscheef --> Mediaan in onderste helft, bovenste whiskers (snorharen) langer
 - o Linksscheef --> Mediaan in bovenste helft
- Zwaarte van de staarten
 - o Veel uitschieters --> Zwaarstaartige verdeling
 - Lengte snorharen gebaseerd op normale verdeling

Detectie Uitschieters (bij normale verdeling)

- Boxplot-Regel
 - o Kans dat observaties whiskers overschrijden is 0.7%
 - o Werkt goed als max 25% uitschieters zijn
- Via Z-Score

- Als $|z_i| > 2.58$
- Uitschieters beïnvloeden $\bar{x}$ en s waardoor niet altijd klopt
- Via Robuuste Z-Score
 - Minder gevoelig voor uitschieters (tot 50%)
 - $$\frac{x_i - \text{med}(x_1, \dots, x_n)}{MAD/0.67}$$

2.7 Transformaties

Soms nuttig om gegevens te transformeren om betere weergave/verdeling te krijgen

2.7.1 Lineaire Transformatie

$$g(x_i) = a + bx_i \quad \text{met } a, b \in \mathbb{R}$$

Lineaire Transformatie

- Gegevens x_i verschoven over afstand a en herschaalt met factor b
- Voorbeeld: Omzetten meeteenheden

$$\overline{g(x)} = a + b\bar{x} = g(\bar{x})$$

$$\text{med}(g_1, \dots, g_n) = a + b \text{med}(x_1, \dots, x_n) = g(\text{med}(x_1, \dots, x_n))$$

$$\hat{Q}_g(p) = a + b\hat{Q}_x(p)$$

$$s_g = |b|s_x$$

$$s_g^2 = b^2 s_x^2$$

$$\bar{z} = \frac{x_i - \bar{x}}{s_x} = -\frac{\bar{x}}{s_x} + \frac{1}{s_x}\bar{x} = 0$$

2.7.2 Monotoon Stijgende Transformaties

Monotoon Stijgend (niet-lineaire) Transformatie

- Als $x_1 < x_2$ dan $g(x_1) < g(x_2)$
 - Orde van gegevens wijzigt niet!
- Voorbeeld: $\log$, $\exp$, x^2 , $\text{wortel}(x)$

$$\overline{g(x)} \neq g(\bar{x})$$

$$s_g \neq g(s_x)$$

$$\hat{Q}_g(p) = g(\hat{Q}_x(p))$$

$$\text{med}(g_1, \dots, g_n) = g(\text{med}(x_1, \dots, x_n)) \quad \text{als } n \text{ oneven}$$

$$\text{med}(g_1, \dots, g_n) \approx g(\text{med}(x_1, \dots, x_n)) \quad \text{als } n \text{ even}$$

2.8 Verbanden tussen twee variabelen

2.8.1 Twee Kwalitatieve Variabelen

Twee kwalitatieve Variabelen

- Kruistabel
 - Tabel met absolute frequenties van voorkomens van combinatie variabelen
 - Totalen van rij- en kolommen toegevoegd
- Bivariate staafdiagrammen
 - Hoogte absolute frequenties
 - 3 dimensies

- Gegroepeerde staafdiagrammen
 - o Staafdiagram gegroepeerd per waarde van een variabele
- Werken met relatieve frequenties in staafdiagram/kruistabel
 - o Handig om verbanden te zoeken/berekenen

2.8.2 Een kwantitatieve en een kwalitatieve variabele

Kantitatieve en kwalitatieve variabele

- Verdeling van kwantitatieve variabele over kwalitatieve variabele
 - o Wordt snel onoverzichtelijk
- Boxplots voor elke waarde van de kwalitatieve variabele

2.8.3 Twee kwantitatieve variabelen

Twee kwantitatieve Variabelen

- Histogram in 3 dimensies (bivariaat histogram)
 - o Twee variabele opdelen in klassen en voor elke combinatie van klassen een absolute frequentie

2.8.4 Covariantie en Correlatiecoëfficiënt

Puntenwolk / Spreidingsdiagram

- Voor elke observatie de meting voor ene variabele op de x-as, meting voor de andere op y-as
- Associatie afleiden
 - o Stijging naar rechts -> Positieve associatie
 - o Stijging naar links -> Negatieve associatie
 - o Hoe dikker de wolk, hoe zwakker de associatie
Sterke associatie: kennis over de ene variabele helpt de andere variabele voorspellen

Steekproefcovariantie

- Maat voor sterkte associatie
- $$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$s_{xy} = s_{yx} \quad s_{xx} = s_x^2$$

- Herschalen naar Pearson Correlatiecoëfficiënt

Pearson Correlatiecoëfficiënt

- Maat voor lineaire associatie
- Enkel nuttig voor **lineaire** associatie!
- Herschaling: afstanden $(y_i - \bar{y})$ uitdrukken in standaardafwijkingen, niet meer in waarden
- $$r_{xy} = r = \frac{s_{xy}}{s_x s_y} \quad \text{en} \quad -1 \leq r \leq 1$$
- Indien geen associatie tussen X en Y ligt r dicht bij 0
 - o Omgekeerd geldt NIET! Als r dicht bij 0 ligt kan er nog steeds een niet-lineair verband zijn.
- Gevoelig aan uitschieters

Spearman Correlatiecoëfficiënt

- Maat voor stijgende curve op puntenwolk
- Geobserveerde waarden vervangen door rangnummers
 - o Observaties ordenen en van klein naar groot rangen toekennen vanaf 1
 - o Bij gelijke observaties de rangen het gemiddelde nemen van de rangen
- Pearson Correlatiecoëfficiënt gebruiken op rangen

2.9 Lineaire Combinaties

Combinatie van meerdere variabelen

$$V = a + bX + cY$$

Als $v_i = a + bx_i + cy_i$, dan is

$$\bar{v} = a + b\bar{x} + c\bar{y}$$

$$s_v^2 = b^2 s_x^2 + c^2 s_y^2 + 2bc s_{xy}$$

Als $v_i = x_i - y_i$, dan is $\bar{v} = \bar{x} - \bar{y}$ en $s_v^2 = s_x^2 + s_y^2 - 2s_{xy}$.

Als $w_i = \frac{1}{2}x_i + \frac{1}{2}y_i$, dan is $\bar{w} = \frac{1}{2}\bar{x} + \frac{1}{2}\bar{y}$, en $s_w^2 = \frac{1}{4}s_x^2 + \frac{1}{4}s_y^2 + \frac{1}{2}s_{xy}$.

2.10 Samenvatting

Zie cursus

3. Kansen en Toevalsvariabelen

3.1 Kansen en Kansregels

Gebeurtenissen

- Deelverzameling van φ
- Eventualiteiten waaraan uitkomsten van een experiment voldoen
- Kunnen beschreven worden met kansen
- Kans op gebeurtenis A: $P(A)$
- Totale kans = 1!
 - o $P(\varphi) = 1$

De relatieve frequentie van een gebeurtenis A onder een groot aantal proeven stabiliseert naar een getal, de kans op die gebeurtenis

Voorstellingen kansen

- Kanstabel
- Venndiagram
 - o $P(A)$ met $A \subset \varphi$: oppervlakte tov φ die gelijk is aan 1

Eigenschappen van kansen A en B

- Complement
 - o A^c en $B^c \rightarrow$ Kans dat A en B niet optreden
 - o $P(A^c) = 1 - P(A)$
- Unie
 - o $A \cup B \rightarrow$ Kans dat A of B optreedt (of beide)
 - o $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Doorsnede
 - o $A \cap B \rightarrow$ Kans dat A en B optreden
 - o $P(A \cap B) = P(A) + P(B) - P(A \cup B)$
- Verschil
 - o $A \setminus B \rightarrow$ Kans dat A maar niet B optreedt
 - o $P(A \setminus B) = P(A) - P(B)$ (als $B \subset A$)

3.2 Voorwaardelijke Kansen

3.2.1 Onafhankelijkheid van Gebeurtenissen

Twee gebeurtenissen A en B zijn onafhankelijk als de kans op het voorkomen van de ene gebeurtenis A niet beïnvloed wordt door het al dan niet optreden van de andere gebeurtenis B, en omgekeerd.

Voorbeeld

Lukraak experiment met teruglegging --> Onafhankelijk

Lukraak experiment zonder teruglegging --> Afhankelijk

Voorwaardelijke kans

- "De kans op A als je weet dat B geldt"
- $P(A | B) = \frac{P(A \cap B)}{P(B)}$ als $P(B) \neq 0$
- $P(A \cap B) = P(A) * P(B | A) = P(B) * P(A | B)$

Onafhankelijkheid Kansen

- Gebeurtenissen $A_1, \dots, A_n$ zijn onderling onafhankelijk als
 $P(A_1 \cap \dots \cap A_n) = P(A_1) * \dots * P(A_n)$

3.2.2 Wet van de Totale Kans en Regel van Bayes

Wet van de Totale Kans

Als $\mathcal{S} = A_1 \cup A_2 \cup \dots \cup A_k$ met $A_i \cap A_j = \emptyset$ voor $i \neq j$ (partitie), dan geldt de **wet van de totale kans**:

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_k) \\ &= P(B | A_1)P(A_1) + P(B | A_2)P(A_2) + \dots + P(B | A_k)P(A_k) \end{aligned}$$

Regel van Bayes

- Omvormen formule wet van totale kans
- Grondlegger voor Bayesiaanse Statistiek
- $P(A | B) = \frac{P(B | A) P(A)}{P(B)}$ als $P(B) \neq 0$
- $P(A_1 | B) = \frac{P(B | A_1) P(A_1)}{P(B)} = \frac{P(B | A_1) P(A_1)}{\sum_{i=1}^k P(B | A_i) P(A_i)}$

3.3 Toevalsvariabele

Een reële toevalsvariabele neemt numerieke waarden aan die overeenstemmen met de uitkomsten van een experiment dat onderhevig is aan het toeval.

3.4 Dichtheidsfunctie

3.4.1 Dichtheid van een discrete toevalsvariabele

Dichtheid

- $f(m_j) = P(X = m_j)$ voor $j = 1, 2, \dots, k$
- Voldoet aan: $\begin{cases} f(m_i) \geq 0 \\ \sum_{m_j \in \Phi} f(m_j) = 1 \end{cases}$ voor $i = 1, 2, \dots, k$

Uniforme Verdeling

- $f(m_j) = \frac{1}{k}$ voor $j = 1, \dots, k$

Dichtheid om kansen te berekenen

$$- P(X \in A) = \sum_{m_j \in A} f(m_j)$$

Belangrijk bij discrete variabele!

$$P(X = x) \neq 0 \Rightarrow P(X < x) \neq P(X \leq x) \text{ en } P(X > x) \neq P(X \geq x)$$

3.4.2 Dichtheid van continue toevalsvariabele

Continue variabele

- Gebaseerd op frequentiedichtheden
- Opsplitsing in k klassen
- Overgang van steekproef naar populatie:
 - o Steekproefgrootte n laten toenemen, klassenbreedte Δ tot nul naderen
 - o ($n \rightarrow \infty$ en $\Delta \rightarrow 0$)

Histogram van Continue Variabele

- $f(x) \geq 0 \rightarrow$ Elke $v \in \varphi$ heeft kans groter of gelijk aan 0
- $\int_{\varphi} f(x) dx = 1 \rightarrow$ Totale oppervlakte onder histogram = 1
- $P(X = x) = 0 \rightarrow$ Kans op één waarde is 0
 - o $P(X < x) = P(X \leq x)$ en $P(X > x) = P(X \geq x)$
- $P(X \in A) = \int_A f(x) dx$

3.5 Verdelings- en Kwantielfunctie

3.5.1 Verdelings- en Kwantielfunctie voor Discrete Variabele

Discrete variabele

- $\varphi = \{m_1, m_2, \dots, m_k\}$
- $P(X \leq x) = \sum_{m_j \leq x} P(X = m_j)$

Cumulatieve Verdelingsfunctie

- Is niet nuttig bij nominale variabelen
- $F(x) = P(X \leq x) = \sum_{m_j \leq x} f(m_j)$ (met $f(m_j)$ de dichtheid)

Discrete Kwantielfunctie $Q(p)$

- Inverse van de verdelingsfunctie
- $Q(p)$ is het kleinste getal x waarvoor $F(x) \geq p$ ($0 < p < 1$)

3.5.2 Verdelings- en Kwantielfunctie voor een continue variabele

Continue Verdelingsfunctie

- $F(x) = P(X \leq x) = \int_{-\infty}^x f(y)dy$
- $F'(x) = \frac{dF(x)}{dx} = f(x)$
- $P(a \leq X \leq b) = \int_a^b f(x)dx$

Continue Kwantielfunctie $Q(p)$

- Inverse van de continue verdelingsfunctie
- $Q(p)$ is de x - waarde waarvoor $F(x) = p$ voor elke $0 < p < 1$

3.6 Centrum- en Spreidingskenmerken

Hoe centrum- en spreidingsmaten definiëren op niveau van populatie?

3.6.1 Verwachtingswaarde

Discreet

Verwachtingswaarde

- Steekproefgemiddelde naar populatieniveau

- $E(X) = \sum_{j=1}^k m_j f(m_j)$

Afleiding uit steekproefgemiddelde

$$\bar{x} = \sum_{j=1}^k m_j f_{j,n} \xrightarrow{f_{j,n} \text{ gaat de kans } f(m_j) \text{ benaderen bij grotere steekproeven}} E(X) = \sum_{j=1}^k m_j f(m_j)$$

Continu

Verwachtingswaarde

- Histogram voorstelling, $\Delta \rightarrow 0$
- Bestaat enkel als de som voor $\Delta \rightarrow 0$ stabiliseert naar een eindige waarde

- $E(X) = \int_{\varphi} x f(x)dx$

3.6.2 Mediaan

Mediaan

- Kwantielfunctie met $q = 0.5$
- Zowel bij continu als discreet bestaat mediaan altijd!
- $Med(X) = Q(0.5)$

3.6.3 Modus

Modus

- Waarde waar de dichtheid maximaal en uniek is
 - o Uniforme dichtheid --> Geen Modus

- Top van histogram

Als dichtheid f een modus heeft en

- Symmetrisch is rond μ
 - o $E(X) = Med(X) = \text{modus}(X) = \mu$
- Rechtsscheef is
 - o $\text{modus}(X) < Med(X) < E(X)$
- Linksscheef is
 - o $\text{modus}(X) > Med(X) > E(X)$

3.6.4 Variantie en Standaardafwijking

Dicreet

Discrete Variantie

$$- \text{Var}(X) = \sum_{j=1}^k (m_j - E(X))^2 f(m_j)$$

Afleiding uit steekproefvariantie

$$s^2 = \frac{n}{n-1} \sum_{j=1}^k (m_j - \bar{x})^2 f_{j,n} \xrightarrow{\text{benaderen bij grotere steekproef } n} \text{Var}(X) = \sum_{j=1}^k (m_j - E(X))^2 f(m_j)$$

Continu

Variantie

$$- \text{Var}(X) = \int_{\varphi} (x - E(X))^2 f(x) dx$$

Standaardafwijking

$$- \sigma = \sqrt{\text{Var}(X)}$$

3.6.5 Interkwartielafstand

Interkwartielafstand

- Variabiliteitsmaat
- $IQR(X) = Q(0.75) - Q(0.25)$

Steekproef	Populatie
Deelverzameling φ	φ
Relatieve Frequentie	Kans
$\bar{x}$	$E(X)$
$med(x_1, x_2, \dots, x_n)$	$Med(X)$
$\text{modus}(x_1, x_2, \dots, x_n)$	$\text{modus}(X)$
s^2	$\text{Var}(X)$
s	$\sigma = \sqrt{\text{Var}(X)}$
$IQR = \hat{Q}_n(0.75) - \hat{Q}_n(0.25)$	$IQR(X) = Q(0.75) - Q(0.25)$

3.7 Transformaties en lineaire combinaties van toevalsvariabelen

3.7.1 Verwachtingswaarde en variantie bij een algemene transformatie

$$E(g(X)) \neq g(E(X))$$

Discrete Variabele

$$E(g(X)) = \sum_{j=1}^k g(m_j) f(m_j)$$

$$\text{Var}(g(X)) = \sum_{j=1}^k \left(g(m_j) - E(g(X)) \right)^2 f(m_j)$$

Continue Variabele

$$E(g(X)) = \int_{\varphi} g(x) f(x) dx$$

$$\text{Var}(g(X)) = \int_{\varphi} \left(g(x) - E(g(X)) \right)^2 f(x) dx$$

3.7.2 Lineaire Transformaties

$$g(X) = a + bX \quad \text{met } a, b \in \mathbb{R}$$

$$\begin{aligned} E(a + bX) &= a + b E(X) \\ \text{Med}(a + bX) &= a + b \text{Med}(X) \\ Q_G(p) &= a + b Q_X(p) \quad \text{als } b > 0 \\ \text{Var}(a + bX) &= b^2 \text{Var}(X) \end{aligned}$$

$$\text{Var}(X) = E(X^2) - (EX)^2$$

Speciale functies

- Denk aan z-waarde

$$\text{Stel } g(X) = \frac{X - EX}{\sqrt{\text{Var}(X)}} = \frac{X - EX}{\sigma}$$

$$\begin{aligned} E(g(X)) &= 0 \\ \text{Var}(g(X)) &= 1 \end{aligned}$$

3.7.3 Monotoon Stijgende Transformaties

Monotoon Stijgend

- $\log(x)$, e^x , ...
- Speciale eigenschappen geldig

- $Q_G(p) = g(Q_G(x))$
- $Med(g(X)) = g(Med(X))$
- $Med(\log(X)) = \log(Med(X))$

3.7.4 Benaderende methode bij een algemene transformatie

Geen leerstof, niet kennen

3.7.5 Lineaire Combinaties van Toevalsvariabelen

Stel $v_i = a + bx_i + cy_i$

$$E(a + bX + cY) = a + bE(X) + cE(Y)$$

$$Var(a + bX + cY) = b^2Var(X) + c^2Var(Y) + 2bc Cov(X, Y)$$

X en Y onafhankelijk

- Als $P((X \in A) \cap (Y \in B)) = P(X \in A) P(Y \in B)$
- $Var(a + bX + cY) = b^2Var(X) + c^2Var(Y)$

3.8 Samenvatting

Zie cursus

4. Univariate Kansmodellen

$$DISCREET: \begin{cases} E(X) = \sum_{j=1}^k m_j f(m_j) \\ Var(X) = \sum_{j=1}^k (m_j - EX)^2 f(m_j) \end{cases} \quad CONTINU: \begin{cases} E(X) = \int_{\varphi} x f(x) dx \\ Var(X) = \int_{\varphi} (x - EX)^2 f(x) dx \end{cases}$$

4.1 Bernoulli Verdeling $X \sim B(1, p)$ **Bernoulli Verdeling**

- Variabele X met slechts 2 mogelijke uitkomsten
- $\varphi = \{0, 1\}$

$$X \sim B(1, p)$$

- Indien $X = 1$ noemen we het een succes
- Voorbeelden: Kop/Munt, twee verschillende kleuren,
- Kansen, $E(X)$ en $Var(X)$

$$f(1) = P(X = 1) = p$$

$$f(0) = P(X = 0) = 1 - P(X = 1) = 1 - p \quad \text{met } p = \text{kans op succes}$$

$$E(X) = 0 * (1 - p) + 1 * p = p$$

$$Var(X) = (0 - p)^2 (1 - p) + (1 - p)^2 p = p(1 - p)$$

4.2 Binomiaalverdeling $X \sim B(n, p)$ **Binomiaalverdeling**

- Wanneer herhaaldelijk Bernoulli-experimenten worden uitgevoerd is het aantal keer dat het succes voorkomt binomiaal verdeeld

$$\varphi = \{0, 1, 2, \dots, n\}$$

$$X \sim B(n, p) \rightarrow \text{Met } n \text{ het aantal experimenten, } p \text{ de kans op succes}$$

- $X \sim B(n, p)$ is binomiaal verdeeld met parameters n en p als $X = \sum_{i=1}^n X_i$ met $X_i \sim B(1, p)$ en onderling onafhankelijk

- Dichtheid bepalen door voor elke n in φ : $P(X = n)$ te kennen
- Aantal manieren om m keer waarde 1 te plaatsen in een rij van n posities
 - o "m uit n"
 - o Binomiaalcoëfficiënt $\binom{n}{m}$

$$\binom{n}{m} = \frac{n(n-1) \dots (n-m+1)}{m!} = \frac{n!}{m!(n-m)!} \quad \text{en} \quad \binom{n}{0} = 1$$

- Kansen, $E(X)$ en $\text{Var}(X)$

$$f(m) = P(X = m) = \binom{n}{m} p^m (1-p)^{n-m}$$

$$E(X) = \sum_{i=1}^n E(X_i) = np$$

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) = np(1-p)$$

$$X \sim B(n, p) \leftrightarrow n - X \sim B(n, 1-p)$$

4.3 Poisson Verdeling $Poisson(\lambda)$

Poisson Verdeling

- Aantal successen geteld binnen vaste tijds- of ruimte-interval
- $\varphi = \mathbb{N}$
- Mogelijk om interval op te splitsen in zeer kleine deelintervallen zodat
 - o Verwaarloosbare kans op meer dan 1 voorkomen
 - o Gelijke kans op voorkomen in elk deelinterval
 - o Onafhankelijkheid tussen deelintervallen
- De deelintervallen benaderen door binomiale verdeling met n zeer groot en p zeer klein

Stel $X \sim B(n, p)$ en $n \rightarrow \infty$ en $p \rightarrow 0$ waarbij $np \rightarrow \lambda > 0$,

★ Dan volgt: $P(X = m) \rightarrow \frac{e^{-\lambda} \lambda^m}{m!}$ **(KUNNEN AFLEIDEN! P166)**

- Kansen, $E(X)$ en $\text{Var}(X)$

$$f_{\lambda}(m) = P(X = m) = \frac{e^{-\lambda} \lambda^m}{m!}$$

$$E(X) = \lambda$$

$$\text{Var}(X) = \lambda \quad \text{kunnen afleiden uit binomiaal en omdat } np \rightarrow \infty \text{ en } p \rightarrow 0$$

Benadering van $B(n, p)$ met $Poisson(\lambda)$

Indien n groot genoeg, en p klein genoeg kunnen we de binomiaalverdeling benaderen met een poissonverdeling

- Gemakkelijker uit te schrijven
- Gemakkelijker te berekenen!

$$Poisson(\lambda) \approx B(n, p) \quad \text{als } n \geq 100 \text{ en } np < 5$$

4.4 Andere Discrete Verdelingen

4.5 Normale Verdeling $X \sim N(\mu, \sigma^2)$

4.5.1 Definitie en kenmerken

Normale Verdeling

- Veelgebruikte verdeling met specifieke dichtheidsfunctie (klokcurve)

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2} \quad \text{met } x \in \mathbb{R}, \mu \in \mathbb{R} \text{ en } \sigma > 0$$

$$X \sim N(\mu, \sigma^2)$$

- Buigpunten in $\mu - \sigma$ en $\mu + \sigma$
- Kansen, $E(X)$ en $\text{Var}(X)$

$$E(X) = \mu$$
$$\text{Var}(X) = \sigma^2$$

Standaardnormale Verdeling

- Normaalverdeling met $\mu = 0$ en $\sigma = 1$

$$Z \sim N(0,1)$$

$$\phi(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad \text{en} \quad \Phi(z) = F_Z(z) = P(Z \leq z)$$

4.5.2 Kansen en Kwantielen onder de standaardnormale verdeling

Kans bij standaardnormale verdeling

- Verzameld in de tabel
- Door symmetrie en omzettingen van kansen kan je tabel gebruiken om kans te berekenen

$$P(Z \in A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Kwantielen

- $Q(0.5) = 0$
 - o Want curve is symmetrisch
- $Q(0.25) = -Q(0.75)$
- Voor elke $0 < \alpha < 1$
$$P(Z > z_\alpha) = \alpha$$
$$P\left(-\frac{z_\alpha}{2} \leq Z \leq \frac{z_\alpha}{2}\right) = 1 - \alpha$$

4.5.3 Kansen en Kwantielen onder een willekeurige normale verdeling

Is de verdeling niet standaardnormaal?

- Je kan het probleem omzetten naar een standaardnormale verdeling

$$\frac{X - \mu}{\sigma} = Z \quad \text{met } X \sim N(\mu, \sigma^2)$$

$$X \sim N(\mu, \sigma^2) \text{ impliceert } Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

★ $Z \sim N(0,1) \text{ impliceert } X = \mu + \sigma Z \sim N(\mu, \sigma^2) \rightarrow \text{Kunnen afleiden}$

Eigenschappen Normale Verdeling

$P(\mu - \sigma \leq X \leq \mu + \sigma) = P(-1 \leq Z \leq 1) = 0.682$
$P(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) = P(-1.96 \leq Z \leq 1.96) = 0.95$
$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(-3 \leq Z \leq 3) = 0.998$

- Een willekeurig kwantiel van een willekeurige normale verdeling staat in lineair verband met het overeenkomstige kwantiel van de standaardnormale verdeling
- $\sigma < IQR < 2\sigma$
- $Q_X(p) = \mu + \phi^{-1}(p)$ voor alle $0 < p < 1$

4.5.4 Lineaire Transformaties en Combinaties

Indien $X \sim N(\mu, \sigma^2)$, dan geldt:

$$a + bX = (a + b\mu) + b\sigma \left(\frac{X - \mu}{\sigma} \right) = (a + b\mu) + b\sigma Z$$

$$X \sim N(\mu, \sigma^2) \Rightarrow a + bX \sim N(a + b\mu, b^2\sigma^2)$$

Als $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ en X en Y zijn onafhankelijk, dan geldt dat

$$a + bX + cY \sim N(a + b\mu_x + c\mu_y, b^2\sigma_x^2 + c^2\sigma_y^2)$$

4.6 Normale Kwantielplot

4.6.1 Definitie

Hoe kunnen we grafisch een normale verdeling herkennen?

- Histogram
 - Lastig, kan je moeilijk inschatten, klassenbreedte bepaald heel hard welke verdeling je ziet
- Kwantielplot
 - Vergelijken van steekproefkwantielen $\hat{Q}_n(p)$ met de verwachte kwantielen $Q_X(p)$
 - $\hat{Q}_n(p)$ op x-as, $Q_X(p)$ op y-as
 - Bij normale verdeling verwachten we dat de grafiek de curve volgt van $y = x$

Probleem: $X \sim N(\mu, \sigma^2)$ maar μ en σ zijn onbekend, dus we kunnen $Q_X(p)$ niet bepalen.

Oplossing: $Q_X(p) = \mu + \sigma Q_Z(p)$, zie (14). Als we een plot maken met $Q_Z(p)$ op de x-as en $\hat{Q}_n(p)$ op de y-as, dan verwachten we dat de punten ongeveer op de rechte liggen met intercept μ en richtingscoëfficiënt σ .

De **normale kwantielplot** wordt gegeven door:

$$\left(\Phi^{-1} \left(\frac{i - 0.5}{n} \right), x_{(i)} \right) \quad \text{voor } i = 1, \dots, n$$

4.6.2 Afwijkingen van normaliteit

Kwantielplot ook goed om niet-normale verdelingen te herkennen.

Zie cursus p191-192 voorbeelden

4.6.3 Transformaties tot normaliteit

Geen leerstof

4.7 Chi-Kwadraatverdeling $X \sim \chi^2$

Chi-Kwadraatverdeling

- Vorm van rechtsscheve verdeling
- Enkel positieve waardes

$$X = \sum_{i=1}^m Z_i^2 \sim \chi_m^2$$

Dichtheid

$$f_m(x) = \frac{2^{-\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} \quad \text{voor } x \geq 0$$

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$$

Kansen, $E(X)$ en $\text{Var}(X)$

$$E(X) = m$$

$$\text{Var}(X) = 2m$$

4.8 F-Verdeling

F-Verdeling

- Als $X \sim \chi_n^2$ en $Y \sim \chi_m^2$

$$F = \frac{X/m}{Y/m} \sim F_{n,m}$$

Aanvullen

4.9 Exponentiële Verdeling

Aanvullen

4.10 Weibull Verdeling

Niet behandeld in de les

4.11 Lognormale Verdeling

Aanvullen

4.12 t-Verdeling

Aanvullen

4.13 Samenvatting

Zie cursus

Verdelingen, dichtheid en verwachtingswaarde

<u>Verdeling</u>	<u>Notatie</u>	<u>$f(m)$</u>	<u>$E(X)$</u>	<u>$Var(X)$</u>
Bernoulli	$X \sim B(1, p)$	$f(1) = p \mid f(0) = 1 - p$	p	$p(1 - p)$
Binomiaal	$X \sim B(n, p)$	$\binom{n}{m} p^m (1 - p)^{n-m}$	np	$np(1 - p)$
Poisson	$Poisson(\lambda)$	$\frac{e^{-\lambda} \lambda^m}{m!}$	λ	λ
Normaal	$X \sim N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2
Chi-Kwadraat				
F				
Exponentieel				
Lognormaal				
t				

5. Schatters en hun Verdeling

5.1 Steekproefgemiddelde als Toevalsvariabele

Omdat geobserveerde waarden in de steekproef onderhevig zijn aan het toeval, is ook de waarde van het steekproefgemiddelde onderhevig aan het toeval

Verdeling van steekproefgemiddelde

- Om kansen over het gemiddelde te kunnen berekenen
- Gemiddelde is benadering van $E(X)$, onderzoeken hoe benadering verschilt
- $\bar{x}$ is schatting van $E(X)$, $\bar{X}$ is schatter van $E(X)$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

5.2 Verdeling van het steekproefgemiddelde

5.2.1 Verwachtingswaarde en Variantie

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) = \frac{1}{n}(E(X_1) + \dots + E(X_n)) = \frac{1}{n}nE(X_1) = E(X_1)$$

Gemiddelde van het gemiddeldes is het gemiddelde

$$Var(\bar{X}) = Var\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) = \frac{1}{n^2}(Var(X_1) + \dots + Var(X_n)) = \frac{n}{n^2}Var(X_1) = \frac{1}{n}Var(X_1)$$

Hoe groter steekproef, hoe dichter het gemiddelde bij verwachtingswaarde komt

$$\sqrt{Var(\bar{X})} \rightarrow \text{Standaardfout van } \bar{X}$$

5.2.2 Verdeling van het gemiddelde van een normale variabele

Als $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ onafhankelijk zijn, dan geldt:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{zodat } E(\bar{X}) = \mu \text{ en } Var(\bar{X}) = \frac{1}{n}Var(X_1)$$

5.3 Centrale Limietstelling

5.3.1 Verdeling van het gemiddelde van een niet-normale variabele

Centrale Limietstelling:

Als $X_1, \dots, X_n$ onafhankelijk, identiek verdeelde toevalsvariabele zijn met $X_i \sim F$ en $Var(X_i) < \infty$, dan geldt voor n voldoende groot:

$$\bar{X} \approx N\left(E(X_1), \frac{Var(X_1)}{n}\right) \quad \text{en} \quad \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\bar{X} - E(X_1)}{\sqrt{\frac{Var(X_1)}{n}}}$$

Wanneer n voldoende groot?

- Hangt af van (scheefheid van) F
- Vaak al goed als $n \geq 30$

5.3.2 Toepassing van de centrale limietstelling

Zie voorbeeld in cursus

5.4 Normale benadering voor binomiaalkansen

Binomiaalverdeling

- $X \rightarrow$ aantal successen in n Bernoulli experimenten $(X_1, \dots, X_n)$

Relatieve Frequentie / Proportie successen

$$\hat{P} = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

met $\hat{P}$ steekproefproportie als toevalsvariabele en $\hat{p}$ als observatie proportie binnen specifieke steekproef

Eigenschappen verdeling

$$E(\hat{P}) = p \quad \text{Var}(\hat{P}) = \frac{p(1-p)}{n}$$

$$\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right) \quad \text{als } np \geq 5 \quad \text{en} \quad n(1-p) \geq 5 \quad (\text{centrale limietstelling toegepast op } \hat{P})$$

Als $X \sim B(n, p)$, dan geldt:

$$B(n, p) \approx N(np, np(1-p)) \quad \text{als } np \geq 5 \quad \text{en} \quad n(1-p) \geq 5$$

Continuïteitscorrectie

Continuïteitscorrectie

- Bij overgang discrete naar continue verdeling
- Nodig voor nauwkeurige benadering

$$P(X \geq a) \approx 1 - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(X \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

5.5 Eigenschappen van Schatters

5.5.1 Steekproefverdeling van puntschatters

Normale Verdeling

- μ meestal onbekend en benaderd door $E(X)$
 - o $E(X)$ is limiet van steekproefgemiddelde als steekproefgrootte n oneindig groot wordt
 - o Steekproefgemiddelde is een puntschatting voor μ
 - o Analoog is de standaardafwijking een puntschatting voor σ
- μ kan bij symmetrie benaderd worden door mediaan
 - o Steekproefmediaan benaderd populatiemedian
 - o Populatiemedian is gelijk aan μ

Welke schatting is de beste in algemeen?

- Voorbeelden schatters: $\bar{X}_n, Med_n, s_n^2, \hat{p}_n, \dots$
- Schatten een onbekende parameter θ
- Hoe kleiner variantie, hoe beter schatting
 - o Daarom verkiezen we gemiddelde boven mediaan (die hogere verdeling heeft)

Zuivere schatter

- Als de schatter gemiddeld gezien gelijk is aan de parameter die je zoekt
- Is belangrijk voor definitie steekproefvariantie

Schatter $\hat{\theta}_n$ voor θ is zuiver of onvertekend als:

$$E(\hat{\theta}_n) = \theta \quad \text{of} \quad E(\hat{\theta}_n - \theta) = 0$$

- Vertekening/bias van een schatter: $E(\hat{\theta}_n) - \theta$

Variantie Schatter

- Willen we zo klein mogelijk houden

$$Var(\hat{\theta}_n) = E\left(\left(\hat{\theta}_n - E(\hat{\theta}_n)\right)^2\right)$$

$$\text{Standaardfout: } \sqrt{Var(\hat{\theta}_n)}$$

Mean Squared Error

- Als schatter zuiver is wordt de variantie gelijk aan MSE

$$MSE(\hat{\theta}_n) = E\left((\hat{\theta}_n - \theta)^2\right) \quad (\text{want } E(\hat{\theta}_n) = \theta \text{ bij zuivere schatter})$$

$$MSE(\hat{\theta}_n) = Var(\hat{\theta}_n) + \left(Bias(\hat{\theta}_n)\right)^2 \quad (\text{algemeen voor elke schatter})$$

5.5.2 Voorbeelden van puntschatters

Schatten gemiddelde van normale verdeling

- $\bar{X}$ is uniform minimumvariantie onvertekende schatter voor μ

Schatten variantie van normale verdeling

- S^2 is onvertkende schatter voor σ^2 met $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- $(n-1) \frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2$

Schatten proportie/kansen

- $\hat{P}_n$ onvertkende schatter is voor p

5.6 Samenvatting

Zie cursus

6. Univariate Inferentie

Puntschatters voor θ

- Op basis van specifieke steekproef één steekproefafhankelijke schatten
- Benaderingsfout van de schatter bovenop de variantie van de schatter zelf?
- Fout/afwijking is zelf afhankelijk van toeval

6.1 Betrouwbaarheidsinterval voor het gemiddelde van een normale variabele

6.1.1 Variantie bekend

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95 \quad \rightarrow \quad P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right] \quad \text{Betrouwbaarheidsinterval 95\%}$$

Betrouwbaarheidsinterval Interpretatie

Indien we zeer vele steekproeven van omvang n uit de populatie zouden trekken en bij elke steekproef het 95%-betrouwbaarheidsinterval voor μ zouden berekenen, dan zouden gemiddeld 95% van de resulterende betrouwbaarheidsintervallen het onbekende gemiddelde μ bevatten

Betrouwbaarheidsniveau

- Kans dat een willekeurig gekozen betrouwbaarheidsinterval μ bevat
- $1 - \alpha$

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] \quad \text{Betrouwbaarheidsinterval } (1 - \alpha)\%$$

Foutmarge

- Helft van breedte betrouwbaarheidsinterval
- $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Eigenschappen Betrouwbaarheidsinterval

- Symmetrisch rond $\bar{x}$
- Als σ groter, dan BI breder
- Als α kleiner, dan BI breder
- Als n groter, dan BI smaller
- Gebaseerd op normaliteit van $\bar{X}$, dus wegens centrale limietstelling ook bruikbaar als X niet normaal verdeeld is en n voldoende groot is

Hoe groot moet steekproef zijn voor maximale foutmarge e ?

$$P(-e \leq \bar{X} - \mu \leq e) = 1 - \alpha$$

$$n = \left(z_{\alpha/2} \frac{\sigma}{e} \right)^2$$

6.1.2 De Variantie is niet bekend

Wat als we σ^2 niet kennen?

- σ schatten met steekproefvariantie schatter S^2

$\frac{\bar{X} - \mu}{S/\sqrt{n}}$ is niet standaardnormaal zoals $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow$ Kan je dus niet gebruiken!

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad (t \text{ verdeling met vrijheidsgraad } n-1)$$

$$P(T > t_{n-1, \alpha/2}) = \frac{\alpha}{2} \quad \text{met } T \sim t_{n-1}$$

$$\left[\bar{x} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \quad \bar{x} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right] \quad \text{Betrouwbaarheidsinterval } 100(1 - \alpha)\%$$

6.2 Testen omtrent het gemiddelde van een normale variabele

6.2.1 Rechtseenzijdige Test

Rechtseenzijdige Test

- Nulhypothese H_0 $H_0: \mu \leq \mu_0$
- Alternatieve hypothese H_1 $H_1: \mu > \mu_0$ met gegeven μ_0
- Alternatieve hypothese ontkennen en kijken hoe waarschijnlijk het is dat bij H_0 steekproefgemiddelde $\bar{x}$ voorkomt.
- Indien kans groot is, is het waarschijnlijke steekproef, dus alternatieve hypothese verwerpen
- Indien kans klein is, zou steekproef zeer ongewoon zijn, of is de alternatieve hypothese waar

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim_{H_0} t_{n-1} \quad \text{en voor gegeven steekproef testwaarde} \quad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Hoe groter t , hoe meer $\bar{x}$ afwijkt van μ_0 , dus hoe meer evidentie dat de alternatieve hypothese aannemelijk is
- Nulhypothese verwerpen en overgaan naar H_1 als $t \gg 0$

P-Waarde

- Kans onder H_0 op een testwaarde die minstens zo extreem ligt in de richting van het alternatief als de waarde $\bar{x}$ die we verkrijgen voor de geobserveerde steekproef
- Kleine P-Waarde
 - o Alternatieve hypothese erg aannemelijk
- Grote P-Waarde
 - o Gevonden testwaarde niet uitzonderlijk onder nulhypothese

- Geen rede om H_0 te verwerpen
- $\bar{x} - \mu_0 \leq t_{n-1, \alpha} \frac{s}{\sqrt{n}}$

Significantieniveau α

- Waarde die beslist of P-Waarde klein genoeg is om meer geloof te hechten aan de alternatieve hypothese
- $P \text{ waarde} < \alpha \rightarrow \text{Verwerpen van } H_0 \text{ op } \alpha \text{ significantieniveau}$ *aanvaarden van H_1*
- $P \text{ waarde} \geq \alpha \rightarrow \text{Verwerpen } H_0 \text{ niet}$ *GEEN BEWIJS VOOR H_0 !*

6.2.2 Aanvaardings- en verwerpsgebied

$$\begin{aligned} P \text{ Waarde} \geq \alpha \quad \text{of} \quad t \leq t_{n-1, \alpha} &\Rightarrow \text{Verwerpen } H_0 \text{ niet op } \alpha \text{ significantieniveau} \\ P \text{ Waarde} < \alpha \quad \text{of} \quad t > t_{n-1, \alpha} &\Rightarrow \text{Verwerpen } H_0 \text{ wel op } \alpha \text{ significantieniveau} \end{aligned}$$

Gebieden

- Verwerps / Kritisch gebied
 - $] t_{n-1, \alpha}, +\infty[$
- Aanvaardingsgebied
 - $] -\infty, t_{n-1, \alpha}]$

6.2.3 Linkseenzijdige Test

Linkseenzijdige Test

- $H_0: \mu \geq \mu_0$
- $H_1: \mu < \mu_0$

Analoog aan rechtseenzijdige test, enkel omkeren van ongelijkheden

P Waarde: $P(T < t)$ met $T \sim t_{n-1}$

6.2.4 Tweezijdige Test

Tweezijdige Test

- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$

P Waarde: $2P(T > |t|)$ met $T \sim t_{n-1}$

6.2.5 Type 1 en Type 2 Fout

Fouten bij Hypothesetesten

- Type 1
 - Nulhypothese verwerpen terwijl ze in werkelijkheid waar is
 - $\alpha = P(\text{type 1 fout}) = P(H_0 \text{ verwerpen} \mid H_0 \text{ waar})$ *(significantieniveau)*
- Type 2
 - Nulhypothese niet verwerpen terwijl ze in werkelijkheid niet waar is
 - $\beta = P(\text{type 2 fout}) = P(H_0 \text{ niet verwerpen} \mid H_0 \text{ niet waar})$

Berekening van type 2 fout is geen leerstof!

Onderscheidingsvermogen Test

- Hoe goed kan de test een alternatief ontdekken?

	verwerp H_0	verwerp H_0 niet
H_0 waar	α = significantieniveau = $P(\text{type I fout})$	$1 - \alpha$ = betrouwbaarheidsniveau
H_0 niet waar	$1 - \beta$ = onderscheidingsvermogen	β = $P(\text{type II fout})$

6.2.6 De Variantie is bekend

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim_{H_0} N(0,1) \quad \text{met } H_0: \mu = \mu_0$$

$H_0: \mu = \mu_0$	Teststatistiek	P-waarde	Aanvaardingsgebied
	$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim_{H_0} N(0,1)$ als σ^2 bekend		
$H_1: \mu \neq \mu_0$		$2P(Z > z)$	$[-z_{\alpha/2}, z_{\alpha/2}]$
$H_1: \mu > \mu_0$		$P(Z > z)$	$] -\infty, z_{\alpha}]$
$H_1: \mu < \mu_0$		$P(Z < z)$	$[-z_{\alpha}, +\infty[$

6.2.7 Algemene procedure en begrippen bij hypothesetesten

Stappenplan hypothesetest

1. Onderzoek voorwaarden voor uitvoering hypothesetest
2. Formuleer H_0 en H_1
3. Kies teststatistiek U , de verdeling van U onder H_0 moet bekend zijn. Bepaal u
4. Voer de test uit
 - o Bereken P-Waarde
 - o Bereken aanvaardingsgebied (als α gegeven is)
 - o Bereken betrouwbaarheidsinterval voor onbekende parameter
5. Formuleer het besluit
 - o P-Waarde $< \alpha$
 - o Testwaarde u niet in aanvaardingsgebied
 - o Bij tweezijdig testen de voorgestelde waarde voor parameter niet tot betrouwbaarheidsinterval

6.3 Testen van Normaliteit

H_0 : toevalsvariabele X is normaal verdeeld

H_1 : toevalsvariabele X is niet normaal verdeeld

Testen of X normaal verdeeld is

- Kwantiplot rechtlijnig?
- Pearson correlatiecoëfficiënt r_Q dicht bij 1?
- Shapiro-Wilk test en W ligt dicht bij 1?

Wat als normaliteit verworpen wordt?

- Verwijderen van uitschieters
 - o Kan normale verdeling bekomen
- Transformatie toepassen
 - o $\log(X)$, $E(X)$, $\text{Med}(X)$, ...
 - o Kan normale verdeling bekomen
- Niet-parametrische testmethode
 - o Zie 6.5

Telkens numerieke test combineren met grafische test via kwantielplot

6.4 Betrouwbaarheidsinterval voor de mediaan van continue variabele

Mediaan $\text{Med}(X)$ bestaat altijd, gemiddelde $E(X)$ niet

95% Betrouwbaarheidsinterval over mediaan

$$[x_{(k)}, x_{(n-k+1)}] \quad \text{met } k \text{ afhankelijk van steekproefgrootte en te vinden in tabel}$$

6.5 Testen omtrent de mediaan van een continue variabele

$H_0 : \text{Med}(X) = m_0$	Teststatistiek	P-waarde
	$A = (\text{aantal } x_i > m_0)$	
	$\sim_{H_0} B(n', 0.5)$	
$H_1 : \text{Med}(X) \neq m_0$		$2P(A \leq a) \text{ als } a \leq n'/2$ $2P(A \geq a) \text{ als } a > n'/2$
$H_1 : \text{Med}(X) > m_0$		$P(A \geq a)$
$H_1 : \text{Med}(X) < m_0$		$P(A \leq a)$

Geen leerstof

6.6 Inferentie omtrent een proportie

6.6.1 Betrouwbaarheidsinterval voor een proportie

Betrouwbaarheidsinterval opstellen voor kans op succes p

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right] \quad \text{bij steekproef voldoende groot en schatter } \hat{P} \text{ voor } p$$

$$\text{Voorwaarden: } n\hat{p} \geq 5 \quad \text{en} \quad n(1-\hat{p}) \geq 5$$

6.6.2 Testen omtrent een proportie

Hypothesetesten als $X_i \sim B(1, p)$:

Voldoende grote steekproef

$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

$$H_0: p \leq p_0$$

$$H_1: p > p_0$$

$$H_0: p \geq p_0$$

$$H_1: p < p_0$$

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx_{H_0} N(0,1) \quad \text{als } np_0 \geq 5 \text{ en } n(1-p_0) \geq 5$$

Kleine steekproef

$$X = n\hat{p} \sim_{H_0} B(n, p_0) \quad \text{als } np_0 < 5 \text{ of } n(1-p_0) < 5$$

6.7 Testen van de verdeling van een discrete variabele

Discrete Variabelen

- Kwantielplots niet handig
 - o Vaak belangrijk aantal observaties gelijk zijn aan dezelfde mogelijke uitkomst waardoor kwantielplot horizontaal zou liggen
- Nakijken of variabele bepaalde verdeling volgt
 - o Staafdiagram van gegevens overeenstemt met de dichtheid van de variabele
 - o Onbekende parameters in formule van dichtheid vervangen door schatting

H_0 : toevalsvariabele X heeft dichtheid f_θ

H_1 : toevalsvariabele X heeft dichtheid f_θ niet

Principe teststatistiek

- Voor elke uitkomst m_j de absolute frequentie n_j vergelijken met het verwachte aantal onder H_0

$$nP(X = m_j) = nf_\theta(m_j) \quad \text{en} \quad \theta \text{ schatten door } \hat{\theta}$$

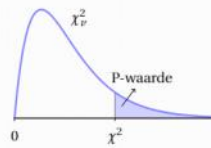
$$X^2 = \sum_{j=1}^k \frac{(n_j - nf_{\hat{\theta}}(m_j))^2}{nf_{\hat{\theta}}(m_j)} \approx_{H_0} \chi_v^2 \quad \text{met } v = k - n_0 - 1$$

(chi kwadraat test) als $n_j \geq 5$ voor elke j en $n_0 = \text{aantal geschatte parameters}$

$$P \text{ Waarde: } P(X^2 > \chi^2) \quad \text{met } X^2 \sim \chi_v^2$$

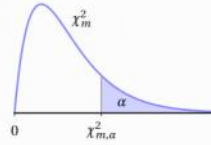
- Enkel grote waarde van X^2 wijzen in richting van alternatieve hypothese

$$\text{P-waarde} = P(X^2 > \chi^2) \quad \text{met } X^2 \sim \chi_v^2$$



Kwantielen:

$$P(X^2 > \chi_{m,\alpha}^2) = \alpha \quad \text{met } X^2 \sim \chi_m^2$$



6.8 Samenvatting

Zie cursus

7. Bivariate Kansmodellen

Wat is het verband tussen 2 variabelen op populatieniveau?

7.1 Gezamenlijke verdeling van twee discrete variabelen

Bivariate / Gezamenlijke dichtheid bij discrete variabelen

- Grafische voorstelling via staafjes/balkjes
- Dichtheidsfunctie

$$f(m_x, m_y) = P(X = m_x, Y = m_y)$$

$$f(m_x, m_y) \geq 0 \text{ voor elke } (m_x, m_y) \in \varphi \quad \text{en} \quad \sum_{m_x} \sum_{m_y} f(m_x, m_y) = 1$$

$$P((X, Y) \in A) = \sum_{(m_x, m_y) \in A} f(m_x, m_y)$$

Marginale Dichtheden

$$f_X(m_x) = P(X = m_x) = \sum_{m_y} P(X = m_x, Y = m_y) = \sum_{m_y} f(m_x, m_y)$$

$$E(X) = \sum_{m_x} m_x f_X(m_x)$$

Voorwaardelijke Dichtheden

$$f_{X|Y}(m_x, m_y) = P(X = m_x \mid Y = m_y) = \frac{P(X = m_x, Y = m_y)}{P(Y = m_y)} = \frac{f(m_x, m_y)}{f_Y(m_y)}$$

$$E(X \mid m_y) = \sum_{m_x} m_x f_{X|Y}(m_x, m_y)$$

7.2 Gezamenlijke Verdeling van twee continue variabelen

Bivariate / Gezamenlijke dichtheid bij continue variabelen

- Grafische voorstelling via histogram
- Dichtheidsfunctie

$$f(x, y) = P(X = x, Y = y)$$

$$f(x, y) \geq 0 \text{ voor elke } (x, y) \in \varphi \quad \text{en} \quad \int \int_A f(x, y) dx dy = 1$$

$$P((X, Y) \in A) = \int \int_A f(x, y) dx dy$$

Marginale Dichtheden

$$f_X(x) = \int_{\varphi_Y} f(x, y) dy$$

Voorwaardelijke Dichtheden

$$f_{X|Y}(x | y) = f(x | y) = \frac{f(x, y)}{f_Y(y)}$$

$$E(X | y) = \int_{\varphi} f(x | y) dx$$

Zie formularium voor overzicht

7.3 Onafhankelijkheid van twee variabelen

Zie formules in formularium en verbanden met hoofdstuk 3

Als X en Y onafhankelijke toevalsvariabelen zijn, dan geldt:

$$\text{Cov}(X, Y) = 0 \quad \text{en} \quad \rho(X, Y) = 0 \quad \text{met Cov de populatiecovariantie}$$

Onafhankelijk $\Rightarrow$ Ongecorreleerd maar Ongecorreleerd $\nRightarrow$ Onafhankelijk

7.4 Bivariate Normale Verdeling

Bivariate Normale Verdeling

- Belangrijke verdeling bij 2 variabelen
- Als zowel X als Y beide normaal verdeeld zijn en onafhankelijk zijn
- Door samenvoeging van 2 normale dichtheden
- $(X, Y) \sim N_2(\mu, \Sigma)$

Stel $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$, en X en Y zijn onafhankelijk. Dan volgt dat voor elke $x, y \in \mathbb{R}$

$$\begin{aligned} f(x, y) &= f_X(x) f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2} \\ &= \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)} \end{aligned}$$

Stellen we

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{en} \quad \Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

dan kunnen we dit herschrijven als

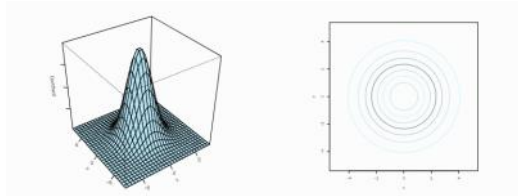
$$f(x, y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu_x, y - \mu_y)\Sigma^{-1}(x - \mu_x, y - \mu_y)^t\right)$$

Standaard Bivariate Normale Verdeling

- Als $\mu = (0, 0)$

Dichtheidscontours

- Bij gelijke varianties: Cirkels



- Bij ongelijke varianties: Ellips

Covariantiematrix Σ

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

Eigenschappen bij Bivariate Normale Verdeling

X en Y onafhankelijk $\Leftrightarrow X$ en Y zijn ongecorreleerd

- Als X normaal verdeeld is en Y normaal verdeeld is, is (X,Y) niet altijd normaal verdeeld!
- Als X of Y niet normaal verdeeld is, dan is (X,Y) nooit biviaat normaal verdeeld!

8. Bivariate Inferentie

8.1 Twee Kwalitatieve Variabelen

8.1.1 Inferentie omtrent twee proporties

Wat als we twee proporties met elkaar willen vergelijken?

- Is arbeid in Noordoosten hoger dan in Zuiden?
- Via steekproef krijg je een idee, maar je wil vergelijken voor populaties

Twee binaire variabele X en Y , waarbij X de variabele is die de metingen van Y in 2 groepen verdeelt:

$$Y_{1,1}, Y_{1,2}, \dots, Y_{1,n_1} \sim B(1, p_1) \text{ als } X = 1 \quad Y_{2,1}, Y_{2,2}, \dots, Y_{2,n_2} \sim B(1, p_2) \text{ als } X = 2$$

als alle variabelen onafhankelijk. Het mag zijn dat $n_1 \neq n_2$

Vergelijken van proporties adhv steekproefproporties

$$\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1,i} \quad \text{en} \quad \hat{p}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{2,i}$$

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(n-p_1)}{n_1} + \frac{p_2(n-p_2)}{n_2}}} \approx N(0,1) \quad \text{wegens centrale limietstelling}$$

Betrouwbaarheidsinterval voor $p_1 - p_2$

$$\left[\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(n - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(n - \hat{p}_2)}{n_2}} \right]$$

$$H_0 : p_1 = p_2$$

$$H_0 : p_1 \leq p_2$$

$$H_0 : p_1 \geq p_2$$

$$H_1 : p_1 \neq p_2$$

$$H_1 : p_1 > p_2$$

$$H_1 : p_1 < p_2$$

Teststatistieken

- Zie formularium
- Telkens voorwaarden controleren!

8.1.2 Onafhankelijkheid van twee variabelen testen

H_0 : toevalsvariabelen X en Y zijn onafhankelijk

H_1 : toevalsvariabelen X en Y zijn afhankelijk

Kruistabel

- Verwachte waarde = $\frac{\text{rijtotaal} * \text{kolomtotaal}}{n}$

Teststatistiek

- Zie formularium
- Chi-Kwadraat

Opmerkingen bij Chi² Teststatistiek

- Indien een verschil wordt gevonden wordt niet aangegeven op welke manier variabelen afhankelijk zijn
- De test meet de sterkte van afhankelijkheid NIET!

8.2 Een Kwantitatieve en een Kwalitatieve Variabele

Kwantitatieve variabele met k uitkomsten verdeelt kwalitatieve variabele in k groepen

8.2.1 Het Gemiddelde van een variabele vergelijken voor twee groepen

Ongepaard = Onafhankelijk

$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \leq \mu_2$	$H_0 : \mu_1 \geq \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 > \mu_2$	$H_1 : \mu_1 < \mu_2$

Teststatistieken & BI's

- Zie formularium

8.2.2 De Variantie van een variabele vergelijken voor twee groepen

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \left(\text{of } \frac{\sigma_1^2}{\sigma_2^2} = 1 \right) \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \quad \left(\text{of } \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \right)$$

Testen via F-Test

- Vuistregel: verwerpen van H_0 als $f < \frac{1}{4}$ of $f > 4$
- Niet met F-tabel kunnen werken, vuistregel gebruiken of via R-code de P-Waarde krijgen

P-Waarde test kleiner dan α ?

- Verwerpen nulhypothese van gelijke varianties en voeren we bij t-test voor vergelijken van gemiddeldes de versie voor ongelijke varianties uit
- Indien niet verwerpen, ervan uitgaan dat varianties gelijk zijn

8.2.3 De Verdeling van een variabele vergelijken voor twee groepen

Eerdere testen van gemiddelde en variantie vereisen normaliteit van beide variabelen

Wat indien geen normaliteit?

- Via transformaties naar normaliteit proberen te gaan
- Niet-parametrische testen uitvoeren

Geen leerstof

8.3 Twee Kwantitatieve Variabelen

8.3.1 Het verschil van twee variabelen vergelijken

Bestuderen van verschillende variabelen in dezelfde groepen/populatie --> Dus AFHANKELIJK/GEPAARD

Gepaarde variabelen

- Voor elke X meting hebben we een Y meting
- $(X_i, Y_i) \sim F$

We doen uitspraak over verschilvariabele $V = X - Y$

Vershilvariabele normaal verdeeld

$$H_0 : \mu_v = 0 \quad H_1 : \mu_v \neq 0$$

We veronderstellen niet de normaliteit van X en Y! Enkel van V

Teststatistiek en BI

- Zie formularium

Vershilvariabele niet normaal verdeeld

$$H_0 : \text{Med}(V) \leq 0 \quad H_1 : \text{Med}(V) > 0$$

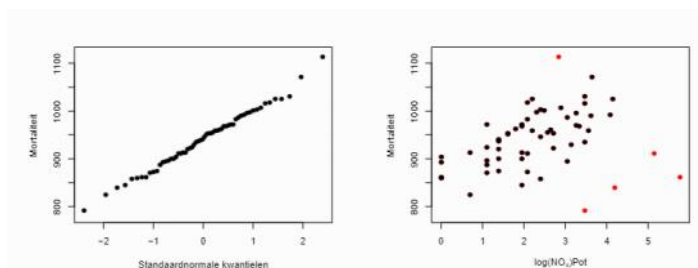
Teststatistiek en BI

- Zie formularium

8.3.2 Bivariate Normaliteit Nagaan

Als (X, Y) biviaat normaal verdeeld is, moet

- X normaal verdeeld zijn
- Y normaal verdeeld zijn
- De scatterplot (x_i, y_i) een elliptische puntenwolk vormen (bij evt verwijderen van uitschieters)



Overige uit deze sectie is geen leerstof

8.3.3 De correlatie tussen twee variabelen testen

Variabelen zijn biviaat normaal verdeeld

Bij bivariate normale verdeling en populatiecorrelatie ρ

$$H_0 : \rho = 0 \quad (X \text{ en } Y \text{ onafhankelijk}) \quad H_1 : \rho \neq 0 \quad (X \text{ en } Y \text{ lineair afhankelijk})$$

Teststatistiek en BI

- Zie formularium
- R = Steekproefcorrelatiecoëfficiënt

Variabelen zijn niet bivariaat normaal verdeeld

Bij niet bivariate normale verdeling en Spearman populatiecorrelatie ρ_s

$$H_0 : \rho_s = 0 \quad (\text{Geen monotoon verband tussen } X \text{ en } Y) \qquad H_1 : \rho_s \neq 0 \quad (\text{monotoon verband tussen } X \text{ en } Y)$$

Teststatistiek en BI

- Zie formularium
- R_s = Spearman Correlatiecoëfficiënt als schatter

8.4 Samenvatting

Zie cursus

9. Lineaire Regressie

Hoe lineair verband tussen variabelen weergeven?

- Pearson Correlatiecoëfficiënt
- Lineaire Regressiemethodes

Puntenwolk

- Voor elke observatie de meting van de ene variabele op de x-as, van de andere op de y-as

Pearson Correlatiecoëfficiënt

- Geeft aan of er een verband is, maar niet in welke mate
- 1 of -1 -> Perfecte rechte en lineair verband

9.1 Enkelvoudige Regressiemodel

$$Y_i = \alpha + \beta x_i + \epsilon_i \quad \text{voor alle } i = 1, 2, \dots, n \quad \text{met } \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

X = Verklarende/Onafhankelijke Variabele (x-as)

Y = Response/Afhankelijke Variabele (y-as)

$$E(Y_i | x_i) = E(\alpha + \beta x_i + \epsilon_i) = \alpha + \beta x_i + E(\epsilon_i) = \alpha + \beta x_i$$

$$\text{Var}(Y_i) = \text{Var}(\epsilon_i) = \sigma_\epsilon^2$$

Er wordt geen normaliteit van X, Y of X-Y verondersteld!

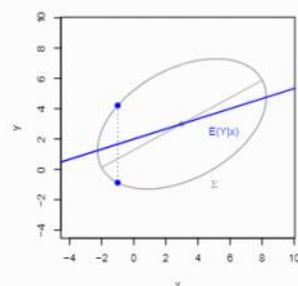
$\alpha \rightarrow$ Intercept, gemiddelde waarde voor Y bij $x = 0$

$\beta \rightarrow$ Richtingscoëfficiënt / slope, hoeveel wijzigt Y als x met 1 eenheid groter wordt?

Als (X, Y) bivariaat normaal verdeeld is, dan komt de regressierechte

NIET overeen met de hoofdas van de ellipsen met constante dichtheid. De rechte draait naar het gemiddelde van Y toe:

'regression towards the mean'.



De modelveronderstellingen zijn soms niet voldaan, maar kunnen wel voldaan zijn wanneer X en/of Y getransformeerd worden.

9.2 Kleinste Kwadratenmethode

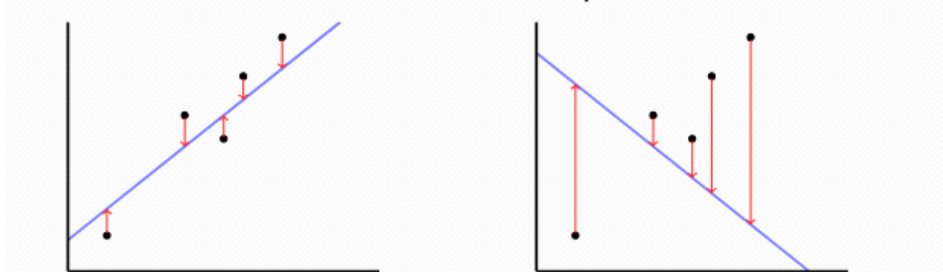
α en β schatten op basis van een steekproef

We maken een schatting $\hat{\alpha}$ en $\hat{\beta}$ voor α en β

- Rechte door puntenwolk: $y = \hat{\alpha} + \hat{\beta}x$
- Zoeken naar rechten met zo klein mogelijke spreiding errond

Residu

- Verticale afstand van observatie tot de rechte
- $e_i = e_i(\hat{\alpha}, \hat{\beta}) = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x)$



Kleinste Kwadratenmethode

- Zoeken wanneer de som van gekwadrateerde residu's minimaal is

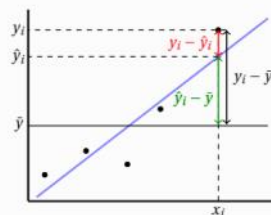
$$\sum_{i=1}^n y_i - (\hat{\alpha} + \hat{\beta}x) \text{ minimaal} \rightarrow \alpha = \bar{y} - \hat{\beta}\bar{x} \text{ en } \beta = r \frac{s_{yg}}{s_x}$$

$\hat{\beta}$ is veelvoud van Pearson Correlatiecoëfficiënt!

- Grote invloed van uitschieters
 - o Nuttig om deze weg te laten in berekeningen!

9.3 ANOVA-tabel en determinatiecoëfficiënt

Er geldt steeds: $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$



Men kan aantonen dat

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

S_{(um of)S_{(quares)T}(otal)} = SSM_(odel) + SSE_(rrors)

Aangezien $SST = (n-1)s_y^2$ mogen we aannemen dat $SST \neq 0$.

Determinatiecoëfficiënt R^2

$$R^2 = \frac{SSM}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

- Ligt steeds tussen 0 en 1
- Geeft aan welk percentage van totale variatie van y verklaard wordt door het regressiemodel
- Als alle residu's 0 zijn $\rightarrow R^2 = 1 \rightarrow$ Alle punten op één rechte

- Verband met Pearson-Coëfficiënt

$$R^2 = r_{xy}^2$$

Mean Squared Error MSE

- Gemiddelde Kwadratische Fout

$$MSE = \frac{SSE}{n-2} = s_{\epsilon}^2$$

Vrijheidsgraden / Degrees of Freedom / Df

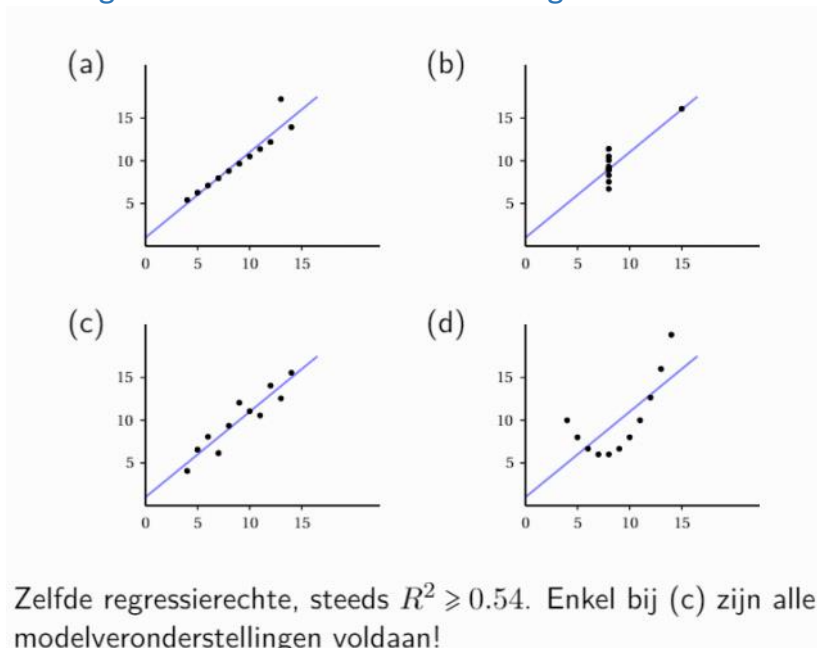
- SSM: 1
- SSR: $n-2$ (verlies van 2 vrijheidsgraden door schatten van regressieparameters α en β)

ANOVA Tabel / Variantieanalyse

ANOVA tabel in R (boek p 408)

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
(Model)	1	SSM	$MSM = \frac{SSM}{1}$	$f = \frac{MSM}{MSE}$	$P(F_{1,n-2} > f)$
Residuals	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$		

9.4 Nagaan van Modelveronderstellingen



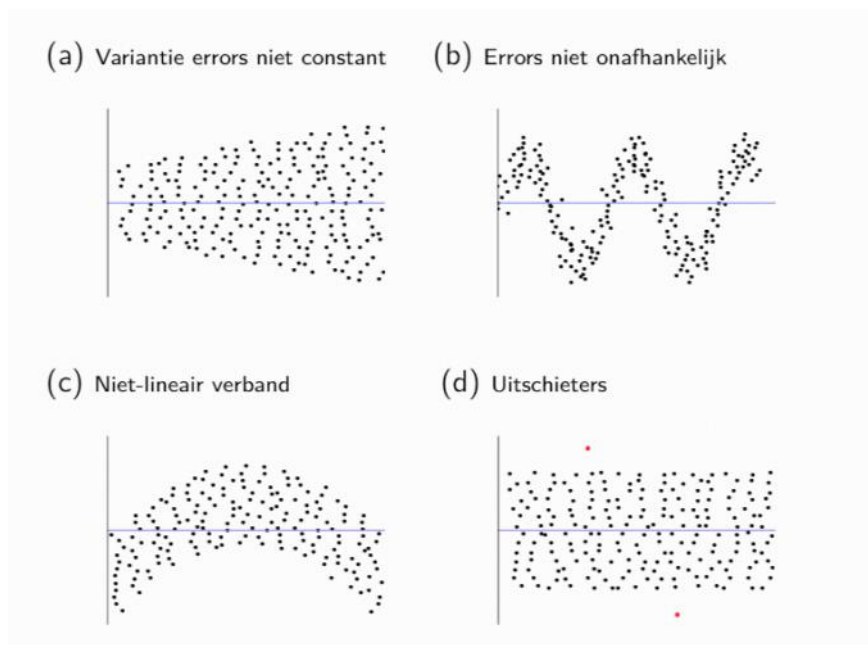
Modelverstelling: $\epsilon_i \sim N(0, \sigma_{\epsilon}^2)$

We kunnen niet controleren of modelveronderstelling voldaan is, want we kennen α en β niet

- We kunnen wel adhv e_i (residu) nagaan of veronderstelling klopt

Residuplot

- Scatterplot waarbij residu's e_i uitgezet worden tegenover de verklarende waarden x_i



9.5 Inferentie omtrent regressieparameters

Coefficients tabel in R (boek p417)

	Estimate	Std. Error	t value	$Pr(> t)$
intercept	$\hat{\alpha}$	$s.e.(\hat{A})$	$t = \frac{\hat{\alpha}}{s.e.(\hat{A})}$	$2P(T_{n-2} > t)$
(slope)	$\hat{\beta}$	$s.e.(\hat{B})$	$t = \frac{\hat{\beta}}{s.e.(\hat{B})}$	$2P(T_{n-2} > t)$

Residual standard error: s_e on $n - 2$ degrees of freedom

Multiple R-squared: R^2

F-statistic: $f = \frac{MSM}{MSE}$ on 1 and $n - 2$ DF, p-value $P(F > f) = 2P(T > |t|)$

Einde :)