HOOFDSTUK 1

TYPISCHE FOUTEN BIJ STATISTIEK

° Foute gegevens
° Fouten in berekening kans
° Foute interpretatie resultaten

Statistiek : de wetenschap van het leren uit data & van het meten, controleren en communiceren van onzekerheid

1. Eigenschappen van variabelen

1.1 Verschillende schaalfamilies

Nominaal	- Namen → Geen hoeveelheid, gewoon identificatie - Waarden kunnen ook getallen zijn	*Variabele = geslacht, waarde = man of vrouw* *Variabele = nummer tram, waarde = 24, 21..* *Variabele = land, waarde = België, Frankrijk..*
Ordinaal	- Geen hoeveelheid - Hiërarchie : ordening! → volgorde belangrijk, waarde zelf niet	*Variabele = uitslag wedstrijd, waarde = goud, zilver, brons* *Variabele = mate van instemming, waarde = volledig oneens, neutraal, volledig eens…*
Interval	- Hiërarchie - Waarde zelf ook belangrijk - Geen absoluut nulpunt - Geen onderlinge verhoudingen - Recht evenredig : onderlinge verschillen blijven even groot	*Variabele = temperatuur, waarde = 10°C, 5°C* *→ 10°C is niet het dubbele van 5°C (want in F is het niet zo op een grafiek)* *Recht evenredig: even groot verschil tss 10° en 20° & 50° en 60°* *0° is geen absoluut nulpunt: 32°F*
Ratio	- Absoluut nulpunt - Verhoudingen	*€10 is dubbel zoveel dan €5* *Variabele = lengte in cm, geldbedrag in euro..*

Dia 49 : vragen 1 = ratio , 2 = interval , 3 = nominaal , 4 = ratio , 5 = ordinaal & ratio , 6 = ordinaal

1.2 Discrete & continue variabelen

Continue variabelen	- Tussenwaarden → tussen elke 2 waarden ligt een 3e → oneindig veel waarden	*Lengte in cm* *Temperatuur in °C* *Tijd in seconden*
Discrete variabelen	Geen tussenwaarden → eindig aantal waarden	*Aantal kinderen, aantal volgers op Twitter, aantal GSMs die je al hebt*

Deel I : beschrijvende statistiek

HOOFDSTUK 2 : VISUALISEREN VAN DATA

Populatie	Volledige verzameling van objecten of personen waarover men info wil
Steekproef	Deelverzameling van de populatie, die ook echt onderzocht wordt → moet representatief zijn : aselect!

CIRKELDIAGRAM	- Nominaal - Relatieve oppervlaktes cirkel = relatieve frequenties

	- Nadeel : geen goed overzicht onderlinge verhoudingen
STAAFDIAGRAM	- Nominaal & ordinaal - Horiz : waarden variabele Vertic : AF of RF - Rechthoeken los van elkaar, breedte en afstand even groot **- Voordeel : snel overzicht onderlinge verhoudingen**
HISTOGRAM	- Interval & ratio - Voordeel : snel overzicht onderlinge verhoudingen - Horiz : waarden variabele - Rechthoeken tegen elkaar → Breedte = klassenbreedte → Oppervlakte = RF → Hoogte = RF / klassenbreedte - Klassenindeling → Beslist onderzoeker zelf → Klassenbreedte niet altijd even groot, dus rechthoeken niet altijd even breed → Als klassen toch even groot zijn, kan.. .. hoogte rechthoek = AF .. hoogte rechthoek = RF .. oppervlakte rechthoek = RF → Uiterste klassen heel kleine AF = samenvoegen → vuistregel aantal klassen : $\sqrt{n}$ - Verdeling → Symmetrische verdeling → Scheef naar rechts : staart naar rechts → Scheef naar links : staart naar links

1. Algemene begrippen & notatie

ABSOLUTE FREQUENTIES (AF)	Het aantal (per waarde)
ABSOLUTE FREQUENTIEVERDELING	Tabel met absolute frequenties
VARIABELE –NOTATIE-	Met een hoofdletter (vaak X) → Waarden die variabele aanneemt : kleine letters, cijfers als subscriptie
STEEKPROEFGROOTTE (N)	
RELATIEVE FREQUENTIES (RF)	
VERDELING VE VARIABELE	Het geheel van mogelijke waarden, samen met de absolute en/of relatieve frequenties
KLASSEN –NOTATIE-	$]a, b]$
GEGROEPEERDE FREQUENTIEVERDELING	tabel, 2 kolommen : klassen & overeenkomstige frequenties

2. Cumulatieve frequentiecurve

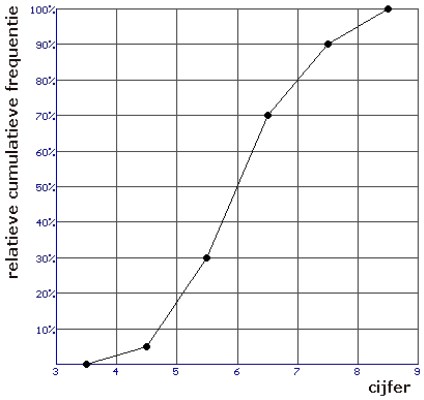
2.1 Ongegroepeerde data

# INFO	Meer dan bij gegroepeerde data (indelen in klassen leidt tot informatieverlies)
CUMULATIEVE ABSOLUTE FREQUENTIE	absolute frequenties optellen - geeft dan het aantal gegevens weer die gelijk aan of kleiner dan de bijhorende waarde zijn - grootste waarde = steekproefgrootte (want alles is kleiner of gelijk aan)

CUMULATIEVE RELATIEVE FREQUENTIE	zelfde systeem als bij rel.abs.fr.
CUMULATIEVE FREQUENTIECURVE	Horizontaal : gegevens Verticaal : cumulatieve frequenties Tekenen - Alle waarden aanduiden - Onderling trapsgewijs verbinden - Laagste waarde & hoogste waarde : horizontale lijn!

2.2 Gegroepeerde data

# INFO	Minder dan bij ongegroepeerde data → Je hebt info over een klasse, niet over een specifieke waarde! → NADEEL
CUMULATIEVE FREQUENTIE	Frequenties optellen
CUMULATIEVE FREQUENTIECURVE	Telkens het klassenmidden gebruiken als punt voor op de grafiek



2.3 Illustratie methoden

Zie cursus pagina 59 – 63

HOOFDSTUK 3 : SAMENVATTEN VAN DATA

(Centrummaten & spreidingsmaten, tot aan variantie en standaarddeviatie: zie geschreven samenvatting)

2.4 De interkwartielafstand

2.4.1 Percentielen p_k

Voor een geheel getal k tussen 0 en 100, is het k-de percentiel (symbool p_k) het getal p_k waarvoor geldt dat:

$$\frac{F(P_k)}{n} = \frac{k}{100}$$

WAT BETEKENT DIT?
° Het k-de percentiel is de waarde waarvan k% van de hele verzameling kleiner of gelijk aan die waarde is
° Komt zowat overeen met de cumulatieve relatieve frequentie

VOORBEELD?
P10: tiende percentiel: de waarden van een variabele, waarvoor 10% van de waarden hetzelfde of kleiner zijn

SPECIAAL PERCENTIEL
mediaan = percentiel P50

2.4.2 Kwartielen

1E KWARTIEL	P25	→ 25% van alle waarden
2E KWARTIEL	P50	→ Mediaan : 50% van alle waarden
3E KWARTIEL	P75	→ 75% van alle waarden
INTERKWARTIELAFSTAND	Q	° 3e kwartiel – 1e kwartiel $(P_{75} - P_{25})$ ° Interval & ratio
INTERKWARTIELINTERVAL	$[P_{25}, P_{75}]$	° Overspant 50% van alle waarden ° Ordinaal, interval & ratio

2.5 De spreidingsmaat d

VARIABELEN	Allemaal Vooral nominaal
UITKOMST	Uitkomst ligt tussen 0 en 1 , geeft op die manier spreiding weer ° 0 = geen spreiding ° 1 = maximale spreiding
FORMULE	$d = \frac{1-\frac{f_{mo}}{n}}{1-\frac{1}{p}}$ KOMT IN FORMULARIUM p aantal unieke waarden die een variabele kan aannemen f_{mo} frequentie van de modus (kan een waarde of een klasse zijn) n steekproefgrootte WAT ALS… .. f_{mo} = n Geen spreiding, want alle waarden zijn gelijk aan de modus .. f_{mo} = n/p d = 1

2.6 Gevoeligheid aan outliers

HOE BEREKEN JE OF EEN SPREIDINGSMAAT GEVOELIG IS OF NIET?	1. Bereken spreidingsmaat met alle waarden (inclusief outliers) 2. Bereken spreidingsmaat zonder outliers 3. Als er een groot verschil is tussen deze waarden : de spreidingsmaat is gevoelig aan outliers!	
GEVOELIGHEID?	Variatiebreedte	Ja
	Gemiddelde absolute afwijking	Ja
	Variantie	Ja
	Standaarddeviatie	Ja
	Interkwartielafstand	Neen
	d	Neen

3. Boxplot

DATA GROEPEREN?	Hoeft niet - Niet gebruikersafhankelijk - Verschil met histogram!		
VOORDEEL	- Handig om overzicht te krijgen over verdeling van data (mediaan, interwartielafstand..) - Makkelijk weten wat de outliers zijn		
OUTLIERS VASTSTELLEN – REKENREGEL	Laagste outliers	$P_{25} - 1.5 * Q$	Alles lager dan deze waarde = outlier
	Hoogste outliers	$P_{75} + 1.5 * Q$	Alles hoger dan deze waarde= outlier
HOE TEKENEN	1. Alle waarden op grafiek tekenen 2. Outliers bepalen (via rekenregel) & aanduiden op grafiek 3. Horizontale lijn bij laagste & hoogste waarde die geen outlier is 4. Horizontale lijn voor P_{25} en P_{75} 5. Lijnen voor kwartielen met elkaar verbinden, als een rechthoek 6. Alle waarden van grafiek wissen, behalve outliers 7. Verticale stippellijn tekenen tussen overblijvende horizontale lijn & grens rechthoek → stippellijnen noemen we **snorharen** of **whiskers** 8. Mediaan aanduiden met horizontale lijn **ILLUSTRATIE: ZIE CURSUS P. 100!**		

HOOFDSTUK 4: SAMENHANG TUSSEN 2 VARIABELEN

Hoofdstuk 2 & 3 : één variabele per keer bekijken → univariate statistiek
Hoofdstuk 4 : twee variabelen tezamen bekijken → bivariate statistiek

1. Bivariate frequentieverdeling

WAT?	Frequentietabel, maar dan voor 2 variabelen (in plaats van voor 1)

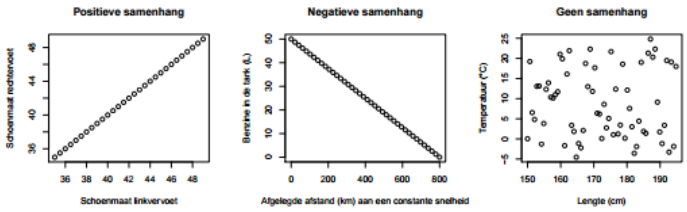
VOORDEEL	° We kunnen vanuit de bivariate tabel de univariate gegevens afleiden - OPGELET! Het werkt niet langs de andere kant! Vanuit univariate gegevens kunnen we geen bivariate gegevens afleiden - Marginale verdeling = andere naam voor univariate verdeling die je kent via de bivariate verdeling
NADEEL	° Conclusies kunnen anders zijn naarmate de data anders gegroepeerd is → subjectief - Oplossing : spreidingsdiagram & correlatiecoëfficiënten

2. Spreidingsdiagram

WAT?
° Geeft samenhang tussen 2 variabelen weer
° Alle waarden als bollen op grafiek

SOORTEN?
(Veel te extreem weergegeven, in realiteit bijna nooit zo)
° Positieve samenhang
° Negatieve samenhang
° Geen samenhang

	Klassen Hersengrootte		
IQ Groep	]790, 886]	]886, 982]	]982, 1080]
gemiddeld IQ	9	9	2
hoog IQ	7	9	4



3. Maten van samenhang

3.1 De covariantie cov_{XY}

NOTATIE	cov_{XY}
FORMULE	$cov_{XY} = \frac{1}{n-1} * \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})$
MEETNIVEAU	Beide variabelen : interval, ratio
SAMENHANG?	Uitkomst positief — positieve samenhang Uitkomst negatief — negatieve samenhang Uitkomst ong 0 — geen samenhang
NADEEL	° grootte van de covariantie hangt af van sterkte van samenhang & meeteenheid - Je kan niet echt zeker weten of je samenhang nu écht groot is of niet - oplossing : correlatiecoëfficiënt

3.2 De correlatiecoëfficiënt r_{XY}

NOTATIE	r_{XY}
FORMULE	$r_{XY} = \frac{cov_{XY}}{s_X * s_Y}$
EIGENSCHAPPEN	° Ligt altijd tussen -1 en 1 * 1 = perfect positieve samenhang * -1 = perfect negatieve samenhang ° Hetzelfde teken als covariantie ° Enkel te gebruiken bij lineaire samenhang
VOORDEEL	Beter dan covariantie, want de uitkomst is onafhankelijk van de meeteenheid → Het is altijd tussen -1 en 1
PROBLEMEN/VRAGEN BIJ OEFENINGEN	Als er geen lineaire samenhang is, kan je r_{XY} nog altijd berekenen zonder probleem, alleen is dat getal dan niet betrouwbaar

3.3 Kendall's Tau τ

FORMULE	$\tau = \frac{2 * (aantal\ concordante\ paren - aantal\ discordante\ paren)}{n * (n-1)}$ *(Komt in formularium)*
CONCORDANTE PAREN	Mathematisch $\frac{y_j - y_i}{x_j - x_i} > 0$ Grafisch Positieve rico → stijgende lijn
DISCORDANTE PAREN	Mathematisch $\frac{y_j - y_i}{x_j - x_i} < 0$ Grafisch Negatieve rico → dalende lijn
FORMULE # PAREN TOTAAL?	Om te weten hoeveel mogelijke paren er zijn (zowel concordant als discordant) $\frac{n * (n-1)}{2}$
WERKWIJZE	° Spreidingsdiagram tekenen ° Alle mogelijke rechten tussen 2 punten trekken ° Aantal concordante en discordante paren tellen ° Formule toepassen (formularium!)
EIGENSCHAPPEN	° Altijd tussen -1 en 1
MEETNIVEAU	Ordinaal, interval, ratio

3.4 Lineaire en niet-lineaire verbanden

LINEAIRE FUNCTIE	° Grafisch kan je een rechte lijn tekenen ° CORRELATIECOEFFICIENT gebruiken
MONOTONE FUNCTIE	° Bewaart de orde : eenmaal stijgen/dalen, blijven stijgen/dalen → Maar dus niet noodzakelijk in een rechte lijn ° KENDALL'S τ gebruiken
NIET-MONOTONE FUNCTIE	° Bewaart de orde niet ° correlatiecoëfficiënt en Kendall's τ allebei niet goed
BELANGRIJKE TIP	Data eerst visualiseren, dan pas weet je welke spreidingsmaat goed is

PROBLEMEN/VRAGEN BIJ OEFENINGEN	Zien aan de grafiek of het een lineaire of monotone samenhang is: echt heel algemeen kijken, niet te gedetailleerd!

3.5 Gevoeligheid aan outliers

GEVOELIG AAN OUTLIERS	Covariantie Correlatiecoëfficiënt
NIET GEVOELIG AAN OUTLIERS	Kendall's τ

4. De regressielijn

WAT?	Regressielijn zorgt ervoor dat we de r_{XY} kunnen visualiseren op een spreidingsdiagram → onze vroegere 'functie' dus
MEETNIVEAU	Interval, ratio
FORMULE (VOOR DE RECHTE)	$Y = b_0 + b_1 * X$ b_1 regressiecoëfficiënt – helling van de rechte (- de rico) $b_1 = \frac{y_j - y_i}{x_j - x_i}$ b_0 Snijpunt met de verticale as - intercept $b_0 = y_i - b_1 * x_i$!!! Deze 2 formules te gebruiken bij perfecte samenhang, anders de 2 hieronder!!!
KLEINSTE KWADRATENMETHODE	° Erg vaak is de samenhang niet perfect : onmogelijk om een regressielijn te tekenen die door alle punten gaat ° We willen uiteindelijk een rechte die toch zo goed mogelijk door de punten gaat ° Methode om dit te bereiken : kleinste kwadratenmethode → $\sum_{i=1}^{n}(y_i - (b_0 + b_1 * x_i))^2$ → Logica erachter : * Je hebt je regressielijn & je eigenlijke punten * De afstand tussen elk punt en de lijn is je fout (want je zit ernaast) * Het kwadraat van deze afstand wil je zo klein mogelijk, op die manier is je fout zo klein mogelijk → Kwadraat van het bolletje (y_i) en de regressielijn ($(b_0 + b_1 * x_i)$ → Via deze 2 formules zijn je b_0 en b_1 het meest geschikt volgens de methode * deze waarden kan je dan integreren in de 1e formule * $b_1 = r_{XY} * \frac{s_Y}{s_X}$ * $b_0 = \bar{y} - b_1 * \bar{x}$

5. Samenhang en causaliteit

Samenhang betekent GEEN causaliteit!
Er kan een derde variabele zijn, die niet bestudeerd is

Deel II : Kansrekening

WAAR GAAT ELK DEEL OVER?

Deel I steekproef
Deel II populatie
Deel III inductieproces

We willen altijd iets weten over de populatie, maar dat is te veel om allemaal te onderzoeken. Daarom doen we aan steekproeftrekking. Maar uiteindelijk willen we deze resultaten veralgemenen naar de gehele populatie.

HOOFDSTUK 5: DE POPULATIE EN VERDELINGSFUNCTIES

VERDELINGSFUNCTIE	*Frequentieverdeling, maar dan voor een populatie (geen steekproef)* Frequentieverdeling – steekproef Verdelingsfunctie – populatie Hoe deze eruit ziet hangt af van het soort variabele: discreet of continu

1. Verdelingsfunctie discrete variabelen

DISCRETE VARIABELEN	° kan geen tussenwaarden aannemen ° Eindig aantal waarden → Aantal mogelijke waarden: p → OPGELET : het is de variabele die een eindig aantal waarden heeft, niet de populatie (want erg vaak heeft populatie zodanig veel waarden, dat het wiskundig makkelijker is om aan te nemen dat er oneindig veel zijn
NOTATIE REL.FR. VAN DE POPULATIE	$P(X = x_i) = \lim_{n \to \infty} \frac{f_i}{n}$ ° De kans dat de variabele X de waarde x_i aanneemt ° $\frac{f_i}{n}$ = de relatieve frequentie
SOORTEN?	° kansverdeling ° Cumulatieve verdelingsfunctie

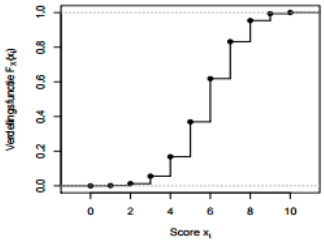
1.1 Kansverdeling

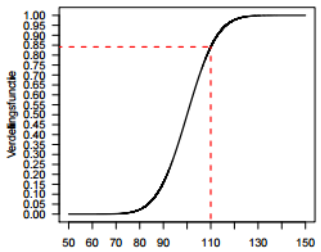
WAT ?	° *relatieve frequentieverdeling van de populatie (geen steekproef)* ° tabel met 2 kolommen : de waarden van x_i & de overeenkomstige kansen

1.2 Cumulatieve verdelingsfunctie / verdelingsfunctie $F_X(x)$

WAT?	° *Cumulatieve relatieve frequentie van de populatie (geen steekproef)* ° Geeft de kans weer dat een waarde van X kleiner of gelijk aan x is
FORMULE?	$F_X(x) = P(X \leq x)$
GRAFIEK OF TABEL?	Kan beide zijn! Het gaat gewoon om het feit dat er aan elke X-waarde een bijhorende Y-waarde wordt gekoppeld, maakt niet uit op welke manier dat weergegeven wordt → OPGELET! Beide zijn alleen mogelijk bij discrete variabelen! Bij continue variabelen is een tabel niet mogelijk, want er zijn oneindig veel punten
GRAFIEK	Trapsgewijs

Score x_i	$F_X(x_i) = P(X \leq x_i)$
0	0.00013
1	0.00171
2	0.01265
3	0.05593
4	0.16798
5	0.36938
6	0.61860
7	0.83201
8	0.95385
9	0.99388
10	1.00000





2. Verdelingsfunctie continue variabelen

Score x_i	$P(X = x_i)$
0	0.00013
1	0.00158
2	0.01094
3	0.04328
4	0.11205
5	0.20140
6	0.24922
7	0.21341
8	0.12184
9	0.04003
10	0.00612

CONTINUE VARIABELEN	° Kan oneindig veel tussenwaarden aannemen ° Kan oneindig veel waarden aannemen

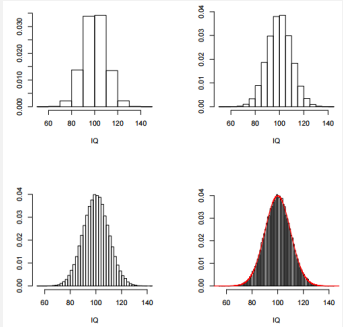
Probleem : doordat er oneindig veel waarden zijn, is de kans dat 1 specifieke waarde voorkomt quasi 0
→ $P(X = x_i) = 0$
→ We gaan een andere manier moeten vinden om kansen te berekenen
(zie volgende titels)

2.1 Cumulatieve verdelingsfunctie $F_X(x)$

WAT?	De kans dat een waarde van X kleiner of gelijk aan x is
FORMULE	$F_X(x) = P(X \leq x)$ of $F_X(x) = P(X < x)$

	$\leq$ of $<$ maakt niet uit, want de kans is uiteindelijk toch 0
GRAFIEK	continu (niet trapsgewijs)

2.2 De dichtheidsfunctie of de kansdichtheid $f_X(x)$

WAT?	° Formele uitleg: de afgeleide van de verdelingsfunctie ° Duidelijkere uitleg: → Histogram tekenen, waarbij oppervlakte rechthoek gelijk is aan relatieve frequentie → Er zijn een oneindig aantal waarden, dus we kunnen histogram opstellen met oneindig aantal klassen → Hoe meer klassen, hoe meer het lijkt op de dichtheidsfunctie 
FORMULE	AFGELEIDEN NIET ZELF BEREKENEN, GEWOON OM WAT TE KUNNEN VATTEN WAT DICHTHEIDSFUNCTIE IS $f_X(x) = \lim_{b \to 0} \frac{F_X(x+b) - F_X(x)}{b}$ → De kans dat X valt binnen het interval $]x, x+b]$ gedeeld door b → b = de breedte van het interval ; gaat richting 0
KANSEN BEREKENEN?	° Welke soort kansen? → Van de vorm $P(x_1 \leq X \leq x_2)$ ° Via integralen (niet zelf uitvoeren, gewoon begrijpen) →Grafisch : de oppervlakte tussen de 2 grenswaarden is wat je zoekt → Algemene formules (niet zelf kennen of gebruiken, gwn begrijpen) $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx$ $P(X \leq x) = \int_{-\infty}^{x} f_X(x)dx$ $P(X > x) = \int_{x}^{+\infty} f_X(x)dx$
EIGENSCHAPPEN	° $P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_2)$ ° Altijd positief, nooit negatieve waarden → functie is gebaseerd op een kans, een kans kan nooit negatief zijn ° Oppervlakte onder dichtheidsfunctie = 1 → Komt altijd overeen met de kans ; de volledige kans is 100% of 1 ° $P(X > x) = 1 - P(X < x)$

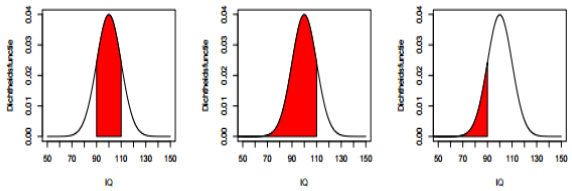
Voorbeeld: in de linker figuur zie je de kans dat X tussen 110 en 90 ligt. Dit is gelijk aan de kans dat X onder 110 ligt (middelste figuur), min de kans de X onder 90 ligt (rechter figuur)

3. Populatieparameters

3.1 Populatiegemiddelde of verwachtingswaarde E(X), μ_x of μ

Weetje : de E komt van 'expectation', vandaar verwachtingswaarde

3.1.1. Discrete variabelen



FORMULE	$E(X) = \sum_{i=1}^{p} P(X = x_i) * x_i$		
WAT IS VERANDERD?	relatieve frequenties	→	kansen $P(X = x_i)$
	x_i^u	→	x_i

3.1.2. Continue variabelen

We kunnen de vorige definitie niet gebruiken, want $P(X = x_i) = 0$
DUS gaan we gebruik maken van integralen (weeral niet zelf uitrekenen)

3.2 Populatievariantie V(X), σ_X^2 of σ^2

3.2.1. Discrete variabelen

FORMULE	$V(X) = \sum_{i=1}^{p} P(X = x_i) * (x_i - E(X))^2$		
WAT IS VERANDERD?	relatieve steekproeffrequenties	→	kansen $P(X = x_i)$
	Steekproefgemiddelde	→	E(X)
	x_i^u	→	x_i
STANDAARDDEVIATIE $\boldsymbol{\sigma}$	$\sigma = \sqrt{V(X)}$		

3.2.2. Continue variabelen

gebruik maken van integralen: niet zelf kunnen, wel begrijpen

4. Bivariate kansverdelingen

- Twee variabelen samen bekijken (in tabel), daar dan kansen over uitspreken
- Vergelijkbaar met hoofdstuk 4 (alleen was H4 over steekproefverdeling)

4.1 Discrete variabelen

MARGINALE VERDELINGEN AFLEIDEN	° Zelfde als normaal : de aparte waarden optellen ° $P(X = x_i) = \sum_{j=1}^{q} P(X = x_i \; en \; Y = y_j)$ → X wordt vastgehouden bij de waarde x_i (verandert dus niet), dan gaan we alle mogelijke y-waarden optellen die passen bij die waarde x_i → q = aantal mogelijke waarden van y ° $P\big(Y = y_j\big) = \sum_{i=1}^{p} P(X = x_i \; en \; Y = y_j)$ → Y wordt vastgehouden bij de waarde y_j (verandert dus niet), dan gaan we alle mogelijke x-waarden optellen die passen bij die waarde y_j
STATISTISCHE ONAFHANKELIJKHEID	° 2 discrete variabelen X en Y zijn onafhankelijk als de gelijkheid $P\big(X = x_i \; en \; Y = y_j\big) = P(X = x_i) * P(Y = y_j)$ geldt voor alle mogelijke combinaties van i en j → De kans dat zowel X als Y een specifieke waarde aannemen moet altijd gelijk zijn aan de kans dat X haar specifieke waarde aanneemt, vermenigvuldigd met de kans dat Y zijn specifieke waarde aanneemt → Als dit inderdaad zo is, zijn de variabelen onafhankelijk. Anders zijn ze afhankelijk.
COVARIANTIE COV(X,Y)	$COV(X,Y) = \sum_{i=1}^{p}\sum_{j=1}^{q} P\big(X = x_i \; en \; Y = y_j\big) * \big(x_i - E(X)\big) * (y_j - E(Y))$
CORRELATIECOËFFICIËNT $\boldsymbol{\rho_{XY}}$	$\rho_{XY} = \frac{COV(X,Y)}{\sigma_X * \sigma_Y}$

Illustratie statistische onafhankelijkheid: uit de tabel lees je af dat
P(X=4 en Y=10) = 0.09119
maar… P(X=4)x P(Y=10) = 0.35461 x 0.35286 = 0.1251277
Deze twee uitkomsten zijn niet hetzelfde, dus zijn X en Y niet onafhankelijk.

4.2 Continue variabelen

- Erg vaak kennis nodig van integralen : deze gebruiken we niet zelf, dus continue variabelen wordt maar zeer beknopt besproken
- Lijkt behoorlijk hard op univariate continue variabelen
  - $P(X = x_i \; en \; Y = y_j) = 0$

CUMULATIEVE BIVARIATE VERDELINGSFUNCTIE $\boldsymbol{F_{X,Y}(x,y)}$	$\boldsymbol{F_{X,Y}(x,y)} = P(X \leq x \; en \; Y \leq y)$
BIVARIATE DICHTHEIDSFUNCTIE $\boldsymbol{f_{X,Y}(x,y)}$	° De afgeleide van de cumulatieve bivariate verdelingsfunctie
STATISTISCHE ONAFHANKELIJKHEID	Continue variabelen zijn onafhankelijk als voor alle mogelijke waarden van x en y geldt dat: $P(X \leq x \; en \; Y \leq y) = P(X \leq x) * P(Y \leq y)$
COVARIANTIE COV(X,Y)	° Maakt gebruik van integralen : zullen we nooit zelf moeten berekenen ° $COV(X,Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y) * \big(x - E(X)\big) * \big(y - E(Y)\big) dxdy$
CORRELATIECOËFFICIËNT $\boldsymbol{\rho_{XY}}$	$\boldsymbol{\rho_{XY}} = \dfrac{COV(X,Y)}{\sigma_X * \sigma_Y}$

	Leeftijd Y	
Score X	10	11
0	0.00341	0.00021
1	0.02730	0.00404
2	0.08275	0.03291
3	0.12110	0.13337
4	0.09119	0.26342
5	0.02711	0.21319

5. Nuttige stellingen

1. **Als X en Y onafhankelijk zijn dan COV(X,Y) = 0**
   a. OPGELET: omgekeerd niet altijd waar! Niet altijd dat als de covariantie nul is, dat ze onafhankelijk zijn
      i. Kan ook bijvoorbeeld door niet-lineaire samenhang komen
   b. Alleen bij populatie, niet bij steekproef
2. **Als Y = X+a dan E(Y) = E(X)+a**
   a. Hierbij is a een constante
      i. Het gemiddelde van a = a
   b. Bij populatie en steekproef
3. **Als Y = a*X dan E(Y) = a*E(X)**
   a. Zelfde logica als stelling 2
   b. Bij populatie en steekproef
4. **E(X+Y) = E(X) + E(Y)**
   E(X-Y) = E(X) - E(Y)
   a. Zowel bij onafhankelijke als afhankelijke variabelen
   b. Bij populatie en steekproef

5. **Als X en Y onafhankelijk zijn**
   dan E(X*Y) = E(X)*E(Y)
   a. Alleen populatie, niet steekproef
6. **Als Y = X+a**
   dan V(Y) = V(X)
   a. Hierbij is a een constante
7. **Als Y = a*X**
   dan V(Y) = a^2*V(X)
   a. Hierbij is a een constante
8. **V(X+Y) = V(X) + V(Y) + 2*COV(X,Y)**
   a. Als X en Y onafhankelijk zijn
      dan V(X+Y) = V(X) + V(Y)
      i. Komt voort uit stelling 1 en 8
   b. Populatie en steekproef
9. **V(X-Y) = V(X) + V(Y) – 2*COV(X,Y)**
   a. Als X en Y onafhankelijk zijn
      dan V(X-Y) = V(X) + V(Y)
      i. Komt voort uit stelling 1 en 9
   b. Bij populatie en steekproef

6. Bijzondere (kans)verdelingen

6.1. De binomiale verdeling

WAT?	° Geeft de kans weer om k successen te halen bij N mogelijkheden → Bv. Een meerkeuze-examen met N vragen, hoe groot is je kans om k correcte antwoorden te hebben?
ILLUSTRATIE NODIG?	(Cursus p.164-166!!) ° Situatie : meerkeuze-examen, elke vraag heeft 4 antwoordmogelijkheden, proefpersonen moeten ad random antwoorden invullen ° Stel dat er maar 1 vraag is: 25% van de proefpersonen zullen A aangeduid hebben, 25% B enzovoort → Stel dat A het juiste antwoord is ; kans op succes = p = 25% ° Er komt een 2e vraag bij: van de mensen die A geantwoord hebben bij vraag 1, gaat weer een kwart A bij de 2e vraag antwoorden, een kwart B enzovoort → Antwoord A is weer het juiste antwoord ° Nu zijn er 3 mogelijkheden → Persoon heeft beide antwoorden fout - Zowel op vraag 1 als 2 gekozen voor antwoord B, C of D - 9 van de 16 groepen - P(X=0) = 9/16 → Persoon heeft 1 van de 2 antwoorden juist - OF op vraag 1 antwoord A en op vraag 2 geen A OF op vraag 1 geen A en op vraag 2 wel antwoord A - 6 van de 16 groepen - P(X=1) = 6/16 → Persoon heeft beide antwoorden juist - op beide vragen antwoord A

	- 1 van de 16 groepen - P(X=2) = 1/16
FORMULE	$P(X=k) = \frac{N!}{k!*(N-k)!} * p^k * (1-p)^{N-k}$ (Formularium!) ° ! = faculteit : vermenigvuldiging, afrollend tot 1 → N! = N x (N-1) x (N-2) x (N-3) x … x 2 x 1 →Bv. 4! = 4x3x2x1 = 24 → 0! = 1 ° p = kans op succes ° k = aantal successen ° N = maximaal aantal successen
NOTATIE & FORMULES	° Variabele met binomiale verdeling X ~ Binom(N,p) ° Verwachtingswaarde E(X) = N*p ° Variantie V(X) = N*p*(1-p)
VOORWAARDEN	° N is vast ° de kans op succes (p) blijft ongewijzigd
VARIABELE?	Altijd bij discrete variabelen
PROBLEMEN/VRAGEN BIJ OEFENINGEN	- Er is geen typische vorm qua grafiek - k en N bepalen: ° N is vaak je steekproefgrootte: hoe vaak de test gedaan wordt ° k is hoeveel je daarvan nodig hebt

6.2. De normale verdeling HEEL BELANGRIJK STUK!!

NORMAAL VERDEELDE VARIABELE	° Notatie: X ~ N(μ,σ²) ° Dichtheidsfunctie, formule
DICHTHEIDSFUNCTIE	° Formule $f_X(x) = \frac{1}{\sigma*\sqrt{2\pi}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ e ≈ 2.71 ; π ≈ 3.14 ° Symmetrisch ° Hoogste punt : ter hoogte van μ (op de x-as) – mediaan = gemiddelde!! ° Hoe groter σ², hoe breder en minder hoog de functie ° De functie wordt nergens 0 ° Volledige oppervlakte is nog steeds gelijk aan 1
VARIABELE?	Altijd bij continue variabelen

6.2.1. De standaardnormale verdeling

- De normale verdeling met μ = 0 en σ² = 1
  - Symmetrisch rond 0 (want dat is het gemiddelde)

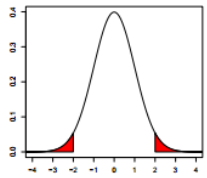
ENKELE EIGENSCHAPPEN..

1) $\boldsymbol{P(X > x) = P(X \leq -x)}$

Voorbeeld: de kans dat X groter is dan 2 (het rechter rode deel), is even groot als de kans dat X kleiner of gelijk aan -2 is (het linker rode deel)

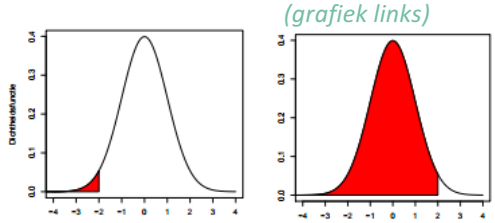
2) $\boldsymbol{P(X \leq -x) = 1 - P(X \leq x)}$

Voorbeeld: de kans dat X kleiner of gelijk aan -2 is (rode deel linker grafiek), is gelijk aan 1 – de kans dat X kleiner of gelijk aan 2 is (het witte deel van de rechter grafiek)



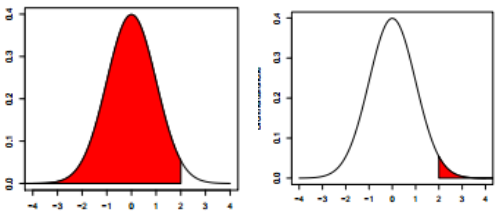
3) $\boldsymbol{P(X > x) = 1 - P(X \leq x)}$

Voorbeeld: de kans dat X groter is dan 2 (grafiek rechts) is gelijk aan 1 – de kans dat X kleiner of gelijk aan 2 is (grafiek links)



6.2.2. Kansen berekenen

- Er is een tabel die geldt voor de standaardnormale verdeling om kansen te berekenen

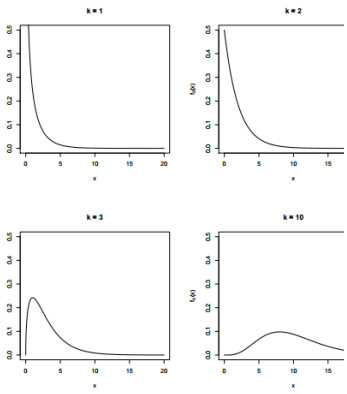


  - Zie formularium
  - OPGELET! Geldt enkel voor de standaardnormale verdeling
  - Het is de cumulatieve verdelingsfunctie van de standaardnormale variabele X
    - de kansen die je uitkomt zijn telkens de kansen dat de variabele waarden aanneemt die <u>kleiner of gelijk</u> aan je specifieke x-waarde zijn : <u>P(X ≤ x)</u>
- Hoe de tabel gebruiken?
  - Je wil de kans berekenen dat X kleiner of gelijk aan x is : P(X ≤ x)
  - Je x-waarde ga je aflezen in de kolommen bovenaan en links
    - Linkerkant: de eerste twee cijfers van x (tot 1 getal na de komma)
    - Bovenaan: het laatste cijfer van x (het 2e getal na de komma)
  - Zoek het kruispunt van deze 2 kolommen in de tabel zelf
  - De waarde die je daar treft, is de kans dat X kleiner of gelijk aan je x-waarde is
- Wat als het geen standaardnormale verdeling is, maar een normale verdeling van een andere vorm?
  - Dan moet je de variabele <u>standaardiseren</u>
  - *Eigenschap*: als X een niet-standaardnormale verdeling is, dan heeft de variabele Z wel een standaardnormale verdeling via deze formule: $Z = \frac{X-\mu}{\sigma}$
  - Als X ~ N(μ,σ²) , dan geldt dat : $P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma})$
    - De waarde die je uitkomt na het uitvoeren van de bewerking $\frac{X-\mu}{\sigma}$ (die dus rechts van het ongelijkheidsteken bij Z staat) is de x-waarde die je wil zoeken in de zijkanten van je tabel in het formularium
    - Van daaruit is het weer hetzelfde: het getal in de tabel dat overeenkomt met die x-waarde, is de kans dat $P(X \leq x)$
  - *Illustratie:*

- $P(X \leq 3) = P\left(Z \leq \frac{3-1}{\sqrt{4}}\right) = P(Z \leq 1)$
- *Je zoekt 1 dan als x-waarde op in de tabel*
- *De bijhorende kans is 0.8413*
- $P(X \leq 3) = 0.8413$

6.3. De χ^2-verdeling

WAT?	° Je neemt k aantal onafhankelijke standaardnormale variabelen ° de χ^2-verdeling is de som van die gekwadrateerde variabelen → som van kwadraten: altijd alleen maar positieve waarden
NOTATIE	Y ~ χ_k^2
FORMULES	°Algemene formule: $Y = X_1^2 + X_2^2 + X_3^2 + \cdots + X_k^2$ ° Verwachtingswaarde: E(Y) = k ° Variantie: V(Y) = 2*k
K?	k = het aantal vrijheidsgraden → hoeveel er zijn, dus.
VARIABELEN?	Altijd bij continue variabelen
PROBLEMEN/VRAGEN BIJ OEFENINGEN	Er is geen typische vorm qua grafiek

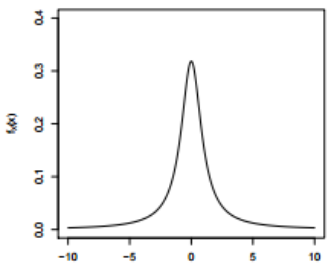


6.3.1. Kansen berekenen

- Zie tabel in formularium
  - Werkt wel wat anders dan tabel voor normale verdeling!
  - Linker kolom: het aantal vrijheidswaarden k zoeken
  - Bovenste kolom: de waarden van de verdelingsfunctie $F_Y(y)$
  - Tabel zelf: de waarden y van de variabele → datgene wat tussen haakjes staat als 'y' bij '$F_Y(y)$'

6.4. De t-verdeling

WAT?	° Er is een standaardnormale variabele en een variabele volgens de χ^2-verdeling. ° De variabelen zijn onderling onafhankelijk. ° De t_k-verdeling is de verdeling van de variabele $T = \frac{X}{\sqrt{\frac{1}{k} * Y}}$
NOTATIE	T ~ t_k
FORMULES	° Verwachtingswaarde E(T) = 0 ° Variantie $V(T) = \frac{k}{k-2}$ $voor\ k > 2$
VARIABELEN?	Altijd bij continue variabelen



PROBLEMEN / VRAGEN BIJ OEFENINGEN	Er is geen typische vorm qua grafiek

6.4.1. Kansen berekenen

- Zie tabel in formularium
  - Zelfde logica als bij chi-kwadraat verdeling
  - Linker kolom: het aantal vrijheidswaarden k zoeken
  - Bovenste kolom: de waarden van de verdelingsfunctie $F_T(t)$
  - Tabel zelf: de waarden van de variabele → datgene wat tussen haakjes staat als 't' bij '$F_T(t)$'

HOOFDSTUK 6: DE STEEKPROEVENVERDELING

ALGEMEEN

Wat bestuderen we?	Eigenschappen van variabelen die we bekomen door op willekeurige wijze een steekproef te trekken uit de populatie
Dus…?	We kijken ook naar een steekproef, alleen is het één van de ontelbaar mogelijke steekproeven uit de populatie. We weten niet exact wat de waarden zijn, het is een stuk abstracter en algemener
Wat belangrijk?	Reproduceerbaarheid: als je zelfde experiment zou uitvoeren met andere steekproef, gelijkaardige resultaten bekomen
→ Probleem	Vaak maar tijd & geld om experiment 1x uit te voeren
→ Oplossing	statistische formules

1. Steekproeftrekking

ASELECTE STEEKPROEFTREKKING	° Uit de populatie worden op random wijze n elementen geselecteerd → Die n elementen zijn onderling afhankelijk (daar gaan we toch van uit) ° Er zijn veel meer soorten steekproeftrekkingen, wij bespreken alleen deze ° Alle methoden & formules die we zien, zijn van toepassing op aselecte steekproef
NOTATIE	Alles met een hoofdletter schrijven ° De variabele: hoofdletter X ° De specifieke waarden: $X_1,\ X_2,\ X_3$ … → Bij één expliciete steekproef: x_1, x_2, x_3 …
TOEVALSVARIABELE	Een variabele die bekomen wordt door op toevallige wijze een steekproef uit de populatie te selecteren
KANSBEREKENING	$P\ (X = x_i)$ kan 2 betekenissen hebben: ° De relatieve frequentie in de populatie ° Kansberekening → Kans op gebeurtenis = rel.fr. van de gebeurtenis als experiment ∞ keer herhaald wordt → In praktijk is dat onmogelijk: we moeten dit proberen te benaderen * Hoe vaker je een experiment uitvoert, hoe dichter bij ∞, hoe dichter bij de echte kans, hoe betrouwbaarder

→ Een experiment meerdere keren herhalen = herhaalde steekproeftrekking

2. Steekproevenverdeling van het gemiddelde

NOTATIE	$\bar{X}$ (van een steekproef in het algemeen dus, niet van specifieke waarden in 1 steekproef)
VARIABELE	° Als je een herhaalde steekproeftrekking doet, dan merk je dat de waarden voor $\bar{X}$ veranderen → Logisch, want de waarden van de steekproef zijn telkens anders → $\bar{X}$ verandert, dus is letterlijk variabel
FORMULE	$\bar{X} = \frac{1}{n} * \sum_{i=1}^{n} X_i$
STEEKPROEFGROOTHEID OF EEN STATISTIEK	° Een bewerking toegepast op de variabelen $X_1, \dots, X_n$ ° $\bar{X}$ is hier een voorbeeld van ° Andere voorbeelden : modus, variantie..
STEEKPROEVENVERDELING	° Uiteindelijk willen we toch weten hoe alle verschillende gemiddelden vd steekproeven vd populatie zich verhouden onder elkaar! ° Stappenplan: → Oneindig aantal steekproeven trekken & gemiddelden berekenen → Histogram opstellen van alle steekproefgemiddelden (is dan een histogram met oneindig veel klassen) → Wat krijgen we: de dichtheidsfunctie van het gemiddelde of de steekproevenverdeling van het gemiddelde OPGELET: verschil frequentieverdeling – steekproevenverdeling ° frequentieverdeling = verdeling van een variabele ° steekproevenverdeling = verdeling van een steekproefgrootheid

2.1 Stellingen

$\boldsymbol{E(\bar{X}) = \mu_X}$	De verwachtingswaarde van het steekproevengemiddelde $\boldsymbol{\bar{X}}$ = het populatiegemiddelde van de variabele X → Het steekproefgemiddelde is een zuivere schatter voor het populatiegemiddelde!
$\boldsymbol{V(\bar{X}) = \frac{\sigma_X^2}{n}}$	De variantie van het steekproevengemiddelde = populatievariantie van de variabele, gedeeld door de steekproefgrootte → Logica: hoe groter de steekproef, hoe minder de gemiddeldes zullen variëren, hoe dichter bij het échte steekproevengemiddelde
$\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$	ALS $X_1, \dots, X_n$ random, onafhankelijk en normaal verdeeld zijn DAN is $\bar{X}$ ook normaal verdeeld: $\bar{X} \sim N(\boldsymbol{\mu_X}, \frac{\boldsymbol{\sigma_X^2}}{\boldsymbol{n}})$ → Logica: als populatie normaal verdeeld is, dan steekproefgemiddelde ook → Gaat op voor elke grootte van n, maar alleen als het normaal verdeeld is
CENTRALE LIMIETSTELLING	Het maakt niet uit hoe de steekproef verdeeld is ALS de steekproefgrootte groot genoeg is (vuistregel: > 30 DAN is het steekproefgemiddelde (+/-) normaal verdeeld → We kunnen deze dan, net als gewone normale verdelingen, standaardiseren & op die manier kansen berekenen! → Gaat op voor elke soort verdeling, maar alleen als n groot genoeg is

2.2 OPGELET!!!

Bij het standaardiseren van $\bar{X}$: je moet de waarden van $\bar{X}$ gebruiken, niet van X!!
→ Het gemiddelde blijft hetzelfde (zie formule)
→ de standaarddeviatie blijft NIET hetzelfde!! Je moet deze berekenen via de formule!

ALGEMENE FORMULE: $Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$

3. Steekproevenverdeling van de variantie

ALGEMEEN	° Ook een voorbeeld van een steekproefgrootheid of een statistiek ° Ook een variabele
NOTATIE	$SD_X^2 \; of \; S_X^2$
FORMULES	$SD_X^2 = \frac{1}{n} * \sum_{i=1}^{n} (X_i - \bar{X})^2$ $S_X^2 = \frac{1}{n-1} * \sum_{i=1}^{n} (X_i - \bar{X})^2$ $E(SD_X^2) = \frac{n-1}{n} * \sigma_X^2$ $E(S_X^2) = \sigma_X^2$

3.1 Stelling

$\frac{(n-1) * S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2$	ALS $X_1, \dots, X_n$ random, onafhankelijk en normaal verdeeld zijn DAN geldt $\frac{(n-1)*S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2$

HOOFDSTUK 7: BETROUWBAARHEIDSINTERVALLEN EN STATISTISCHE TOETSEN VOOR HET POPULATIEGEMIDDELDE

ALGEMEEN – Waar gaat het hoofdstuk over?

- We gaan proberen om op basis van een steekproef een uitspraak te formuleren over de populatie.
  - Eerst moeten we populatieparameter schatten op basis van resultaten uit steekproef
  - Dan Betrouwbaarheidsinterval opstellen & statistische toets gebruiken
  - Je kan nooit 100% zeker zijn dat je uitspraak over de populatie correct is
- Hier gaan we dat enkel doen bij het populatiegemiddelde. (Bij andere populatieparameters kan dat ook, maar dat is voor latere cursussen)

1. Schatters

NOTATIE	- Een populatieparameter: θ → een populatieparameter is een heel algemeen woord voor iets wat iets zegt over de

	toestand van uw populatie zelf. Bijvoorbeeld: het gemiddelde, de variantie.. - Een schatter van een populatieparameter: $\hat{\theta}$
WAT MAAKT EEN GOEIE SCHATTER?	- Ze moet zuiver zijn ° De verwachtingswaarde van de schatter = de populatieparameter ° $E(\hat{\theta}) = \theta$ ° Dat geeft aan dat de populatieparameter niet systematisch te groot of te klein wordt geschat - De variantie van de schatter wordt kleiner naarmate de steekproefgrootte groter wordt ° $V(\hat{\theta}) \downarrow \; als \; n \uparrow$ ° Nauwkeuriger, want meer info
$\sqrt{V(\hat{\theta})}$	° Standaarddeviatie van de schatter ° Standaardfout van de schatter → De schatter met de kleinste standaardfout is de beste: het efficiëntste
SCHATTER VS SCHATTING?	Schatter: algemeen, veranderlijk Schatting: van 1 specifieke steekproef, vaste waarde

1.1 Het gemiddelde

HOE GAAN WE POPULATIEGEMIDDELDE SCHATTEN?	Een logische optie: het steekproefgemiddelde gebruiken
IS HET STEEKPROEFGEMIDDELDE WEL EEN GOEDE SCHATTER?	Voldoet het aan de 2 voorwaarden van een goede schatter? ° Is ze zuiver? → $E(\bar{X}) = \mu_X$ → Dat was een stelling uit het vorige hoofdstuk ° Daalt de variantie als de steekproefgrootte stijgt? → Stelling uit vorig hoofdstuk: $V(\bar{X}) = \frac{\sigma_X^2}{n}$ → Als n groter wordt, is $\frac{\sigma_X^2}{n}$ kleiner, dus minder variantie
CONCLUSIE	Ja, het steekproefgemiddelde is een goede schatter! → Je wilt populatiegemiddelde schatten: gebruik steekproefgemiddelde

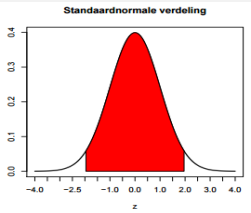
1.2 De variantie

HOE GAAN WE POPULATIEVARIANTIE SCHATTEN?	Een logische optie: de steekproefvariantie gebruiken
PROBLEEM – WELKE FORMULE GEBRUIKEN?	We hebben voor de steekproefvariantie 2 formules: S_X^2 en SD_X^2 → Stelling uit vorig hoofdstuk: $E(S_X^2) = \sigma_X^2$
CONCLUSIE	S_X^2 is een goede schatter voor de populatievariantie SD_X^2 is GEEN goede schatter voor de populatievariantie

2. Betrouwbaarheidsintervallen

- De steekproevenverdeling laat ons toe betrouwbaarheidsintervallen te construeren
- Via een betrouwbaarheidsinterval kunnen we met een bepaalde zekerheid een uitspraak doen over het populatiegemiddelde
- Er zijn verschillende werkwijzen, afhangende van de verdeling en de kennis over de populatievariantie

2.1 X normaal verdeeld en gekende populatievariantie

$z_{\propto}$	° De waarde van de standaardnormale verdeling, waarvoor de oppervlakte onder de curve rechts ervan gelijk is aan α ° $P(Z > z_{\propto}) = \propto \; met \; Z \sim N(0,1)$ ° OPGELET: als je in de tabel waarden wilt aflezen, gaat het altijd over 'inclusief alle waarden rechts ervan' → je moet gebruik maken van $1 - P(Z \leq z_{\propto})$ ° De standaardnormale verdeling is symmetrisch rond 0: → $P(-z_{\propto/2} \leq Z \leq z_{\propto/2}) = 1-\propto$ → of $P(-z_{\propto/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\propto/2}) = 1-\propto$ * We gaan standaardiseren, zodat het van toepassing is op een standaardnormale verdeling (is vereist, zie bovenste zin) → Waarom α/2?  Het is $z_{\propto/2}$ want hier is ∝ de som van de 2 witte gebieden, zowel links als rechts! Dus uw boven- of onderwaarde apart moet je delen door 2
BETROUWBAARHEIDS-INTERVAL BI	$P\left(\bar{X} - z_{\propto/2} * \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\propto/2} * \sigma/\sqrt{n}\right) = 1-\propto$ → Logica van de formule: ° Zie cursus p.211 ° Je hebt alle elementen van de Z-score aan alle kanten van de ongelijkheidstekens toegevoegd, zodat je uiteindelijk in het midden alleen μ uitkomt, want dat is uiteindelijk wat je zoekt → Uiteindelijk wil je dat je getal dat voor μ staat tussen die 2 grenzen ligt Grenzen van het betrouwbaarheidsinterval: $\left[\bar{X} - z_{\propto/2} * \sigma/\sqrt{n}\,, \bar{X} + z_{\propto/2} * \sigma/\sqrt{n}\right]$ → Deze formule ga ik vooral gebruiken bij oefeningen Dat is dan het (1-∝)*100% betrouwbaarheidsinterval → Bv. ∝ = 0.05 : het (1-0.05)*100% BI : het 95%-BI → <u>Wat betekent het? Dat is 95% van de gevallen van de steekproeftrekkingen het reële populatiegemiddelde er zich ook echt in bevindt</u> Het steekproefgemiddelde gaat altijd exact in het midden van het BI liggen
INTERPRETATIE	Wat gebeurt er als het experiment herhaalt op basis van een nieuwe steekproef? → σ en $z_{\propto/2}$ zijn vaste waarden → Maar nieuwe steekproef = nieuwe gegevens = nieuw gemiddelde Resultaat: gezien het gemiddelde variabel is, zal het betrouwbaarheidsinterval ook variabel zijn → de grenzen van het betrouwbaarheidsinterval zullen verschillen per steekproef → een 95%-betrouwbaarheidsinterval garandeert dat 95% van al die variabele intervallen het reële populatiegemiddelde zullen bevatten als we het experiment een oneindig aantal keer zullen herhalen
EIGENSCHAPPEN	Breedte van het interval $[a, b] = 2 * z_{\propto/2} * \sigma/\sqrt{n}$ → Hoe smaller het interval, hoe nauwkeuriger → Het kleinste betrouwbaarheidsinterval mogelijk: α = 0.01 ° interval is dan 99% betrouwbaar

2.2 X normaal verdeeld en ongekende populatievariantie

FORMULE	° $P\left(\bar{X} - t_{n-1;\propto/2} * S_X/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1;\propto/2} * S_X/\sqrt{n}\right) = 1-\propto$ ° Grenzen betrouwbaarheidsinterval: $\left[\bar{X} - t_{n-1;\propto/2} * S_X/\sqrt{n}\ ,\ \bar{X} + t_{n-1;\propto/2} * S_X/\sqrt{n}\right]$
LOGICA ACHTER FORMULE	° Combinatie van 2 stellingen: → $\frac{(n-1)*S_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$ → $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ ° Een χ^2-verdeling en een standaardnormale verdeling vormen samen een t-verdeling ° Het is uiteindelijk een t_{n-1}-verdeling, omdat het in de eerste stelling ook een χ^2_{n-1}-verdeling is
EIGENSCHAPPEN $\boldsymbol{t_{n-1}}$-VERDELING	° Een grotere variantie dan een standaardnormale verdeling ° $t_{n-1;\propto/2}$-waarde is groter dan $z_{\propto/2}$-waarde → OPGELET: ook hier is de $t_{n-1;\propto/2}$-waarde de waarde voor alles rechts ervan, niet links! Moet je weer rekening mee houden bij het aflezen van de tabel ° Het betrouwbaarheidsinterval is groter, dus extra variabiliteit → Logisch, want je het de populatiestandaarddeviatie moeten schatten, waardoor een grotere kans op fouten ° OPGELET: die 'n-1' betekent niet dat je voor het aflezen van de tabel er eentje moet bijtellen bij het aantal vrijheidswaarden! 'n-1'=k!

2.3 X niet normaal verdeeld & populatievariantie niet gekend

GROTE STEEKPROEF	° Centrale limietstelling gebruiken: $\bar{X}$ is normaal verdeeld ° Zelfde werkwijze vorige titel: → interval BI: $\left[\bar{X} - t_{n-1;\propto/2} * S_X/\sqrt{n}\ ,\ \bar{X} + t_{n-1;\propto/2} * S_X/\sqrt{n}\right]$
KLEINE STEEKPROEF	Wordt niks over gezegd, moeten we niet kunnen..

2.4 Algemeen overzicht werkwijzen

X NORMAAL VERDEELD σ_X^2 GEKEND	Grenzen betrouwbaarheidsinterval: $\left[\bar{X} - z_{\propto/2} * \sigma/\sqrt{n}, \bar{X} + z_{\propto/2} * \sigma/\sqrt{n}\right]$
X NORMAAL VERDEELD σ_X^2 ONGEKEND	Grenzen BI : $\left[\bar{X} - t_{n-1;\propto/2} * S_X/\sqrt{n}\ ,\ \bar{X} + t_{n-1;\propto/2} * S_X/\sqrt{n}\right]$
X NIET-NORMAAL VERDEELD σ_X^2 ONGEKEND	Grote steekproef: centrale limietstelling Grenzen BI: $\left[\bar{X} - t_{n-1;\propto/2} * S_X/\sqrt{n}\ ,\ \bar{X} + t_{n-1;\propto/2} * S_X/\sqrt{n}\right]$ (Kleine steekproef: niet kunnen)

3. Statistische toetsen

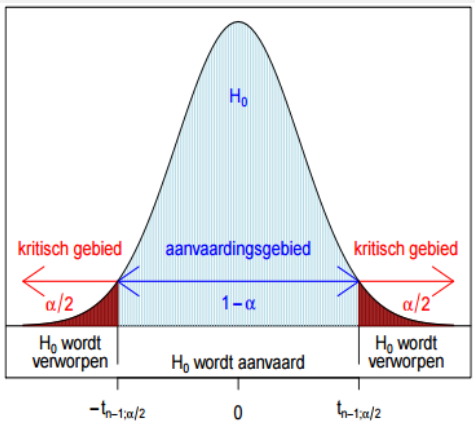
ALGEMENE VOORWAARDEN VOOR HET GEBRUIK VAN STATISTISCHE TOETSEN	° Het moet over <u>1 grote steekproef</u> gaan ° De variabele is <u>normaal verdeeld</u> ° <u>σ_x^2 is ongekend</u>
DE 2 HYPOTHESEN	Er zijn telkens 2 stellingen ivm de hypothese, 1 van de 2 is juist: ° Nulhypothese H_0 → $H_0: \mu = \mu_0$ → We bewijzen dat het reële populatiegemiddelde gelijk is aan het geschatte populatiegemiddelde → In praktijk is het erg vaak dat we deze hypothese proberen te verwerpen ° Alternatieve hypothese H_a → $H_a: \mu \neq \mu_0$ → We bewijzen dat het reële populatiegemiddelde niet gelijk is aan het geschatte populatiegemiddelde → Ook wel de tweezijdig alternatieve hypothese genoemd: μ_0 kan uiteindelijk langs beide kanten van μ liggen, erboven of eronder
EN DAN?	° Als $\bar{x}$ ongeveer gelijk is aan μ_0, dan gaan we H_0 aanvaarden & H_a verwerpen ° Als $\bar{x}$ ver afwijkt van μ_0, dan gaan we H_0 verwerpen & H_a aanvaarden ° MAAR nu is de vraag: wat is 'ongeveer gelijk aan' en 'ver afwijken van'?? → Gaan we bewijzen via een statistische toets: een toetsingsgrootheid → We gaan proberen voldoende bewijs te vinden om H_0 te verwerpen

3.1 Toetsingsgrootheid

EXTRA VARIABELE	We gaan een extra variabele G invoeren:

	$G = \frac{\bar{X} - \mu_0}{S_X/\sqrt{n}}$ → Als H_0 correct is, dan volgt G een t_{n-1}-verdeling

3.2 Beslissingsregels

WAT?	Regels die je moet volgen om te beslissen of we H_0 gaan aanvaarden of verwerpen
HOE GELDEN ZE?	Als H_0 waar is, dan ligt G rond 0 Als H_0 niet waar is, dan ligt G ver van 0 (zowel positief als negatief kan) Werkt ook langs de andere kant: - Als g rond 0 ligt, dan is H_0 waar - als g ver van 0 ligt, dan is H_0 fout
BESLISSINGSINTERVAL & KRITISCHE WAARDEN	$-t_{n-1;\propto/2} \leq g \leq t_{n-1;\propto/2}$ → Als je g-waarde (die je bekomt via formule $G = \frac{\bar{X}-\mu_0}{s_X/\sqrt{n}}$) tussen deze 2 waarden ligt, dan aanvaarden we H_0 Kritische waarden: $-t_{n-1;\propto/2}$ en $t_{n-1;\propto/2}$ 

7.3 Type I en type II fout

TYPE I FOUT	° H_0 verwerpen, maar eigenlijk is die juist ° De kans hiertoe : $P(verwerp\ H_0\ \mathrm{I}\ \mu = \mu_0) = \alpha$ → H_0 correct aanvaarden: $P(aanvaard\ H_0\ \mathrm{I}\ \mu = \mu_0) = 1 - \alpha$
TYPE II FOUT	° H_0 aanvaarden, maar eigenlijk is die fout ° De kans hiertoe: : $P(aanvaard\ H_0\ \mathrm{I}\ \mu \neq \mu_0) = \beta$ → H_0 correct verwerpen: $P(verwerp\ H_0\ \mathrm{I}\ \mu \neq \mu_0) = 1 - \beta$
ONDERLING VERBAND	° α daalt = β stijgt ° n stijgt = β daalt

	In werkelijkheid is H_0 juist	In werkelijkheid is H_0 fout
We aanvaarden H_0	Juiste beslissing (A) Betrouwbaarheid $(1-\alpha)$	Foute beslissing (C) Type II fout β
We verwerpen H_0	Foute beslissing (B) Type I fout α	Juiste beslissing (D) Onderscheidingsvermogen $(1-\beta)$

3.4 Beslissingsregels op basis van het betrouwbaarheidsinterval

μ_0 ligt binnen betrouwbaarheidsinterval	H_0 aanvaarden
μ_0 ligt buiten betrouwbaarheidsinterval	H_0 verwerpen

Extra info: rekenmachine

Je kan redelijk wat zaken ook ingeven in de GRM en daarna info hierover aflezen
!! OPGELET Ik denk niet dat we een grafische rekenmachine mogen gebruiken op het examen !!

- Info ingeven in lijsten
    - 1-Var-Stat = in verband met 1 variabele
    - 2-Var-Stat = in verband met 2 variabelen
        - 'a' = b_1
        - 'b' = b_0
        - Deze twee kan je gebruiken om het functievoorschrift van de regressielijn te noteren
- 's_X' = de standaarddeviatie die in de formule deelt door 'n – 1'
- 'σ_X' = de standaarddeviatie die in de formule deelt door 'n'