Inleiding

1 Enkele misvattingen

1.1 Met statistiek kan je alles bewijzen

- door statistische analyses verkeerdelijk toe te passen, kan je de impressie wekken dat je kan aantonen wat je wil
- fouten zijn snel gemaakt en moeilijk te detecteren

1.2 Statistiek is nutteloos voor de gedragswetenschappen

Gedragswetenschappen gebaseerd op: empirisch onderzoek → kennis door middel van observaties en metingen

Observaties en metingen → data

Statistisch analyseren van data → inzicht in de procesen die bestudeerd worden

1.3 Statistiek is enkel wiskunde

Cursus bestaat uit drie componenten:

- wiskunde: methodes in de taal van de wiskunde
- software: wiskundige methodes op data toepassen via statistische software
- interpretatie en besluitvorming

2 De betekenis van statistiek

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

Populatie: de volledige verzameling van objecten of personen waarover informatie wordt gewenst

Elementen: de individuele leden van de populatie

Steekproef: een deelverzameling van de populatie die feitelijk zal onderzocht worden om informatie te bekomen

Variabele: een eigenschap die bij de elementen van de populatie of steekproef varieert

Data: de verzameling van gegevens die wordt bekomen door de variabelen te meten

Verdeling: welke waarden worden aangenomen en hoe vaak

Inductie: uitgaande van het bijzondere het algemene besluiten

3 Eigenschappen van variabelen

3.1 Schaalfamilies

Vier meetschalen:

- nominale schaal: identificatie zonder dat ze een hoeveelheid aanduiden
- ordinale schaal: waarden duiden een volgorde aan
- intervalschaal: verschillen tussen waarden hebben een betekenis, maar er is geen absoluut nulpunt
- ratioschaal: absoluut nulpunt

3.2 Discrete en continue variabelen

Continue variabelen: kunnen tussenwaarden aannemen

Discrete variabelen: steeds twee waarden waar geen derde waarde kan tussen liggen, eindig aantal

Visualiseren van data

1 Cirkeldiagram

Gebruikt voor: variabelen van nominaal meetniveau

Notatie:
- variabele: hoofdletter X
- waarden: kleine letter met cijfers als subscript x_n
- aantal elementen in de steekproef: n
- één van de mogelijke waarden van X: x

Absolute frequentie: het aantal keer dat de waarde x in de steekproef voorkomt

Absolute frequentieverdeling: tabel met twee rijen waar de eerste rij de mogelijke waarden van X weergeeft en de tweede rij overeenkomstige absolute frequenties

Steekproefgrootte: het aantal elementen in de steekproef

Relatieve frequentie: $\frac{absolute\ frequentie}{steekproefgrootte\ n}$ → som moet gelijk zijn aan 1

Verdeling: geheel van mogelijke waarden

Cirkeldiagram: relatieve oppervlaktes stukken zijn gelijk aan de relatieve frequenties

Probleem: menselijk oog is niet goed in staat om de oppervlaktes van een cirkeldiagram te beoordelen

2 Staafdiagram

Opstellen: rechthoek waarbij de hoogte gelijk is aan de frequentie en alle breedtes gelijk zijn, afstand tussen rechthoeken moet ook gelijk zijn

Meetniveau: nominaal of ordinaal

3 histogram

Klassenbreedte $]a, b]$: b – a

Gegroepeerde frequentieverdeling: tabel met twee kolommen of rijen waar de eerste kolom / rij de klassen van X weergeeft en de tweede de overeenkomstige frequenties

Opstellen: waarden variabele op horizontale as → boven elke klasse een rechthoek met breedte = klassenbreedte en hoogte = relatieve frequentie gedeeld door klassenbreedte zodat oppervlakte rechthoek gelijk is aan relatieve frequentie

Alle klassen dezelfde breedte: hoogte kan gelijk zijn aan absolute frequentie

Verschillen staafdiagram – histogram:
- bij histogram raken rechthoeken elkaar en kunnen breedtes verschillen
- staafdiagram vooral voor ordinale en nominale variabelen (beperkt aantal waarden)
- histogram vaak gebruikt voor interval- en ratioschaalvariabelen (groot aantal waarden)

Vuistregel: data indelen in ongeveer $\sqrt{n}$ klassen

Nadeel gebruikersafhankelijkheid: door een slechte keuze te maken van de klassen kan de figuur een vertekend beeld geven

Verdeling:

- scheef naar rechts: meeste massa van histogram ligt links en staart rechts
- scheef naar links: uitlopende linkerstaart
- symmetrisch: linker- en rechterstaarten ongeveer gelijk

4 Cumulatieve frequentiecurve

4.1 Ongegroepeerde data

Cumulatieve absolute frequentie: het aantal elementen in de steekproef kleiner dan of gelijk aan x → F(x)

Cumulatieve absolute frequentieverdeling: tabel met twee kolommen / rijen waar in de eerste kolom / rij de waarden van de variabele X worden gegeven en in de tweede de overeenkomstige cumulatieve absolute frequenties

4.2 Gegroepeerde data

Cumulatieve absolute frequentie: het aantal elementen in die klasse plus het aantal elementen in lagere klassen

Groeperen van data: informatieverlies

Samenvatten van data

1 Centrummaten

1.1 Het gemiddelde

Berekenen op basis van:

- waarden van een variabele
- frequentieverdeling
- gegroepeerde data

1.1.1 Het gemiddelde op basis van de waarden van een variabele

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

Meetniveau: interval- en ratiovariabelen

Klassieke afrondingsregels:

- derde cijfer na de komma kleinder dan 5 → naar beneden
- derde cijfer na de komma gelijk aan of groter dan 5 → naar boven

Bij symmetrische verdeling: gemiddelde meer in het midden

Scheef naar rechts: gemiddelde meer naar links

1.1.2 Het gemiddelde berekenen op basis van de frequentieverdeling

Unieke waarden: x_i^u

Absolute frequentie horende bij unieke waarde: f_i

Gemiddelde: $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} f_i\, x_i^u$

Aantal unieke waarden van X: p

1.1.3 Het gemiddelde voor gegroepeerde data

Klassenmidden: klassenmidden van interval $]a, b]$ is $\frac{a+b}{2}$

Gemiddelde: $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} f_i \frac{(a_1+b_1)}{2}$

1.2 De mediaan

Symbool: md_x

Mediaan: waarde md_x waarvoor geldt dat

- niet meer dan de helft van de elementen in de steekproef een waarde kleiner dan md_x hebben
- niet meer dan de helft van de elementen in de steekproef een waarde groter dan md_x hebben

Meetniveau: ordinale, interval- en ratiovariabelen

Meetniveau mediaan door rekenkundig gemiddelde: interval- en ratiovariabelen
Voor gegroepeerde data: klasse bepalen waartoe de mediaan behoort = mediane klasse, daarna $md_x = a + \frac{\left(\frac{n}{2}-c\right)(b-a)}{d}$ met

- a: de ondergrens van de mediane klasse
- b: de bovengrans van de mediane klasse
- c: cumulatieve absolute frequentie van de klasse net kleiner dan de mediane klasse
- d: absolute frequentie van de mediane klasse
- n: steekproefgrootte

1.3 De modus

Modus mo: de klasse of de waarde met de grootste frequentie

Meetniveau: nominale, ordinale, interval- en ratiovariabelen

1.4 Gevoeligheid aan outliers

Outlier: waarde die ver verwijderd is van de overige waarden van een variabele

Gemiddelde: gevoelig

Mediaan: niet gevoelig

Modus: niet gevoelig

2 Spreidingsmaten

2.1 De variatiebreedte

Variatiebreedte v_x:

- de grootste min de kleinste waarde
- de bovengrens van de laatste klasse min de ondergrens van de eerste klasse

Variatiebreedte gelijk aan 0: grootste en kleinste waarde gelijk → geen spreiding

Meetniveau: interval- en ratiovariabelen

2.2 De gemiddelde absolute afwijking

Gemiddelde absolute afwijking: $ga_x = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

Meetniveau: interval- en ratiovariabelen

2.3 De variantie en de standaarddeviatie

Nadeel gemiddelde absolute afwijking: absolute waarden gebruikt

Variantie: $sn_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ of $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Meetniveau: interval- en ratiovariabelen

Standaarddeviatie: $sn_x = \sqrt{sn_x^2}$ of $s_x = \sqrt{s_x^2}$

Variantie op basis van frequentieverdeling: $sn_x^2 = \frac{1}{n}\sum_{i=1}^{p} f_i(x_i^u - \bar{x})^2$

2.4 De interkwartielafstand

k-de percentiel: $\frac{F(P_k)}{n} = \frac{k}{100}$

Drie belangrijke kwartielen:

- eerste kwartiel P_{25}
- tweede kwartiel / mediaan P_{50}
- derde kwartiel P_{75}

Interkwartielafstand: $Q = P_{75} - P_{25}$

Meetniveau: ordinale, interval- en ratiovariabelen

2.5 De spreidingsmaat d

Spreidingsmaat: $d = \frac{1-\frac{fmo}{n}}{1-\frac{1}{p}}$

Geen spreiding: d = 0

Maximale spreiding: d = 1

Meetniveau: nominale, ordinale, interval- en ratiovariabelen

2.6 Gevoeligheid aan outliers

Niet gevoelig: enkel interkwartielafstand

3 Boxplot

- bepalen van outliers
- stip voor alle waarden
- outliers inkleuren
- horizontale lijn bij grootste en kleinste waarde die geen outlier is
- twee horizontalen bij eerste en derde kwartiel
- box maken
- stippenlijn tussen box en kwartiellijnen
- horizontale mediaan

Samenhang tussen twee variabelen

1 Bivariate frequentieverdeling

Univariate absolute frequentieverdeling: tabel bevat enkel informatie over één variabele → conclusies over iedere variabele afzonderlijk

Bivariate absolute frequentieverdeling: bevat informatie over twee variabelen → gezamenlijk bestuderen

Marginale verdelingen: univariate verdelingen bepalen op basis van bivariate verdeling

Bivariate frequentieverdeling:

- bevat meer informatie dan univariate
- inzicht in samenhang tussen twee variabelen
- conclusies kunnen wijzigen door data te hergroeperen → subjectiviteit

2 Spreidingsdiagram

Spreidingsdiagram: figuur die ons zal toelaten de samenhang tussen deze twee variabelen te visualiseren

Verschillende soorten samenhang:

- positieve samenhang: punten van linksonder tot rechtsboven
- negatieve samenhang: punten van linksboven tot rechtsonder
- geen samenhang: geen patroon

Interpreteren is subjectief: verschillende personen kunnen andere conclusies trekken op basis van dezelfde figuur → samenhang kwantificeren

3 Maten van samenhang

3.1 De covariantie

Covariantie: $cov_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Meetniveau: beiden variabelen van tenminste intervalniveau

Er geldt dat:

- $cov_{XY} > 0$ bij een positieve samenhang
- $cov_{XY} < 0$ bij een negatieve samenhang
- $cov_{XY} \approx 0$ indien er geen samenhang is

Spreidingsdiagram in kwadranten:

- verticale lijn bij gemiddelde van x
- horizontale lijn bij gemiddelde van y

Nadeel: de grootte van de covariantie hangt niet enkel af van de sterkte van de samenhang, maar ook van de meeteenheid

3.2 De correlatiecoëfficiënt

Correlatiecoëfficiënt: $r_{XY} = \frac{cov_{XY}}{s_X s_Y}$ met $-1 \leq r_{XY} \leq 1$

Er geldt dat:

- bij een perfecte positieve samenhang: $r_{XY} = 1$

- bij een perfecte negatieve samenhang: $r_{XY} = -1$
- indien er geen samenhang is: $r_{XY} \approx 0$

3.3 Kendall's τ

Berekenen door: concordante en discordante paren te tellen

Concordant: $\frac{y_j - y_i}{x_j - x_i} > 0$

Discordant: $\frac{y_j - y_i}{x_j - x_i} < 0$

Kendall's τ: $\tau = \frac{2(aantal\ concordante\ paren - aantal\ cordante\ paren)}{n(n-1)}$ met $-1 \leq \tau \leq 1$

Meetniveau: ordinale, interval- en ratiovariabelen

Concordante en discordante paren visueel voorstellen: alle punten paarsgewijs verbinden → rechten met positieve rico zijn concordant

3.4 Lineaire en niet-lineaire verbanden

Correlatiecoëfficiënt: maat voor de lineaire samenhang

Kendall's τ: maat voor monotone samenhang

Lineaire functie: functie die kan voorgesteld worden door een rechte lijn

Monotone functie: functie die de orde bewaart → ofwel stijgend, ofwel dalend

3.5 Gevoeligheid aan outliers

Gevoelig: covariantie en correlatiecoëfficiënt (maken gebruik van waarden van variabelen)

Niet gevoelig: Kendall's τ (maakt enkel gebruik van volgorde van variabelen)

4 De regressielijn

Regressielijn: correlatiecoëfficiënt visualiseren op een spreidingsdiagram

Formule: $Y = b_0 + b_1 X$

- b_1 = regressiecoëfficiënt (helling)
- b_0 = intercept (snijpunt met verticale as)

4.1 Formules indien het lineair verband perfect is

Perfect lineair verband: precies één rechte door alle punten

Regressiecoëfficiënt: $b_1 = \frac{y_j - y_1}{x_j - x_i}$

Intercept: $b_0 = y_i - b_1 x_i$

4.2 Formules indien het lineair verband niet perfect is

Regressiecoëfficiënt: $b_1 = r_{XY} \frac{s_Y}{s_X}$

Intercept: $b_0 = \bar{y} - b_1\bar{x}$

Best passende rechte: hoeft niet door de punten te gaan, maar haar gekwadrateerde afstand tot de punten is het kleinst

Meetniveau: beiden variabelen tenminste van intervalniveau

Regressielijn tekenen:
- twee willekeurige waarden van X nemen
- formule van regressielijn invullen
- punten tekenen
- punten verbinden

5 Samenhang en causaliteit

Causaal verband: oorzakelijk verband

De populatie en de verdelingsfuncties

1 Verdelingsfunctie discrete variabelen

Verdelingsfunctie: tegenhanger van de frequentieverdeling maar voor een populatie

p = aantal waarden dat een variabele kan aannemen

P(X = x_i): kans dat de variabele X de waarde x_i aanneemt → $P(X = x_i) = \lim_{n\to\infty} \frac{f_i}{n}$ → relatieve frequentie van x_i in de populatie

1.1 De kansverdeling

Kansverdeling: tabel met twee kolommen (of rijen) waarbij de eerste kolom (of rij) de waarden x_i weergeeft en de tweede kolom (of rij) de overeenkomstige kansen P(X=x_i)

Kans ligt in interval: $[0,1]$

1.2 De cumulatieve verdelingsfunctie

Cumulatieve verdelingsfunctie: de kans dat de waarde van een variabele X kleiner dan of gelijk is aan $F_x(x) = P(X \leq x)$

2 Verdelingsfunctie continue variabelen

Continue variabelen kan oneindig veel waarden aannemen: kans P(X = x) = 0 voor elke waarde x

2.1 De cumulatieve verdelingsfunctie

Er zijn wel kansen die verschillend zijn van 0: $P(X \leq x)$

Cumulatieve verdelingsfunctie: de kans dat de waarde van een variabele X kleiner dan of gelijk is aan x → $F_X(x) = P(X \leq x)$

Opgelet: bij continue variabelen maakt het niet uit of we < of ≤ gebruiken omdat P(X = x) = 0

2.2 De dichtheidsfunctie

Dichtheidsfunctie: $f_X(x) = \lim_{b\to 0} \frac{F_X(x+b)-F_X(x)}{b}$

Histogram: oppervlaktes rechthoeken zijn gelijk aan de relatieve frequenties

Naarmate aantal klassen toeneemt: histogram kan meer en meer benaderd worden door een continue functie

Algemeen: $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx$

Er geldt dat: $P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F_X(x_2) - F_X(x_1)$

Eigenschappen:

- de dichtheidsfunctie is een positieve functie
- de volledige oppervlakte onder de dichtheidsfunctie is gelijk aan 1
- er geldt dat P(X > x) = 1 – P(X ≤ x)

3 Populatieparameters

3.1 Populatiegemiddelde

3.1.1 Discrete variabelen

Gemiddelde: $E(X) = \sum_{i=1}^{p} P(X = x_i)x_I$

3.1.2 Continue variabelen

Gemiddelde: $E(X) = \int_{-\infty}^{+\infty} f_x(x)dx$

3.2 Populatievariantie

3.2.1 Discrete variabelen

Variantie: $V(X) = \sum_{i=1}^{p} P(X = x_i)\,(x_i - E(X))^2$

Standaarddeviatie: $\sigma_X = \sqrt{V(X)}$

3.2.2 Continue variabelen

Variantie: $V(X) = \int_{-\infty}^{+\infty} f_X(x)(x - E(X))^2\,dx$

4 Bivariate kansverdelingen

4.1 Discrete variabelen

Kans: $P(X = x_i\ en\ Y = y_j$

Univariate verdeling van X: $P(X = x_i) = \sum_{j=1}^{q} P(X = x_i\ en\ Y = y_j)$

Univariate verdeling van Y: $P\big(Y = y_j\big) = \sum_{i=1}^{q} P(X = x_i\ en\ Y = y_j)$

Statistische onafhankelijkheid: twee 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

Covariantie: $COV(X,Y) = \sum_{i=1}^{p}\sum_{j=1}^{q} P(X = x_i\ en\ Y = y_j)(x_i - E(X))\big(y_j - E(Y)\big)$

Correlatiecoëfficiënt: $\rho_{XY} = \frac{COV(X,Y)}{\sigma_X\sigma_Y}$

4.2 Continue variabelen

Cumulatieve bivariate verdelingsfunctie: $F_{X,Y}(x,y) = P(X \le x\ en\ Y \le y)$

Bivariate dichtheidsfunctie: $f_{X,Y}(x,y)$

Twee continue variabelen X en Y zijn onafhankelijk als geldt dat: $P(X \le x\ en\ Y \le y) = P(X \le x)P(Y \le y)$

Covariantie: $COV(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\big(x - E(X)\big)\big(y - E(Y)\big)\,dxdy$

Correlatiecoëfficiënt: $\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X\sigma_Y}$

5 Nuttige stellingen

Stelling 1: als X en Y onafhankelijke variabelen zijn, dan geldt dat COV(X,Y) = 0

Stelling 2: voor een variabele Y = X + a geldt dat E(Y) = E(X) + a waarbij a een constante is

Stelling 3: voor een variabele Y = aX geldt dat E(Y) = aE(X) waarbij a een constante is

Stelling 4: voor twee variabelen X en Y (die onafhankelijk of afhankelijk kunnen zijn) geldt dat E(X + Y) = E(X) + E(Y) en E(X – Y) = E(X) – E(Y)

Stelling 5: voor twee onafhankelijke variabelen X en Y geldt dat E(XY) = E(X)E(Y)

Stelling 6: voor een variabele X + a geldt dat V(Y) = V(X) waarbij a een constante is

Stelling 7: voor een variabele Y = aX geldt dat $V(Y) = a^2V(X)$ waarbij a een constante is

Stelling 8: voor twee afhankelijke variabelen X en Y geldt dat V(X + Y) = V(X) + V(Y) + 2COV(X,Y) en voor twee onafhankelijke variabelen X en Y geldt dat V(X + Y) = V(X) + V(Y)

Stelling 9: voor twee afhankelijke variabelen X en Y geldt dat V(X – Y) = V(X) + V(Y) – 2COV(X,Y) en voor twee onafhankelijke variabelen X en Y geldt dat V(X – Y) = V(X) + V(Y)

6 Bijzondere verdelingen

6.1 De binomiale verdeling

Binomiale kansverdeling: $P(X = k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$

Binomiale verdeling: X ~ Binom(N,p)

Verwachtingswaarde: E(X) = Np

Variantie: V(X) = Np(1 – p)

Kan enkel gebruikt worden als:

- N vast is
- de kans op succes p ongewijzigd blijft

6.2 De normale verdeling

Dichtheidsfunctie: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Verwachtingswaarde: E(X) = μ

Variantie: V(X) = σ^2
Hoogste punt: in het gemiddelde

Bij grotere variantie: dichtheidsfunctie breder en minder hoog

Standaardnormale verdeling: normale verdeling met $\mu = 0$ en $\sigma^2 = 1$

Voor standaardnormale verdeling: $P(X > x) = P(X \leq -x)$

Verband tussen kansen: $P(X \leq -x) = 1 - P(X \leq x)$

Stelling 10: als X een normale verdeling heeft met gemiddelde μ en variantie σ^2, dus $X \sim N(\mu, \sigma^2)$, dan heeft de variabele $Z = \frac{X-\mu}{\sigma}$, een standaardnormale verdeling, dus $Z \sim N(0,1)$

Standaardiseren van X: $P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$

6.3 De χ^2-verdeling

χ^2-verdeling: verdeling van de variabele $Y = X_1^2 + X_2^2 + \cdots + X_k^2$

Parameter k: aantal vrijheidsgraden

Verwachtingswaarde: E(Y) = k

Variantie: V(Y) = 2k

6.4 De t-verdeling

t_k-verdeling: $T = \frac{X}{\sqrt{\frac{1}{k}Y}}$

Naarmate k toeneemt: t-verdeling gelijkt meer en meer op de dichtheid van een standaardnormale

Verwachtingswaarde: E(T) = 0

Variantie: V(T) = $\frac{k}{k-2}$, voor k > 2

De steekproevenverdeling

Reproduceerbaarheid: we verwachten gelijkaardige conclusies wanneer we het experiment opnieuw uitvoeren

1 Steekproeftrekking

Aselecte steekproeftrekking:
- op willekeurige wijze
- elementen onafhankelijk van elkaar

Kans: relatieve frequentie van de gebeurtenis indien we het experiment een oneindig aantal keer herhalen

Variabele X is toevalsvariabele:
- ze duidt het resultaat aan van een toevallige trekking van een element uit de populatie
- ze is veranderlijk omdat niet alle elementen in de populatie dezelfde waarde hebben

2 Steekproevenverdeling van het gemiddelde

Steekproefgemiddelde is variabel: de waarde hangt af van de frequentieverdeling van de scores in de steekproef EN verschillende steekproeven hebben verschillende fresuentieverdelingen → variabele

Steekproefgemiddelde: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Steekproefgrootheid / statistiek: bewerking toegepast op de variabelen $X_1, \ldots, X_n$

Steekproevenverdeling: verdeling van een steekproefgrootheid

Stelling 11: de verwachtingswaarde van het steekproefgemiddelde $\bar{X}$ is gelijk aan het populatiegemiddelde van de variabele X → $E(\bar{X}) = \mu_X$

Stelling 12: de variantie van het steekproefgemiddelde is gelijk aan de populatievariantie van de variabele gedeeld door de steekproefgrootte → $V(\bar{X}) = \frac{\sigma_X^2}{n}$

De wet van de grote aantallen: het steekproefgemiddelde zal met hoge waarschijnlijkheid weinig verschillen van het populatiegemiddelde indien de steekproef groot is

Stelling 13: Stel dat $X_1, \ldots, X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met een normale verdeling $N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$, dan zal $\bar{X}$ ook normaal verdeeld zijn → $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$

Stelling 14 / centrale limietstelling: stel dat $X_1, \ldots, X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met gemiddelde μ_X en variantie σ_X^2, dan wordt de verdeling van het steekproefgemiddelde $\bar{X}$ naarmate n groter wordt, steeds beter benaderd door de normale verdeling gemiddelde μ_X en variantie $\frac{\sigma_X^2}{n}$

Steekproefgemiddelde standaardiseren: $P(\bar{X} \leq x) = P\left(Z \leq \frac{x-\mu_X}{\sqrt{\frac{\sigma_X^2}{n}}}\right), Z \sim N(0,1)$

Indien X niet uit normale verdeling: enkel geldig voor grote n

3 Steekproevenverdeling van de variantie

Steekproefvariantie: $SN_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ en $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Verwachtingswaarde: $E(SN_X^2) = \frac{n-1}{n}\sigma_X^2$ en $E(S_X^2) = \sigma_X^2$

Stelling 15: stel dat X_1, …, X_n n onafhankelijke lukrake trekkingen zijn uit een populatie met normale verdeling $N(\mu_X, \sigma_X^Z)$, dan geldt $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2$

Betrouwbaarheidsintervallen en statistische toetsen voor het populatiegemiddelde

1 Schatters

Schatter voor populatieparameter θ: $\hat{\theta}$ met $\hat{\theta}$ een steekproefgrootheid

$\hat{\theta}$ is een goede schatter van θ indien:

- ze zuiver is: verwachtingswaarde van de schatter is gelijk aan de populatieparameter → $E(\hat{\theta}) = \theta$
- de variantie van de schatter kleiner wordt naarmate de steekproefgrootte toeneemt

Standaarddeviatie van de schatter: standaardfout → schatter met kleinste standaardfout is het efficiëntst

1.1 Het gemiddelde

Steekproefgemiddelde als schatter: $\hat{\theta} = \bar{X}$ als $\theta = \mu$

Uit stelling 11: steekproefgemiddelde is zuivere schatter → $E(\bar{X}) = \mu$

Uit stelling 12: $V(\bar{X}) = \frac{\sigma^2}{n}$, dus als n toeneemt, wordt de variantie kleiner

Standaardfout steekproefgemiddelde: $\frac{\sigma}{\sqrt{n}}$

Schatting: de waarde van een schatter op basis van één steekproef → $\bar{x}$

1.2 De variantie

Probleem met SN^2: geen zuivere schatter → gemiddelde varianties niet gelijk aan populatievariantie omdat $\frac{n-1}{n}\sigma^2 < \sigma^2$ en de populatieparameter dus systematisch te klein geschat wordt

Populatievariantie: $E(S_X^2) = \sigma^2$

2 Betrouwbaarheidsintervallen

2.1 X normaal verdeeld en gekende populatievariantie

z_α: de waarde van de standaardnormale verdeling zodat de oppervlakte onder de curve rechts van de waarde gelijk is aan α → $P(Z > z_\alpha) = \alpha$

Omdat standaardnormale symmetrisch is rond 0: $P(-z_{\alpha/2} \leq Z \leq -z_{\alpha/2}) = 1 - \alpha$ dus $P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq -z_{\alpha/2}\right) = 1 - \alpha$

Na herwerken: $P\left(\bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

Betrouwbaarheidsinterval: $\left[\bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right]$=(1-$\alpha$)100%betrouwbaarheidsinterval

Betrouwbaarheidsinterval is variabel: hangt af van het steekproefgemiddelde $\bar{x}$
Breedte interval: $b - a = \left(\bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 2 * z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$

Breedte hangt af van:
- steekproefgrootte: als de steekproef groter wordt, dan verkleint de intervalbreedte
- waarde $z_{\frac{\alpha}{2}}$: als α toeneemt zal $z_{\frac{\alpha}{2}}$ afnemen, zal 1 - α afnemen en zal breedte afnemen

- populatiestandaarddeviatie

2.2 X normaal verdeeld en ongekende populatievariantie

We kunnen σ niet zomaar vervangen door S_X: S_X is een variabele en σ is een constante

Eigenschap 1: als X normaal verdeeld is → $\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}$

Eigenschap 2: Als X normaal verdeeld is → $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Eigenschap 1 en 2 gecombineerd: $\frac{\bar{X}-\mu}{S_X/\sqrt{N}} \sim t_{n-1}$

$P\left(T > t_{n-1;\frac{\alpha}{2}}\right) = \frac{\alpha}{2}, T \sim t_{n-1}$

Verdeling is symmetrisch rond 0: $P\left(-t_{n-1;\frac{\alpha}{2}} \leq T \leq t_{n-1;\frac{\alpha}{2}}\right) = 1-\alpha$ dus $P\left(-t_{n-1;\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{S_X}{\sqrt{n}}} \leq t_{n-1;\frac{\alpha}{2}}\right) = 1-\alpha$

(1-α)100%betrouwbaarheidsinterval: $\left[\bar{X} - t_{n-1;\frac{\alpha}{2}} * \frac{S_X}{\sqrt{n}},\ \bar{X} + t_{n-1;\frac{\alpha}{2}} * \frac{S_X}{\sqrt{n}}\right]$

Verschillen met populatiestandaardvariatie:

- de t_{n-1}-verdeling heeft een grotere variantie dan de standaardnormale verdeling
- de $t_{n-1;\alpha/2}$-waarde van een t_{n-1}-verdeling is groter dan de $z_{\alpha/2}$-waarde van een standaardnormale verdeling

Naarmate n groter wordt: t_{n-1}-verdeling benadert standaardnormale verdeling steeds beter

2.3 X niet normaal verdeeld en ongekende populatievariantie

Bij grote steekproef: centrale limietstelling → garandeert dat $\left[\bar{X} - t_{n-1;\frac{\alpha}{2}} * \frac{S_X}{\sqrt{n}},\ \bar{X} + t_{n-1;\frac{\alpha}{2}} * \frac{S_X}{\sqrt{n}}\right]$ bij benadering een (1 - α)100% betrouwbaarheidsinterval voor het populatiegemiddelde

3 Statistische toetsen

Nulhypothese: H_0

Alternatieve hypothese: H_a

Statistische toets: H_0: $\mu = \mu_0$ en H_a: $\mu \neq \mu_0$ → trachten H_0 te verwerpen

3.1 Toetsingsgrootheid

Toetsingsgrootheid: $G = \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}}$

- aanname dat H_0 klopt
- volgt t_{n-1}-verdeling als H_0 correct is

- g = waarde van G op basis van één steekproef

Waarden dat G aanneemt:
- g rond 0 wanneer H_0 waar is
- g positief wanneer $\mu > \mu_0$
- g negatief wanneer $\mu < \mu_0$

3.2 Beslissingsregels

Conclusies i.v.m. toetsingsgrootheid:
- als H_0 waar is, verwachten we dat G waarden zal aannemen rond 0
- als H_0 niet waar is, verwachten we dat G waardan zal aannemen sterk verschillen van 0

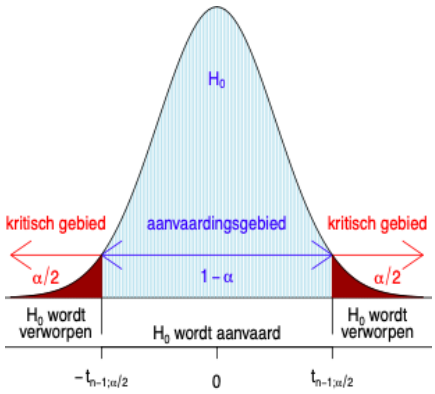
Regels op basis van g:
- als g rond 0 ligt, verwerpen we H_0 niet
- als g sterk verschilt van 0, verwerpen we H_0 en besluiten we H_a

Beslissingsregels:
- als $-t_{n-1;\alpha/2} \leq g \leq t_{n-1;\alpha/2}$ verwerpen we H_0 niet
- als $g < -t_{n-1;\alpha/2}$ of $g > t_{n-1;\alpha/2}$ verwerpen we H_0 en besluiten we H_a

Kritische waarden: $t_{n-1;\alpha/2}$ en $-t_{n-1;\alpha/2}$

Aanvaardingsgebied: gebied tussen kritische waarden

Kritisch gebied: gebied buiten twee kritische waarden



Beslissingsregels meer compact geschreven:
- als $|g| \leq t_{n-1;\alpha/2}$ verwerpen we H_0 niet
- als $|g| > t_{n-1;\alpha/2}$ verwerpen we H_0 en besluiten we H_a

3.3 Type I en type II fout

4 mogelijke scenario's:
- de nulhypothese is waar en we verwerpen H_0 niet
- de nulhypothese is waar en we verwerpen H_0 → type I
- de alternatieve hypothese is waar en we verwerpen H_0 niet → type II
- de alternatieve hypothese is waar en we verwerpen H_0

	In werkelijkheid is H_0 juist	In werkelijkheid is H_0 fout
We verwerpen H_0 niet	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 $(1-\beta)$

P(verwerp $H_0 \mid \mu = \mu_0$) = α → kans op type I-fout = significantieniveau

Kans op correcte conclusie: P(verwerp H_0 niet $\mid \mu = \mu_0$) = 1 - α = betrouwbaarheid

Opgelet: kans op type I fout is exact gelijk aan α als X uit een normale verdeling komt (anders centrale limietstelling)

P(verwerp H_0 niet | $\mu \neq \mu_0$) = β → kans op type II

Kans op correcte conclusie: P(verwerp H_0 | $\mu \neq \mu_0$) = 1 - β = onderscheidingskans / power

Kans op type II-fout hangt af van:
- significantieniveau: β stijgt als α daalt
- steekproefgrootte: β daalt als n stijgt

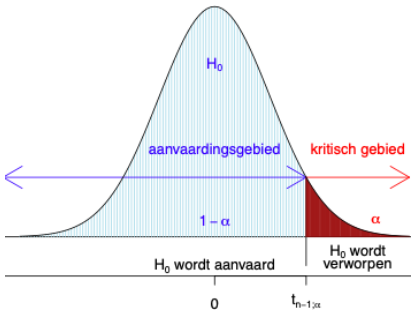
3.4 Beslissingsregels op basis van het betrouwbaardheidsinterval

Beslissingsregels:
- als μ_0 in het betrouwbaarheidsinterval ligt, verwerpen we H_0 niet
- als μ_0 niet in het betrouwbaarheidsinterval ligt, verwerpen we H_0 en besluiten we H_a

3.5 Eenzijdige en tweezijdige toetsen

3.5.1 Rechtszijdig

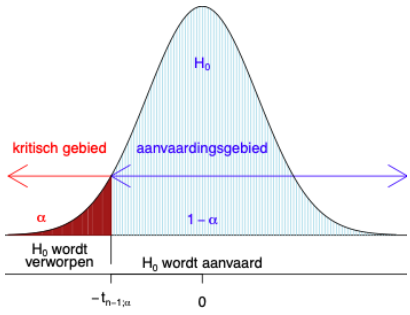
Uitvoeren eenzijdige toets: gelijkaardige start (berekenen toetsingsgrootheid), maar andere beslissingsregels



Beslissingsregels:
- als $g \leq t_{n-1;\alpha}$ verwerpen we H_0 niet
- als $g > t_{n-1;\alpha}$ verwerpen we H_0 en besluiten H_a

3.5.2 Linkszijdig

Beslissingsregels:
- als $g > t_{n-1;\alpha}$ verwerpen we H_0 niet
- als $g < t_{n-1;\alpha}$ verwerpen we H_0 en besluiten we H_a



3.5.3 Eenzijdig of tweezijdig toetsen?

In werkelijkheid is $\mu < \mu_0$:
- de tweezijdige toets zal een specifiek alternatief kunnen detecteren met een bepaalde power
- de linkszijdige toets zal een specifiek alternatief kunnen detecteren met een hogere power dan de tweezijdige toets
- de rechtszijdige toets heeft een power van maximaal α (zeer lage power)

In werkelijkheid is $\mu > \mu_0$:
- de tweezijdige toets zal een specifiek alternatief kunnen detecteren met een bepaalde power
- de linkszijdige toets heeft een power van maximaal α (zeer lage power)
- de rechtszijdige toets zal een specifiek alternatief kunnen detecteren met een hogere power dan de tweezijdige toets

Bij eenzijdige toetsen: mogelijkheid dat je toets quasi geen power zal hebben

3.6 p-waarde

$H_0 : \mu = \mu_0$		
Kies de alternatieve hypothese H_a		
Bepaal het significantieniveau α		
Bereken de toetsingsgrootheid g		
Besluit op basis van de gekozen H_a:		
Indien linkszijdig	Indien rechtszijdig	Indien tweezijdig
$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu \neq \mu_0$
Verwerp H_0 als	Verwerp H_0 als	Verwerp H_0 als
$g < -t_{n-1;\alpha}$	$g > t_{n-1;\alpha}$	$\lvert g \rvert > t_{n-1;\alpha/2}$

p-waarde: overschrijdingskans

Beslissingsregels:
- als p ≥ α verwerpen we H_0 niet
- als p < α verwerpen we H_0 en beslissen we H_a

Formele omschrijving: de p-waarde is de kans om een toetsingsgrootheid te observeren die minstens even extreem is als deze die waargenomen is, berekend in de veronderstelling dat de nulhypothese waar is

Minder abstract:
- de p-waarde is een kans → kan nooit kleiner dan 0 of groter dan 1 zijn
- de p-waarde wordt berekend in de veronderstelling dat H_0 waar is
- de p-waarde hangt af van de alternatieve hypothese

3.6.1 Linkszijdige alternatieve hypothese

Toetsingsgrootheid: neemt negatieve waarden aan

p-waarde: $P(G < g \mid \mu \neq \mu_0)$

3.6.2 Rechtszijdige alternatieve hypothese

Toetsingsgrootheid: neemt positieve waarden aan

p-waarde: $P(G > g \mid \mu \neq \mu_0)$

3.6.3 Tweezijdige alternatieve hypothese

Berekeken van p-waarde hangt af van het teken van g:
- als g > 0 dan is de p-waarde gelijk aan $2 * P(G > g \mid \mu = \mu_0)$
- als g < 0 dan is de p-waarde gelijk aan $2 * P(G < g \mid \mu = \mu_0)$

3.6.4 Interpretatie van de p-waarde

Hoe kleiner de p-waarde: hoe meer bewijskracht tegen de nulhypothese

3.7 Overzicht en opmerkingen

Overzicht te volgen stappen toetsingsprocedure:
- formuleer $H_0 : \mu = \mu_0$ en kies een alternatieve hypothese H_a
- leg het significantieniveau vast
- bereken de toetsingsgrootheid g

- formuleer een beslissing met behulp van
    - de kritieke waarden
    - de p-waarde
    - het betrouwbaarheidsinterval

Misvattingen rond de p-waarde:

- "De p-waarde is de kans dat H_0 waar is en 1 – p is de kans dat H_a waar is." → p-waarde is de overschrijdingskans die we bekomen op basis van de geobserveerde data op voorwaarde dat H_0 waar is
- "In het algemeen: hoe kleiner de p-waarde, hoe groter het verschil tussen μ en μ_0." → enkel indien steekproefgrootte en variabiliteit constant blijven
- "Een statistisch significant verschil tussen μ en μ_0 is voor de theorie of voor de praktijk ook significant." → klein verschil is theoretisch significant, maar vaak praktisch niet
- "Als ik geen significant verschil vind, is mijn onderzoek nutteloos." → het niet vinden van een significant verschil kan bijzonder informatief zijn op voorwaarde dat een gepast onderzoeksopzet werd gehanteerd wat vervolgens correct werd uitgevoerd en geanalyseerd