Statistiek

Visualiseren van data

De populatie en de steekproef

Steekproef -> afspiegeling van de populatie

- Moet representatief zijn
    - Aselecte steekproef
        - Mensen selecteren op willekeurige wijze en per toeval

Het IAT experiment

=Impliciete associatie test

- Gebruik gevoelsthermometer
    - Aangeven hoe warm of koud ze zich voelen tegenover iets

Congruent = overeenstemmend

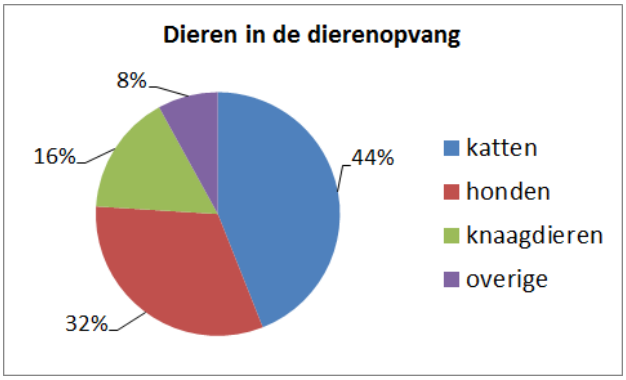
Incongruent = niet overeenstemmend

De data

- Nominaal
    - Identificatie zonder dat ze een hoeveelheid aanduiden
- Ordinaal
    - Waarden duiden een volgorde aan
- Ratio
    - Er is een 0-waarde : absoluut nulpunt
- Intervalschaal
    - Verschillen tussen waarden hebben een betekenis, maar er is geen absoluut nulpunt
- Continue variabelen
    - Kunnen tussenwaarden aannemen
- Discrete variabelen
    - Steeds twee waarden waar geen derde waarde kan tussen liggen, eindig aantal

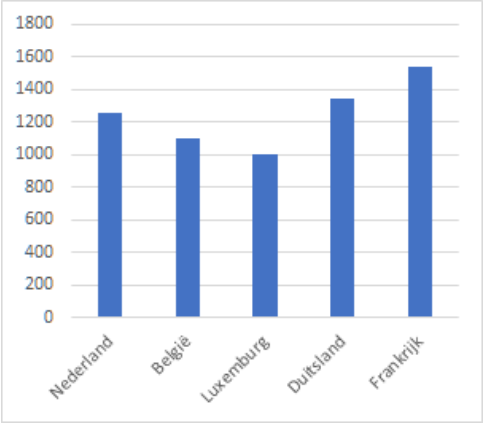
Cirkeldiagram

- **Absolute frequenties**
    - Het aantal keer dat een waarde voorkomt in de steekproef
- **Absolute frequentieverdeling**
    - Een tabel met twee rijen waar de 1^{e} rij de mogelijke waarden van X weergeeft en de 2^{e} rij de overeenkomstige absolute frequenties
- **Relatieve frequenties**
    - De absolute frequenties gedeeld door de steekproefgrootte
- **Steekproefgrootte (n)**
    - Gelijk aan het aantal elementen in de steekproef
- **Verdeling van een variabele**



- Het geheel van mogelijke waarden

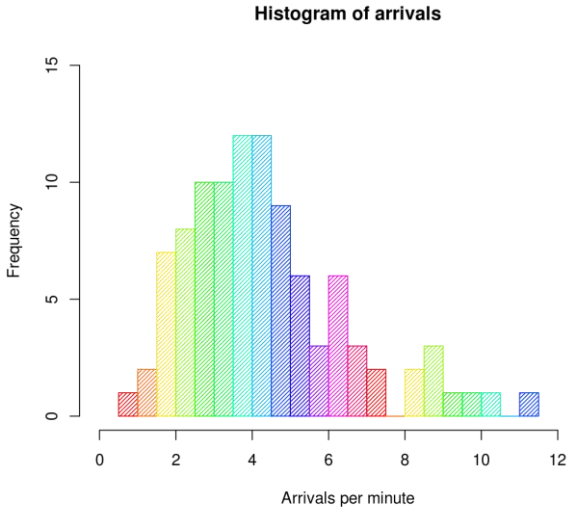
Staafdiagram



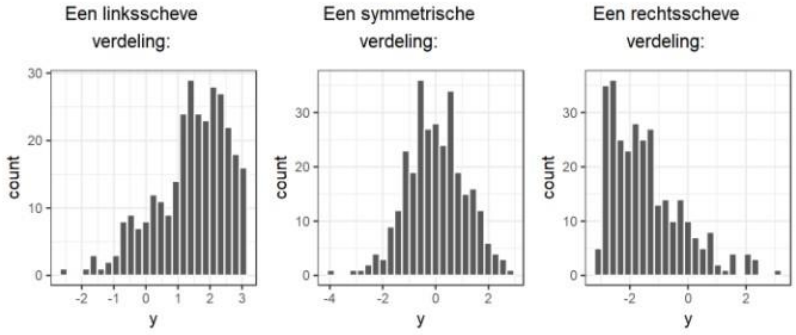
- Gebruikt bij nominale of ordinale meetniveaus
- Op basis van absolute frequenties

Histogram

- Groeperen van data
- De klassenbreedtes van de intervallen }a,b} zijn gelijk
- **Klassenbreedtes**
  - Ondergrens bovengrens
- **Gegroepeerde frequentieverdeling**
  - Een tabel met twee kolommen waar de eerste kolom de klassen van X weergeeft en de tweede de overeenkomstige frequenties
- <u>Bij een histogram raken de rechthoeken elkaar en kunnen de breedtes van de rechthoeken verschillen</u>
  - Klassenbreedte x hoogte = relatieve frequentie
- <u>Voor interval en ratioschaal variabelen</u>



Verdeling van histogrammen



Cumulatieve frequentiecurve

Ongegroepeerde data

- **Cumulatieve absolute frequentie**
  - Is gelijk aan het aantal elementen in de steekproef die kleiner dan of gelijk aan x zijn. We duiden dit aan met het symbool F(x)
- **Cumulatieve absolute frequentieverdeling**
  - Een tabel met 2 kolommen waar de 1e kolom de waarden van de variabele X worden weergegeven en in de 2e kolom de overeenkomstige cumulatieve absolute frequenties

cijfer	freq.	cum. freq.	rel. cum. freq.
4	1	1	5%
5	5	6	30%
6	8	14	70%
7	4	18	90%
8	2	20	100%

Gegroepeerde data

- **Cumulatieve absolute frequentie van een klasse**
  - Gelijk aan het aantal elementen in die klasse plus het aantal elementen in lagere klassen
- **Cumulatieve absolute gegroepeerde frequentieverdeling**
  - Een tabel met twee kolommen waar de 1e kolom de klassen van X weergeven en de 2e kolom de overeenkomstige cumulatieve absolute frequenties

Samenvatten van data

Het gemiddelde

- Alle waarden van een variabele op tellen en delen door de steekproefgrootte

Gemiddelde en gemiddelde gegroepeerde data

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{p} f_i \frac{(a_i + b_i)}{2}$$

- Is alleen zinnig voor interval en ratiovariabelen

Het gemiddelde voor gegroepeerde data:

- We maken gebruik van het klassenmidden

De mediaan

- mdx
- de middelste waarde nadat we de waarden van een variabele van klein naar groot hebben geordend
  - in een histogram het oppervlakte delen door twee
- enkel zinnig voor een ordinale, interval en ratiovariabele

De modus

- Symbool mo
    - Klasse of waarde met de grootste frequentie
- Indien een verdeling twee modi heeft = bimodaal
- Enkel zinnig voor nominale, ordinale, interval en ratiovariabelen

Gevoeligheid aan outliers

- Zijn waarden die ver verwijderd zijn van de overige waarden van een variabele
    - Is wel soms subjectief
- Outliers kunnen bepaalde centrummaten sterk beïnvloeden

Spreidingsmaten

Variatiebreedte v_x

Afstand tussen de grootste en de kleinste waarde

- **Variatiebreedte**
    - De grootste – de kleinste waarde
    - Bovengrens van de laatste klasse – ondergrens van de eerste klasse
- Enkel zinnig voor interval en ratiovariabelen

De gemiddelde absolute afwijking

- Som =0
    - Geen afwijking

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

- Voor interval- en ratiovariabelen

De variantie en de standaarddeviatie

De variantie van de variabele X in een steekproef wordt gegeven door:

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

De variantie op basis van een frequentieverdeling wordt gegeven door:

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

De standaarddeviatie van een variabele X in een steekproef wordt gegeven door:

$$sn_x = \sqrt{sn_x^2} \text{ of } s_x = \sqrt{s_x^2}$$

- Voor interval- en ratiovariabelen

De interkwartielafstand

- Op basis van percentielen
    - P50 = de mediaan

        - Alles van 50% of lager
    - P30
        - Alles van 30% of lager
    - …
- De interkwartielafstand (Q)
    - Is gelijk aan P75 – P25
        - Het 3e kwartiel – het 1e kwartiel
- Voor ordinale, interval- en ratiovariabelen

De spreidingsmaat 'd'

- p: aantal unieke waarden
- fmo: frequentie van de modus
- n: steekproefgrootte

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

- geen spreiding -> d=0

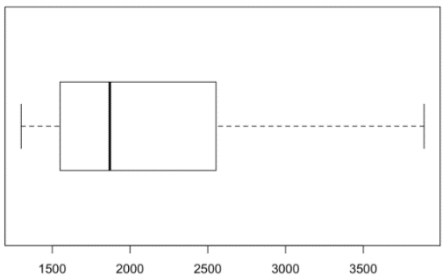
Gevoeligheid aan outliers

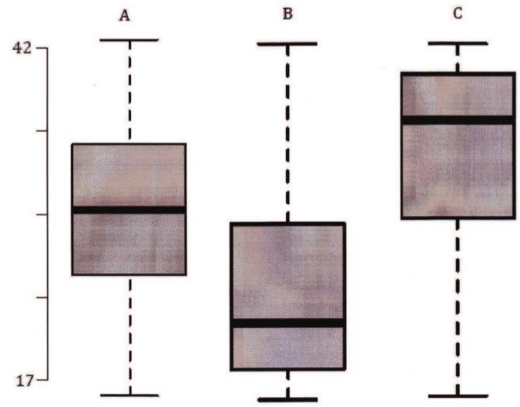
- spreidingsmaat d is niet gevoelig aan outliers
    - omdat het afhangt van de aantal unieke waarden en de steekproefgrootte

Boxplot

- het bevat visueel veer informatie:
    - de mediaan
    - de interkwartielafstand
    - de outliers

1. alle gegevens noteren met een stip
2. outliers inkleuren
3. streep bij grootste en kleinste waarde dat geen outlier is
4. bredere streep bij 1e en 3e kwartier
5. de box tekenen
6. alle niet-ingekleurde stippen wegdoen
7. stippellijn tekenen van boven naar onderen
8. streep bij de mediaan zetten
9. de box naar rechts draaien





Symmetrisch rechts links

Samenhang tussen 2 variabelen

Bivariate frequentieverdeling

- **univariate frequentieverdelingen**
  - over een variabele afzonderlijk
- **bivariate frequentieverdelingen**
  - twee variabelen gezamenlijk bestuderen
- **marginale verdeling**
  - de univariate verdeling bepalen via de bivariate verdeling
- de bivariate frequentieverdeling:
  - bevat meer info dan univariate
  - inzicht in samenhang tussen twee variabelen
  - conclusies kunnen wijzigen door data te hergroeperen -> subjectiviteit

Opleidingsjaar * Hoe vaak flirt je? Crosstabulation

			Hoe vaak flirt je?					
			Nooit	Zelden	Af en toe	Vaak	Heel vaak	Total
Opleidingsjaar	Eerstejaars	Count	93	222	931	1573	715	3534
	Tweedejaars	Count	307	2168	5779	2323	199	10776
	Derdejaars	Count	11811	20464	9019	703	66	42063
Total		Count	12211	22854	15729	4599	980	56373

I-> marginale verdeling

Spreidingsdiagram

- **perfecte positieve samenhang**
  - de punten gaan van linksonder tot rechtsboven en liggen op een rechte
- **perfecte negatieve samenhang**
  - de punten gaan van linksboven naar rechtsonder en liggen op een rechte
- **geen samenhang**
  - puntenwolk
- het interpreteren van een spreidingsdiagram is subjectief

Maten van samenhang

De covariantie

- **cov is groter dan 0** positieve samenhang
- **cov is kleiner dan 0** negatieve samenhang
- **cov is 0** geen samenhang
    - de covariantie geeft ons echter geen info over hoe sterk de samenhang is
        - daarvoor gebruiken we de correlatiecoëfficiënt

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

De correlatiecoëfficiënt

- rxy is 1 perfecte positieve samenhang
- rxy is -1 perfecte negatieve samenhang
- rxy is 0 geen samenhang
    - hoe dichter bij het ene, hoe sterker de samenhang

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

kendall's T

- een paar is **concordant** als
    - de waarde groter is dan 0 / het verband stijgend is
- een paar is **discordant** als
    - de waarde kleiner is dan 0 / het verband dalend is
        - tip -> teken de waarden in een grafiek en verbind

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

Lineaire en niet-lineaire verbanden

- correlatie en covariatie
    - lineaire samenhang
- Kendall's T
    - Monotone samenhang
        - Functie die ofwel alleen stijgt of alleen daalt
- Lineaire functie:
    - Functie die kan voorgesteld worden door een rechte lijn

Gevoeligheid aan outliers

- Covariantie en correlatie
    - Gevoelig aan outliers
- Kendall's T
    - Niet gevoelig aan outliers

De regressielijn

- De correlatie visualiseren op een spreidingsdiagram

B1: de regressiecoëfficiënt (helling van de rechte)

B0: het intercept (het snijpunt met de verticale as)

$$Y = b_0 + b_1 X$$

$$b_1 = r_{XY}\frac{s_Y}{s_X} \text{ en } b_0 = \overline{y} - b_1\overline{x}.$$

- Hoe teken je een regressielijn?
  - 2 willekeurige waarden van X nemen
  - Formule van regressielijn invullen
  - Punten tekenen
  - Punten verbinden

Samenhang en causaliteit

- Er kan samenhang zijn, maar is niet noodzakelijk causaal verband
  - Mogelijk door derde variabele

De populatie en verdelingsfuncties

Verdelingsfuncties van discrete variabelen

- Verdelingsfunctie
  - Voor een populatie en niet voor een steekproef
- Discrete variabalen
  - Kunnen maar een eindig aantal waarden aannemen
- **Kans dat een variabele X de waarde xi aanneemt**

$P(X = x_i) = \lim_{n\to\infty} \frac{f_i}{n}$ relatieve frequentie van xi in de populatie

p= aantal waarden dat een variabele kan aannemen

De kansverdeling

- Een tabel met twee kolommen waarbij de eerste kolom de waarden xi weergeeft en de tweede kolom de overeenkomstige kansen P(X=xi)
- Kans ligt in interval [0,1]

De cumulatieve verdelingsfunctie

- De kans dat de waarde van een variabele X kleiner dan of gelijk is aan x
  - (door de waarden steeds op te tellen)

Verdelingsfuncties van continue variabelen

- Continue variabelen
  - Kan oneindig veel verschillende waarden aannemen
    - -> P(X=x) = 0

De cumulatieve verdelingsfunctie

- De kans dat de waarde van een variabele X kleiner of gelijk is aan x
    - Bij continue variabelen maakt het niet uit of we < of ≤ gebruiken omdat de kans altijd 0 is

De dichtheidsfunctie

- fx geeft de kans weer dat X valt binnen het interval (x, x+b) gedeeld door b, waar b de breedte van het interval voorstelt en naar nul convergeert. Omdat we delen door b is het eigenlijk geen kans

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

- de kans visualiseren door:

$$P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F_X(x_2) - F_X(x_1)$$

➔ x<X = je zoekt een kans groter dan -> 1 – kans uit R
➔ x>X = je zoekt een kans kleiner dan -> kans uit R aflezen

- **drie interessante eigenschappen:**
    - dichtheidsfunctie is een positieve functie
    - de volledige oppervlakte van de dichtheidsfunctie is gelijk aan 1
    - er geldt dat : P(X > x) = 1 – P(X ≤ x)

populatieparameters

populatiegemiddelde

E= verwachtingswaarde

E(X) / μx / μ = populatiegemiddelde

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

Populatievariantie

V(X) / $\sigma^2 x$ / σ^2 = populatievariantie

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

- de standaarddeviatie (σ)

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

Bivariate kansverdeling

Discrete variabelen

- de univariate (marginale) verdelingen
    - p is het aantal mogelijke waarden dat X kan aannemen
    - q is het aantal mogelijke waarden dat Y kan aannemen

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

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

- **statistische onafhankelijkheid is een belangrijk begrip binnen bivariate kansverdelingen**
    - twee discrete variabelen X en Y zijn onafhankelijk als de gelijkheid

$P(X = xi\ en\ Y = yj) = P(X = xi\ )P(Y = yj)$

    - geldt voor alle mogelijke combinaties van i en j
- **we kunnen ook de covariantie en de correlatie voor de populatie berekenen**
    - covariantie:

$COV(X, Y) = \sum \sum P(X = xi\ en\ Y = yj)(xi - E(X))(yj - E(Y))$

    - correlatiecoëfficiënt

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

Nuttige stellingen

- als X en Y onafhankelijke variabelen zijn, dan geldt dat: 1
    - COV(X,Y) = 0
- Voor een variabele Y = X = a geldt dat: 2
    - E(Y) = E(X) + a (a is een constante)
- Voor een variabele Y = aX geldt dat: 3
    - E(Y)= aE(X) (a is een constante)
- Voor twee variabelen X en Y (die onafhankelijk of afhankelijk kunnen zijn) geldt dat: 4
    - E(X + Y) = E(X) + E(Y)
    - E(X – Y) = E(X) – E(Y)
- Voor twee onafhankelijke variabelen X en Y geldt dat: 5
    - E(XY) = E(X)E(Y)
- Voor een variabele Y = X + a geldt dat: 6
    - V(Y) = V(X) (a is een constante)
- Voor twee afhankelijke variabelen X en Y geldt dat: 7
    - V(X + Y) = V(X) + V(Y) + COV(X,Y)
- Voor twee afhankelijke variabelen X en Y geldt dat: 8
    - V(X+Y) = V(X) + V(Y) + 2COV(X,Y)
- Voor twee onafhankelijke variabelen X en Y geldt dat: 8
    - V(X+Y) = V(X) + V(Y)
- Voor twee variabelen X en Y geldt dat: 9
    - V(X -Y) = V(X) + V(Y) – 2COV(X,Y)

Bijzondere verdelingen

De binominale verdeling

- k: aantal successen
- N: maximum aantal successen
- p: kans op succes
    - P(X = 0) = 9/16
    - P(X = 1) = 6/16
    - P(X = 2) = 1/16
- **De binominale kansverdeling**

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

- N! staat voor N faculteit

- De verwachtingswaarde van een binominale variabele X

E(X) = Np

- De variantie van een binominale variabele X

V(X) = Np(1 – p)

- Kan enkel gebruikt worden als N vast is en de kans op succes p ongewijzigd blijft

De normale verdeling

E(X) = μ

V(X) = σ^2

- Een normaal verdeelde variabele is continu en de dichtheidsfunctie wordt gegeven door:

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

- Algemeen geldt voor de standaardnormale verdeling dat:

$P(X > x) = P(X \leq -x)$

- Het standaardiseren:

$$Z = \frac{X-\mu}{\sigma},$$ standaardiseren met een steekproefgemiddelde: $$Z \leq \frac{x - \mu_X}{\sqrt{\sigma_X^2/n}}$$

- Hierbij hoort de volgende stelling: 10
  - Als X een normale verdeling heeft met een gemiddelde μ en een variantie σ^2, dus X˜N(μ,σ^2), dan heeft de variabele

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

  - Een standaardnormale verdeling Z˜N (0,1)

De x2 verdeling

- Altijd positief
- Totale kans = 1 (100%)
  - k= aantal vrijheidsgraden = E(Y) = populatiegemiddelde
    - 2k = V(Y)

De t-verdeling

$$T = \frac{X}{\sqrt{\frac{1}{k}Y}}$$

- Totale kans = 1 (100%)
- Als k oneindig is = standaardnormale
- E(T) = 0 (altijd!!)
- V(T) = k / k – 2
  - Als nk groter is dan 2

De steekproevenverdeling

De steekproeftrekking

- Aselecte steekproef
  - Volledig willekeurig
  - Elementen zijn onafhankelijk van elkaar
- Een variabele X die bekomen wordt door op toevallige wijze een element uit de populatie te trekken, wordt ook een toetsvariabele genoemd omdat:
  - Ze het resultaat aanduidt van een toevallige trekken van een element uit de populatie
  - Ze veranderlijk (variabelen) is omdat niet alle elementen in de populatie dezelfde waarde hebben

Steekproevenverdeling van het gemiddelde

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

- Het steekproefgemiddelde is een variabele
  - Het is een bewerking toegepast op de variabelen X1, X2, …., Xn
- **Hierbij horen de volgende stellingen:**
  - De verwachtingswaarde (E) van het steekproefgemiddelde $\bar{X}$ is gelijk aan het populatiegemiddelde van de variabele X (stelling 11)
    - E($\bar{X}$)= μx
  - De variantie van het steekproefgemiddelde is gelijk aan de populatievariantie van de variabele gedeeld door de steekproefgrootte (stelling 12)
    - V($\bar{X}$) = σ^2x/ n
- **De wet van de grote aantallen**
  - Stelt dat het steekproefgemiddelde met grote waarschijnlijkheid weinig zal verschillen van het populatiegemiddelde indien de steekproef oneindig / groot is
- **De verdelingsfunctie berekenen aan de hand van deze stellingen:**
  - Stel dat X1, X2, …, Xn n onafhankelijke lukrake trekkingen zijn uit een populatie met een normale verdeling N, dan zal $\bar{X}$ ook normaal verdeeld zijn (stelling 13)
    - $\bar{X}$~N (μx, σ^2x / n)
  - Stel dat X1, X2, …., Xn n onafhankelijke lukrake trekkingen zijn uit een populatie met gemiddelde μx en een variantie σ^2x, dan wordt de verdeling van het steekproefgemiddelde $\bar{X}$ naarmate n groter wordt, steeds beter benaderd door de normale verdeling met een gemiddelde μx en variantie σ^2x
    - Centrale limietstelling (stelling 14)
- **Er zijn twee mogelijkheden om de kans van een populatie te berekenen**
  - Het experiment vele malen herhalen
  - Het experiment maar 1 keer uitvoeren en het standaardiseren

Steekproevenverdeling van de variantie

- Steekproefvariantie

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

En

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

- De verwachtingswaarde van de variantie

$$E(SN_X^2) = \frac{n-1}{n}\sigma_X^2,$$

  - Deze is niet gelijk aan de populatievariantie
    - Voor S2x wel

$$E(S_X^2) = \sigma_X^2.$$

- Hierbij hoort de stelling (stelling 15)
  - Stel dat X1, X2, …, Xn n onafhankelijke lukrake trekkingen zijn uit een populatie met normale verdeling N, dan geldt:
    - $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2.$

Betrouwbaarheidsintervallen en statistische toetsen voor het populatiegemiddelde

Schatters

- Schatter voor populatieparameter θ noteren we als $\hat{\theta}$, waarbij $\hat{\theta}$ een steekproefgrootheid is
  - $\hat{\theta}$ is een goede schatter van θ indien:
    - Ze zuiver is : $E(\hat{\theta}) = \theta$
    - De variantie van de schatter V($\hat{\theta}$) kleiner wordt naarmate de steekproefgrootte toeneemt
      - Zal meer nauwkeurig zijn

Gemiddelde

- $\hat{\theta} = \bar{X}$ als θ = μ
  - E($\bar{X}$) = μ (stemming 11)
  - V($\bar{X}$) = σ^2 / n
    - Als n toeneemt -> σ^2 / n wordt kleiner
    - Hoe groter de steekproef, hoe nauwkeuriger we het populatiegemiddelde kunnen schatten via het steekproefgemiddelde
- Standaardfout van het steekproefgemiddelde => σ / √ n

De variantie

- Populatievarianie ($\theta = \sigma^2$)

Zuivere schatter:

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

- $E(SN^2x)$ is geen zuivere schatter

X normaal verdeeld en gekende populatievariantie

- $z\alpha$ = de waarde van de standaardnormale verdeling
  - zodat de oppervlakte onder de curve rechts van de waarde gelijk is aan α
    - $P(Z > z\alpha) = \alpha$
- Is volgens standaardnormale dus:
  - $$P\left(\bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$
- **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**
  - **Is variabel:**
    - <u>Hangt af van steekproefgemiddelde en breedte interval</u>
      - *Breedte hangt af van:*
        - steekproefgrootte: als de steekproef groter wordt, dan verkleint de intervalbreedte
        - waarde $z\alpha$ / 2 : als α toeneemt zal $z\alpha$ / 2 afnemen, zal 1 - α afnemen en zal breedte afnemen
        - populatiestandaarddeviatie

werkproces:

- we kunnen $z\alpha$ berekenen via R:
  - doen we altijd met tweezijdige toetsing : $z\alpha/2$
    - stel alfa is 5% -> 0,05 0,05/2 = z0,025 = 1,96 (zie je in R)
- nu gaan we standaardiseren:
  - $$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
  - => deze uitdrukking herschrijven zodat enkel μ overblijft in het midden
- $$P(-z_{\alpha/2}\sigma/\sqrt{n} \leq \bar{X} - \mu \leq z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha.$$
- We draaien de uitdrukking om om de negatieven te verwerken
  - $$P(\bar{X} + z_{\alpha/2}\sigma/\sqrt{n} \geq \mu \geq \bar{X} - z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha.$$
- We draaien de uitdrukking terug om
  - $$P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha.$$
- Als laatste stap noteren we het interval (μ wegwerken)
  - $$[\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}]$$
    - De kans dat het populatiegemiddelde in het interval ligt is gelijk aan 1-α

X is normaal verdeeld en ongekende populatievariantie

- We houden rekening met eigenschap 1 en 2

Eigenschap 1. Als X normaal verdeeld is, dan:

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Eigenschap 2. Als X normaal verdeeld is, dan volgt:

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

  -
  - Dit betekend dat we nu gebruik maken van een t-verdeling
    - $\frac{\bar{X}-\mu}{S_X/\sqrt{n}} \sim t_{n-1}.$
- We gebruiken dus de R code : qt

X is niet normaal verdeeld en ongekende populatievariant

- We maken gebruik van de centrale limietstelling
  - Deze garandeert dat $[\bar{X} - t_{n-1;\alpha/2}S_X/\sqrt{n}, \bar{X} + t_{n-1;\alpha/2}S_X/\sqrt{n}]$, bij benadering een 100% BI is

Statistische toetsen

- H_0 : de nulhypothese
- H_a : de alternatieve hypothese
  - Ofwel is H_0 waar ofwel is H_0 niet waar
- $H_0 = \mu$ en Ha = μ > / < x

De toetsingsgrootheid berekenen

- $G = \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}}.$
  - Volgt een t-verdeling => $G \overset{H_0}{\sim} t_{n-1}.$

Beslissingsregels

- H_0 is waar => toetsingsgrootheid ligt rond 0
  - H_0 niet verwerpen
- H_0 is niet waar => toetsingsgrootheid ligt ver van 0
  - H_0 verwerpen

Type I en Type II fouten

4 scenario's ;

- H_0 is waar -> we verwerpen H_0 niet
  - Juist -> betrouwbaarheid
- H_0 is waar -> we verwerpen H_0
  - Fout -> type I fout
- H_0 is niet waar -> we verwerpen H_0 niet
  - Fout -> type II fout
- H_0 is niet waar -> we verwerpen H_0
  - Juist -> onderscheidingsvermogen

Eenzijdig en tweezijdig toetsen

- $H_a : \mu \neq \mu_0$
  - Tweezijdig
- $H_a : \mu < \mu_0$
  - Linkszijdig
- $H_a : \mu > \mu_0$
  - Rechtszijdig

De p-waarde

- *Of de overschrijdingskans*
- Beslissingsregels:
  - $p > \alpha$
    - we verwerpen H_0 niet
  - $p > \alpha$
    - we verwerpen H_0 en besluiten H_a

hoe werkt de toetsingsprocedure?

STAP 0: -gegevens noteren

-is variabele normaal verdeeld ? (n>/= 30)

STAP 1: $H_0 = \mu$ en $Ha = \mu > / < x$

Dit bepaald de richting

STAP 2: α bepalen

STAP 3: bereken de toetsingsgrootheid -> $G = \frac{\bar{X} - \mu_0}{S_X/\sqrt{n}}$.

STAP 4: <u>beslissing nemen; (3 opties)</u>

- *kritieke waarde:* => volgens $t_{n-1,\,\alpha}$

  ⇨ Rcode qt(1- α,n-1)

  = _____!!

  ⇨ Ligt g in deze kritieke waarde?
  - Nee : H_0 verwerpen

- *p-waarde (=overschrijdingskans)*
  - ⇨ H_0 verwerpen indien p< α
  - ⇨ H_0 niet verwerpen indien p> α
  - ⇨ Welke toets? (welke kant van de grafiek)
  - ⇨ Pt(x,n-1)
    =_____!!!
  - ⇨ Is de p-waarde kleiner dan α?
    - H_0 verwerpen
- *Betrouwbaarheidsinterval*
  - ⇨ Alleen bij 2-zijdige toets
  - ⇨ Ligt μ_0 in het interval?
    - Niet verwerpen
  - ⇨ Ligt μ_0 niet in het interval?
    - H_0 verwerpen

Misvattingen rond de p-waarde

- **"De p-waarde is de kans dat H0 waar is en 1−p is de kans dat Ha waar is."**
  - Fout: p-waarde is de overschrijdingskans die we bekomen op basis van de g-waarde.
- **"Hoe kleiner de p-waarde, hoe groter het verschil tussen μ en μ0"**
  - Fout: Dit geldt enkel indien de steekproefgrootte en variabiliteit constant blijven. Echter met een grote steekproef en weinig variabiliteit kan zelfs een klein verschil een kleine p-waarde opleveren.
- **"Een statistisch significant verschil tussen μ en μ0 is voor de theorie of voor de praktijk ook significant."**
  - Niet noodzakelijk: zelf een klein verschil is significant, daarom niet van praktische waarde.
- **"Als ik geen significant verschil vind, is mijn onderzoek nutteloos."**
  - Fout: Dit onderzoek kan heel informatief zijn als het op de juiste manier uitgevoerd werd.