Statistiek I

Hoofdstuk 1: Inleiding

1. Enkele misvattingen

1.1. "Met statistiek kan je alles bewijzen"

De uitspraak "met statistiek kan je alles bewijzen" is volkomen incorrect. Echter, door statistische analyses verkeerdelijk toe te passen, kan je wel de impressie wekken dat je kan aantonen wat je wil en dit kan **nefaste gevolgen hebben**.

1.2. "Statistiek is nutteloos voor de gedragswetenschappen"

Statistiek vormt een belangrijkste schakel in de totstandkoming en het begrijpen van vele inzichten binnen de gedragswetenschappen.

Enkele voorbeelden:

- **Het visueel geheugen onderzoeken** (bv. Benton Visual Retention Test).
- **Impliciete voorkeuren bepalen**:
  De Impliciete Associatie Test (IAT) is een reactietaak die gebruikt wordt om impliciete voorkeuren te meten.
- **Intelligentie en hersengrootte bestuderen**.

1.3. "Statistiek is enkel wiskunde"

Volgende aspecten komen aan bod in de cursus:

- **Wiskunde**: vooral algebra en kansrekening.
- **Software**: R en Rstudio.
- **Interpretatie en besluitvorming**: het beantwoorden van een onderzoeksvraag.

2. De betekenis van statistiek

Statistiek = de **wetenschap** van het **leren uit data** en van het **meten, controleren en communiceren van onzekerheid**.

2.1. Een voorbeeld rond intelligentie

2.2. Enkele definities

Populatie = de **volledige verzameling van objecten of personen** waarover informatie wordt gewenst.

Elementen = de **individuele leden van de populatie** (de objecten of personen).

Steekproef = een **deelverzameling van de populatie** die feitelijk zal onderzocht worden om informatie te bekomen. Vaak is het onmogelijk om de volledige populatie te bestuderen, vandaar dat men dan een steekproef uit de populatie zal nemen voor verder onderzoek.

Variabele = een **eigenschap die bij de elementen** van de populatie of steekproef **varieert**. Vaak worden er bij een steekproef verschillende variabelen gemeten.

Data = de **verzameling van gegevens** die wordt bekomen door de variabelen te meten.

Verdeling = geeft aan **welke waarden worden aangenomen en hoe vaak**.

Inductie = **uitgaande van het bijzondere het algemene besluiten**. Bij inductie proberen we op basis van een aantal waarnemingen tot een algemeen besluit te komen.

3. Eigenschappen van variabelen

3.1. Schaalfamilies

- **Nominale schaal**:
  De waarden van de variabele worden gebruikt voor **identificatie zonder dat ze een hoeveelheid aanduiden**.
  Geen numerieke betekenis, of zelfs geen getallen.
  Woorden kunnen wel **numeriek gecodeerd** worden.
  Voorbeelden:
  - **Geslacht**: man, vrouw of andere.
  - **Land van herkomst**: België, Frankrijk, Spanje, …

- **Ordinale schaal**:
  Erft alle eigenschappen van de nominale schaal.
  + de waarden **duiden een volgorde aan**.
  Woorden kunnen **numeriek gecodeerd** worden (volgorde behouden).
  Voorbeelden:
  - **Uitslag wedstrijd**: goud, zilver of brons.
  - **Mate van instemming**: volledig mee oneens, mee oneens, neutraal, mee eens, volledig mee eens.

- **Interval schaal**:
  Erft alle eigenschappen van de ordinale schaal.
  + **verschillen tussen waarden hebben een betekenis**.
  Geen absoluut nulpunt.
  Voorbeelden:
  - **Temperatuur in graden Celsius**: 0, 10, -30, …
  - **IQ**: 96, 100, 130, …

- **Ratio schaal**:
  Erft alle eigenschappen van de interval schaal.
  + heeft **absoluut nulpunt**.
  Verhoudingen (ratio's) hebben een betekenis (10 is dubbel van 5).
  Voorbeelden:
  - **Lengte in cm**: 0, 1, 354, …
  - **Geldberg in euro**: 0, 5, 2400, …

3.2. Discrete en continue variabelen

- **Continue variabelen**:
  Kunnen **tussenwaarden** aannemen.
  Voorbeelden:
  - Lengte in cm: tussen 2 en 3 cm ligt nog 2.5 cm, 2.35 cm, …
  - Temperatuur in °C: tussen 25.8°C en 25.9°C ligt 25.82147°C, …

- **Discrete variabelen**:
  Bestaan steeds uit **2 waarden**.
  Kan maar **eindig aantal waarden** aannemen.
  Voorbeelden:
  - Aantal kinderen: tussen 0 en 1 kind ligt geen derde waarde.
  - Aantal keer dat 'munt' wordt geworpen bij 4 worpen.

- **Bijna-continue variabelen**:
  Variabelen die **zeer veel verschillende waarden** kunnen aannemen.

DEEL 1: Beschriivende statistiek

Hoofdstuk 2: Visualiseren van data

1. Onderzoek naar raciale voorkeur
1.1. De onderzoeksvraag
1.2. De populatie en de steekproef
1.3. Het IAT-experiment
1.4. De data
2. Cirkeldiagram

Een cirkeldiagram is een grafische voorstelling die voornamelijk gebruikt wordt voor variabelen van **nominaal meetniveau**.

Verdere notatie:

- **Variabele**: hoofdletter, vaak X.
- **Waarden van variabele**: kleine letter met cijfer als subscript, xn (bv. x1).
- **Aantal elementen in steekproef**: kleine letter n.

Definities:

- **Absolute frequentie van x** = het **aantal keer** dat de **waarde** x in de steekproef **voorkomt**.
- **Absolute frequentieverdeling van X** = een **tabel** met 2 rijen waar de 1ste rij de **mogelijke waarden van X** weergeeft en de 2de rij **de overeenkomstige absolute frequenties**. Het kan i.p.v. 2 rijen ook 2 kolommen zijn.
- **Steekproefgrootte (n)** = het **aantal elementen** in de **steekproef**.
- **Relatieve frequentie van x** = **absolute frequentie gedeeld** door de **steekproefgrootte n.**

De **som van de relatieve frequenties** moet 1 zijn!

Op basis van de relatieve frequenties kunnen we **data visualiseren d.m.v. een cirkeldiagram**. De relatieve oppervlaktes van de stukken zijn gelijk aan de relatieve frequenties. Cirkeldiagrammen worden echter afgeraden om te gebruiken.

3. Staafdiagram

- De hoogte is gelijk aan de **relatieve frequentie**.
- De breedte kan **vrij gekozen** worden, zolang ze maar even breed zijn.
- **De afstand** tussen de verschillende rechthoeken moet ook dezelfde zijn.
- Voornamelijk voor variabelen van **nominaal of ordinaal meetniveau**.
- Staafdiagram op basis van **absolute frequenties is ook mogelijk**.

4. Histogram

- **Eerst data groeperen**:
  Onderverdelen in **klassen of intervallen**.
  Onderverdelen is **subjectief**.
- **De klassenbreedte vastleggen**:
  De klassenbreedtes van de intervallen]a, b], [a, b], [a, b[en]a, b[**zijn gelijk**.
  Verschillende vuistregels om het aantal klassen te bepalen: data indelen in ongeveer $\sqrt{n}$ klassen.
- De **waarden van de variabele** liggen op de horizontale as.

- De **breedte van de rechthoek** is gelijk aan de breedte van de klasse.
- **De hoogte** is gelijk aan de relatieve frequentie gedeeld door de breedte van de klasse, zodat de oppervlakte gelijk is aan de relatieve frequentie.
- Indien alle klassen dezelfde breedte hebben, is het ook mogelijk om een histogram op te stellen waar **de hoogte gelijk is aan de absolute frequentie**.

Notatie:
- **]a, b]**: alle leeftijden groter dan a (a niet meegerekend), maar kleiner dan of gelijk aan b (b wel meegerekend), waarbij a en b getallen voorstellen.

Definitie:
- **Klassenbreedte** = de klassenbreedte van een interval]a, b] wordt gegeven door **b-a**.
- **Gegroepeerde frequentieverdeling van X** = een **tabel** met 2 kolommen (of 2 rijen) waar de 1ste kolom **de klassen van X** weergeeft en de 2de de **overeenkomstige frequenties**.

Verschillen staafdiagram en histogram:
- Bij **een histogram** raken de rechthoeken elkaar en kunnen de breedtes van de rechthoeken verschillen.
- Een staafdiagram wordt vooral gebruikt voor **ordinale en nominale variabelen** (omdat ze vaak een beperkt aantal waarden hebben).
- Een histogram wordt vaak gebruikt voor **interval- en ratioschaal variabelen** (omdat ze vaak een groot aantal waarden hebben).

Verdeling:
- **Scheef naar rechts**: indien de meeste massa links ligt & het uiteinde rechts uitloopt
- **Scheef naar links**: indien de meeste massa rechts ligt en het uiteinde links uitloopt.
- **Symmetrisch**: de linker- en rechterstaarten zijn ongeveer gelijk.

Staarten: de uiteinden van een verdeling.

5. Cumulatieve frequentie
5.1. Ongegroepeerde data
Definitie:
- **Cumulatieve absolute frequentie van x (F(x))** = het **aantal elementen** in de **steekproef** die **kleiner dan of gelijk aan x** zijn.
- **Cumulatieve absolute frequentieverdeling van X** = een **tabel** met 2 kolommen (of 2 rijen), waar in de 1ste kolom **de waarden van de variabele X** worden weergegeven en in de 2de kolom **de overeenkomstige cumulatieve absolute frequenties**.

Hoofdstuk 3: Samenvatten van data

1. Onderzoek naar raciale voorkeur
1.1. Het gemiddelde
A: Het gemiddelde op basis van de waarden van een variabele:
- **Alle waarden** van een variabele **optellen** en **delen** door de **steekproefgrootte**.
- Enkel voor **interval- en ratiovariabelen**.
- We noemen het rekenkundig gemiddelde **kortweg het gemiddelde**.
- **Centrummaat** (maat van centrale tendentie): ligt vaak in centrum van de verdeling.
  - **Symmetrisch**: mooi in het midden.
  - **Scheef naar rechts**: gemiddelde schuift op naar links.

Scheef naar links: gemiddelde schuift op naar rechts.

Definitie:

- **Rekenkundig gemiddelde van X** = $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

B: Het gemiddelde berekenen op basis van de frequentieverdeling:

- Eerst **elke frequentie vermenigvuldigen met de unieke waarden**.
- Vervolgens deze **getallen optellen en delen door de steekproefgrootte**.
- **p** staat voor het aantal unieke waarden van de variabele X in de steekproef.

Extra notatie:

- **Unieke waarden**: x_i^u
  Laat toe om **alle unieke waarden** van X weer te geven.
- **Absolute frequentie** horende bij de unieke waarde: f_i

Definitie:

- **Gemiddelde van frequentieverdeling** = $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} f_i x_i^u$

1.2. Mediaan

- **Mediaan**: $md(x)$
- **De middelste waarde** nadat we de waarden van een variabele van klein naar groot geordend hebben.
- **Meetniveau**: ordinaal, interval – en ratiovariabelen.
- Indien verschillende waarden voldoen aan de definitie van de mediaan, wordt de **mediaan gelijkgesteld aan het rekenkundig gemiddelde** van deze waarden.
  Dan enkel nuttig voor **interval- en ratiovariabelen**.
- Mediaan kan je ook **afleiden uit de cumulatieve frequentieverdeling**.
- **Centrummaat**: ligt vaak in het midden van de verdeling.
  Symmetrisch: mooi in het midden.
  Scheef naar rechts: mediaan schuift op naar links.
  Scheef naar links: mediaan schuift op naar rechts.
- **Middelste klasse**: mediane klasse.

Definitie:

- **Mediaan van variabele X** = de **waarde** $md(x)$ waarvoor geldt dat:
  Niet meer dan de helft van de elementen in de steekproef een waarde **kleiner** dan $md(x)$ hebben EN…
  Niet meer dan de helft van de elementen in de steekproef een waarde **groter** dan $md(x)$ hebben.

1.3. De modus

- **Unimodale verdeling**: 1 modus.
- **Bimodale verdeling**: 2 modi.
- **Modale klasse**: klasse met de meeste frequenties.
- **Meetniveau**: nominale, ordinale, interval- en ratiovariabelen.

Definitie:

- **Modus (mo)** = de **klasse** of de **waarde met de grootste frequentie**. Als er meerdere dergelijke klassen of waarden zijn, dan zijn er meerdere modi.

1.4. Gevoeligheid aan outliers

Definitie:

- **Outlier/uitschieter** = **waarden** die **ver verwijderd zijn van de overige waarden** van een variabele.

Mate van beïnvloeding:

- **Gevoelig voor outliers**: gemiddelde.
- **Niet gevoelig voor outliers**: mediaan en modus.

2. Spreidingsmaten

2.1. De variatiebreedte

- De **afstand tussen de grootste en de kleinste waarde**.
- Als de **variatiebreedte gelijk is aan 0**, wil dit zeggen dat de grootste en kleinste waarde gelijk zijn en er geen spreiding is.
- **Meetniveau**: interval- en rationiveau.

Definitie:

- **Variatiebreedte ν_x** =
  De **grootste min de kleinste waarde** voor ongegroepeerde data.
  De **bovengrens van de laatste klasse min de ondergrens van de eerste klasse** voor gegroepeerde data (wanneer de klassen van klein naar groot geordend zijn).

2.2. De gemiddelde absolute afwijking

- Het **verschil tussen de xi-waarde en het gemiddelde**.
  We nemen hiervan **de absolute waarde** (geen min).
  Deze waarden worden **opgeteld en gedeeld door de steekproefgrootte**.
- **Meetniveau**: interval- of ratioschaal.

Definitie:

- **Gemiddelde absolute afwijking (ga_x) van variabele X** = $\mathbf{ga_x} = \frac{1}{n}\sum_{i=1}^{n}\left|x_1 - \bar{x}\right|$

2.3. De variantie en de standaarddeviatie

- **Variantie** = sn_x^2
  Variantie wordt **groter als er meer spreiding** is.
  In plaats van absolute waarden, zoals bij de gemiddelde absolute afwijking, wordt hier gebruikgemaakt van **kwadraten**.
  Meetniveau: interval- en ratioschaal.
- De **standaarddeviatie** is de vierkantswortel van de variantie.

Definitie:

- **Variantie van variabele X** = $\mathbf{sn_x^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ of $s^2x = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$
- **Standaarddeviatie van variabele X** = $\mathbf{sn_x} = \sqrt{sn_x^2}$

2.4. De interkwartielafstand

- Een maat van spreiding op basis van **percentielen**.
  Bv. het 10^e percentiel is de waarde van een variabele, waarvoor 10% van de waarden hetzelfde of kleiner is.
  Indien we dit delen door n, bekomen we de **cumulatieve frequentie**.
- **Bijzonder percentiel**: mediaan is gelijk aan 50^e percentiel.
- **Het eerste kwartiel P25**: een kwart van alle waarden is hetzelfde of kleiner.
- **Het tweede kwartiel P50**: de helft van alle waarden is hetzelfde of kleiner.
- **Het derde kwartiel P75**: driekwart van alle waarden is hetzelfde of kleiner.
- Afleiden van 1^{ste} en 3^{de} kwartiel kan a.d.h.v. **cumulatieve relatieve frequentiecurve**.
- **Meetniveau**: interval- en ratiovariabelen.
  Interkwartielinterval: ordinale, interval- en ratiovariabelen.

Definitie:
- **Interkwartielafstand (Q)** = **P75 – P25**.
- **Interkwartielinterval** = [P25, P75] en bevat **50% van alle waarden**.

2.5. De spreidingsmaat d

- **Meetniveau**: nominale, ordinale, interval- ratiovariabelen.
  Vooral gebruikt voor **nominale variabelen**.
- De waarde d ligt altijd **tussen 0 (geen spreiding) en 1 (veel spreiding)**.

Definitie:
- **Spreidingsmaat d** = $\mathbf{d} = \frac{1 - \mathbf{f}_{\frac{\mathbf{m}_0}{\mathbf{n}}}}{1 - \frac{1}{\mathbf{p}}}$
  P stelt het aantal unieke waarden voor.

2.6. Gevoeligheid aan outliers

Mate van beïnvloeding:
- **Gevoelig voor outliers**: variatiebreedte, de gemiddelde absolute afwijking, de variantie en de standaarddeviatie.
- **Niet gevoelig voor outliers**: interkwartielafstand en de spreidingsmaat d.

3. Boxplot

- Geeft een idee over **de verdeling van de data en kan outliers vaststellen**.
- **Outliers bepalen**:
  Interkwartielafstand berekenen en het verschil nemen:
  - **P25 – 1.5 x Q**
  - **P75 + 1.5 x Q**
- **Boxplot opstellen**:
  Eerst **as tekenen** (verticaal of horizontaal).
  Naast de as telkens **een stip zetten voor elke persoon**.
  Outliers aangeven (bv. rood kleuren).
  Lijn tekenen bij **de laagste en hoogste stip** die geen outlier is.
  Lijn tekenen bij **het 1^{ste} en 3^{de} kwartiel**.
  Lijnen ter hoogte van **de kwartielen verbinden**.
  Alle stippen verwijderen die geen outlier zijn.
  Stippenlijnen tekenen vanaf de laagste en hoogste stip naar de rechthoek van de kwartielen (= whiskers of snorharen).
  Lijn tekenen ter hoogte van **de mediaan**.

- **Boxplot bevat**:
  De mediaan (centrummaat).

De interkwartielafstand (spreidingsmaat): hoogte van rechthoek.
De outliers: de observaties die door bolletjes zijn aangeduid.

Hoofdstuk 4: Samenhang tussen 2 variabelen

1. Onderzoek naar intelligentie en hersengrootte

1.1. De onderzoeksvraag

1.2. De populatie en de steekproef

1.3. De data

2. Bivariate frequentieverdeling

- Tabel met info over 1 variabele: **univariate absolute frequentieverdeling**
  2 aparte tabellen laten niet toe conclusies te formuleren over de **gezamenlijke verdeling**.

- Tabel met info over 2 variabelen: **bivariate absolute frequentieverdeling**
  Laten toe **2 variabelen gezamenlijk** te bestuderen.
  Uit de bivariate verdeling kunnen we **altijd de univariate afleiden**.
  - **Marginale verdeling**: de univariate verdeling bepalen op basis van de bivariate verdeling.

  De **conclusies kunnen wijzigen** door de data te hergroeperen.
  - Deze **subjectiviteit is vaak onwenselijk** en kan vermeden worden door de samenhang te bestuderen.

3. Spreidingsdiagram

- Een figuur die toelaat **de samenhang tussen 2 variabelen te visualiseren**.
- **3 vormen van samenhang**:
  Perfecte positieve samenhang: de punten gaan van linksonder tot rechtsboven en liggen op een rechte.
  Perfecte negatieve samenhang: de punten gaan van linksboven tot rechtsonder en liggen op een rechte.
  Geen samenhang: geen patroon, de punten(wolk) zijn willekeurig verspreid.
- De samenhang hoeft niet altijd perfect op de rechte te liggen, er kan **spreiding** zijn.
- **Het interpreteren van een spreidingsdiagram is subjectief**: sommigen interpreteren het als een sterke samenhang, anderen als eerder zwak.
  Het is daarom handig om de samenhang te **kwantificeren via maten van samenhang**.

4. Maten van samenhang

Extra notatie:

- **Tweede variabele**: Y en kan de waarden y1, y2, … yn aannemen.

4.1. De covariantie

- **Meetniveau**: beide variabelen interval- of rationiveau.
- **Lineaire samenhang**.
- **Samenhang**:
  Positieve (lineaire) samenhang: Cov xy > 0.
  Negatieve (lineaire) samenhang: Cov xy < 0.
  Geen samenhang: Cov xy ≈ 0.

- Of het gaat om een **sterke of zwakke samenhang**, daar kan de covariantie ons geen antwoord op geven.

Definitie:

- **De covariantie** = $\mathbf{covxy} = \frac{1}{\mathbf{n}-1}\sum_{\mathbf{i}=1}^{\mathbf{n}}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{y_i}-\bar{\mathbf{y}})$

4.2. De correlatiecoëfficiënt

- Bekomen we door **de covariantie te delen door de standaarddeviaties**.
- De correlatiecoëfficiënt heeft **hetzelfde teken als de covariantie**.
- **Meetniveau**: interval- en ratiovariabelen.
- **Lineaire samenhang**.
- **Samenhang**:
  - **Positieve (lineaire) samenhang**: r xy > 0.
  - **Negatieve (lineaire) samenhang**: r xy < 0.
  - **Geen samenhang**: r xy ≈ 0.

Definitie:

- **De correlatiecoëfficiënt** = $\mathbf{r_{xy}} = \frac{\mathbf{covxy}}{\mathbf{s_x s_y}}$

4.3. Kendall's T (tau)

- Berekend door **concordante en discordante paren te tellen**.

  Paar is **concordant** als $\frac{y_j - y_i}{x_j - x_i} > 0$
  - **Wanneer** ($x_i < x_j$ en $y_i < y_j$) of ($x_i > x_j$ en $y_i > y_j$)

  Paar is **discordant** als $\frac{y_j - y_i}{x_j - x_i} < 0$
  - **Wanneer** ($x_i < x_j$ en $y_i > y_j$) of ($x_i > x_j$ en $y_i < y_j$)

  Paar is **niet concordant of discordant** als $x_i = x_j$ of $y_i = y_j$
- **Samenhang**:
  - **Positieve (lineaire) samenhang**: T > 0.
  - **Negatieve (lineaire) samenhang**: T < 0.
  - **Geen samenhang**: T ≈ 0.
- **Monotone samenhang**.
- **Meetniveau**: ordinale, interval- en ratiovariabelen.

Definitie:

- **Kendall's T** = $\boldsymbol{\tau} = \frac{2\,(\mathbf{aantal\ concordante\ paren} - \mathbf{aantal\ discordante\ paren})}{\mathbf{n(n-1)}}$

4.4. Lineaire en niet-lineaire verbanden

- Een **lineaire functie** is een functie die kan voorgesteld worden door een rechte lijn.
- Een **monotone functie** is:
  - Een **functie die de orde bewaart**.
  - De functie moet **ofwel stijgen, ofwel dalen**, maar niet beiden.

Moet geen rechte lijn zijn.

- **Een lineaire functie is een monotone functie**, maar er bestaan ook functies die monotoon zijn zonder lineair te zijn.
- Het is belangrijk de data **eerst te visualiseren** d.m.v. een spreidingsdiagram en dan pas te beslissen welke maat van samenhang geschikt is!

4.5. Gevoeligheid aan outliers

Mate van beïnvloeding:

- **Gevoelig voor outliers**: de covariantie en de correlatiecoëfficiënt.
- **Niet gevoelig voor outliers**: Kendall's T.

5. De regressielijn

- De regressielijn kan **het spreidingsdiagram visualiseren**.
- **De rechte bij een lineair verband** noemen we de regressielijn.
- **b1**: de regressiecoëfficiënt (helling van de rechte).
- **b0:** het intercept (het snijpunt met de verticale as).

Definitie:

- **Regressielijn** = $\mathbf{Y = b_0 + b_1 x}$

5.1. Formules indien het lineair verband perfect is

- We kiezen **2 willekeurige punten** (x_i, y_i) en (x_j, y_j).
- Pas de **formule van b1** toe: $b_1 = \frac{y_j - y_i}{x_j - x_i}$
- We kunnen dan **b0 vinden via**: $b_0 = y_i - b_1 x_i$

5.2. Formules indien het lineair verband niet perfect is

- We tekenen een rechte die het best door de puntenwolk zal gaan via **de kleinste-kwadrantenmethode**.
- We vinden **b1 via**: $r_{xy} = \frac{s_y}{s_x}$

  Bekomen we door **de correlatiecoëfficiënt te vermenigvuldigen met de sd van Y en te delen door de sd van X**.

  b1 zal **altijd hetzelfde teken** hebben als de correlatiecoëfficiënt.

- We vinden **b0 via**: $\bar{y} - b_1 \bar{x}$
- **Meetniveau**: interval- en ratiovariabelen.
- **De regressielijn kan je als volgt tekenen** op het spreidingsdiagram:

  Neem **2 willekeurige waarden** voor X.

  Vul voor elk van deze waarden **de formule van de regressielijn in**.

  Teken deze punten op het spreidingsdiagram.

  Als we deze **2 punten verbinden** met een rechte bekomen we de regressielijn.

6. Samenhang en causaliteit

Als we besluiten dat er een samenhang is, vermijden we best een formulering die en **causaal verband** impliceert.

DEEL 2: Kansrekening

Hoofdstuk 5: De populatie en de verdelingsfuncties

1. Verdelingsfunctie discrete variabelen

Een populatie kan beschreven worden a.d.h.v. **een verdelingsfunctie**. Een verdelingsfunctie kan worden gezien als de tegenhanger van de frequentieverdeling, maar nu gedefinieerd voor een populatie i.p.v. een steekproef.

- Discrete variabelen kunnen een **eindig aantal waarden (p)** aannemen.
- Het **aantal elementen** in de populatie kan men best als **oneindig** beschouwen.

Met $P(X = x_i)$ duiden we de kans aan dat de variabele X de waarde x_i aanneemt. De betekenis hangt nauw samen met de frequentieverdeling in een steekproef. Als we met f_i de absolute frequentie voorstellen van x_i in een steekproef van grootte n, dan kan de kans formeel gedefinieerd worden als: $P(X = x_i) = \lim_{n \to \infty} \frac{f_i}{n}$

Het is **de limiet van de relatieve frequentie** in de steekproef wanneer de steekproef oneindig groot wordt. Informeel kunnen we dit interpreteren als **de relatieve frequentie van x_i in de populatie**.

1.1. De kansverdeling

Definitie:

- **Kansverdeling van een discrete variabele X** = een **tabel met 2 kolommen** (of rijen) waarbij de 1ste kolom **de waarden x_i** weergeeft en de 2de kolom de **overeenkomstige kansen P(X = x_i).**

De kansverdeling kan gezien worden als de tegenhanger van **de relatieve frequentieverdeling op populatieniveau**. De kans ligt hierbij in het interval [0.1].

1.2. De cumulatieve verdelingsfunctie

De cumulatieve verdelingsfunctie is de **tegenhanger van de cumulatieve relatieve frequentie**. Soms spreken we kortweg over de verdelingsfunctie.

Definitie:

- **Cumulatieve verdelingsfunctie Fx (x)** = de **kans** dat de waarde van een variabele X **kleiner dan of gelijk is aan x** $\rightarrow F_X(x) = P(X \leq x)$
  Ziet er **trapsgewijs** uit.

De cumulatieve verdelingsfunctie Fx (x) kan je bekomen door de kansen P(X = x_i) uit de kansverdeling waarvoor $x_i \leq x$ op te tellen.

2. Verdelingsfunctie continue variabelen

- Continue variabelen kunnen **oneindig veel verschillende waarden** aannemen.
- Dit impliceert dat: **de kans P(X = x) = 0 voor elke waarde x**.

Om kansen te kunnen berekenen bij continue variabele, moeten we beroep doen op de **dichtheidsfunctie**.

2.1. De cumulatieve verdelingsfunctie

Definitie:

- **Cumulatieve verdelingsfunctie Fx (x)** = de **kans** dat de waarde van een variabele X **kleiner dan of gelijk is aan x** → $\mathbf{F_X(x) = P(X \leq x)}$
  Ziet er **continue** uit.

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

2.2. De dichtheidsfunctie

Definitie:
- **Dichtheidsfunctie fx(x)** = wordt gegeven door de **afgeleide van de verdelingsfunctie**: $\mathbf{f_X(x) = \lim_{b \to 0} \frac{F_X(x+b) - F_X(x)}{b}}$

Uitgedrukt in woorden geeft **fx(x) de kans weer dat X valt binnen het interval [x, x + b] gedeeld door b**, waar b de breedte van het interval voortelt en naar nul convergeert.

(In deze cursus zal deze formule niet worden berekend, aangezien we afgeleiden niet zullen berekenen.)

De dichtheidsfunctie kan bekomen worden **door middel van histogrammen**. De continue functie die wordt bekomen door het histogram op te delen in oneindig veel klassen, wordt de **dichtheidsfunctie** genoemd.

Via de dichtheidsfunctie kunnen we de kansen van de vorm P ($x_1 \leq X \leq x_2$) berekenen. Om deze kansen te bekomen, moeten we **de dichtheidsfunctie integreren**.

Voorbeeld: x_1 = 90 en x_2 = 110. P (90 ≤ X ≤ 110) = $\int_{90}^{110} f_X(x)\,dx$

De integraal kan benaderd worden door de oppervlaktes van de rechthoeken tussen de grenzen 90 en 110 op te tellen. (De integraal zullen we niet berekenen maar gebruiken we formeel om kansen te formuleren).

Algemeen kunnen we stellen dat: P ($x_1 \leq X \leq x_2$) = $\int_{x1}^{x2} f_X(x)\,dx$

Definitie:
- **De kans dat een variabele X in het interval [x_1, x_2] ligt** = de **oppervlakte onder de dichtheidsfunctie fx (x)** tussen de waarden x_1 en x_2.

Volgende kansen worden **zo voorgesteld**:
- **P (X ≤ x)** = $\int_{-\infty}^{x} f_X(x)\,dx$
- **P (X > x)** = $\int_{x}^{+\infty} f_X(x)\,dx$

Om tot **de numerieke waarde te komen** van de kans, zouden we de integraal moeten berekenen. We kunnen dit echter ook berekenen als we de verdelingsfunctie hebben.
Er geldt dat: **$P(x_1 \leq X \leq x_2) = Fx(x_2) - Fx(x_1)$.**

Er zijn nog **3 interessante eigenschappen**:

- **De dichtheidsfunctie is een positieve functie voor alle waarden x**:
  $fx(x) \geq 0$.

- **De volledige oppervlakte onder de dichtheidsfunctie is gelijk aan 1**:
  $$\int_{-\infty}^{+\infty} f_X(x)\,dx = 1.$$

- **Er geldt dat**:
  $P(X > x) = 1 - P(X \leq x)$.

3. Populatieparameters

3.1. Populatiegemiddelde

Discrete variabelen:

Definitie:

- **Het gemiddelde van een discrete variabele X** in een populatie E(X) =
  $$\mathbf{E(x)} = \sum_{i=1}^{p} \mathbf{P(X = x_i)x_i}$$

- Het populatiegemiddelde wordt ook de **verwachtingswaarde** genoemd.
- In plaats van symbool E(X) gebruikt men soms ***μx* of kortweg *μ*** (mu).

Continue variabelen:

Definitie:

- **Het gemiddelde (of verwachtingswaarde) van een continue variabele X** in een populatie = $\mathbf{E(X)} = \int_{-\infty}^{+\infty} \mathbf{fX(x)x}\,d\mathbf{x}$

3.2. Populatievariantie

Discrete variabelen:

Definitie:

- **De variantie van een discrete variabele X** in een populatie V(X) =
  $$\mathbf{V(X)} = \sum_{i=1}^{p} \mathbf{P(X = x_i)\left(x_i - E(X)\right)^2}$$
- **De standaarddeviatie van een variabele X** in een populatie (symbool σ_x) =
  $$\sigma_x = \sqrt{\mathbf{V}(X)} = \sqrt{\sum_{i=1}^{p} \mathbf{P(X = x_i)\left(x_i - E(X)\right)^2}}$$

- In plaats van het symbool V(X) gebruikt men soms ook **σ_x^2 of kortweg σ_x (sigma).**
- De standaarddeviatie kan worden bekomen door **de vierkantswortel te nemen** van de variantie.

Continue variabelen:

Definitie:

- **De variantie van een continue variabele X** in een populatie = $$V(X) = \int_{-\infty}^{+\infty} f_X(x)(x - E(X))^2 \, dx$$

4. Bivariate kansverdelingen

4.1. Discrete variabelen

Op basis van de bivariate verdeling kunnen we **de marginale (univariate) verdelingen afleiden door kansen op te tellen**.

- Het aantal mogelijke waarden van X is **p**.
- Het aantal mogelijke waarden van Y is **q**.

De univariate verdeling van X wordt bekomen via: $$P(X = x_i) = \sum_{j=1}^{q} P(X = x_i \text{ en } Y = y_j).$$

De univariate verdeling van y wordt bekomen via: $$P(Y = y_j) = \sum_{i=1}^{p} P(X = x_i \text{ en } Y = y_j).$$

Statistische onafhankelijkheid is een belangrijk begrip hierbij:

- Twee discrete variabelen X en Y zijn **onafhankelijk** als de gelijkheid $P(X = x_i \text{ en } Y = y_j) = P(X = x_i)\, P(Y = y_j)$

Definitie:

- **De covariantie voor 2 discrete variabelen X en Y** in een populatie = $$COV(X, Y) = \sum_{i=1}^{p} \sum_{j=1}^{q} P(X = x_i \text{ en } Y = y_j)(x_i - E(X))\,(y_j - E(Y))$$
- **De correlatiecoëfficiënt** = $\rho_{XY} = \dfrac{COV(X, Y)}{\sigma_X \sigma_Y}$

4.2. Continue variabelen (formules niet kunnen uitwerken)

We kunnen **de cumulatieve bivariate verdelingsfunctie** definiëren als: $F_{X,Y}(x, y) = P(X \leq x \text{ en } Y \leq y)$.

De bivariate dichtheidsfunctie bekomen we door $F_{X,Y}(x, y)$ af te leiden en noteren we symbolisch als $f_{X,Y}(x, y)$.

Statistische onafhankelijkheid is een belangrijk begrip hierbij:

- Twee continue variabelen X en Y zijn **onafhankelijk** als de gelijkheid $P(X \leq x \text{ en } Y \leq y) = P(X \leq x)\, P(Y \leq y)$.

Definitie:

- **De covariantie voor 2 continue variabelen X en Y** in een populatie =

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

- **De correlatiecoëfficiënt** = $\boldsymbol{\rho_{XY}} = \dfrac{\mathbf{COV}(\mathbf{X},\mathbf{Y})}{\boldsymbol{\sigma_X\sigma_Y}}$

5. Nuttige stellingen

De stellingen gelden voor **zowel discrete als continue variabelen**.

- **STELLING 1**: als X en Y onafhankelijke variabelen zijn, dan geldt dat:
  COV (X, Y) = 0.

  Het omgekeerde geldt niet: een covariantie van 0 impliceert niet dat de variabelen onafhankelijk zijn.

- **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 2 variabelen X en Y (die onafhankelijk of afhankelijk kunnen zijn) geldt dat:
  E(X + Y) = E(X) + E(Y) of **E(X – Y) = E(X) – E(Y).**

- **STELLING 5**: voor 2 onafhankelijke variabelen X en Y geldt dat:
  E(XY) = E(X) E(Y).

Stellingen 2, 3 en 4 **gaan ook op voor het steekproefgemiddelde**, terwijl dit niet zo is voor stellingen 1 en 5.

- **STELLING 6**: voor een variabele Y = 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 2 variabelen X en Y geldt dat:
  V(X + Y) = V(X) + V(Y) + 2COV(X, Y).

  Indien X en Y **onafhankelijke variabelen** zijn, dan volgt uit stelling 1 en 8 dat:
  V (X+Y) = V (X) + V(Y)

- **STELLING 9**: voor 2 variabelen X en Y geldt dat:
  V (X-Y) = V(X) – V(Y) – 2COV(X, Y).

  Indien X en Y **onafhankelijke variabelen** zijn, dan volgt uit stelling 1 en 9 dat:
  V (X-Y) = V(X) + V(Y).

Stellingen 6, 7, 8 en 9 **gaan ook op voor de steekproefvariantie**.

6. Bijzondere verdelingen

6.1. De binominale verdeling

De binominale verdeling zal de kansverdeling weergeven om k correcte antwoorden te hebben op een examen met N vragen. Symbolisch schrijven we de kans als p. **De kans p wordt ook de kans op 'succes' genoemd**.

We maken gebruik van **de binominale kansverdeling** om de kansverdeling van X te berekenen. Dit doen we door: $\mathbf{P(X = k) = \frac{N!}{k!(N-k)!} P^k (1-p)^{N-k}}$

Het symbool N! staat voor N faculteit, p voor de kans op succes, k is een gegeven geheel getal en N voor het maximaal aantal successen.

- Bv. N = 4, 4! = 4 x 3 x 2 x 1 = 24.

Een variabelen die een binominale verdeling heeft, noemen we **een binominale variabele** en noteren dit symbolisch als **X ~ Binom (N, p)**.

- **Verwachtingswaarde X** ~ Binom (N, p) = E(X) = Np.
- **Variantie X** ~ Binom (N, p) = V(X) = Np (1-p).

De binominale verdeling kan enkel gebruikt worden als **N vast is** en indien **de kans op succes p ongewijzigd blijft**.

6.2. De normale verdeling

Een normaal verdeelde variabele in **continue** en **de dichtheidsfunctie** wordt gegeven door:

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

Een variabele die normaal verdeeld is noteren we als **X ~ N (μ, σ²)**.

- **E(X) = *μ*** → populatiegemiddelde.
- **V(X) = σ²** → populatievariantie.

Voor elke keuze van *μ* en σ² bekomen we een **andere dichtheidsfunctie**.

- Het hoogste punt van de dichtheidsfunctie komt overeen met het **gemiddelde**.
- Bij een **grotere variantie σ²** wordt de dichtheidsfunctie breder en minder hoog.
- **De dichtheidsfunctie wordt nergens 0**.

De normale verdeling met *μ* = 0 en σ² = 1 wordt **de standaardnormale verdeling** genoemd. Er geldt dat: P (X ≤ -x) = 1 – P (X ≤ x).

❖ **STELLING 10**: als X een normale verdeling heeft met een gemiddelde *μ* en variantie σ², dus X ~ N (μ, σ²), dan heeft de variabele:
Z = $\mathbf{\frac{X-\mu}{\sigma}}$, een standaardnormale verdeling, dus Z ~ N(0,1).

De stelling impliceert de volgende vergelijking: **als X ~ N(μ, σ²) dan geldt dat:**
P (X ≤ x) = P ($\mathbf{\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}}$**) = P (Z ≤** $\mathbf{\frac{x-\mu}{\sigma}}$**), waarbij Z ~ N(0.1).**

⇒ Dit noemen we **standaardiseren van X**!

6.3. De X²-verdeling

De X^2_k- verdeling (chi-kwadraat) is de verdeling van de variabele: **$Y = X^2_k$** .

De X²-verdeling is bijgevolg de verdeling van de som van de k gekwaddrateerde standaardnormale variabelen. De parameter k wordt het **aantal vrijheidsgraden** genoemd.

- **E(Y) = k**.
- **V(Y) = 2k**.

Een variabele Y die een X^2_k- verdeling heeft, noteren we als **$Y \sim X^2_k$**.

6.4. De t-verdeling

De t_k-verdeling is de verdeling van de variabele: $T = \dfrac{X}{\sqrt{\frac{1}{k} Y}}$

- De parameter k wordt ook het **aantal vrijheidsgraden** genoemd.
- **De dichtheidsfunctie gelijkt op die van een normale verdeling** maar is niet volledig gelijk (naarmate k toeneemt, gelijkt het meer op een standaardnormale verdeling).

Als **$T \sim t_k$** dan geldt dat:

- **E(T) = 0**.
- **$V(T) = \dfrac{k}{k-2}$** voor $k > 2$.

DEEL 3: Inductieve statistiek

Hoofdstuk 6: De steekproevenverdeling

1. Steekproeftrekking

Aselecte steekproeftrekking = op **volledig willekeurige wijze** worden n elementen geselecteerd uit de populatie.

We veronderstellen verder dat deze n elementen **onafhankelijk** zijn van elkaar.

- **X_i:** de variabele X van element i in een steekproef zonder dat we de steekproef effectief getrokken hebben.
- x_i: de variabele X bij element i voor een specifiek getrokken steekproef.
- **$P(X = x_i)$:** de relatieve frequentie van x_i in de populatie.
  De notatie van P komt van **probabiliteit**.

Intermezzo: de betekenis van een kans

Er bestaan verschillende interpretaties van een kans. In deze cursus bespreken we de **frequentistische kans**.

De kans op een gebeurtenis is dus gelijk aan de relatieve frequentie van de gebeurtenis indien we het experiment een oneindig aantal keer herhalen.

Terugkeer naar de Benton Visual Retention Test

We kunnen de waarde $P(X = x_i)$ ook interpreteren als een kans via **de herhaalde steekproeftrekking**.

Deze steekproeftrekking dienen we een **oneindig** aantal keer te herhalen.

Toevalsvariabele = een variabele X die bekomen wordt door **op toevallige wijze** een element **uit de populatie te trekken**.

- Ze duidt het resultaat aan van een **toevallige trekking** van een element uit de populatie.
- Ze is **veranderlijk** (variabel) omdat niet alle elementen in de populatie dezelfde waarde hebben.
- Er zal echter kortweg gesproken worden over de **variabele**.

2. Steekproevenverdeling van het gemiddelde

Het steekproefgemiddelde is variabel: de waarde hangt af van de frequentieverdeling van de scores in de steekproef en verschillende steekproeven hebben verschillende frequentieverdelingen. Het steekproefgemiddelde is bijgevolg een **variabele**.

Een variabele schrijven we met een **hoofdletter**: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Hier stelt $\bar{X}$ het steekproefgemiddelde voor van een **steekproef in het algemeen**.

Het steekproefgemiddelde is een voorbeeld van een **steekproefgrootheid**, het is een bewerking toegepast op de variabelen. Andere voorbeelden van een steekproefgrootheid zijn de mediaan, de modus, de variantie, … Een steekproefgrootheid wordt soms ook **een statistiek** genoemd.

Indien we een oneindig aantal steekproeven trekken en we een histogram opstellen van het steekproefgemiddelde, gekomen we de **dichtheidsfunctie van het gemiddelde**. Deze dichtheidsfunctie wordt ook de steekproevenverdeling van het gemiddelde genoemd.

Definitie:

- **Steekproevenverdeling** = de **verdeling** van een **steekproefgrootheid**.

Opgelet: de frequentieverdeling geeft de verdeling van de variabele weer, terwijl de steekproevenverdeling de verdeling van een steekproefgrootheid weergeeft.

- **STELLING 11**: de verwachtingswaarde van het steekproefgemiddelde $\bar{X}$ is gelijk aan het populatiegemiddelde van de variabele X:
  $E(\bar{X}) = \mu_X$.

  Voor één steekproef is het steekproefgemiddelde over het algemeen **niet gelijk aan** het populatiegemiddelde. Bij een oneindig aantal steekproeven wel.

- **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 variantie van het steekproefgemiddelde is dus **niet gelijk** aan de populatievariantie van de variabele. Indien we 'oneindig' veel steekproeven nemen, zullen deze waarden wel dichter bij elkaar komen te liggen.

De steekproefgemiddelden verschillen meer bij de kleine steekproeven en bijgevolg is hun **variantie groter**. Bij grotere steekproeven zal het steekproefgemiddelde 'dichter' bij het populatiegemiddelde liggen en is de **variantie kleiner**.

Als we beide stellingen combineren, krijgen we **de wet van de grote aantallen**. Die stelt dat het steekproefgemiddelde met hoge waarschijnlijkheid weinig zal verschillen van het populatiegemiddelde indien de steekproef 'groot' is.

- **STELLING 13**: Stel dat $X_i, \ldots X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met een normale verdeling N (μx, $\frac{\sigma_x^2}{n}$), dan zal $\bar{X}$ ook normaal verdeeld zijn:
  $\bar{X} \sim N(\mu x, \frac{\sigma_x^2}{n})$.

  De normale verdeling gaat echter enkel op voor continue variabelen.

- **STELLING 14** (centrale limietstelling): Stel dat $X_i, \ldots X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met gemiddelde μx 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 met gemiddelde μx en variantie $\frac{\sigma_x^2}{n}$.

  In de praktijk wordt de regel $n \geq 30$ gebruikt om aan te geven of een steekproef 'groot' is.

Normaal verdeelde variabelen standaardiseren:

$$P(\bar{X} \leq x) = P\left(Z \leq \frac{x - \mu x}{\sqrt{\sigma_x^2 / n}}\right), Z \sim N(0,1)$$

Indien X uit een normale verdeling komt, geldt deze formule voor alle keuzes van n. Indien X niet uit een normale verdeling komt, geldt deze eigenschap enkel maar voor grote n.

Stelling 13 & 14, en het standaardiseren is belangrijk om conclusies te trekken op basis van 1 experiment. De resultaten moeten namelijk **reproduceerbaar** zijn.

3. Steekproevenverdeling van de variantie

Twee formules voor de steekproefvariantie:

- $SN^2x = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- $S^2x = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

De verwachtingswaarde:

- $E(SN^2x) = \frac{n-1}{n}\sigma_x^2$
- $E(S^2x) = \sigma_x^2$

De verwachtingswaarde van de steekproefvariantie SN^2x is dus niet gelijk aan de populatievariantie. Voor S^2x is dit echter wel zo. Dit is een gunstige eigenschap en daarom zal men in de praktijk vaak **de variantie bereken via S^2x**.

- **STELLING 15**: Stel dat $X_i, \ldots X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met een normale verdeling N (μx, σ_x^2), dan geldt:
  $\frac{(n-1)S_x^2}{\sigma^2} \sim X^2 n-1$.

Hoofdstuk 7: Betrouwbaarheidsintervallen en statistische toetsen voor het

1. Schatters

Indien we een uitspraak wensen te doen over het populatiegemiddelde op basis van een steekproef, zullen we dit **gemiddelde moeten schatten** op basis van informatie in de steekproef.

$\hat{\theta}$ is een **steekproefgrootheid**.

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

- **Ze zuiver is**, wat wil zeggen dat de verwachtingswaarde van de schatter gelijk is aan de populatieparameter.
  $E(\hat{\theta}) = \theta$.

  Dit houdt in dat de populatieparameters niet systematisch te klein of te groot wordt geschat.

- **De variantie van de schatter, $V(\hat{\theta})$, kleiner wordt** naarmate de steekproefgrootte toeneemt. Dit drukt uit dat de schatter meer nauwkeurig zal zijn wanneer de steekproef groter wordt.

De standaarddeviatie van de schatter, $\sqrt{V(\hat{\theta})}$, wordt ook de **standaardfout** genoemd. De schatter met de kleinste standaardfout is het efficiëntste.

1.1. Het gemiddelde

Het steekproefgemiddelde is een logische keuze om het populatiegemiddelde te schatten, dus $\hat{\theta} = \bar{X}$ als $\Theta = \mu$.

- Uit stelling 11 volgt dat: **$E(\bar{X}) = \mu$**.
  Indien we vele steekproeven trekken, zal het gemiddelde van deze steekproefgemiddelden gelijk zijn aan het populatiegemiddelde.

- Uit stelling 12 volgt dat: **$V(\bar{X}) = \frac{\sigma^2}{n}$**.
  Als n toeneemt, dan wordt σ^2/n kleiner: hoe groter de steekproef, hoe nauwkeuriger we het populatiegemiddelde kunnen schatten via het steekproefgemiddelde.
  De **standaardfout** van het steekproefgemiddelde is gelijk aan $\sigma^2/\sqrt{n}$.

Aan beide voorwaarden van een schatter is voldaan, waardoor het steekproefgemiddelde **een goede schatter** is voor het populatiegemiddelde.

Toegepast op het gemiddelde is $\bar{X}$ (variabele) de **schatter** en $\bar{x}$ (waarde van de variabele in 1 steekproef) de **schatting**.

1.2. De variantie

Om de populatievariantie te schatten lijkt het meest logische om beroep te doen op de steekproefvariantie, maar er zijn **2 formules** hiervoor. Welke is de beste schatter?

- **S_N^2X is geen zuivere schatter**: $\frac{n-1}{n}\sigma^2 < \sigma^2$
  De populatieparameter zal systematisch te klein geschat worden.
- **S^2x is wel een zuivere schatter**: $\sigma^2 = \sigma^2$

De variantie van zowel Sn^2x als S^2x zal afnamen naarmate de steekproef groter wordt.

2. Betrouwbaarheidsintervallen

Een **betrouwbaarheidsinterval** zal ons in staat stellen om met een bepaalde zekerheid uitspraak te doen over het populatiegemiddelde.

2.1. X normaal verdeeld en gekende populatievariantie

Extra notatie:

- z_α: de waarde van de standaardnormale verdeling.
  Dit zodat **de oppervlakte onder de curve rechts** van de waarde gelijk is aan α (alfa).
  $P(Z > z_\alpha) = \alpha$, met $Z \sim N(0.1)$ (oppervlakte rechts = α)

Omdat de standaardnormale verdeling **symmetrisch is rond 0**, kan men aantonen dat:
$P(-z_\alpha/2 \leq Z \leq z_\alpha/2) = 1 - \alpha$.
⇒ De kans dat een standaardnormale verdeling een waarde aanneemt tussen … en … is … %.

We kunnen Z **vervangen door het gestandaardiseerde**:

$$P(-z_\alpha/2 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_\alpha/2) = 1 - \alpha$$

Dit zal de basis vormen voor het construeren van een betrouwbaarheidsinterval. Om dit betrouwbaarheidsinterval te bekomen moeten we **de uitdrukking herschrijven zodat enkel *μ* overblijft**.

- **STAP 1**: de drie termen van de ongelijkheid vermenigvuldigen met $\sigma^2/\sqrt{n}$.
  $P(-\frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

- **STAP 2**: bij iedere term $\bar{X}$ aftrekken.
  $P(-\bar{X} - \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

- **STAP 3**: de drie termen vermenigvuldigen met -1.
  $P(\bar{X} + \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

- **STAP 4**: de ongelijkheid herschrijven zodat de kleinste waarde links staat en de grootste rechts.
  $P(\bar{X} - \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

De kan dat het populatiegemiddelde in het interval $\mathbf{[\bar{X} - \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}]}$ ligt is gelijk aan 1-α. Dit interval wordt het **(1-α) 100% betrouwbaarheidsinterval (BI)** genoemd.

We zijn dus in staat om een uitspraak te doen over het populatiegemiddelde op basis van een steekproef. We zijn echter **niet 100% zeker** maar (meestal) 95% zeker.
Merk op: het steekproefgemiddelde ligt altijd in het midden van het interval.

Interpretatie van het betrouwbaarheidsinterval
Het betrouwbaarheidsinterval **hangt af van het steekproefgemiddelde** $\bar{x}$, en verschillende steekproeven zullen verschillende gemiddelden hebben. Dit resulteert bijgevolg in verschillende betrouwbaarheidsintervallen. Het betrouwbaarheidsinterval is dus **variabel**.

De theorie van het betrouwbaarheidsinterval garandeert dat er exact 95% van alle intervallen het populatiegemiddelde zullen bevatten indien we het experiment een oneindig aantal keer herhalen.

Eigenschappen van het betrouwbaarheidsinterval
De breedte van een interval [a, b] is gelijk aan b-a.

$$(\bar{X} + \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}) - (\bar{X} - \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}) = 2 \text{ x } \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}}$$

De breedte hangt af van de standaardgrootte n, de waarde $\frac{z_\alpha}{2}$ en de populatiestandaarddeviatie σ.

- **Als de steekproef groter wordt**, dan zal de breedte van het interval kleiner worden.
  Minder betrouwbaar.
  Meer nauwkeurig.
- **Als α toeneemt**, dan zal de breedte van het interval afnemen.
  Meer betrouwbaar.
  Minder nauwkeurig.

2.2. X normaal verdeeld en ongekende populatievariantie

Het betrouwbaarheidsinterval zoals hierboven opgesteld kunnen we hier niet gebruiken. Op basis van 2 gekende eigenschappen kunnen we echter **een nieuw betrouwbaarheidsinterval opstellen** die gebruik maakt van Sx in plaats van σ.

- **Eigenschap 1**: Als X normaal verdeeld is, dan volgt uit stelling 15 dat:
  $\frac{(n-1)S_x^2}{\sigma^2} \sim X^2 n-1$.
- **Eigenschap 2**: Als X normaal verdeeld is, dan volgt:
  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$.

Door deze **eigenschappen te combineren** krijgen we volgende vereenvoudigde uitdrukking:

$$\frac{\bar{X} - \mu}{S_x / \sqrt{n}} \sim t_{n-1}$$

De verdeling wijzigt van een standaardnormale naar een t_{n-1}-verdeling.

- **$P(T > t_{n-1;\alpha/2}) = \alpha/2$**, $T \sim t_{n-1}$ (oppervlakte rechts = α/2)

Omdat de t_{n-1}-verdeling **symmetrisch is rond 0**, kan men aantonen dat:
$P(-t_{n-1;\alpha/2} \leq T \leq t_{n-1;\alpha/2}) = 1 - \alpha$, waar $T \sim t_{n-1}$

We kunnen dit herschrijven zodat het **(1-α) 100% betrouwbaarheidsinterval** gelijk is aan:

$$[\bar{X} - t_{n-1;\alpha/2} \frac{S_x}{\sqrt{n}}, \bar{X} + t_{n-1;\alpha/2} \frac{S_x}{\sqrt{n}}]$$

Verschillen tussen de standaardnormale en t_{n-1}- verdeling:

- 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.

Dit impliceert dat het betrouwbaarheidsinterval berekend op basis van de t_{n-1}- verdeling **vaak breder zal zijn** dan het betrouwbaarheidsinterval van de standaardnormale verdeling.
Naarmate n groter wordt, zal de t_{n-1}- verdeling steeds beter de standaardnormale verdeling benaderen.

2.3. X niet normaal verdeeld en ongekende populatievariantie

Als X niet normaal verdeeld is, kunnen we voor een grote steekproef beroep doen op de **centrale limietstelling**. Daarbij geldt dat het betrouwbaarheidsinterval van de t_{n-1}- verdeling.

Indien de steekproef te klein is, zullen we dit niet behandelen in de cursus.

3. Statistische toetsen

We zullen ons beperken tot **een t-toets voor één steekproef wanneer X normaal verdeeld is** of wanneer de stekproef groot is ($n \geq 30$).

- H_0: nulhypothese.
- H_a: alternatieve hypothese.

Ofwel is de nulhypothese correct, ofwel de alternatieve hypothese, en op basis van 1 steekproef willen we dit antwoord bekomen.

- Als steekproefgemiddelde ($\bar{x}$) '**ongeveer' gelijk is aan** het getal zullen we H_0 niet verwerpen.
- Als steekproefgemiddelde **verschilt** van het getal, zullen we H_a besluiten.

Bovenstaande redenering is duidelijk, maar we moeten **objectief vastleggen** wat 'ongeveer' gelijk aan en wat verschillend van is.
Om formeel de toets in te voeren, gebruiken we meer **algemene notatie**. We schrijven de hypotheses als: **H_0: $\mu = \mu 0$ en H_a: $\mu \neq \mu 0$**, waar $\mu 0$ de gegeven waarde is.

Deze alternatieve hypothese wordt de **tweezijdige alternatieve hypothese** genoemd. De **eenzijdige hypotheses** zijn: H_a: $\mu > \mu 0$ en H_a: $\mu < \mu 0$.

Bij een statistische toets zullen we **trachten H_0 te gaan verwerpen**, en dus bewijs te gaan zoeken om deze te verwerpen. Het bewijs tegen H_0 zullen we samenvatten d.m.v. een **toetsingsgrootheid**.

3.1. Toetsingsgrootheid

De **steekproefgrootheid** noteren we kort als G: **G =** $\dfrac{\bar{X} - \mu_0}{S_x / \sqrt{n}}$

Om te benadrukken dat G een t_{n-1}- verdeling volgt op voorwaarde dat H_0 correct is, noteren we als volgt: $G \overset{H_0}{\sim} t_{n-1}$

De steekproefgrootheid G wordt de **toetsingsgrootheid** genoemd en t_{n-1} is de verdeling van de toetsingsgrootheid wanneer de nulhypothese waar is. De waarde van G die we bekomen op basis van één steekproef noteren we als **g**.

Welke waarden kan G aannemen, naargelang H_0 waar is of niet?

- **H_0 is waar ($\mu = \mu 0$)**: meeste waarden zijn klein en liggen rond nul.
- **H_0 is niet waar ($\mu > \mu 0$)**: G neemt enkel positieve waarden aan.
- **H_0 is niet waar ($\mu < \mu 0$)**; G neemt enkel negatieve waarden aan.

3.2. Beslissingsregels

Op basis van één steekproef kunnen we **g berekenen**. Hierbij gelden volgende regels:

- **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.

We zullen hierbij volgende **beslissingsregels** gebruiken:

- 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.

De waarden $-t_{n-1;\alpha/2}$ en $t_{n-1;\alpha/2}$ worden **de kritische waarden** van de toets genoemd. Het gebied tussen deze waarden noemt men het **aanvaardingsgebied**, het gebied erbuiten wordt het **kritisch gebied** genoemd.

3.3. Type I en type II fout

Er zijn **4 scenario's mogelijk** waarvan er 2 resulteren in een correcte conclusie en 2 in een foutieve conclusie.

	In werkelijkheid is H0	
	Juist	**Fout**
We verwerpen H0 niet	Juiste beslissing (A) Betrouwbaarheid (1-α)	Foute beslissing (C) Type II fout β
We verwerpen H0	Foute beslissing (B) Type I fout α	Juiste beslissing (D) Onderscheidingsvermogen (1-β)

- De kans om fout B te maken noteren we als: **P (verwerp H_0| $\mu = \mu 0$) = α**.
  Kans op een **type I fout** met **α als significantieniveau**.
  Enkel exact gelijk aan alfa in een **normaal verdeling**.
  Volgens de centrale limietstelling is het bij benadering gelijk aan alfa bij een **grote steekproef**.

- De kans om de juiste beslissing A te maken is: **P (verwerp H_0 niet| $\mu = \mu 0$) = 1-α**.
  Wordt de **betrouwbaarheid** genoemd.

- De kans om fout C te maken noteren we als: **P (verwerp H_0 niet| $\mu \neq \mu 0$) = β**.
  Kans op een **type II fout**, en hangt af van het volgende:
  - **Het significantieniveau α**: β stijgt als a daalt.
  - **De steekproefgrootte n**: β stijgt als n daalt.

- De kans om de juiste beslissing D te maken is: **P (verwerp H_0| $\mu \neq \mu 0$) = 1-β**.
  Wordt de **onderscheidingskans (of power)** genoemd.

We spreken over **het niet verwerpen van H_0**, wat minder sterk geformuleerd is dan bv. het besluiten van H_0 .

3.4. Beslissingsregels op basis van het betrouwbaarheidsinterval

Beslissingsregels voor een toets op het alfa significantieniveau kunnen equivalent worden uitgedrukt door gebruik te maken van **een (1-α) 100% betrouwbaarheidsinterval**.

$$[\bar{x} - t_{n-1;\alpha/2}\frac{S_x}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2}\frac{S_x}{\sqrt{n}}]$$

Deze regels zijn:
- **Als $\mu 0$ in het betrouwbaarheidsinterval ligt**, verwerpen we H_0 niet.
- **Als $\mu 0$ niet in het betrouwbaarheidsinterval ligt**, verwerpen we H_0 en besluiten we H_a: $\mu \neq \mu 0$.

Bij een besluit vermelden we ook steeds het significantieniveau.

3.5. Eenzijdige en tweezijdige toetsen

Rechtszijdig:
- **Nulhypothese**: $\mu = \mu 0$.
- **Alternatieve hypothese**: $\mu > \mu 0$.

Uitvoeren van een eenzijdige toets is zeer gelijkaardig:
- Het **berekenen van de toetsingsgrootheid G (g)**.
- **Beslissingsregels** om H_0 al dan niet te verwerpen:
  Als $g \leq t_{n-1;\alpha}$ verwerpen we H_0 niet.
  Als $g > t_{n-1;\alpha}$ verwerpen we H_0 en besluiten we H_a.

Merk op: het aanvaardingsgebied bepaald door de beslissingsregels heeft een geschatte bovengrens, terwijl het betrouwbaarheidsinterval een geschatte ondergrens heeft.

Linkszijdig:
- **Nulhypothese**: $\mu = \mu 0$.
- **Alternatieve hypothese**: $\mu < \mu 0$.

Uitvoeren van een eenzijdige toets is zeer gelijkaardig:

- Het **berekenen van de toetsingsgrootheid G (g)**.
- **Beslissingsregels** om H_0 al dan niet te verwerpen:
  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.

Eenzijdig of tweezijdig toetsen?

Eenzijdige en tweezijdige toetsen hebben elk hun voordelen. Het type I fout blijft hetzelfde, maar de kans op een type II fout varieert.

- **In werkelijkheid is $\mu < \mu 0$**:
  De tweezijdige toets met H_a: $\mu \neq \mu 0$ zal een specifiek alternatief kunnen detecteren met een bepaalde power.
  De linkszijdige toets met H_a: $\mu < \mu 0$ zal ook een specifiek alternatief kunnen detecteren met een hogere power dan de tweezijdige toets.
  De rechtszijdige toets met H_a: $\mu > \mu 0$ heeft een power van maximaal α en dus een zeer lage power.

- **In werkelijkheid is $\mu > \mu 0$**:
  De tweezijdige toets met H_a: $\mu \neq \mu 0$ zal een specifiek alternatief kunnen detecteren met een bepaalde power.
  De linkszijdige toets met H_a: $\mu < \mu 0$ heeft een power van maximaal α (dus een zeer lage power).
  De rechtszijdige toets met H_a: $\mu > \mu 0$ zal ook een specifiek alternatief kunnen detecteren met een hogere power dan de tweezijdige toets.

In de praktijk zal vaak worden gekozen voor de tweezijdige toets.

3.6. P-waarde

Om tot een besluit te komen kunnen we ook gebruik maken van **de p-waarde of de overschrijdingskans** (p).

Beslissingsregels:

- Als **p ≥ α** verwerpen we H_0 niet.
- Als **p < α** verwerpen we H_0 en besluiten we H_a.

Formeel kunnen we de p-waarde als volgt omschrijven:

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.

Deze omschrijving omvat **volgende informatie**:

- **De p-waarde is een kans**: ze kan bijgevolg nooit kleiner zijn dan 0 en nooit groter dan 1.
- De p-waarde wordt berekend in de **veronderstelling dat H_0 waar is**.
- De p-waarde **hangt af van de alternatieve hypothese**.

Linkszijdige alternatieve hypothese: H_a: $\mu < \mu 0$

De p-waarden zijn de waarden **links van de geobserveerde toetsingsgrootheid**:
P (G < g | $\mu = \mu 0$).

Rechtszijdige alternatieve hypothese: H_a: $\mu > \mu 0$
De p-waarden zijn de waarden **rechts van de geobserveerde toetsingsgrootheid**:
P (G > g | $\mu = \mu 0$).

Tweezijdige alternatieve hypothese: H_a: $\mu \neq \mu 0$
Het berekenen van de p-waarde **hangt af van het teken van g**:

- **Als g > 0** dan is de p-waarde gelijk aan 2 x P (G > g | $\mu = \mu 0$).
- **Als g ≤ 0** dan is de p-waarde gelijk aan 2 x P (G < g | $\mu = \mu 0$).

Interpretatie van de p-waarde
De p-waarde is een kans die we kunnen **interpreteren via de herhaalde steekproeftrekking**.
Bv. een p-waarde van 0.17 drukt uit dat 17% van de toetsingsgrootheden kleiner/groter zullen zijn dan g.

Hoe **kleiner de p-waarde**, hoe meer bewijskracht tegen de nulhypothese.

3.7. Overzicht en opmerkingen

Belangrijke opmerkingen/misvattingen rond gebruik van statistische toetsen:

- **De logica van statistische toetsen is tot op zeker hoogte gelijkaardig met diegene in een rechtbank**: men vertrekt vanuit de assumptie dat een beklaagde onschuldig is (nulhypothese) totdat voldoende bewijs van schuld kan worden aangetoond.

- **De p-waarde kunnen we symbolisch noteren als**: P (Data | H_0): de kans om een bepaalde waarde voor een steekproefgrootheid G te observeren, onder de assumptie dat de nulhypothese waar is.

- **Het strikte onderscheid tussen H_0 niet dan wel verwerpen is misschien te scherp** (bv. p = 0.049). Hierbij wordt gesproken van een '**trend-effect**'.

- De p-waarden worden bekomen **op basis van theoretische verdelingen** die enkel geldig zijn wanneer aan alle assumpties is voldaan.

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**.
  → Allebei fout, de p-waarde bekomen we op voorwaarde dat de nulhypothese waar is, wat niet hetzelfde is.

- **In het algemeen: hoe kleiner de p-waarde, hoe groter het verschil tussen μ en $\mu 0$**.
  → Fout: enkel indien n en de variabiliteit constant blijven.

- **Een statistisch significant verschil tussen μ en $\mu 0$ is voor de theorie of voor de praktijk ook significant**.
  → Niet noodzakelijk. Het omgekeerde geldt ook niet (indien er geen significant verschil gevonden wordt, is er ook gene theoretisch of praktisch verschil.

- **Als ik geen significant verschil vind, is mijn onderzoek nutteloos**.
  → Fout: het is bijzonder informatie op voorwaarde dat er een gepast onderzoeksopzet werd gehanteerd.

DEEL 1: Beschrijvende statistiek

Hoofdstuk 2: Visualiseren van data

Illustratie in R:

- **Dimensie: dim (…)**

  Het 1ste getal geeft het aantal rijen weer.
  Het 2de getal geeft het aantal kolommen weer.
  Voorbeeld: dim (DataIAT) [1] 90 7

- **Bovenste rijen en kolommen: head (…)**

  Voorbeeld: head (DataIAT) 6 rijen en kolommen

- **Namen van variabelen: names (…)**
  Voorbeeld: names (DataIAT) “Geslacht”, “Leeftijd”, “Ras”

- **Waarde van variabelen: $-teken**
  Voorbeeld: DataIAT$Geslacht

- **Absolute frequenties: table (…)**
  Voorbeeld: table (DataIAT$Geslacht) man 36 vrouw 54

- **Relatieve frequenties: table (…) /n**
  Voorbeeld: table (DataIAT$Geslacht)/90 man 0.4 vrouw 0.6

- **Relatieve frequenties in percentage: table ((…)/n) * 100**
  Voorbeeld: (table (DataIAT$Geslacht)/90) * 100 man 40 vrouw 60

- **Cirkeldiagram: pie (…)**
  Voorbeeld: pie (rel.freq.perc.geslacht)

- **Staafdiagram: barplot (…)**
  Voorbeeld: barplot (rel.freq.geslacht).

- **Klassen: cut (…)**
  Voorbeeld: cut (DataIAT$Leeftijd)

- **Breedte van klassen: breaks**

- **Histogram: hist(…)**
  Voorbeeld: hist (DataIAT$Leeftijd)

Hoofdstuk 3: Samenvatten van data

Illustratie in R:

- **Rekenkundig gemiddelde: mean (…)**
  Voorbeeld: mean (DataIAT$Leeftijd) 31.31111

- **Harmonisch gemiddelde: harmonic.mean (…)**
  Voorbeeld: harmonic.mean (DataIAT$Leeftijd) 27.00577

- **Geometrisch gemiddelde: geometric.mean (…)**
  Voorbeeld: geometric.mean (DataIAT$Leeftijd) 28.9237

- **Mediaan: median (…)**

Median (DataIAT$Leeftijd) 26

- **Modus**:
  Frequenties bekijken via table (…).
  Hoogste frequentie is de modus.

- **Minimum: min (…)**
  Voorbeeld: min (DataIAT$Leeftijd) 16

- **Maximum: max (…)**
  Voorbeeld: max (DataIAT$Leeftijd) 74

- **Variatiebreedte: max (…) – min (…)**
  Voorbeeld: max (DataIAT$Leeftijd) – min (DataIAT$Leeftijd) 58

- **Gemiddelde absolute afwijking: aad (…)**
  Voorbeeld: aad(DataIAT$Leeftijd) 10.72741

- **Variantie: var (…)**
  Voorbeeld: var (DataIAT$Leeftijd) 181.565

- **Standaarddeviatie: sd (…)**
  Voorbeeld: sd (DataIAT$Leeftijd) 13.47461
  Voorbeeld: sqrt (var(DataIAT$Leeftijd)) 13.47461

- **Kwartielen (percentielen): quantile (…)**
  Voorbeeld: quantile (DataIAT$Leeftijd)

- **Interkwartielafstand: IQR (…)**
  Voorbeeld: IQR (DataIAT$Leeftijd) 16

- **Boxplot: boxplot (…)**
  Voorbeeld: boxplot (DataIAT$Leeftijd)

Hoofdstuk 4: Samenhang tussen 2 variabelen

- **Spreidingsdiagram: plot (…)**
  Voorbeeld: plot (DataIQ$Hersengrootte, DataIQ$VIQ)

- **Covariantie: cov (…)**
  Voorbeeld: cov (DataIQ$Hersengrootte, DataIQ$VIQ) 575.9717

- **Correlatiecoëfficiënt: cor (…)**
  Voorbeeld: cor (DataIQ$Hersengrootte, DataIQ$VIQ) 0.3374119

- **Kendall's T: cor (…) optie kendall**

Voorbeeld: cor (DataIQ$Hersengrootte, DataIQ$VIQ, method = “kendall”).

DEEL 2: Kansrekening

Hoofdstuk 5: De populatie en de verdelingsfuncties

- **Kansverdeling P (X = k): dbinom (k, N, p)**
  Voorbeeld: dbinom (0, 2, 0.25) 0.5625
- **Cumulatieve verdelingsfunctie P (X ≤ k): pbinom (k, N, p)**
  Voorbeeld: pbinom (0, 2, 0.25) 0.5625
- **Kansen P (X ≤ x) voor standaardnormale verdeling: pnorm (x)**
  Voorbeeld: pnorm (1.55) 0.9394292
- **Kansdichtheid fx (x): dnorm (x)**
  Voorbeeld: dnorm (1.55) 0.120009
- **Kansen P (Y ≤ y) voor variabele Y ~ X^2_k: pchisq (y, k)**
  Voorbeeld: pchisq (16.93, 28)0.05004119
- **Kansen P (T ≤ t) voor variabele T ~ t_k: pt (t, k)**
  Voorbeeld: pt (1, 2) 0.7886751

DEEL 3: Inductieve statistiek

Hoofdstuk 7: Betrouwbaarheidsintervallen en statistische toetsen voor het

- **Kansen P (Z ≤ z) voor standaardnormale verdeling links: qnorm (…)**
  Voorbeeld: qnorm (0.025)
- **Kansen P (Z ≤ z) voor standaardnormale verdeling rechts: qnorm (…) lower.tail = false**.
  Voorbeeld: qnorm (0.025) lower tails = false

- **Kwantiel van een t-verdeling: qt (…) vrijheidsgraden** (oppervlakte links)
- **Kritische waarde van t-verdeling: t.waarde ← qt (…)** (oppervlakte links)

- **Betrouwbaarheidsinterval**:

  Ondergrens ← gem – t.waarde*st.dev/sqrt(n).

  Bovengrens ← gem + t.waarde*st.dev/sqrt(n).

- **t-toets voor één steekproef, waarbij mu (…) staat voor $\mu 0$: t.test (data, mu (…))**.

  Voorbeeld: t.test(IQ, mu = 115).

- **De toetsingsgrootheid g: t = …**

  Voorbeeld: t = -0.9709.

- **Enkelzijdige p-waarde: pt (…) (links)**

  Voorbeeld: pt (-0.98, 30-1) 0.1675957

- **Tweezijdige p-waarde: 2*pt (…)**

  Voorbeeld: 2*pt (-0.98, 30-1) 0.3351915

- **t-toets (linkszijdig) voor één steekproef: t.test (data, mu (…), alternative = "less")**

  Voorbeeld: t.test (IQ, mu = 115, alternative = "less").

- **t-toets (rechtszijdig) voor één steekproef: t.test (data, mu (…), alternative = "greater")**

  Voorbeeld: t.test (IQ, mu = 115, alternative = "greater").