Statistiek

1 Wat is statistiek en opstart PASW

1.1 Wat is beschrijvende statistiek

Er zijn twee typen van statistiek. Deze zijn een ondersteuning voor wetenschappelijk onderzoek (data verzamelen en verwerken). De twee typen liggen in elkaars verlengde.

- **Beschrijvende statistiek (jaar 1)**
  - verzamelen van gegevens bij een steekproef
  - samenvatten van deze gegevens
    - welk zijn de kengetallen?
    - middels een grafische voorstelling?
  - analyseren van de resultaten
    - bestaat er een samenhang?
  - Kortom een overzicht van de resultaten
- **Inductieve statistiek (jaar 2)**
  - Wat betekenen deze resultaten in relatie tot de populatie?
  - schattingsprobleem: hoe kunnen we μ schatten op grond van het steekproefgemiddelde?

Beiden bieden methodologische ondersteuning van onderzoek.
Parameters: kengetallen van een populatie.
Steekproefstatistieken: kengetallen van de steekproef zelf.

1.2 Het onderzoek

- Het gaat over onderzoek dat op basis van *waarnemingen* probeert ware en *algemene uitspraken* te doen over de *werkelijkheid*. (Brinkman, 2006)
- Een uitspraak is een bewering waarin een of meerdere objecten een eigenschap wordt toegeschreven.
- Voorbeelden van uitspraken:
  - Jan is ziek
  - Assepoester is lang ongelukkig geweest
  - Kabouters zijn kleiner dan mensen
  - Bomen hebben een stam
  - Mannen zijn gemiddeld intelligenter dan vrouwen
- Wanneer spreken we van wetenschappelijk verantwoord onderzoek?
  - **Objectiviteit:** als het een objectief onderzoek is
  - **Controleerbaarheid:** als het onderzoek en de resultaten controleerbaar is.
  - **Herhaalbaarheid:** als het onderzoek herhaalbaar is.
  - **Systematiek:** als er een systematiek in het onderzoek zit.
- Er zijn verschillende typen van uitspraken:

  - **Deterministische uitspraken:** wet van de zwaartekracht
  - **Probabilistische uitspraken:** frustratie bevordert agressie.

- Wat is een variabele?
  - Is een eigenschap/kenmerk van een onderzoekseenheid (persoon/huishouden…) waarin mensen verschillen.
    - *Bv.* geslacht
  - Kan diverse waarden (uitkomsten) aannemen. Mensen verschillen op het vlak van deze eigenschappen
    - *Bv.* man/vrouw
  - Tegengestelde van een constante
  - Men gaat de samenhang hiertussen onderzoeken.
  - **Onafhankelijke variabele:** verschillen in deze variabelen worden gezien als oorzaak (?) van verschillen in de afhankelijke variabele. Deze ligt meestal vast, er is een onrechtstreekse interesse, wegens de beïnvloeidng zelf.
  - **Afhankelijke variabele = variabele ter studie:** verschillen in deze variabelen worden gezien als gevolg (?) van verschillen in de onafhankelijke variabele. Rechtstreekse interesse.
    - **IQ (onafhankelijke) te maken met schooluitslag (afhankelijke) ?**
  - Er zijn diverse niveaus van meting (zie verder)
    - Categorieën
    - Meetwaarden

- Twee typen van onderzoek:
  - **Het experiment:** Doelbewust worden één of enkele variabelen gemanipuleerd en we onderzoeken de effecten hiervan op de afhankelijke variabele. Dit heeft ook een beperkt aantal proefpersonen.
    - *Bv.* welk is de relatie tussen de attitude en het gedrag? Het experiment van de verloren brief (Nuttin en Beckers)
  - **Het veldonderzoek/ surveyonderzoek/ enquêteonderzoek:** Hier worden geen variabelen gemanipuleerd.
    - *Bv.* hebben babyboomers en busters een andere levensstijl en koopgedrag?

1.3 Fasen in een onderzoeksproces

- De vraagstelling
- De operationalisering: hoe doen ?
- Steekproefopzet: wie? Selecteren of nadenken?
- Verzamelen van de gegevens
- Beschrijven van de resultaten
- Formulering van conclusies/rapportage

- = inleiding, methode, resultaten en discussie.

1.3.1 De vraagstelling

- De vraagstelling vloeit voort uit de theorie ofwel uit een concreet problee
  - **Vraag naar secundaire gegevens:** wat hebben andere onderzoekers reeds vastgesteld en wat is de waarde hiervan?
  - **Vraag naar primaire gegevens:** via eigen experiment of veldonderzoek
- Relatie tussen milieubesef (als attitude) en het milieuvriendelijk consumentengedrag (effectief toepassen van)?
  - **Bv.** besef van de opwarming van de aarde en het gebruik van een terreinwagen.
  - **Voorbeeld van een experiment:**
    - Nuttin stelt zich de vraag wat het verband is tussen de attitude en het gedrag door het experiment van de verloren brief.
    - *Vraagstelling:* zullen de externe omstandigheden bij het vinden van de brief een verschil in reacties veroorzaken?
  - **Voorbeeld van een veldonderzoek:**
    - *Vraagstelling*:Bestaat er een verschil in levensstijl en koopgedrag bij de busters versus de babyboomers?
    - *Onderzoekshypothese*:Busters en babyboomers vertonen een verschillende levensstijl en koopgedrag
- Het verzamelen van de gegevens behoort tot de beschrijvende statistiek.
- De analyse van de resultaten tot de inductieve statistiek. Wat betekenen deze resultaten voor de populatie?
  - Hierbij werken we met twee tegengestelde veronderstellingen.

- Onderzoekshypothese wordt gesteld in termen van meetbare kenmerken, d.i. variabelen.
  - **Afhankelijke variabele:** de verschillen in deze variabele dienen we te verklaren (bv. vatbaarheid voor dyslexie)
  - **Onafhankelijke variabele**: deze verschillen kunnen een verklaring bieden. (bv. geslacht)
  - Hebben verschillen in de onafhankelijke variabelen effect op de verschillen in de afhankelijke variabele?

- **Gebruik een vraagvorm:**
  - *Bv.* Niet: het eetgedrag van jongeren
- **Specificatie van de begrippen:**
  - *Bv.* Niet: wat is het medicijngebruik in de psychiatrie?
  - *Beter:* wat is het % van de in 2009 opgenomen psychiatrische patiënten in Vlaanderen die gedurende de eerste maand van de

observatieperiode ten minste dagelijks een antidepressivum voorgeschreven kregen?

- **Geen oordelende vragen:**
  - *Bv.* Niet: zijn er voldoende psychiaters in Vlaanderen?
  - *Bv.* Niet: hoeveel % van de jongeren eet gezond?
- **Een rijtje in plaats van een volzin; Hoofdvraag en deelvragen :**
  - *Bv.* Hebben kinderen van handarbeiders lagere schoolcijfers in het basisonderwijs dan deze van hoofdarbeiders? Krijgen deze kinderen als ze dezelfde cijfers behalen een ander advies?

- **Drie typen van vragen:**
  - *Het voorkomen van iets:* Bv. hoeveel % van de Vlamingen is depressief?
  - *De verschillen tussen groepen:* Bv. zijn vrouwen meer depressief dan mannen?
  - *De samenhang:* Bv. bestaat er een samenhang tussen de leefsituatie en al dan niet depressief zijn?

1.3.2 De operationalisering

- Operationaliseren van een begrip tot meetbare variabele = hoe kunnen we dit begrip concreet meten?
  - Voorbeelden van operationalisering:
    - **Gemakkelijk:** geslacht, leeftijd, diploma
    - **Moeilijker:** intelligentie, aanleg voor wiskunde, arbeidstevredenheid

1.3.3 De steekproefopzet

- **Populatie (N):** alle individuen waar we een uitspraak over willen doen en die bijgevolg in aanmerking komen voor het onderzoek. Alle individuen testen?
- **Steekproef (n):** selectie van individuen uit de populatie. Je gaat niet iedereen ondervragen. Hoe selecteren?
  - Kunnen we alle personen ondervragen? Wellicht niet haalbaar tenzij:
    - Zeer kleine populaties: Wie rijdt in Vlaanderen met een Ferrari? Dit is wellicht een kleine groep van mensen die we kunnen opsporen.
    - Ofwel uit noodzaak: bv. Bij medewerkerstevredenheidsonderzoek
  - We spreken van een census.
  - In andere gevallen gebruiken we een steekproef. We selecteren een aantal proefpersonen en trachten op grond van de gegevens iets te zeggen over de populatie.

- Waarom een steekproef?
  - Is sneller af te handelen
  - Is goedkoper
- Hoe moeten we een steekproef trekken? Twee typen van steekproeftrekking
  - **Aselect (at random):** elk individu van de populatie heeft evenveel kans om in de steekproef terecht te komen.
    - *Vereisten:* Aselecte opzet vereist een lijst van de deelnemers in de populatie = steekproefkader (hoe?); elke persoon krijgt een nummer. De computer gaat at random nummers trekken.
    - *Gevolgen:* op grond van gegevens van dergelijke steekproef kan ik iets vertellen over de populatie. Schatting van de parameters van de populatie, op grond van resultaten van de steekproef.
    - *Voordelen:*
      - generalisering naar de populatie is mogelijk
      - veel statistische technieken zijn mogelijk
      - hoeveel proefpersonen?
      - Je kan makkelijk een schatting maken door aantal proefpersonen, de parameters zijn makkelijk inschatbaar.
    - *Nadelen:*
      - Levert geografisch verspreide individuen op met bijkomende kosten en tijdsinvestering
      - Kan wel of niet leiden tot representatieve steekproef. Zeker bij kleine steekproeven kan representativiteit een probleem geven.
      - De samenstelling van de steekproef moet gelijk lopen met de populatie (alleen zo is ze representatief)
    - *At random sampling:*
      - <u>Volledig aselecte steekproef:</u> computer kiest aantal nummers van personen.
      - <u>Systematische aselecte steekproef:</u> kies een (willekeurig) eerste persoon, en de volgende ppn. heeft telkens een nummer dat x aantal hoger ligt. Je kiest 1 willekeurig nr en per zoveel/interval verdergaan.
      - <u>Gestratificeerde steekproef:</u> de populatie wordt verdeeld in deelpopulaties (= strata) en uit elke deelpopulatie trekken we aselect een steekproef. (voorbeeld van strata: ASO, TSO, BSO) Hoe is de samenstelling in de populatie, zijn er subgroepen?
        - Proportionele(eisen representatieve steekproef) of disproportioneel(100 A, 100 T en 100 B) getrokken steekproef

- Clustersteekproef: de populatie wordt ingedeeld in subgroepen die zo heterogeen mogelijk zijn. Vervolgens worden een aantal subgroepen geselecteerd en elk individu van de geselecteerde groepen wordt bevraagd. Dit zijn groepen met interne heterogene subgroepen.
    - Bv. de behoefte aan schoolmateriaal van middelbare scholieren. We hebben een lijst van secundaire scholen en kiezen (at random) een aantal hiervan en zullen vervolgens alle leerlingen van de gekozen scholen bevragen.
- Getrapte steekproef: dwz een combinatie van voorgaande procedures wordt toegepast.

– **Niet aselecte steekproef:** elk individu heeft niet evenveel kans om in de steekproef terecht te komen.
- Geschikt voor verkennend onderzoek
- Geschikt om meetinstrumenten te testen
- Goedkoop en snel
- Veel gebruikt in de psychologie (bachelors en masters)
- Maar: resultaten kunnen niet veralgemeend worden naar de populatie.
- Exploratief onderzoek doen.
- Niet aselect:
    - Convenience sampling: kies de individuen die 'voor het grijpen liggen' de eerste 100 willekeurig.
    - Judgement sampling: kies deze individuen die als bevoorrechte getuigen kunnen fungeren, bv. groep van zware gebruikers. Specialisten in een bepaald gebied.
    - Snowball sampling: de eerste respondent levert volgende proefpersoon op, enz… Te gebruiken bij 'moeilijk vindbare' of toegangkelijke doelgroepen
    - Quota sampling: verdeel de populatie in een aantal subgroepen (= strata) en kies uit elke subgroep willekeurig een aantal proefpersonen. M.a.w. uit elke subgroep trekken we een convenience sampling. Belangrijk voor het representatief karakter. Men heeft oog voor het feit dat de samenstelling hetzelfde is = niet aselect.
    - Random Walk: de onderzoeker selecteert proefpersonen volgens een vooraf vastgesteld wandeltraject. Toeval bij de keuze van proefpersonen speelt een rol. Deze procedure benadert het beste de aselecte opzet en wordt gebruikt wanneer een steekproefkader niet voor handen is. =toevalsteekproef

- Vb: stratenplan maken en dat volgen, het is per toeval er is geen lijst.

1.3.4 Het verzamelen van gegevens

- Diverse methoden voor verrichten van metingen:
  - Vragenlijst
  - Test
  - ….
- Deze gegevens worden systematisch weergegeven in een datamatrix in PASW

- **De datamatrix:**
  - Gebruik van Excel: zeer beperkte mogelijkheden voor inductieve statistiek
  - Gebruik van PASW: zeer handig. Meest gebruikte sofware voor statistische analyse.
    - SPSS: Statistical Package for the Social Science
    - PASW: Predictive Analytics SoftWare

- **In PASW:**
  - *Twee windows:*
    - Data Editor
    - Viewer (= uitkomst)
  - *Twee tabs:*
    - Variable View (eerst invullen) → alle variabelen gaan omschrijven, naam, type, width, decimals, label, values, missing value, measure.
    - Data View (dan pas invullen)

- Variable view:
  - **Name:** naam van de variabele; dit gegeven komt in de data file boven de variabele staan, beperkt aantal karakters, geen spaties,…
  - **Type:** kies standaard voor ‘numeriek’ en u kunt voor deze variabele getallen invoeren. U kunt voor ‘string’(alles) kiezen, om bv. de namen van de proefpersonen in te geven, of om bemerkingen te noteren, bv. persoon geeuwde vaak, vermoeidheid? Alleen cijfers, string → niet meer rekenen
  - **Width:** hoeveel posities heeft u nodig voor deze variabele? Standaard is 8. Grote getallen moet je dus aanpassen.
  - **Decimals:** indien u geen gebruik maakt van decimale getallen kiest u voor nul.

  - **Label:** de naam van de variabele die u nu ingeeft zal gebruikt worden als titel boven een tabel, grafiek… minder beperkingen dan bij name. In jaren (leeftijd)
  - **Values:** hier geeft u aan welke getallen voor welke groepen staan. (niet te gebruiken bij scale metingen), bv. 1 is man en 2 staat voor vrouw. Alleen bij kwalitatieve variabelen, als er 2 verschillende soorten zijn.
  - **Missing value:** op welke wijze geeft u aan dat een persoon dit niet ingevuld heeft? U gebruikt hiervoor een getal dat onmogelijk als waarde gescoord kan worden. Open laten? Of user missing value vb: 9, een waarde die niet effectief voorkomt.
  - **Measure:** welk is het niveau van meting: nominal(subgroepen), ordinal (volgorde waarden), scale(meten).

Begin ALTIJD met het aanmaken van de Variable View

- Data view:
  - Resultaten van één proefpersoon vinden we in één en slechts één rij. Alle gegevens van één persoon staan in één rij. In één rij vinden we de gegevens van slechts één proefpersoon.
  - Resultaten van één variabele vinden we in één en slechts één kolom. Alle gegevens van één variabele vinden we in één kolom en de gegevens in één kolom hebben betrekking op slechts één variabele.
  - User missing value moet je zelf invullen.

- Via analyze kan je een opdracht geven aan PASW
- Bij frequencies kan je je frequenties bekijken in de viewer.
- Gegevens kunnen bewaard worden via File, Save as…
- Datafile krijgt automatisch extensie .sav
  - Bv. busters.sav
- Output gegevens krijgen een extensie .spv
  - Bv. busters.spv
- Gegevens kunnen opgevraagd worden via File, Open, Data…

1.3.5 De beschrijving van gegevens

- Datamatrix is onoverzichtelijk, hoe kunnen we de gegevens overzichtelijk voorstellen?
  - Middels een frequentieverdeling
  - Middels een grafische voorstelling van de resultaten
  - Middels bepaling van centrale tendens en variabiliteit.

1.3.6 Het formuleren van conclusies

- Toetsende/inductieve statistiek.

- Wat zeggen de gegevens uit de steekproef over de populatie? (Jaar II: inductieve statistiek) Hoe werkt dit?
  - Deductief versus inductief redenering
    - **Deductief redeneren:** vanuit een logische redenering, kunnen we conclusies opbouwen(zonder contact met werkelijkheid of ervaringen).
      - Alle mensen zijn sterfelijk
      - Jan is een mens
      - Dus jan is sterfelijk
      - Staat model voor de wiskunde
    - **Inductieve redenering:** vanuit de empirische observaties trachten we een wet te formuleren.
- Hume ziet een probleem. We kunnen niet alle mensen onderzoeken om te besluiten dat mensen sterfelijk zijn.

- **Statistische / inductieve redenering**
  - Twee tegengestelde beweringen/hypothesen worden t.o.v. mekaar afgewogen:
    - mensen zijn onsterfelijk (nulhypothese) (tegengestelde van verwachting onderuit halen)
    - mensen zijn sterfelijk (alternatieve hypothese) (verwachting aantonen)
    - bewijzen dat het andere niet kan.
    - Indien ik uit de empirie gegevens vind die de eerste stelling onderuit halen, is de tweede stelling van toepassing. Inductieve statistiek (Jaar II)

- Uitgewerkte voorbeelden
  - Experiment van de verloren brief (Nuttin en Beckers) → variabelen beïnvloed.
    - Factorieel design → 2X2X2X3 opzet (niveau, variabele) Meerdere onafhankelijke variabelen, alle mogelijke combinaties komen voor ←→fractioneel design.
  - Veldonderzoek Busters versus babyboomers (Apers, et al.) variabelen niet beïnvloed.

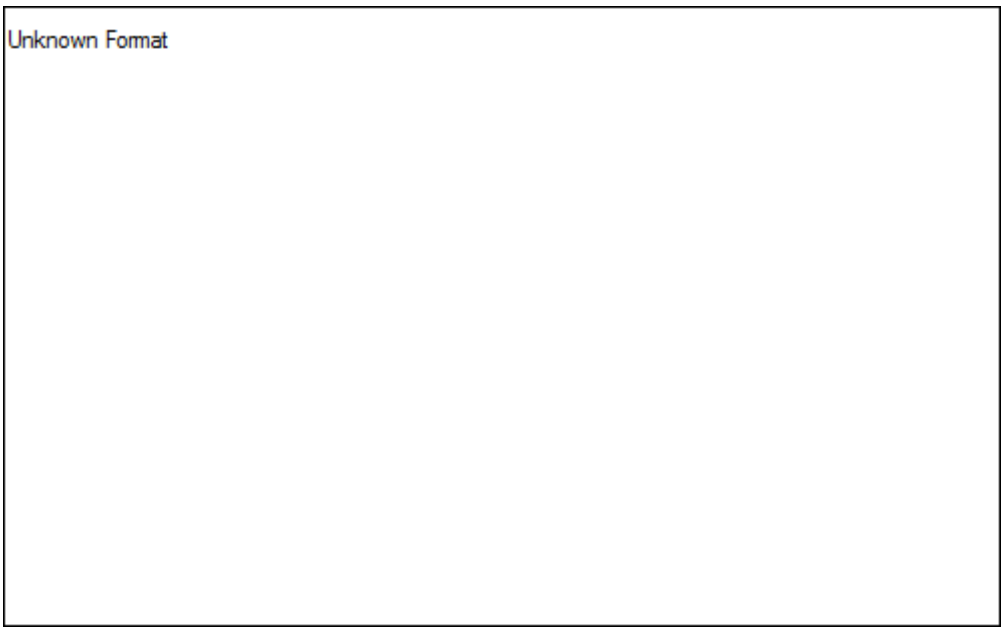
2 Operationaliseren en meten & PASW Transform.

- Transform = de transformatie van variabelen
- De vooropgestelde hypothesen bevatten begrippen. Hoe gaan we deze meten? = operationaliseren.

- bv. Hoe meten we de attitude t.o.v. de therapie?
- bv. Hoe meten we de vooruitgang in therapie?
- bv. Hoe meten we de intelligentie?

2.1 Variabelen

- = zijn kenmerken van de proefpersonen
  - bv. geslacht
- Variabelen kunnen diverse waarden aannemen
- Waarden zijn de individuele uitslagen op een variabele,
  - bv. vrouw,
  - = scores



- **Kwalitatief:** opdelen in subgroepen → geen getalsbetekenis
  - *Vb:* wel of niet in therapie
  - *Vb:* militaire rang
- **Kwantitatief:** waarden van deze variabelen = cijfers
  - **Discreet:** beperkt en geen tussenwaarden
    - *Vb:* # kinderen
    - *Vb:* aantal inschrijvingen voor een cursus
    - *Vb:* aantal kinderen in de huishouding
  - **Continu:** gradaties en onderafdelingen, met oneindig veel tussenschakeringen.
    - *Vb:* lengte, loon, …
    - *Vb:* lichaamslengte
    - *Vb:* tijd nodig om het examen op te lossen

2.2 Meetniveaus

2.2.1 Nominaal

- Alle categorische variabelen zijn nominaal
  - bv. variabele geslacht kent twee waarden: jongen en meisje.
- Kunnen getallen aan gekoppeld worden, maar deze hebben geen numerieke betekenis.
  - waarden kennen geen rangorde
  - er is geen meeteenheid (geen afstand)
  - er is geen nulpunt
- bv. geslacht is een nominale variabele
- alleen proefgroep in subgroepen indelen.
- Speciaal geval van nominale schaal: dichotomie: dwz er zijn slechts twee niveaus mogelijk voor deze variabele
  - bv. Volgt u logopedie? Ja of neen?
- De onafhankelijke variabelen zijn vaak nominaal van niveau.
  - De afhankelijke zoveel mogelijk scale maken
- Voorbeelden:
  - Geslacht:
    - Man
    - Vrouw
  - Burgerlijke stand
    - Alleenstaand
    - Gehuwd
    - Gescheiden
    - weduwe/weduwnaar
    - …….
  - Nationaliteit
    - Belg
    - Nederlander
    - Duitser
    - …….
- Wanneer voldoende alternatieven?
  - Ze moeten elkaar uitsluiten
  - Iedereen kan zijn antwoord terugvinden = exhaustief alternatieven hebben.
  - 2 antwoorden aanduidbaar is niet mutueel exclusief.

2.2.2 Ordinaal

- Categorieën hebben een bepaalde volgorde, kunnen geordend worden.
  - Afstand ertussen is niet te meten.

- Er kunnen getallen aan gekoppeld worden, maar het verschil tussen 2 opeenvolgende waarde heeft geen betekenis, enkel de volgorde.
  - er is wel een rangorde
  - er is geen meeteenheid
  - er is geen nulpunt
- **bv.** hoogst behaalde diploma is ordinale variabele
  - geen/lager secundair/hoger secundair/hoger onderwijs/unief
- **Voorbeeld:**
  - Welk is uw evaluatie van deze cursus?
    - zeer slecht
    - slecht
    - neutraal
    - goed
    - zeer goed
  - Welk is uw hoogst verworven diploma?
    - lager onderwijs
    - lager secundair
    - hoger secundair
    - bachelor
    - master

2.2.3 Interval

- Rangorde is belangrijk.
- Er is info over afstand tss opeenvolgende waarden, die is gelijk en zo is er een meetpunt.
- Maar geen verhouding door geen nulpunt.
- Verschil tussen twee opeenvolgende waarden is gelijk; er is een meeteenheid
- Geen absoluut nulpunt.
  - bv. IQ meting
  - volgorde is belangrijk; en er is een meeteenheid, maar IQ 120 is niet dubbel zo slim als IQ 60
- **Voorbeelden**:
  - Temperatuur in graden Celsius
    - het verschil tussen 6° en 30° is 24°
    - maar 30° is niet het vijfvoud van 6°
  - De jaartelling
  - IQ coëfficiënt

2.2.4 Ratio

- Rangorde is belangrijk
- Er zijn verhoudingen mogelijk
- Er is een meeteenheid

- Er is een absoluut nulpunt
  - **Bv.** Lichaamslengte, gewicht, tijd om de test op te lossen, enz..
  - **Bv.** Iemand van 1,8 meter is dubbel zo groot als iemand van 0,9 meter.

2.3 Betekenis van meetniveau

- In de psychologie vooral nominale, ordinale en interval variabelen, weinig ratio schalen.
  - Onderscheid tussen interval en niet interval is het belangrijkste onderscheid, met oog op de analyse van de gegevens.
- PASW maakt geen onderscheid tussen interval en ratio: nominal(subgroepen), ordinal(volgorde) en scale(meten)
  - Kan zinloze waarden berekenen…
- Een variabele kan van aard veranderen al naargelang de vraagstelling,
  - bv. de leeftijd
    - Hoe oud bent u? …… Dit is ratio niveau
    - In welk jaar bent u geboren?.... Dit is interval
    - Hoe oud bent u, maak uw keuze:
      - 20 à 30 jaar
      - 31 à 40 jaar
      - 41 à 50 jaar
      - ouder dan 51….. Dit is ordinaal niveau
- Naarmate het meetniveau hoger is kunnen we meer bewerkingen uitvoeren
- Zorg voor variabelen – indien mogelijk - met een zo hoog mogelijk meetniveau, bv. leeftijd.
  - Altijd zo ene hoog mogelijk niveau nemen, want de vraagstelling beïnvloedt het niveau.
- Diverse variabelen met intervalniveau kunnen opgeteld worden… Zorg ervoor dat de afhankelijke variabele zo veel mogelijk op interval niveau gemeten wordt. Dit heeft eveneens consequenties voor de analyse van de gegevens
- De vraag naar houdingen, overtuigingen, gevoelens, etc… kan het best gebeuren aan de hand van een likertschaal(ipv ja of nee):
  - Wel genoeg proefpersonen en subgroepen nemen en zoveel mogelijk op scale.
    1. helemaal akkoord
    2. akkoord
    3. neutraal
    4. niet akkoord
    5. helemaal niet akkoord
- De vraag naar feitelijkheden, bv. Bent u een man/vrouw? Woont u in Antwerpen? kan uiteraard niet met een likertschaal bevraagd worden.

2.4 Betrouwbaarheid en validiteit

- Betrouwbaarheid: meet de test iets? Heeft te maken met de stabiliteit van de meting:
  - Hertesting
  - Halvering
  - parallel vorm
  - interne homogeniteit (Chronbach's Alpha)
    - test in zoveel onderdelen als er items zijn.
    - CBA: betrouwbaarheid of dat items zelfde meten bij personen.
- Validiteit: meet de test wat hij behoort te meten?
  Welk is de relatie van de testuitslag met een andere meting van dit begrip (bv. relatie intelligentie en schooluitslag)
- Betrouwbaarheid is een voorwaarde voor validiteit

2.5 Voorbereidende bewerkingen in PASW

- U kunt variabelen bewerken via het menu Transform
  - Compute maakt berekeningen op variabelen, bv. optellen, keer, etc…
  - via Recode kunt u variabelen hercoderen (via intervals)
  - via Count u bepaalde scores tellen per persoon. (hoe vaka dit, hoe vaak dat bepaald antwoord)

2.5.1 Recode

- Op deze wijze ontstaat een nieuwe variabele met als titel 'rec19L'. Deze variabele is precies het omgekeerde van item 19L.
- Tip: maak gebruik van recode 'into a different var';
- Voor de totaalscore van gezondheidsbesef moeten we gebruik maken van 'rec19L' en niet van het item 19L
- Recode IF kan ook gebruikt worden. U definieert een conditie waaraan voldaan moet worden om de recode uit te voeren.
- Men gaat 1 waarde hercoderen (nieuwe naam), code meegeven, hoe te coderen, omgekeerde variabele
  - Links oud, rechts nieuw.

2.5.2 Compute

- Voor het samentellen van items tot een schaal gebruikt u compute. Opdracht tel de items van de schaal gezondheidsbesef samen in het bestand busters.
  - Rekenkundige bewerking, scores met elkaar optellen.
    - ik voel me graag fit
    - Ik doe geregeld aan lichaamsbeweging

- Ik vind mijn gezondheid belangrijk
- Ik eet regelmatig groenten en fruit
- Ik neem een gezond en gevarieerd ontbijt
- Ik let op mijn lichaamsgewicht
- rec19L (Ik eet vaak vettig eten)

- We zouden ook gebruik kunnen maken van de som van de items.
- Op deze wijze ontstaat een nieuwe variabele 'gezondheidsbesef', waarmee we zullen verder werken, en niet meer met de afzonderlijke items.
  - Werken met functies, altijd uitslag, ook al zijn er missing values.
- U kunt ook gebruik maken van de mogelijkheid om een conditie aan te geven (compute IF) om deze compute uit te voeren
- Variabelen omkeren, 1e is pos beoordeelt, de 2e neg, je kan alleen pos en pos optellen, dus moet je de negatieve omzetten.
- Mean = gemiddelde, nodig voor de missing value.

2.5.3 Count

- Op deze wijze ontstaat een nieuwe variabele waarin het aantal missing values per persoon samengeteld wordt.
- Hoe vaak ene bepaalde waarde.
  - **Vb:** geen mening, ik weet het niet.
- U kunt hierbij ook gebruik maken van de optie Count IF.

2.6 Chronbach's Alpha

- Maat van betrouwbaarheid
- Allemaal bij elkaar zetten en dan berekenen
- 2 cijfers na komma → 6-9 omhoog, 1-4 beneden, 5 → afronden naar even getal.
- 70 is ideaal.

3 Frequentieverdelingen en PASW Descriptives

= tabel, overzicht scores visueel voorstellen.

3.1 De frequentietabel

- Datamatrix bevat de resultaten van het onderzoek; dit is onoverzichtelijk
- Een frequentieverdeling geeft een beter overzicht.

- Één dimensionale tabel versus meer dimensionale tabel
  - Univariabele tabel of contingentie- of kruistabel.

3.1.1 Één dimensionale tabel

- Verschillende waarden oplijsten, laten zien hoeveel keer het voorkomt.
- Frequenties en percentages.
- Frequentie is het aantal keren dat een bepaalde waarde voorkomt
- Soorten frequenties:
  - **Absolute frequentie:** hoeveel keer komt de waarde voor
  - **Relatieve frequentie (%):** procentuele aanwezigheid
  - **Absolute cumulatieve frequentie:** hoeveel uitslagen dit en lager
  - **Cumulatieve proportie (%):** hoeveel uitslagen dit en lager in %
    - = percentiel
    - Moeten beide ordinaal zijn, nominaal gaat niet.
  - (beide laatste worden niet gebruikt bij nominale waarden)

Geslacht

		Frequency	Percent
Valid	man	142	35,3
	vrouw	260	64,7
	Total	402	100,0

Lft	Freq	Prop.	Cum freq	Cum prop.
8	25	33%	25	33%
9	30	39%	55	72%
10	21	28%	76	100%

PAWS: prop, cum prop en freq.

3.1.2 Meer dimensionale tabel (zgn. kruistabel)

- Samenhang tussen 2 nominale variabelen
  - Hoe vaak is er een combinatie van de 2 waarden = frequentie.
- Bovensyte 2 rijen en 1e 2 kolommen = cellen, de rest zijn totalen
- = univariabele tabel.

Geslacht * nieuwleeftijd Crosstabulation

Count

		nieuwleeftijd		
		buster	boomer	Total
Geslacht	man	38	99	137
	vrouw	79	177	256
Total		117	276	393

3.2 Grafieken

- Nominale waarden: geen histogram
- Ordinale waarden
- Interval/ratio niveau(scale)
- Visueel voorstellen via grafiek, wel rekening houden met meetniveau
  - Grafische voorstelling hangt af van deze niveaus

3.2.1 Nominale waarden

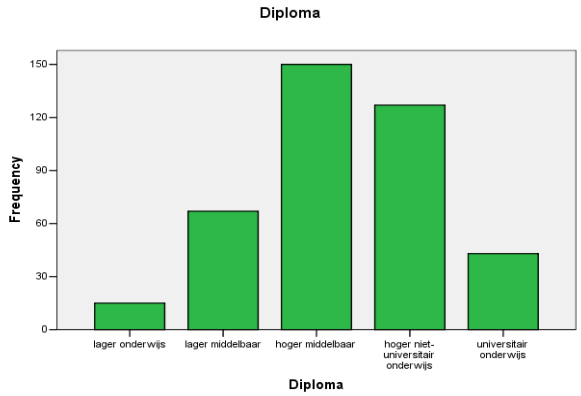
- Gebruik een taartdiagram
- Geen ordening
- Geen volgorde
- = pie chart
- = cirkeldiagram

3.2.2 Ordinale gegevens

- Wel volgorde
- Geen onderlinge afstand
- gebruik een staafdiagram = histogram, wel ruimte tussen de balken
- volgorde in categorieën
- = bar chart

Diploma

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	lager onderwijs	15	3,7	3,7	3,7
	lager middelbaar	67	16,7	16,7	20,4
	hoger middelbaar	150	37,3	37,3	57,7
	hoger niet-universitair onderwijs	127	31,6	31,6	89,3
	universitair onderwijs	43	10,7	10,7	100,0
	Total	402	100,0	100,0	



3.2.3 Scale

- Van datamatrix een histogram maken
- PASW biedt een overzicht van de resultaten, middels een frequentietabel.
- **Grafische voorstelling:** histogram
  - Uitslagen groeperen
  - Gegroepeerde frequentietabel → in interval's zetten.
- **Gegroepeerde frequentietabel:**
  - Enkel om de gegevens overzichtelijk voor te stellen;veel informatie gaat verloren
  - Voor de komst van SPSS gebruikelijke wijze van voorstelling; Hoeveel klassen? Turven van aantallen, enz… verwijzen we naar het pre-SPSS tijdperk.
    - Breedte klassen?
  - GEEN VERDERE ANALYSE VAN DE GEGEVENS AAN DE HAND VAN DERGELIJKE TABEL.
    - Alleen goed voor rapport, geen gemiddelde, mediaan, …
- **Frequentietabel en histogram met PASW**
  - PASW zal voor een frequentietabel niet automatisch een gegroepeerde tabel maken. Dit is wel mogelijk via recode van de gegevens. Eventueel een stem-and-leaf plot.
  - Als histogram zal PASW bij een grote diversiteit van waarden automatisch de waarden in klassen indelen
  - *Aanmaken tabel:* descriptive statistics→ frequencies →(display aanvinken) een of meerdere variabelen selecteren → output tabel
  - *Aanmaken kruistabel:* descriptive statistics→ crosstabs → rij en kolom kiezen → display clustered … aanvinken →grafische voorstelling → output kruistabel
  - *Aanmaken taartdiagram:* descriptive statistics→ frequencies → charts → pie chart → alles selecteren → output pie chart.

    - Alles is aanpasbaar, nominaal, geen absolute getallen, maar proporties.
- *Aanmaken staafdiagram:* descriptive statistics→ frequencies → charts → bar chart → alles selecteren → output bar chart.
    - Ordinaal
    - Absolute frequenties
- *Aanmaken histogram:* descriptive statistics→ frequencies → charts → histogram → (with normal curve →alles selecteren → output histogram.
    - Via mean → gem. Lijn → normaalverdeling.
- *Aanmaken stem-and-leaf plot:* descriptive statistics→ explore → toevoegen → alles in 1 x, display plots.
    - =stamdiagram; helft resultaten en voorbereidende bewerkingen, reeks gegevens, veel info.
    - Individuele scores grafisch voorstellen
    - Stem width geeft aan hoe breed
    - Leaf: observaties per regel, elk blad is 1 observatie
    - Berekeningen mogelijk → individuele gegevens.
    - Zelden in rapport, wel goed voor eerste exploratie en oriëntatie.
- *Aanmaken boxplot:* via explore of graphs → simple → 2 kiezen
    - Mooi voor bachelorproef
    - Meerdere subgroepen vergelijken
    - Outlier zijn de puntjes

- Interactieve grafiek → samenhang gegevens.
    - Hoe zou de verdeling zijn van de diploma's bij dames en heren in deze steekproef?
- *Aanmaken geclusterd staafdiagram:* graphs → legacy dialogs → bar → clustered → selecteren
    - 2 nominale variabelen
    - = geclusterd of gestapeld
    - Identieek zelfde infi, maar ander voorstellen (clustered of stacked)
    - Kan alletwee voor bachelor proef.
- *Selectie cases:* je kan cases aan en uit zetten, voor bepaalde criteria kan je er bepaalde selecteren.

3.3 Positie van een score in een verdeling van uitslagen.

- Het percentiel P van een ruwe score is het percentage metingen dat kleiner is (of gelijk aan) dan deze ruwe score.
    - = cumulatieve proportie
    - = of < aan/dan de score.
    - Zoveel % is < of = en zoveel % >
    - Dus hoeveel procent van de observaties ligt beneden deze score?

- *Voorbeeld:* op een taaltest behaalde Jan een score van 112/120. Is dat een goede score? Kijk hiervoor naar het percentiel.
  - Als 20% van de leerlingen een betere score behaalde, zeggen we dat de uitslag 112 het 80^{ste} percentiel is, ofwel $P_{80}=112$
  - Als 70% van de leerlingen een betere score behaalde, zeggen we dat deze uitslag het 30^{ste} percentiel is, ofwel $P_{30}=112$

- Cumulatief percent biedt inzicht in percentiel
  - Ertussen bestaat niet of komt niet voor, het kan wel ergens inzitten.
- Welk is het percentiel bij een bepaalde score? D.i. hoeveel procent van de observaties zijn lager (of gelijk aan) dan deze score?
- Welke score komt overeen met een bepaald percentiel? Beneden welke waarde situeren zich een bepaald percentage observaties?
  - Gebruik steeds de frequentietabel
  - GEEN berekeningen uit een gegroepeerde tabel.
- 100 % → 1 % = percentieel
- 100 % → 10 % = decentieel
- 100 % / 4 = kwartiel = grenzen boxplot.

- **Varianten van de percentielscore**
  - *Decielscore:* We verdelen de uitslagen in 10 delen, zodanig dat in elk onderdeel 10% van de observaties zich situeren;
    dus D_1 = Pc10, D_2 = Pc20, enz…
  - *Kwartielscores:* We verdelen de uitslagen in vier onderdelen, die elk 25% van de observaties bevatten.
    dus Q_1 = Pc25; Q_2 = Pc 50 en Q_3 = Pc75

4 Centrummaten & PASW Descriptives

4.1 De modus

- Is de waarde met de hoogste frequentie, komt het meeste voor
- ≠ de meerderheid
- Altijd berekenbaar.
  - **Bijvoorbeeld scores op een Likertschaal (1-5)** Ik vind de opwarming van de aarde een groot probleem (helemaal akk….. helemaal niet akk)
    - score frequentie

1 helemaal akk	13
2 akkoord	12
3 weet niet	3
5 helemaal niet akk	1

Welk is de modus? Score 1 ‘helemaal akkoord’

- Zal vooral gebruikt worden voor nominale waarden; maar kan in principe altijd bepaald worden. Is meteen duidelijk in de frequentietabel
- Meer dan één modus is mogelijk, bij een bimodale verdeling zijn er twee modi.
- Gebruikt weinig informatie uit de gegevens.

4.2 De mediaan

- De mediaan is de middelste waarde wanneer de observaties in volgorde van laag naar hoog zijn gezet. (niet mogelijk voor nominale waarden)
  - Alle waarden op ene rij, van klein naar groot..
  - Bij nominaal gaat dit niet, minstens ordinaal nodig
  - Precies in het midden
  - Weinig info over het geheel van de waarde.
- Bij een oneven aantal observaties precies de middelste, en bij een even aantal observaties het midden tussen de twee middelste scores;
- Komt dus overeen met percentiel 50.
  - Welk is de mediaanwaarde van 2, 4, 6, 8, 10? De mediaanwaarde is 6, als middelste waarde
  - Welk is de mediaanwaarde van 2, 4, 6, 7, 8, 10? De mediaan is 6,5 zijnde het midden tussen 6 en7.
  - Welk is de impact van een wijzing van de laatste observatie 10 in 20? Verandert hierdoor de mediaan? Neen, want ligt nog altijd in het midden
- Kan ook op de grens liggen
- Kan niet gebruikt worden bij nominale waarden;
- Is niet afhankelijk van extreem hoge of lage uitslagen.
  Gebruikt weinig info uit de gegevens;
- Kan gezien worden in vergelijking met het rekenkundig gemiddelde;
- Is gemakkelijk te begrijpen/uit te leggen/grafisch voor te stellen.
- Kan grafisch voorgesteld worden via een boxplot. PASW kan een verdeling van uitslagen voorstellen middels een boxplot. In dergelijke boxplot worden PC25, PC50 en PC75 grafisch voorgesteld middels een ‘doos’

4.3 Rekenkundig gemiddelde

- Het gemiddelde is de som van alle scores gedeeld door het aantal scores.
  $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$

- Is enkel mogelijk voor interval en ratio meetniveaus, bv. IQ, schooluitslagen, testuitslagen, leeftijd,…
  - Alleen met kwalitatieve variabelen (ratio en intverval)

  - Niet bij gegroepeerde frequentietabel!
- **Voorstelling van gemiddelde:**
  - In de steekproef: X
  - In de populatie: μ
  - Gem. steekproef basis μ en gem populatie willen we weten
  - *Vb:*

Score Frequentie
4 9
6 15
8 21
gemiddelde: (9*4 + 15*6 + 21*8)/45 = 6,53

4.3.1 Het gemiddelde bij een samengestelde steekproef:

- Veronderstel je beschikt over twee steekproeven n_1 en n_2 met een respectievelijk gemiddelde X_1 en X_2, welk is dan het zgn. gewogen gemiddelde? = rekenkundig gemiddelde van 2 groepen samen, uitschieters hebben hier ferme invloed op.

- Een analoge eigenschap voor de mediaan bestaat niet. Om de mediaan van de samengestelde steekproef te kennen, moet je alle metingen kennen
- = gewogen gemiddelde

4.3.2 Het getrimde gemiddelde:

- Het rekenkundig gemiddelde van het deel van de waarnemingsgetallen dat overblijft na weglating van de P% kleinste en P% grootste.
- **Eigenschappen:**
  - Som van de afwijkingen van de waarnemingsgetallen tot het rekenkundig gemiddelde is gelijk aan 0.

X_i	
18	18-14=4
13	13-14=-1
17	17-14=3
16	16-14=2
10	10-14=-4
09	9-14=-5
15	15-14=1
	SOM=0

  - Bij een lineaire transformatie van de scores, wordt het rekenkundig gemiddelde op dezelfde wijze getransformeerd, d.w.z. als je alle waarnemingsgetallen met b vermenigvuldigt en daar een constante a bijtelt, dan wordt het rekenkundig gemiddelde op dezelfde manier getransformeerd.
    - Dit heeft een invloed op het rekenkundig gemiddelde en niet op de vormverdeling.

  - Het rekenkundig gemiddelde van een aselecte steekproef is een zuivere schatter van het populatiegemiddelde (μ).
    - D.w.z. dat wanneer we van een oneindig aantal steekproeven (met hetzelfde aantal n) steeds het steekproefgemiddelde berekenen, het rekenkundig gemiddelde van alle steekproefgemiddelden gelijk is aan het populatiegemiddelde.
    - d.i.Centrale limietstelling → het gemiddelde van de gemiddelden
- Snel te berekenen en eenvoudig te begrijpen
- In dezelfde meeteenheid als de waarden
- Alle waarden worden bij de berekening betrokken
- Gevoelig voor extreme waarden
- Steeds berekenen bij interval en ratio waarden
- Eventueel vergelijken met mediaan

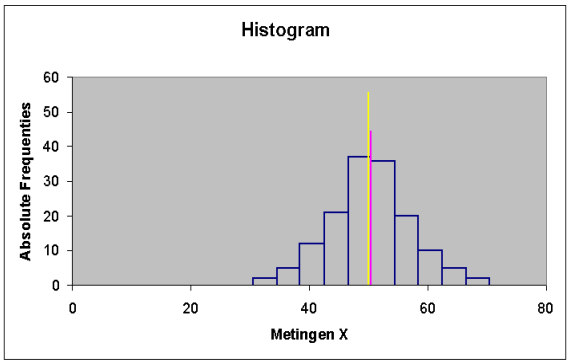
4.4 Gebruik van centrummaten

- **Modus:** bij nominale, ordinale, interval en ratio waarden;
  - Modus vooral bij nominale waarden
- **Mediaan:** bij ordinale, interval en ratio waarden;
- **Gemiddelde:** bij interval en ratio waarden.
- **Gemiddelde versus mediaan?**
  - Gemiddelde gebruikt meer informatie dan de mediaan; de mediaan gebruikt enkel de rangorde van de getallen, dus bij interval waarden….
  - Invloed van ‘uitbijters’/’’outliers’? Uitbijters hebben geen invloed op de mediaan, wel op het gemiddelde.
    - Bij de mogelijkheid van extreme waarden kan getrimde gemiddelde een oplossing bieden. Getrimde gemiddelden worden berekend zonder rekening te houden met bv. de 5% hoogste en 5% laagste waarden.
- **Gemiddelde versus mediaan:**
  - Het gemiddelde varieert minder van steekproef tot steekproef t.o.v. de mediaan.
  - Dus het gemiddelde wordt meer gebruikt in de toetsende statistiek om het centrum van de populatie te schatten.

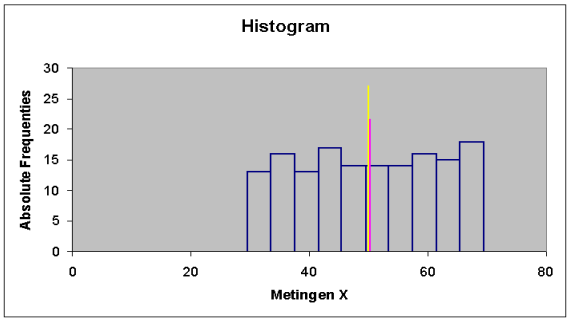
- Gemiddelde is algebraïsch aardiger. We kunnen gegevens van subgroepen samenvoegen om gewogen gemiddelde te berekenen, … dit kan niet bij een mediaan.
- Het gemiddelde verdient de voorkeur bij interval/ratio schalen.
- Onderlinge positie van gemiddelde en mediaan zegt iets over de mate van scheefheid van de verdeling.

4.5 Vergelijking van centrummaten

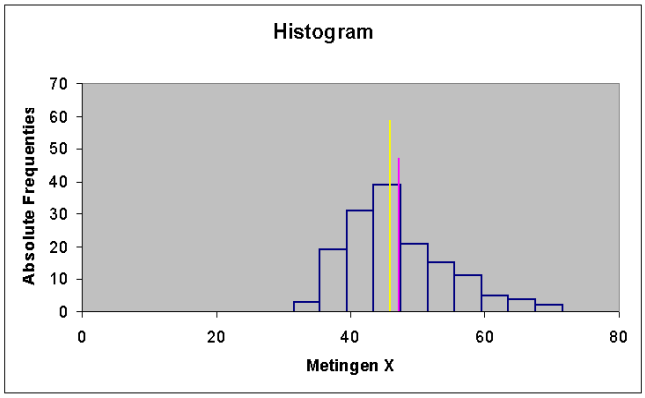
- **Voor symmetrische verdelingen:**
  - Bij een normaalachtige verdeling is MO=Me=Gem.
    - Verdeling van de IQ's



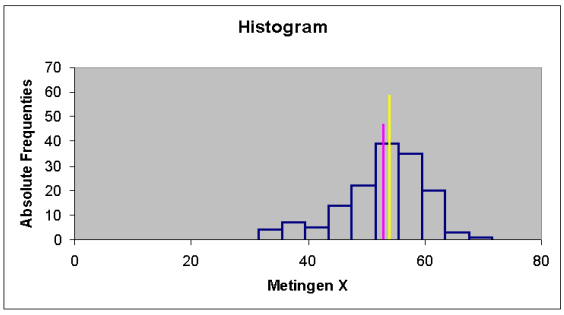
  - Bij een uniforme verdeling is Me=gemid. Modus?
    - Verdeling van leeftijd, van 20 t/m 50 jaar



- **Bij asymmetrische verdelingen(hangt af van volgorde med., mod. en gem.):**
  - Voor een rechts scheve verdeling (scheefheid pos.) Mo<Me<gemid
    - Rechtscheef is concentratie lage waarden en beperkte hoge, eerst modus dan gem.
    - **Bv.** Verdeling van inkomens



- Voor een links scheve verdeling (neg. Scheefheid) mo>me>gemid
  - Linksscheef is concentratie van hgoe waarde, beperkte lage, gem. dan modus.
  - **Bv.** Een gemakkelijke toets



- **Besluit:**
  1. De vorm van de verdeling heeft invloed op de onderlinge positie van de centrummaten.
  2. Indien mogelijk maak gebruik van het rekenkundig gemiddelde als maat van centrale tendens.
- De uitersten hebben geen invloed op de modus.

4.6 PASW

Analyze → descirptive statistics → frequencies → format bepaalde var → Mean, mode, median → output.

PASW en het rekenkundig gemiddelde:

- Om subgroepen te vergelijken maken we vaak gebruik van het rekenkundig gemiddelde.
- Analyze → compare means → Means → 2 kiezen → output.

5 Spreidingsmaten & PASW Descriptives

- **Spreidingsmaten:**
  - Naast de centrale tendens vormt de mate van verscheidenheid van de uitslagen een belangrijk gegeven.
    - Hoe situeren de uitslagen zich rond het gemiddelde?

- Spreidingsmaten worden enkel bepaald bij interval en ratio schalen (bij ordinale maten ook wel interkwartielbereik).

5.1 Variatiebreedte of bereik

- De range of het bereik is de hoogste minus de laagste waarde.
- =variatiebreedte
  - Verschil tussen hoogste en laagste uitslag.
- **Voorwaarde:**
  - Minimaal ordinaal niveau
- **Enkele voorbeelden ter illustratie:**

Voorbeeld 1. Ordinaal niveau

Oordeel	Absolute frequentie
Zeer slecht	15
Slecht	20
Neutraal	18
Goed	10
Zeer goed	00
TOTAAL	**63**

Bereik = 'Zeer slecht tot goed'

Voorbeeld 2. Interval/ratio niveau

Reeks 1	Reeks 2
9	2
11	10
11	11
12	11
12	12
13	15
14	16
15	17
16	17
17	24

$\text{Bereik} = 17 - 9 = 8$

$\text{Bereik} = 24 - 2 = 22$

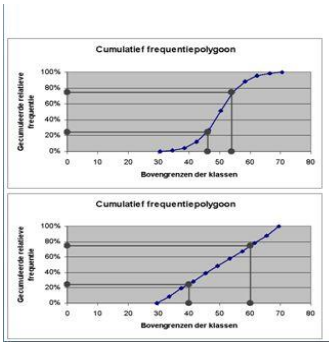
Niet berekenbaar, maar bereik wel te geven

- **Probleem:**
  - Wordt enkel beïnvloed door twee (extreme) uitslagen; we houden geen rekening met de tussen liggende observaties.
  - Outliers te veel invloed op bereik.
- **Zinvol:**
  - Outliers kunnen eruit gehaald worden, maar wordt weinig gebruikt.

5.2 Interkwartielafstand

- Het gebied op de X-as waartussen de middelste helft (50%) van alle waarnemingen valt, is de interkwartielafstand (Q).
  - Outliers spelen hier geen rol
  - Wel rekening houden met tussenliggende observaties
- Welk is het verschil tussen Q_3 en Q_1?

- Biedt een goed inzicht in de variabiliteit van de uitslagen.
- **Vereiste:**
  - Min ordivaal niveau, vooral bij dat niveau gebruikt.
- Speelt geen rol in de inductieve statistiek, enkel in beschrijvende.

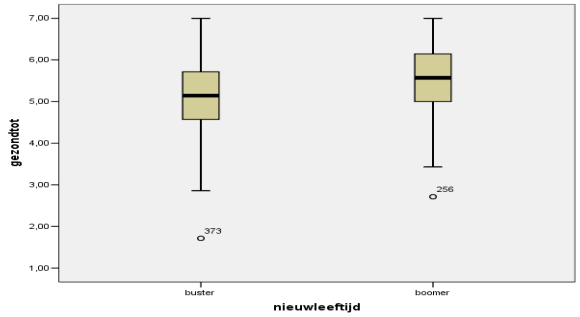


Bij een normaal verdeling
Me = 50
Q = 8

Bij een uniforme verdeling
Me = 50
Q = 20

percentielen vs grenzen van de klassen (= interkwartielafstand) 1e = geconcentreerde verdeling, 2e = niet geconcentreerde verdeling.

- Grafische voorstelling interkwartielafstand
  - De boxplot biedt een grafische voorstelling van alle observaties. De ‘doos’ van deze boxplot bevat de 50% middelste observaties en geeft derhalve een beeld van het interkwartielbereik



- Rekenkundig gemiddelde voor centrum van de scores
- Deviatiescores → afwijkingscores: verschil in de uitslag en het gem.
- Daar gemiddelde vn bepalen

5.3 Gemiddelde afwijkingscore

- In hoeverre wijkt elke individuele uitslag af van het rekenkundig gemiddelde?
  - Negatieve en positieve afwijkingsscores.

  - **Eigenschap:** De som van de afwijkingscores bedraagt 0; gemiddelde afwijkingscore dus ook

  - Dus deze maat kunnen we niet gebruiken als maat van variabiliteit, door het rekenkundig gemiddeld wordt dit bekomen.

Voorbeeld

Reeks 1	Reeks 2	$X_i - \bar{X}_1$	$X_i - \bar{X}_2$
9	2	-4	-11
11	5	-2	-8
11	7	-2	-6
12	9	-1	-4
12	10	-1	-3
13	15	0	2
14	16	1	3
15	20	2	7
16	22	3	9
17	24	4	11

$$\sum_{i=1}^{k}\left(X_i - \bar{X}\right) = 0$$

$$\bar{X}_1 = 13$$

$$\bar{X}_2 = 13$$

2e kolom = deviatiescores

5.4 Gemiddelde absolute afwijking

De gemiddelde absolute afwijking komt overeen met het rekenkundig gemiddelde van de absolute (geen rekening houden met het teken) waarde van de afwijking van elke score t.o.v. het gemiddelde.

- Absolute scores gebruiken → alleen naar verschil kijken = goede maat voor variabiliteit.

- Geeft een goed idee van de variabiliteit van de scores.
- Maar deze maat wordt in de praktijk niet gebruikt.
- Ligt aan de basis van het begrip variantie.

Voorbeeld (ter illustratie)

Reeks 1	Reeks 2	$X_i - \bar{X}_1$	$X_i - \bar{X}_2$	$\lvert X_i - \bar{X}_1 \rvert$	$\lvert x_i - \bar{x}_2 \rvert$
9	2	-4	-11	4	11
11	5	-2	-8	2	8
11	7	-2	-6	2	6
12	9	-1	-4	1	4
12	10	-1	-3	1	3
13	15	0	2	0	2
14	16	1	3	1	3
15	20	2	7	2	7
16	22	3	9	3	9
17	24	4	11	4	11

$\bar{X}_1 = 13 \quad \bar{X}_2 = 13$ g.a. $= 2$ g.a. $= 6,4$

5.5 De variantie (s^2)

- We gaan uit van het kwadraat van de afwijking t.o.v. het gemiddelde. Hiervan maken we het gemiddelde.
  - Probleem met variantie: vb: met IQ
- Aldus ontstaat de gemiddelde gekwadrateerde afwijkingsscore = variantie
- 2e kolom ², dan de som / # observaties = variantie

Voorbeeld

Reeks 1	Reeks 2	$X_i - \bar{X}_1$	$X_i - \bar{X}_2$	$(X_i - \bar{X}_1)^2$	$(X_i - \bar{X}_2)^2$
9	2	-4	-11	16	121
11	5	-2	-8	4	64
11	7	-2	-6	4	36
12	9	-1	-4	1	16
12	10	-1	-3	1	9
13	15	0	2	0	4
14	16	1	3	1	9
15	20	2	7	4	49
16	22	3	9	9	81
17	24	4	11	16	121

$\bar{X}_1 = 13 \quad \bar{X}_2 = 13$

$s_X^2 = 5,6 \quad s_X^2 = 51$

PASW & Variantie: Frequencies → variabele selecteren → statistics → variance → output

5.6 De standaardafwijking (s)

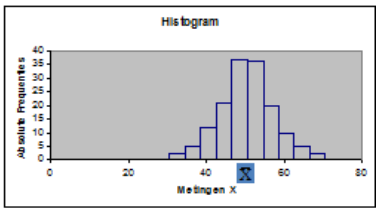
- Standaardafwijking of standaarddeviatie is gelijk aan de wortel uit de variantie
  - Wanneer s=0? Als alle deviatiescores 0 zijn → als alle uitslagen = rekenkundig gem. → alle uitslagen gelijk.
  - Meer variabiliteit, meer afwijking.

- **Betekenis standaardafwijking:**

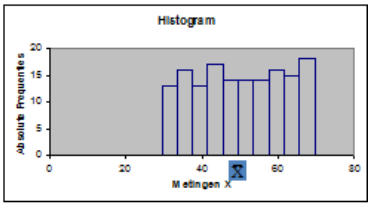
	Reeks 1	Reeks 2	Reeks 3
	0	99	100
	100	100	100
	200	101	100
$\bar{X}$	100	100	100
S	81,65	0,82	0,00

Normaalachtige verdeling
gemiddelde = 50,49

s = 6,84



Uniformachtige verdeling
gemiddelde = 50,25
s = 11,68



Sd = s

S^2 = variantie

- **Opmerking:**
  - Standaardafwijking van de populatie (sigma: σ)
  - Standaardafwijking van de steekproef (s)
  - Schatting van sigma:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

– Analoog voor de variantie; PASW maakt altijd gebruik van N-1

- Standaarddeviatie en variantie zijn beiden steeds positief;
  - Kleinste = 0
- Standaarddeviatie wordt steeds uitgedrukt in dezelfde eenheid als de scores; variantie als gekwadrateerde eenheid van de scores.
- Variantie wordt (samen met het rekenkundig gemiddelde) zeer veel gebruikt in de inductieve statistiek.

  - Variantie-analyse
- **Probleem:** relatratie nodig aan rekenkundig gemiddelde → hoe groepen vergelijken → oplossing!

5.7 De variatiecoëfficiënt

- De standaarddeviatie kan, naast het rekenkundig gemiddelde gebruikt worden om groepen met elkaar te vergelijken
  - Subgroepen vergelijken → % variabiliteit
- De grootte van de s hangt ook af van het gemiddelde.
  - **Oplossing:** de variatiecoëfficiënt
- Door deze coëfficiënt wordt de s gecorrigeerd voor het gemiddelde. Variatie wordt dan gezien als % van gemiddelde
- **Let op:** enkel bij ratio niveau van meting.
- S aan x koppelen
- **Vb:**

5.8 Lineaire transformaties

- Lineaire transformatie = alle waarden worden met bepaald getal gedeeld/vermenigvuldigd/opgeteld/afgetrokken.
  - Geen wortels of kwadraten

- Wat is het rekenkundig gemiddelde en s van de nieuwe verdeling van uitslagen?
  - Wat is het effect op het rekenkundig gemiddelde als elke waarde met b vermenigvuldigd wordt en a bij opgeteld?
    - Uitslagen vergelijken → lineaire transformatie nodig.
    - Bij elke waarde a optellen/aftrekken → geen invloed op s, blijft zelfde.
    - Bij elke waarde b vermenigvuldigen/delen → s verandert (stijgt of daalt), de verschillen in de uitslagen worden groter of kleiner, s gaat met b omhoog.
    - Variantie gaat ook met b omhoog.

- Wat is het effect op de standaarddeviatie en de variantie van de nieuwe verdeling als elke waarde met b vermenigvuldigd wordt en er a bij opgeteld?
  - A geen invloed, b wel.

Besluit:

- Als je alle waarden (X) met eender welke constante vergroot of verkleint, dan wordt het gemiddelde eveneens met die waarde vergroot of verkleind, maar dit heeft geen effect op de standaardafwijking, noch op de variantie.
- Als je alle waarden (X) met 3 vermenigvuldigt, dan wordt het gemiddelde 3 keer groter en de standaardafwijking ook 3 keer groter. De variantie wordt 9 keer groter.
- Als je alle waarden (X) met -3 vermenigvuldigt, dan wordt het gemiddelde met -3 keer vermenigvuldigd, maar de standaardafwijking wordt met 3 vermenigvuldigd. De variantie wordt 9 keer groter.

Voorbeeld Lessius

X	Y= -5+2X
9	13
11	17
11	17
12	19
12	19
13	21
14	23
15	25
16	27
17	29

$(X_i - X)^2$	$(Y_i-Y)^2$
16	64
4	16
4	16
1	4
1	4
0	0
1	4
4	16
9	36
16	64

$\bar{X} = 13 \quad \bar{Y} = -5 + 2 \cdot 13 = 21$ $\quad s_X^2 = 5,6$ $\quad s_Y^2 = \frac{224}{10} = 22,4$

2e kolom = gekwadrateerde deviatiescores

3e kolom = deviatiescores Y

$S^2_X = . 2^2$

S^2_Y = nieuwe variantie → 5 telt ni mee, 2 wel impact

5.8.1 Belangrijke transformatie

- Standaardscores of Z-scores geven aan hoeveel standaardafwijkingen een score van het gemiddelde ligt.
  - Individuele score – gem. /sd

X	Z
1	-1,5
3	-0,5
4	0,0
5	0,5
7	1,5

$\overline{X} = 4$ $\quad \overline{Z} = 0$

$s_X = 2$ $\quad s_Z = 1$

$$\frac{1-4}{2} = -1,5$$

$$\frac{3-4}{2} = -0,5$$

gem Z : altijd 0 en S_Z ook.

- Wat wordt het rekenkundig gemiddelde en variantie van deze Z-waarden?
  - het gemiddelde zal altijd nul zijn
  - de variantie en standaarddeviatie zullen altijd 1 zijn.
- Deze omzetting noemen we standaardiseren.
- Zinvolheid?
  - **Voorbeeld:**
    - Arie behaalt een 7 voor Frans, terwijl het gemiddelde 6 is en de standaarddeviatie 2
    - Arie behaalt voor Engels 6, terwijl het gemiddelde 5 was en de standaarddeviatie 1,5 bedroeg.
    - Welke prestatie is het beste?
      Z_F= (7 - 6)/2 = 0,50
      Z_E = (6 – 5)/1,5 = 0,67
    - Dus de prestatie voor Engels is beter dan deze voor Frans.
  - Z-waarden zijn een dimensieloos getal, en kunnen zowel positief als negatief zijn.
  - Een negatieve Z-waarde betekent dat deze uitslag zich links van het gemiddelde bevindt.
    - 1 SD vn gem verwijdert, lager.
  - Een positieve Z-waarde betekent dat deze uitslag zich rechts van het gemiddelde bevindt.
    - 1 sd vn gem. verwijdert, hoger.

- De uitslagen kunnen omgezet worden in Z-waarden, maar ook omgekeerd, indien we het rekenkundig gemiddelde hebben en de standaarddeviatie van de oorspronkelijke gegevens kunnen we elke Z-waarde terug plaatsen in de oorspronkelijke verdeling.

Score voor taal

$X_i = 84$

$\overline{X} = 76$

$s_X = 10$

Score voor rekenen

$Y_i = 90$

$\overline{Y} = 82$

$s_Y = 16$

Welke van beide scores 84 of 90 is de beste prestatie voor deze persoon?

$$z_i = \frac{84-76}{10} = 0,8$$

$$z_i = \frac{90-82}{16} = 0,5$$

taal = beste score, terugredenen gaat ook.

- Uit de Z-waarden kunnen de oorspronkelijke scores bepaald worden.

$$z_i = \frac{X_i - \overline{X}}{s_X} \qquad \Rightarrow X_i = \overline{X} + z_i \cdot s_X$$

$z_i = 0,8$

$\overline{X} = 76$

$s_X = 10$

$$\Rightarrow X_i = 76 + 0,8 \cdot 10 = 84$$

Score voor taal is 84. En voor rekenen?

- Standaardscores kunnen omgezet worden naar een verdeling met een bepaald gemiddelde, en dat via lineaire transformatie
  - Omzetting naar T scores met gemiddelde van 50 en een standaarddeviatie van 10
  - Of een omzetting naar C scores, met gemiddelde 5 en standaarddeviatie van 2.
  - Hoe?
- Let wel: het standaardiseren heeft geen effect op de vorm van de verdeling, m.a.w. scheef blijft scheef, symmetrisch blijft symmetrisch.
  - Enkel het gemiddelde wordt nul en de standaarddeviatie wordt één.
- PAWS: descriptives → variabele, save → output

5.9 Samenvatting

Meetniveau	Centrummaat	Spreidingsmaat	Andere
Nominaal	Modus		
Ordinaal	Modus	Interkwartielafstand	

	Mediaan	Bereik	
Interval	Mediaan Gemiddelde	Bereik Interkwartielafstand Standaardafwijking	Gemiddelde absolute afwijking
Ratio	Mediaan Gemiddelde	Bereik Interkwartielafstand Standaardafwijking	Gemiddelde absolute afwijking

5.10 Besluit

- Een spreidingsmaat geeft aan hoe de scores verschillen t.o.v. het rekenkundig gemiddelde:
  - Variatiebreedte (range)
  - interkwartielafstand
  - Gemiddelde absolute afwijking
  - Variantie
  - Standaardafwijking
- Een lineaire transformatie heeft geen invloed op de vorm van de verdeling, maar wel op het nieuwe rekenkundige gemiddelde en mogelijk op de standaarddeviatie.
- Meest gebruikte transformatie is de omzetting in Z-waarden.

6 De normale verdeling

- Een intervalwaarde die afhankelijke is van een oneindig aantal onafhankelijke factoren, die los van elkaar inwerken, zal in de populatie een normale verdeling vertonen (Gausscurve); (gebruikt voor kansberekening)
    - **bv.** Intelligentie.
  - Gemiddelde in steekproef: X
  - Gemiddelde in de populatie: μ
  - Standaarddeviatie in de steekproef: s
  - Standaarddeviatie in de populatie: σ

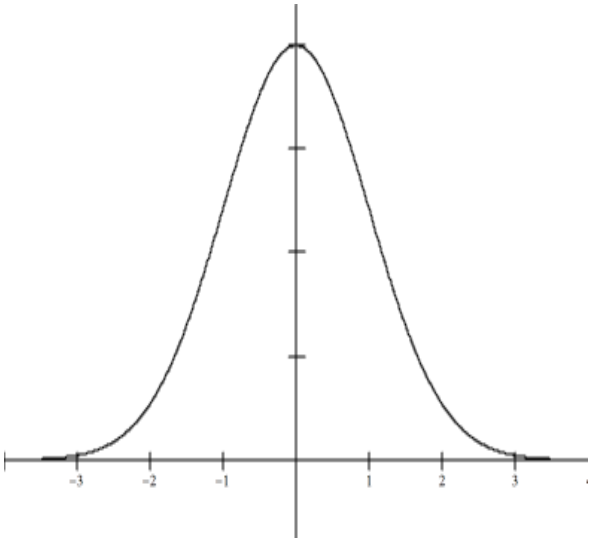
6.1 Kenmerken van de normale verdeling

- Niet bij scheve verdeling, alleen symmetrische.
- Dergelijke verdeling heeft 1 maximum
  - MO = Me = Gemiddelde

  - Twee buigpunten

- Op (-) 1 σ van rekenkundig gem. Ertussne liggen de meeste observaties (altijd zo)

– Symmetrie
  - De oppervlakte links en rechts van het gemiddelde zijn gelijk (skw = 0)
  - Kurtosis: kurt = 0



- De grafische weergave van de normale verdeling is klokvormig;
  De uitslagen liggen vooral geconcentreerd rond het gemiddelde, naarmate scores afwijken t.o.v. het gemiddelde wordt de frequentie kleiner.

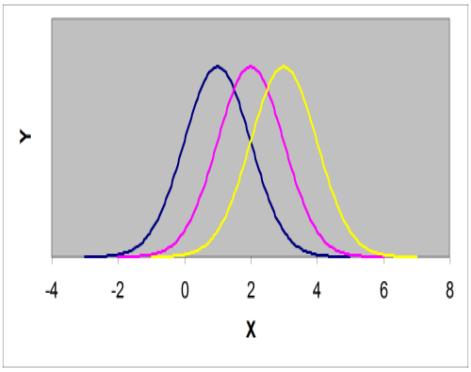
- Parameters van de normale verdeling
  – **Gemiddelde:** μ (mu)
    - In μ bereikt de Gausscurve zijn maximum
  – **Standaarddeviatie:** σ (sigma)
    - De twee buigpunten van de Gausscurve bevinden zich op μ - σ en op μ + σ

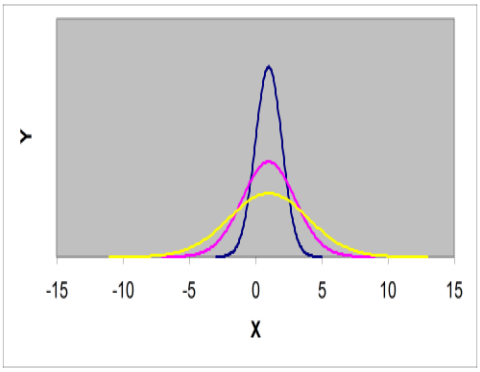
- **Verdeling in de Gausscurve:**
  – Tussen beide buigpunten (μ - σ en μ + σ)bevinden zich +/- 68% van de observaties;
  – Tussen μ - 2 σ en μ + 2 σ situeren zich +/- 95% van de observaties;
  – Tussen μ - 3 σ en μ + 3 σ bevinden zich +/- 99% à 100 van alle waarnemingen.
    - Erboven kan bijna niet.

- Als we van een normale verdeling het gemiddelde en de standaarddeviatie kennen, is deze verdeling gedefinieerd.
  – **Notatie:** (de variabele) X ~ n(μ,σ)

De normale verdeling:

- Verschillende μ, zelfde σ
- Verschillende σ, zelfde μ

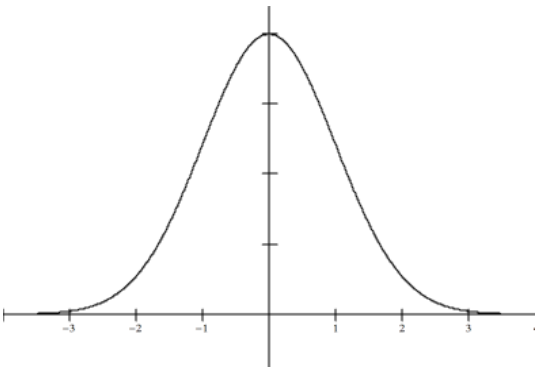




6.2 Speciale normale verdeling

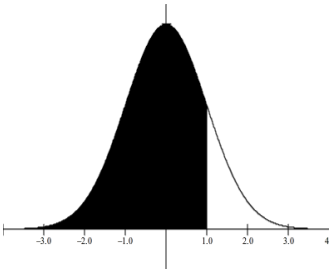
6.2.1 De Standaardnormale verdeling

- Een normale verdeling met gemiddelde nul en standaarddeviatie 1, noemen we een standaardnormale verdeling.
  - Z-uitslag, moeder alle normaal verdelingen
  - **Notatie:** Z ~ n(0,1)

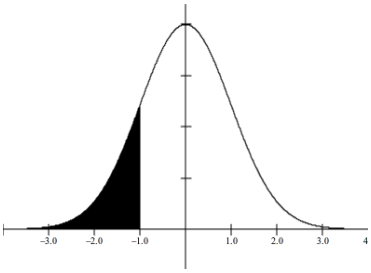


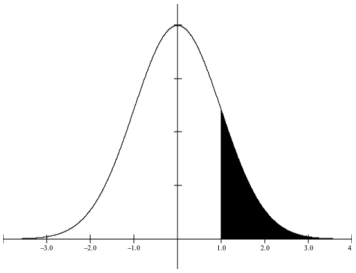
- Vanuit de eigenschappen van de normale verdeling kunnen we vaste relaties vinden tussen de proportie van uitslagen en Z-waarden.
- Welk is de kans om in een standaardnormale verdeling een uitslag te vinden die groter of kleiner is een bepaalde Z-waarde?
  - Tabel met standaard normale verdeling (komt op exaam, kennen en kunnen combineren)
  - Rij met kolom combineren 1,77 (1.7 = 17^{e} rij + 7^{e} kolom)
  - Kans dat uitslag kleiner is dan 1.0 → 84.13 % kans kleiner en 15.87 % kans groter.

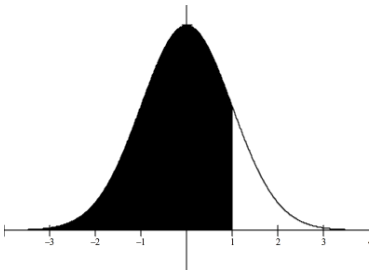
- **Eerste geval:** Z is positief.
  - Hoeveel percent van de scores verwachten we beneden deze Z-waarde?
  - In dat geval kunnen we het percentage aflezen uit de tabel, bv. Z = 1
    - (kans dat Z-waarde kleiner is dan 1)



- **Geval twee:** de Z-waarde is negatief.
  - Hoeveel % van de observaties liggen in een standaard normaal verdeling onder de Z-waarde -1?



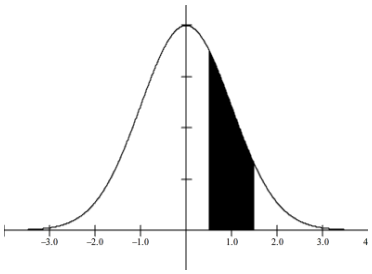




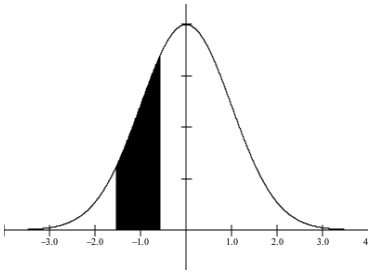
Methode: 1 – percentiel |Z|

- In een standaardnormale verdeling, hoeveel observaties situeren zich tussen twee Z-waarden?
  - Grootste Z-waarde – kleinste Z-waarde
  - In principe maak het verschil van % beneden de hoogste waarde, verminderd met % beneden de laagste waarde;

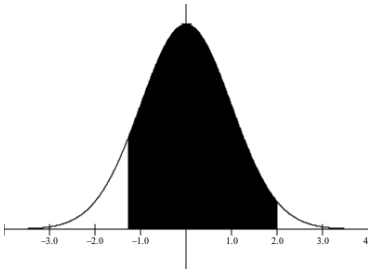
- **1e geval:** Beide Z-waarden zijn positief
  - Concreet hoeveel % van de uitslagen verwacht je in een standaardnormale verdeling tussen Z = 0,5 en Z = 1,5? (links en links
    overschrijdingskans)



- **2e geval:** Beide Z-waarden zijn negatief
  - Bepaal de oppervlakte tussen Z = -1,53 en Z= - 0,56 (linker hoogste – linker laagste) → (1- …) – (1-…)



- **3e geval:** 1 van beide Z-waarden is negatief, de ander is positief.
  - Bepaal in een standaardnormale verdeling het aantal ppn tussen de Z-waarden -1,26 en 2,01 (linker hoogste – linker laagste) → (… - (1-…)



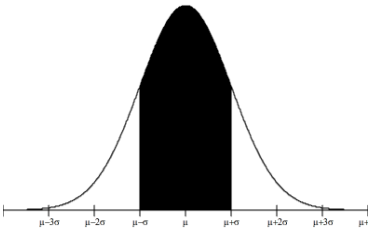
- Uitbreiding van de resultaten van de standaardnormale verdeling naar om het even welke normaalverdeling met μ als gemiddelde en sigma als standaarddeviatie
  - bv. Hoeveel procent van de bevolking heeft een IQ lager dan 80? In de veronderstelling dat μ = 100 en sigma = 15. Werk het probleem eerst uit via Z-waarden en zet deze om naar X-waarden.

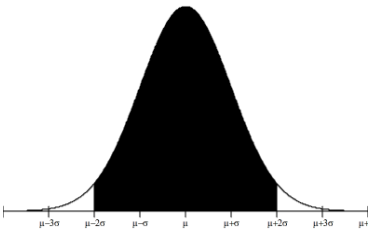
→

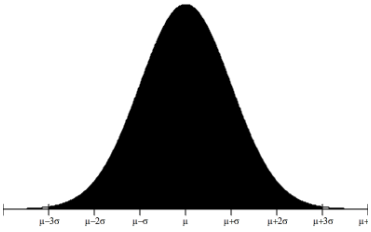
→ →

    - 80 - μ / σ
  - Ander vraagstuk: Stel normale verdeling en μ =100 en sigma = 15. Hoeveel procent van de observaties in een normale verdeling verwacht u boven de 130? (opl. 2%)
  - Andere zie apart docu ☺

- **Belangrijk:**

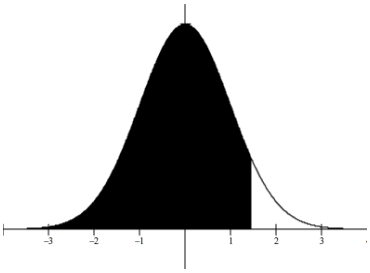






- Bepaal de Z-waarde behorende bij een bepaald percentiel in een normale verdeling

- **Geval 1:** Het % is groter als 0,50 Deze Z-waarde is rechtstreeks af te lezen uit de tabel Beneden welke Z-waarde situeert zich 87,29% van de uitslagen?

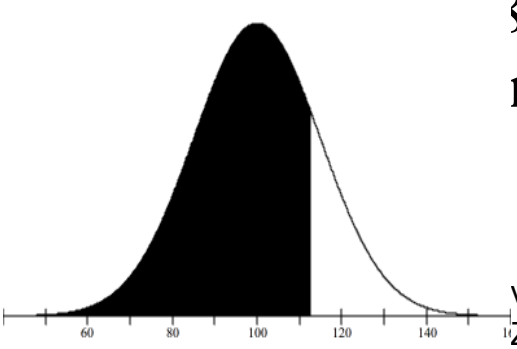


- **Geval 2:** De proportie is kleiner dan 50% In dit geval verwachten we een Z-waarde die negatief is.Bv. Beneden welke Z-waarde situeert zich 20,05% van de observaties in een normale verdeling?

- Methode voor de oplossing als het percentage kleiner is dan 0,50
  - Bereken 1 – P
  - Zoek de Z-waarde behorende bij 1 – P
  - Plaats een minteken voor de gevonden Z-waarde.

- Uitbreiding naar om het even welke normale verdeling. Beneden welke score liggen een bepaald percentage van de observaties, gegeven μ en gegeven sigma?
  - Bijvoorbeeld bepaal het IQ waarvan je weet dat 80% van de mensen een lager IQ heeft.
    - Eerst naar standaardnormen, dan naar de gegeven en dan uitrekenen.

Bepaal de IQ-score met percentiel 80 (m.a.w. bepaal de
30% van de
r zit)



$$z_{80} = 0,84$$

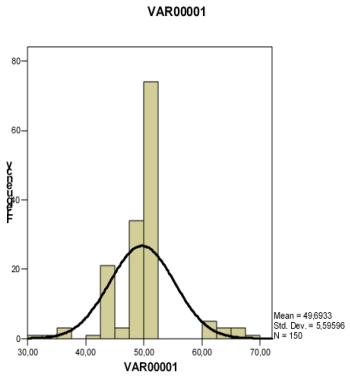
$$\Rightarrow x_{80} = \mu + z_P \cdot \sigma = 100 + 0,84 \cdot 15 = 112,6$$

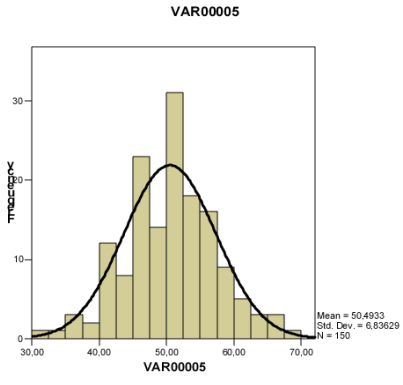
7 Vorm van de verdeling en invloed van transformaties

- Veel frequentieverdelingen hebben niet de vorm van de normale verdeling. Ze vertonen niet de gelijkmatige welving van de klokfunctie, of zijn scheef.
- Skewness of scheefheid (skw) en welving of kurtosis (kur) kunnen via PASW berekend worden.
  - **Skw:** afwijking van symmetrische vorm
  - **Kur:** afwijking klokvorm, welving
- Deze skw en kur dient gerelateerd te worden aan de betreffende standaardfout. (de uitslag dient tweemaal zo groot of groter te zijn om betekenisvol te zijn)
  - Kleine = verwaarloosbaar
  - **Standaardfout :** mate van onnauwkeurigheid

7.1 Kurtosis

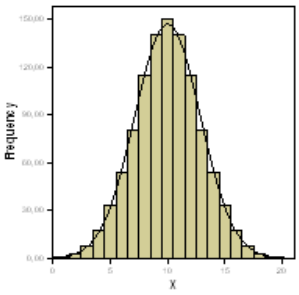
- Kurt > 0 wijst op een scherpe top (positief) (tgo normaalverdeling)





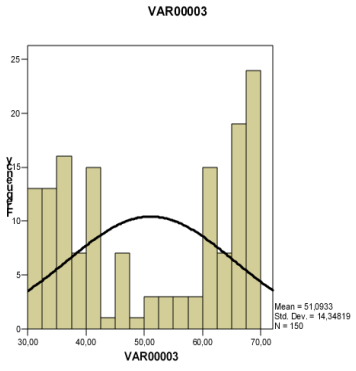
Kur = 3,45 Kur = 0,30

- Kurt = 0 wijst op een welving die vergelijkbaar is met de normaalverdeling (perfect, geen afwijking)

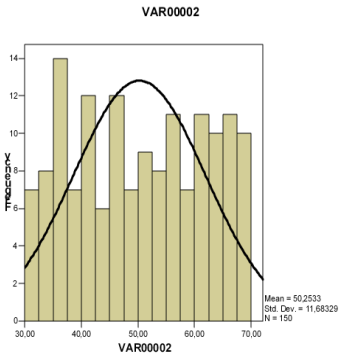


Kur = 0,05

- Kurt < 0 wijst op een afgeplatte top (negatief)

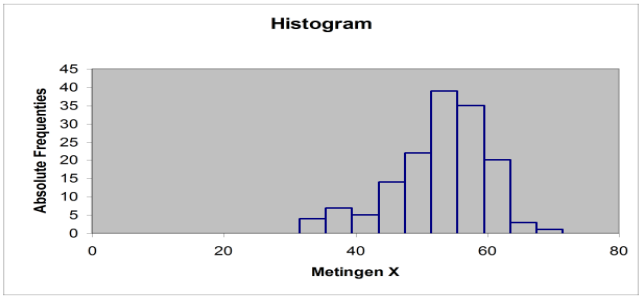


Kur = - 1,69 Kur = - 1,25



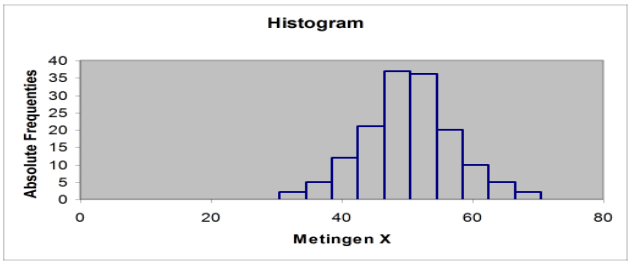
7.2 Scheefheid van de verdeling

- Skw < 0 wijst op een links scheve verdeling, staart naar links.
  - Mo, Me, Gem
  - Minder dan 50% onder rek. gem.
  - **bv.** score voor een test met heel gemakkelijke items (zgn. plafondeffect)



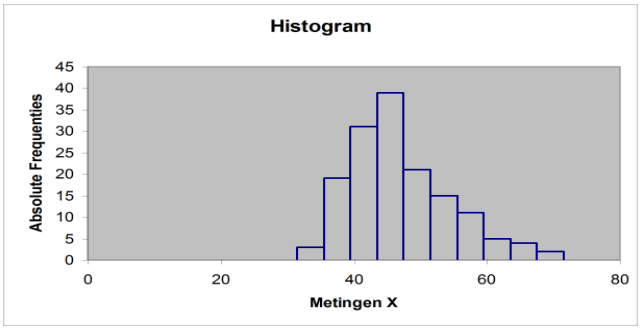
Skw = - 0,75

- Skw = 0 wijst op een symmetrische verdeling
  - Geen afwijking
  - Qua symmetrie zoals normaalverdeling



Skw = 0,02

- Skw > 0 wijst op een rechts scheve verdeling, staart naar rechts
  - **bv.** de scores op een test bestaande uit veel te moeilijke items (zgn. vloereffect)



Skw = 0,75

7.3 Berekening via PASW

Desciptive statistics → Explore → dependent list en gaan !

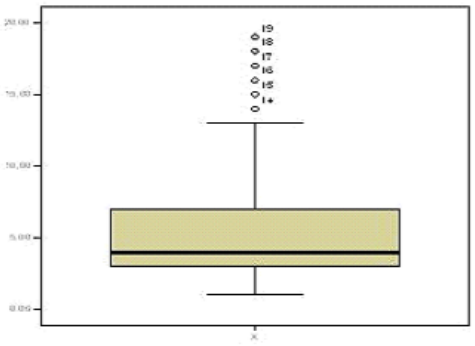
7.4 Lineaire transformatie

Model: Y = a + bX

- Het gemiddelde wordt op dezelfde wijze getransformeerd.
- De standaarddeviatie wordt met |b| vermenigvuldigd, de variantie met b^2.
- De 'skewness' blijft onveranderd indien b>0. (van teken verandereden als --
  - Linksscheef wordt rechtsscheef
  - Rechtsscheef wordt linksscheef
- De kurtosis blijft onveranderd.
- Heeft een omzetting in Z-waarden een invloed op de vorm van de verdeling.
  - Deze omzetting heeft GEEN invloed op de scheefheid en kurtosis van de verdeling; m.a.w. scheef blijft scheef.

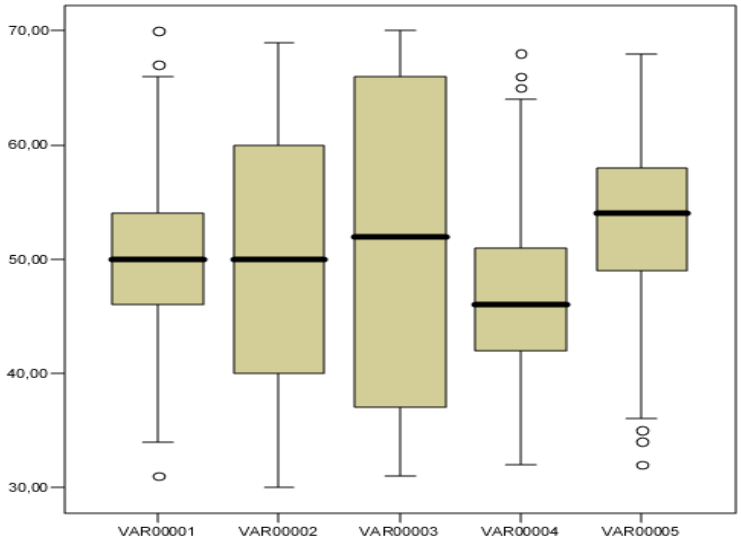
- Deze omzetting heeft wel een invloed op het rekenkundig gemiddelde (altijd nul) en de s (altijd 1).

7.5 De boxplot

- In de boxplot wordt in een doos de mediaan, het 25ste en het 75ste percentiel geplaatst, waardoor de doos in feite het interkwatielafstand voorstelt.
- Daarnaast worden de extreme en uiterst extreme waarden (=uitbijter) afgebeeld.
- Uitbijter ligt op meer dan 1,5 dooslengte van het 25ste of 75ste percentiel
- Extreme uitbijters liggen op meer dan 3 dooslengtes van het 25ste of 75ste percentiel
- Rechts scheve  verdeling:

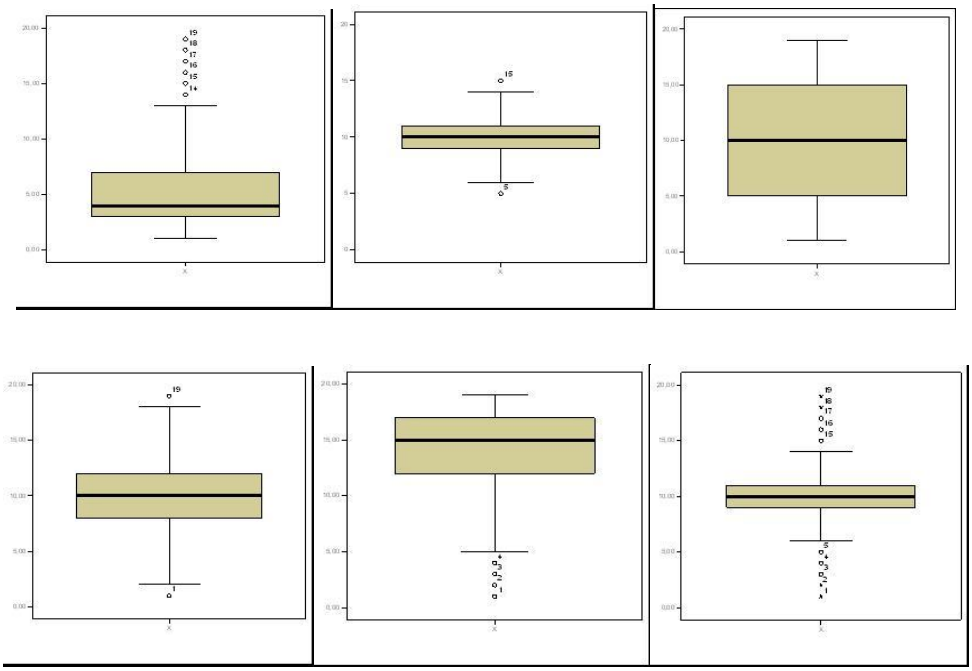
- Bij welke variabele is de mediaan het grootst?
  - 5
- Welke variabele heeft de grootste interkwartielafstand?
  - 3
- Welke variabele(n) is (zijn) ongeveer linksscheef?
  - 5
- Welke variabele(n) is (zijn) ongeveer rechtsscheef?
  - 4
- Welke variabele(n) is (zijn) ongeveer symmetrisch?
  - 2,3
- Welke variabelen hebben outliers?
  - 1,4,5

Statistiek



	V1	V2	V3	V4	V5	V6
Mean	9,93	10,00	14,37	10,00	5,10	10,00
Median	10,00	10,00	15,00	10,00	4,00	10,00
Mode	10,00	10,00	17,00	10,00	2,00	10,00
Std. Deviation	3,20	2,86	3,32	1,49	3,28	5,48
Variance	10,27	8,16	11,02	2,22	10,75	29,98
Skewness	0,04	0,00	-0,94	0,00	1,07	0,01
Kurtosis	1,10	-0,05	0,89	-0,04	1,21	-1,20
Range	18,00	18,00	18,00	10,00	18,00	18,00
Percentiles 25	9,00	8,00	12,00	9,00	3,00	5,00
Percentiles 75	11,00	12,00	17,00	11,00	7,00	15,00

V1: F, V2: D, V3: E, V4: B, V5: A, V6: C



8 Kruistabellen & PASW Crosstabs

- **Analyse van de samenhang:**
  - Samenhang tussen variabelen = contigentietabel.
  - *Vereiste:* twee uitslagen per persoon
  - Twee nominale variabelen stellen we voor in een kruistabel. Analyse met Chi-kwadraat en de associatiematen.
    - Ook met variabele die nominaal is gemaakt.
  - *Twee interval variabelen:* gebruik de Pearson correlatiecoëfficiënt en de regressietechniek.
  - *Twee ordinale variabelen:* gebruik de correlatiecoëfficiënt van Spearman.
- Bestaat er een betekenisvol verband tussen twee nominale variabelen?
  - Gebruik de Chi-kwadraat test van onafhankelijkheid.
  - Zowel in steekproef als populaite
- Hoe sterk is dit verband?
  - Phi-coëfficiënt (enkel in een vierveldentabel)
  - Contingentiecoëfficiënt (kan nooit 1 worden)
  - Cramér's V
  - =associatiematen, kan – zijn, maar geen betekenis

8.1 Voorbeeld van onderzoek

- Bestaat er een verband tussen het al dan niet roken en het voorkomen van hart- en vaatziekten?
- Operationalisering?

- Wie gaan we bevragen?
- Onafhankelijke/afhankelijke variabele?
- Hoe samenhang berekenen?

8.1.1 Verband tussen roken en hart- en vaatziekten.

- Veronderstel een perfect verband:

	Hart- en vaatziekten	
	ja	**neen**
Roker	20	0
Niet roker	0	100

- Veronderstel dat er geen verband bestaat

	Hart- en vaatziekten		
	Ja	**neen**	Totaal:
Roker	1	19	20
Niet roker	5	95	100
Totaal:	6	114	120

- Feitelijke observaties

	Hart- en vaatziekten		
	Ja	**neen**	Totaal:
Roker	5	15	20
Niet roker	1	99	100
Totaal:	6	114	120

- Welke zijn de celfrequenties?
  - Roker - niet roker, ja – neen.
- Welke de randtotalen/marginale totalen?
  - Totalen aan buitenkant.

8.1.2 Voorstelling via percentages

- In functie van het totaal aantal ppn. Dit heeft geen zin
  - Verticaal percenteren?
  - Horizontaal percenteren?
- Maak een keuze in de richting van de onafhankelijke variabele.

- Foutieve manier van voorstelling

	Hart- en vaatziekten	
	Ja	**neen**
Roker	83%	13%

Niet roker	17%	87%
Totaal:	100%	100%

- Juiste voorstelling van zaken

	Hart- en vaatziekten		
	Ja	**neen**	
Roker	25%	75%	100%
Niet roker	1%	99%	100%

- Let op in welke richting percenteren.
  - Meer kans op hart- en vaatziekten als roken.

8.2 Samenhang tussen twee nominale variabelen

- Is het verband betekenisvol?
- **Chi-kwadraat waarde:** welk is de afstand tussen de geobserveerde waarden en de verwachte waarden in de veronderstelling dat er geen verband zou zijn. Is dit significant?
  - Nooit negatief
  - 0 - # ppn in steekproef
  - f_o : geobserveerde celfrequentie
  - f_e : verwachte celfrequentie

8.2.1 Voorbeeld

- Er bestaan slechts drie politieke partijen:
  - NOTAX, NOVA, en Nieuw groen.
  - Welk is uw voorkeur?
- We bevragen de houding t.o.v. een belastingsvermeerdering voor milieuonvriendelijke producten. (voor/weet niet/tegen)

	Voor	Weet niet	Tegen	Totaal
Nieuw Groen	12	5	8	25
NOTAX	5	12	23	40
NOVA	12	7	1	20
Totaal:	29	24	32	85

- Geen samenhang politieke partij en KB → nulhypothese.

- Bestaat er een betekenisvol verband?

- Gebruik de Chi-kwadraat waarde
- Opstellen van de verwachte frequenties
- Bereken de Chi-kwadraat waarde
- Test de significantie

- **Berekening:** F_e
- Voor de bepaling van de verwachte frequenties dient u het product van de overeenkomstige randtotalen te delen door het totaal aantal ppn.
  - F_e = 29*25/85 = 8,53

	Voor	Weet niet	Tegen
Nieuw Groen	8,53	7,05	9,41
NOTAX	13,65	11,29	15,05
NOVA	6,82	5,64	7,53

- **Berekening** (Chi-kwadraat)

Deze waarde is gelijk aan: 1,41 + 0,60 + 0,21 + 5,48 + 0,04 + 4,20 + 3,93 + 0,33 + 5,66 = 21,86

→ formule van het begin

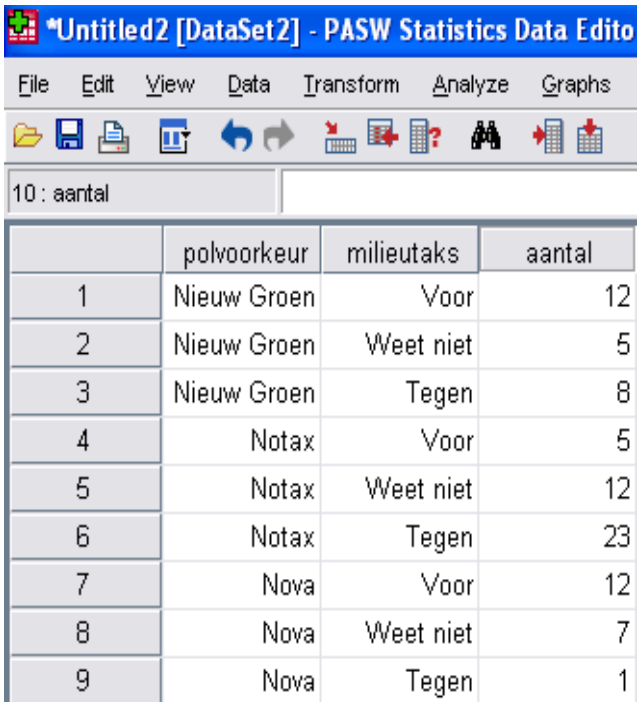
  - Toets vervolgens of deze waarde significant is, rekening houdend met het aantal vrijheidsgraden.
    - Vrijheidsgraden ? (r-1)*(k-1) → # vrij in te voelen cellen, zodat de randtotalen wel constant blijven.(afh grootte tabel)
    - Kritische waarde bij 5% niveau (df=4): 9,49. Derhalve bestaat er een significant verband tussen beide variabelen.

- **Het begrip vrijheidsgraden (df)**
  - Hoeveel waarden kunnen er in een kruistabel vrij variëren wanneer de randtotalen ingevuld werden?
  - Dit is in een viervelden tabel slechts 1. Na het invullen van één van de celfrequenties liggen de overige drie cellenfrequenties ook vast.
  - df: (r-1)*(k-1)
- **Het begrip significantie**

  - Wat is de kans om deze waarde van Chi-kwadraat te vinden indien de nulhypothese waar zou zijn?
  - Deze nulhypothese is bij een Chi-kwadraat waarde: er is geen verband tussen beide variabelen. Met de alternatieve hypothese is er wel een verband.
  - Deze kans is uiterst klein, vandaar dat we de nulhypothese verwerpen.
  - Ofwel Deze kans is redelijk aanwezig, vandaar dat we de nulhypothese niet verwerpen.
- **Handmatige berekening van significantie:**
  - Vergelijk de gevonden Chi-kwadraat met de kritische waarden.
  - De vastgestelde waarde van 21,86 is groter dan de kritische waarde 9,49, derhalve is de kans om dergelijke Chi-kwadraat te vinden indien er geen samenhang was, kleiner dan 5%, dus weinig waarschijnlijk. Vandaar …
  - Kritische waarden van de Chi-kwadraat: maak gebruik van de tabel

df	0,05
1	3,84
2	5,99
3	7,82
4	9,49
5	11,07
…	….

- **Snelle invoer kruistabel in PASW**
  - Rechtstreeks invoeren, tegen PASW zeggen, niet # ppn, maar alle mogelijke combi's vn kruistabel
  - # regels = # cellen van kruistabel
  - 3e kolom = # van die cel
  - In PASW → weight cases → weight cases by



	polvoorkeur	milieutaks	aantal
1	Nieuw Groen	Voor	12
2	Nieuw Groen	Weet niet	5
3	Nieuw Groen	Tegen	8
4	Notax	Voor	5
5	Notax	Weet niet	12
6	Notax	Tegen	23
7	Nova	Voor	12
8	Nova	Weet niet	7
9	Nova	Tegen	1

- **Opvragen:** → frequency variabele → crosstabs → invullen vn variabelen→ output.
- **Opvragen chi-kwadraat**
  - Via statistics → Chi-kwadraat aanvinken
  - Associatiematen hetzelfde

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	21,849[a]	4	,000
Likelihood Ratio	25,559	4	,000
Linear-by-Linear Association	1,437	1	,231
N of Valid Cases	85		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 5,65.

Alleen 1e lijntje is Chi-kwadraat, nulphypothese verwerpen → 0 kans om die chi² te vinden als nulhypothese waar zou zijn.
(a = vrijheidswaarden)

- Opvragen verwachte frequentie
  - Crosstabs → cells
- Ovpragen geobserveerde en verwachte aantallen
  - In display aanvinken

- **Chi-kwadraat beperkingen**

- Gebruik enkel absolute getallen, geen proporties
- Verwachte frequentie mag in max. 20% van de gevallen kleiner zijn dan 5; geen enkele Fe waarde mag kleiner zijn dan 1.
- Niet gebruiken bij herhaalde meting;
  - *vb.* voor- en nameting gebruik hiervoor de Mc Nemar test
- Is sterk afhankelijk van het aantal ppn. (zeker als te weinig)

	Milieubesef	
	Hoog	**Laag**
Vrouw	14	6
Man	9	11

Chi-kwadraat = 2,56 (p=.110)

	Milieubesef	
	Hoog	**Laag**
Vrouw	28	12
Man	18	22

Chi-kwadraat = 5,12 (p=.024)

- Besluit de Chi-kwadraat is sterk onderhevig aan het aantal proefpersonen.
- Vandaar belang van associatiematen.
- Gebruik Chi-kwadraat niet bij kleine proefgroepen.
- Tip: controle via PASW

8.2.2 Mc Nemar toets

- Kan enkel gebruikt worden voor herhaalde metingen
  - **Bv.** voor en na de therapie.
- Let op welke celfrequenties u gebruikt als A en D versus C en B.
  - A en D hebben betrekking op de proefpersonen die veranderen in functie van de behandeling.

		Na	
		Niet geslaagd	*Geslaagd*
Voor	*Geslaagd*	5(A)	35(B)
	Niet geslaagd	40(C)	20(D)

Mc Nemar test: $(|A-D|-1)^2/(A+D)$
Uitwerking: 14*14/25 = 7,84, hetgeen getest wordt als Chi-kwadraat waarde
→ AD van positive verandert → herhaalde metingen + nominale → variant Chi² = McNemar → alleen cellen met positieverandering belangrijk.

- **Let op:** enkel de groep van proefpersonen die van positie gewisseld zijn worden opgenomen in de formule.
- **Let op:** enkel te gebruiken bij dichotome variabelen.

- Input via weight cases (voor, na, aantal)
- Analyze – descriptive stat – crosstabs - statistics → Mc Nemar

8.2.3 Chi-kwadraat goodness-of-fit

- Niets te maken met analyse van een kruistabel
- Voldoet een verdeling van nominale waarden aan bepaalde verwachtingen/verdeling? In dit geval hebben we niet te maken met een kruistabel !!!
- Gebruik de goodness-of-fit Chi-kwadraat test.
- Onderzoek de aanhang bij 90 studenten voor drie politieke partijen

NOTAX	60
NOVA	5
Nieuw Groen	25
totaal:	90

- Betekenisvolle voorkeur voor één van de drie partijen?

Fo	60	5	25 (geobserveerd)
Fe	30	30	30 (verwacht)
Verschil	30	25	5
Kwadraat	900	625	25
Kw/Fe	30	20,83	0,83 (aantal mogelijke waarden -1, def GOF

- Bestaat er een betekenisvolle voorkeur voor een van de drie partijen?
  - Chi-kwadraat = 51,66
  - Kritische Chi-kwadraatwaarde =5,99
  - Dus er bestaat een significante voorkeur voor de partij NOTAX.
- Non-parametic-tests → Chi square → variabele erin

- Wordt vaak gebruikt om aan te tonen dat de samenstelling van de steekproef een weerspiegeling is van de populatie.
- **Let op:** gebruik geen percentages, beide verdelingen dienen eenzelfde aantal ppn. te hebben. Eventueel downscalen.

8.2.4 Associatiematen, gebaseerd op de Chi-kwadraat

- Niet afhankelijk van # ppn, berekend op chi² → afh. Sterkte samenhang
- **De Cramér's V**
  - Deze index geeft de sterkte van het verband aan en varieert van 0 tot 1. Kan in elk type kruistabel gebruikt worden.
  - K = aantal rijden, of kolommen → kleinste van de twee nemen. $V = \sqrt{\frac{\chi^2}{N(k-1)}}$
- **De Contingentiecoëfficiënt**

- In vergelijking met hetgeen max. haalbaar is in functie van de grootte van de tabel

- **De Phi-coëfficiënt (**enkel bij een 2X2 tabel, enkel bij een viervelden tabel) (-gene betekenis, pasw berekent zo)
  - a,b,c, en d vormen de celfrequenties
  - e,f,g,h vormen de randtotalen

$$\phi = \frac{ad - bc}{\sqrt{efgh}}$$

- Kan beschouwd worden als een standaardisering van de Chi-kwadraat waarde. (tweede wijze van berekening)

- Kan enkel gelijk worden aan 1 indien verhouding tussen rijtotalen en kolomtotalen gelijk is, in het andere geval blijft deze waarde kleiner dan 1.
- Het teken van deze coëfficiënt speelt geen rol: we kunnen de waarden van plaats wisselen, waardoor het teken verandert.
  Het is niet zinvol te spreken over een negatief of positief verband bij nominale waarden.

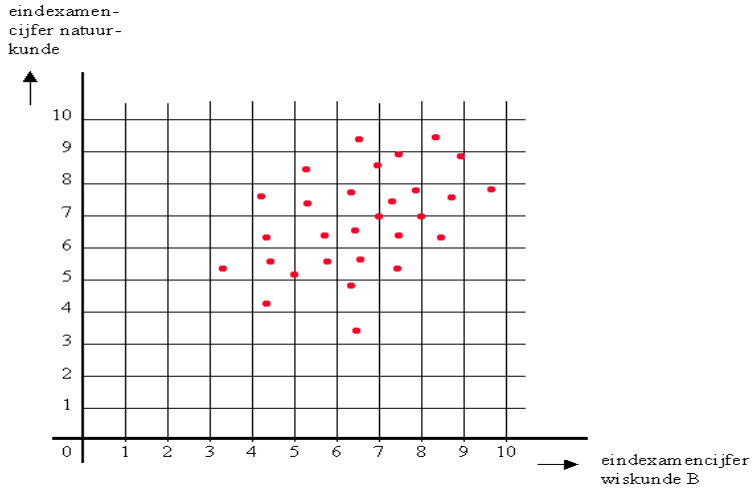
9 Het correlatievraagstuk & PASW toepassing

- Vooral via PASW en output.
- **Analyse van de samenhang(herhaling):**
  - Vereiste: ***twee*** uitslagen per persoon
  - Twee nominale variabelen stellen we voor in een kruistabel. Analyse met Chi-kwadraat en de associatiematen.
  - Twee interval variabelen: gebruik de Pearson correlatiecoëfficiënt en de regressietechniek (niet symmetrisch).
  - Bv. samenhang tussen IQ en schooluitslag
- Twee ordinale variabelen: gebruik de correlatiecoëfficiënt van Spearman.
- **Samenhang tussen twee interval variabelen:**
  - Bestaat er een (lineair) verband? Gebruik de Pearson correlatiecoëfficiënt → a en b van plaats wisselen (symmetrisch)
  - Hoe kunnen we de Y variabele voorspellen op grond van de X variabele? Gebruik de regressielijn van Y op X (anders dan X op Y).

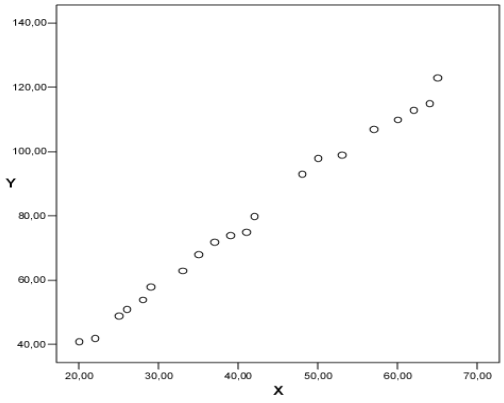
9.1 De correlatiecoëfficiënt

9.1.1 Spreidingsdiagram (scatterplot)

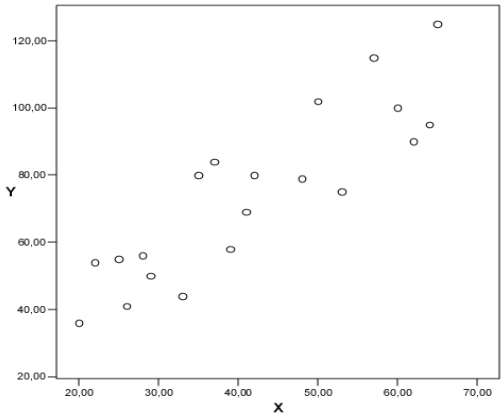
- Elk punt 1 ppn met gecombineerde score → bepaalde combinatie uitslagen.
- 2 zelfde variablen (min. Interval)
- = puntenwolk, tendens
- Geen perfecte samenhang
- Breedte wolk geeft gebrek correlatie aan.



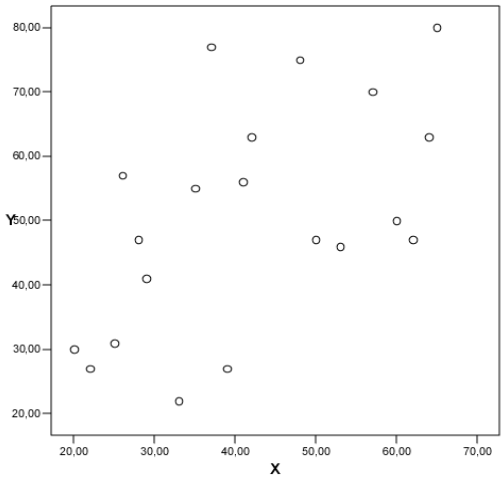
- Voorbeeld van een zeer hoge positieve correlatie
  - Zeer smalle wolk
  - lage X-as, lage Y-as
  - Hoge X-as, hoge Y-as
  - Afh y, onaf. x



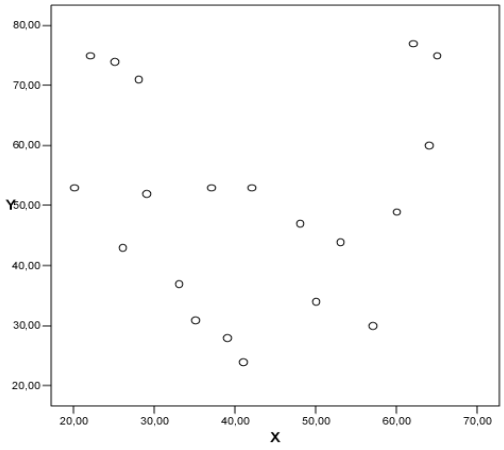
- Voorbeeld van een hoge positieve correlatie



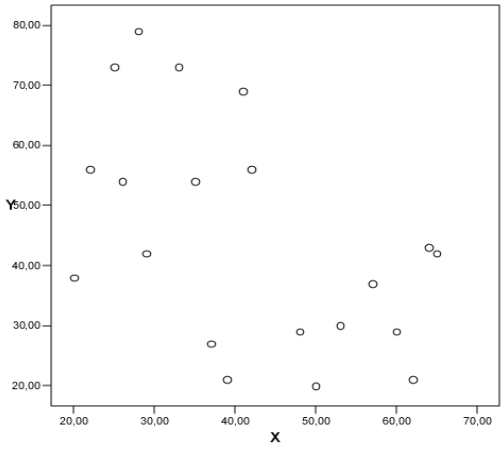
- Voorbeeld van een geringe positieve correlatie



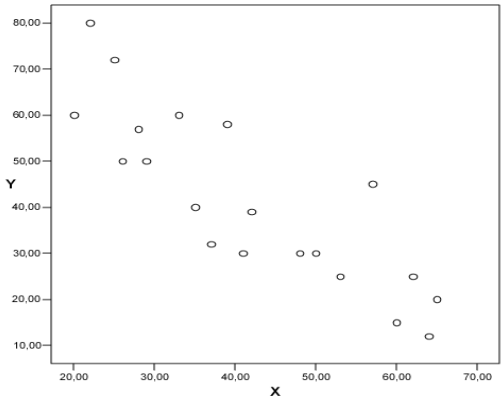
- Voorbeeld van geen correlatie



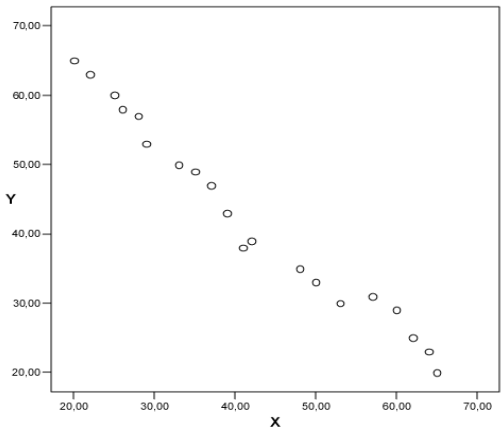
- Voorbeeld van een negatieve matige correlatie



- Voorbeeld van hoge negatieve correlatie

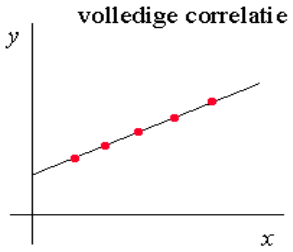


- Voorbeeld van zeer hoge negatieve correlatie

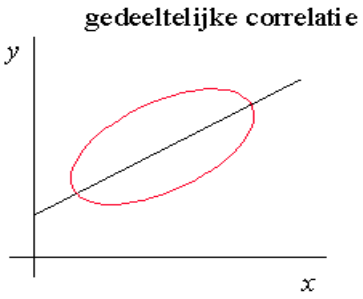


- Pearson = gestandariseerde maat
- Is het verband negatief? Of positief?
- Hoe sterk is het verband?

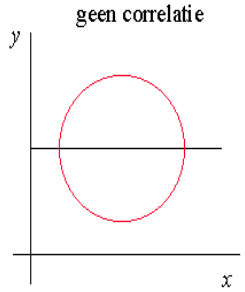
  $-1 \leq r \leq 1$
- Enkele concrete voorbeelden



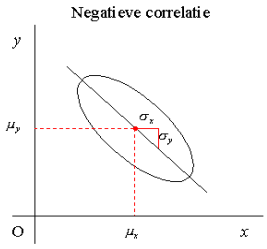
Bv. aantal juiste oplossingen en punt voor examen



Bv. intelligentie en schooluitslag



Bv. lichaamslengte en schooluitslag (geen tendens)



Bv. faalangst en schooluitslag

Correlatie =

- Is de mate waarin elk individu eenzelfde relatieve positie inneemt op de twee variabelen.
- Is positief als hoge score voor een variabele samengaat met hoge score voor tweede variabele.
- Is negatief als hoge score voor een variabele samengaat met lage score voor tweede variabele.

Y-a

$Z_x<0$ en $Z_y>0$

$Z_X>0$ en $Z_y>0$

$Z_x<0$ en $Z_Y<0$

$Z_x>0$ en $Z_y<0$

X-as

9.1.2 Pearson correlatie

- Is het gemiddelde product van de bij X en Y horende z-scores. (productmomentcorrelatie van Karl Pearson)
- Steekproef (populatie N-1)
- Product Z-waarden optellen / # observaties

  - $R_{XY} = \sum Z_X * Z_Y / N$
  - $R_{YX} = \sum Z_X * Z_Y / N$
  - D.w.z. de correlatie is symmetrisch
- **Covariantie:**
  - = niet gestandaardiseerde maat van samenhang tussen twee interval variabelen.
  - Gemiddelde product van de afwijking t.o.v. het rekenkundig gemiddelde
  - Pearson correlatie= een gestandaardiseerde maat van samenhang, varieert van – 1 tot + 1 (covariantie standariseren)
  - Correlatie kan gedefinieerd worden als de covariantie van de twee variabelen gedeeld door het product van de bijbehorende standaarddeviaties.

- Blijft constant als de X en/of de Y waarden vermenigvuldigd, gedeeld worden door een bepaald getal. Let wel op het teken van de correlatie. (absolute waarden gelijk, maar kan van teken veranderen)
- Blijft constant als de X en/of de Y waarden opgeteld of verminderd worden met een bepaald getal.
- Dus r (Pearson) is *invariant* onder lineaire transformaties (afgezien van het teken).

- Invariant van de correlatie

X	Y	(X+3)/4	(Y+2)/6
1	5	1,00	1,17
2	2	1,25	0,67
3	3	1,50	0,83
4	4	1,75	1,00
5	1	2,00	0,50
	r = -.60		r = -.60

- Invariant van de correlatie? (teken zal dus omdraaien daardoor)

X	Y	(X+3)/4	-(Y+2)/6
1	5	1,00	- 1,17
2	2	1,25	- 0,67
3	3	1,50	- 0,83
4	4	1,75	- 1,00
5	1	2,00	- 0,50
	r = -.60		r = .60

- $r_{X',Y'} = \frac{b \cdot d}{|b| \cdot |d|} \cdot r_{X,Y}$ Correlatie bij lineaire transformatie

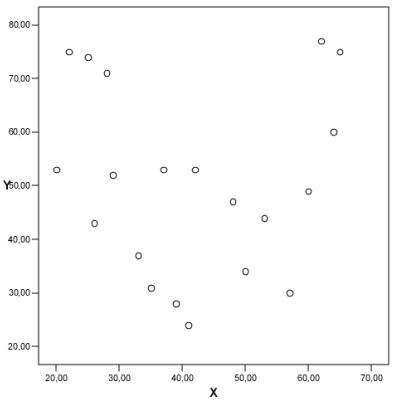
  En dan is + en -, naargelang teken van b en d.

- Is niet invariant voor niet-lineaire transformaties, zoals bv. omzetting in percentielscores, of bv. worteltrekking of kwadratering.
- Niet lineaire transformaties wijzigen de vorm van de verdeling en tevens de correlatie met andere variabelen.

9.1.3 Lage correlaties?

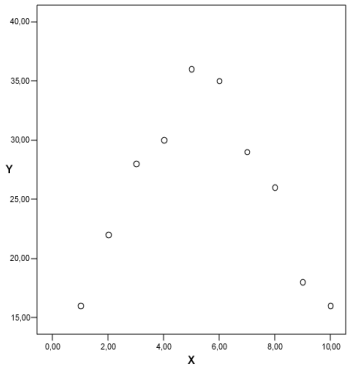
- De variabelen hangen niet met elkaar samen (bv. lichaamslengte en schooluitslag)
- Het verband tussen de beide variabelen is niet lineair (bv. relatie tussen angst en prestaties)
- Er is sprake van ‘restriction of range’. Eén van de variabelen heeft onvoldoende bereik, waardoor de correlatie gedrukt wordt.

1. Geen verband



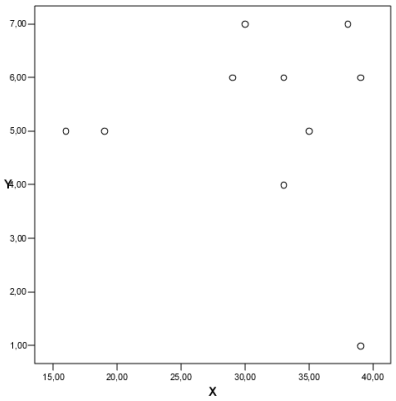
2. Niet-lineair verband

a. Negatieve samenhang, volgens Pearson niet, kan alleen linear.



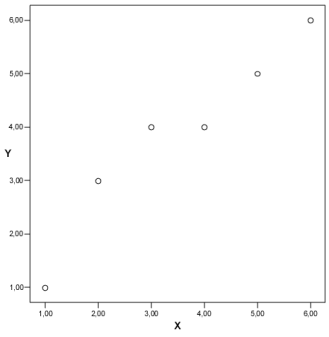
$$r_{X,Y} = -0,10$$

3. Restriction of range



$$r_{X,Y} = 0,959$$

Let op bij gering aantal metingen (100 ppn nodig ong)



3.a..4 Correlatie en causaliteit

- Als er een samenhang bestaat tussen twee variabelen, betekent dit een causaal verband? Misschien
  - X veroorzaakt Y

- Y veroorzaakt X
- Z veroorzaakt X, maar ook Y
- Andere ..;
- bv. medewerkerstevredenheid en productiviteit
- bv. ooievaarsnesten en aantal geboorten

Het onderzoeksbureau 'Reason Foundation' publiceerde een opmerkelijke studie. Drinkers verdienen ruim 10% meer dan geheelonthouders. Iemand die buitenshuis zijn pintje drinkt, verdient op zijn beurt meer dan een thuisdrinker. De reden lijkt logisch: mensen die drinken, onderhouden meestal meer contacten. Contacten – en dus netwerking – kunnen zorgen voor een nieuwe of betere baan en snellere loonsverhogingen. Vrouwelijke drinkers verdienen gemiddeld 14% meer dan vrouwelijke niet drinkers. Het verschil bij de mannen bedraagt maar 10%, maar bij hen kan een regelmatig toogbezoek daar nog 7% aan toevoegen. Vanzelfsprekend geldt voor dit onderzoek ook de bekende slogan: overdaad schaadt.

Verband IQ en schooluitslag → 2 scale nodig!

- IQ Schooluitslag
- 100 70
- 120 80
- 130 85
- 140 85
- 112 82
- 90 60
- 97 65
- 111 70
- PASW input: kolomeke IQ en kolommeke schooluitslag. Graphs legacy → scatter plot, simple. Afhankelijke variabele op de Y-as, onafhankelijke variabele op de X-as, hoogste meetniveau bovenaan.
- PASW Opvragen van correlatie → analyze, correlate, bivariate, twee variabelen, Pearson(interval, k&s: ordinaal).

Correlations

		Intelligentie quotient	schooluitslag
Intelligentie quotient	Pearson Correlation	1	,914**
	Sig. (2-tailed)		,001
	N	8	8
schooluitslag	Pearson Correlation	,914**	1
	Sig. (2-tailed)	,001	
	N	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

Nulhypothese waar, 1/100 kans juist, dus 0-hypothese weg.

- Correlatie tussen meer dan twee interval geschaalde variabelen (scale)
  - Samenhang is in de werkelijkheid vaak complex.

- Bijvoorbeeld. Er is een samenhang tussen het schoolresultaat en de intelligentie, maar ook tussen het schoolresultaat en de studietijd.
- Multiple correlatie: samenhang tussen één afhankelijke variabele en meerdere onafhankelijke variabelen
- Partiële correlatie: samenhang tussen één afhankelijke variabele en één onafhankelijke variabele onder constant houding van derde variabele.

3..2 De rangcorrelatiecoëfficiënt

- Het betreft het verband tussen twee ordinale variabelen.
- Correlatiecoëfficiënt van Spearman
- Formule:

me

Deze correlatie varieert van -1 tot + 1

Een voorbeeld

- Welk is het verband tussen intelligentie en leiderschap bij kinderen?
- Geef voor elke ppn. een rangorde voor beide variabelen
- Dit geeft volgende beeld

Leerling	Leiderschap	Intelligentie	D	D^2
A	1	4	3	9
B	3	2	1	1
C	5	1	4	16
D	2	3	1	1
E	4	6	2	4
F	7	5	2	4
G	6	8	2	4
H	9	7	2	4
I	10	6	4	16
J	8	10	2	4
			totaal	63 → in formule)

r_S= 1-6*63/10(99)= 0,618

Besluit? = Spearmancorrelatie → redelijke samenhang.

Correlations

			Leiderschap	Intelligentie
Spearman's rho	Leiderschap	Correlation Coefficient	1,000	,614
		Sig. (2-tailed)	.	,059
		N	10	10
	Intelligentie	Correlation Coefficient	,614	1,000
		Sig. (2-tailed)	,059	.
		N	10	10

Geen significante samenhang.

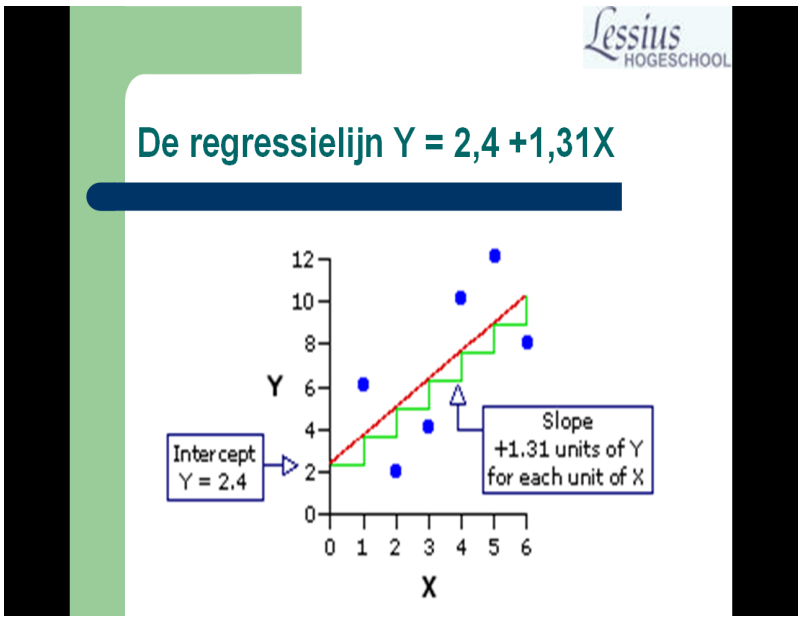
10 De regressieanalyse

(verlengde HS 9, alle var = interval)

- Samenhang tussen twee interval variabelen
  - Bestaat er een (lineair) verband? Gebruik de Pearson correlatiecoëfficiënt
  - Hoe kunnen we de Y variabele voorspellen op grond van de X variabele? Gebruik de regressielijn van Y op X.

Lineaire regressie (enkelvoudige)

- Welke rechte past het best bij een gevonden spreidingsdiagram? Welke rechte lijn biedt een zo goed mogelijke samenvatting van de trend in de puntenwolk?
- Zoeken de vergelijking van deze rechte, op basis waarvan we op grond van een X waarde de Y waarde kunnen voorspellen. Waarde regressielijn ook zoeken. Omgekeerd evenredig met breedte wolk.



- Cte waarde voor Y als X = 0
- B = regressiecoëfficiënt = helling regrssielijn
  - Hoe Y verandert als X met 1 eenheid toeneemt

Statistiek

- **Y = a + bX**
- Waarbij:
  - **X** = de onafhankelijke (predictor) variabele (horizontale as)
  - **Y** = de afhankelijke (criterium) variabele (verticale as)
  - **a** = de constante, die het snijpunt (*intercept*) met de Y-as vormt
  - **b** = de hellingscoëfficiënt (*slope*, of richtingscoëfficiënt)
- De regressielijn een eenvoudig voorbeeld: het salaris (jaarsalaris schatten via maandsalaris)

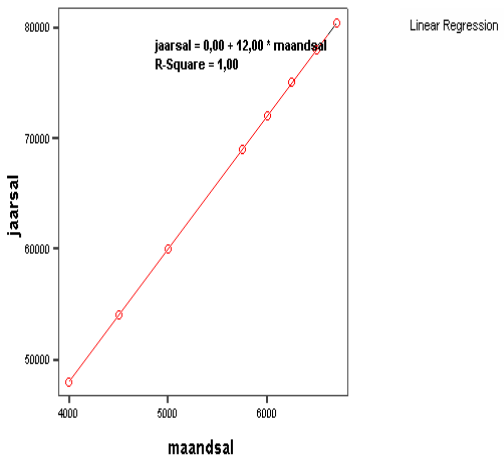
Maandelijks salaris	Jaarlijks inkomen
4.000	48.000
4.500	54.000
5.000	60.000
5.750	69.000
6.000	72.000
6.250	75.000
6.500	78.000
6.700	80.400

Correlations

		maandsal	jaarsalaris
maandsal	Pearson Correlation	1	1,000**
	Sig. (2-tailed)		,000
	N	8	8
jaarsalaris	Pearson Correlation	1,000**	1
	Sig. (2-tailed)	,000	
	N	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

Scatterplot maken



De regressielijn: Y= a + bX
Y = 12*X
Fouten bij de voorspelling?

- Voorbeeld, nu met eindejaarspremie € 1000

Maandelijks salaris	Jaarlijks inkomen
4.000	49.000
4.500	55.000
5.000	61.000
5.750	70.000
6.000	73.000
6.250	76.000
6.500	79.000
6.700	82.400

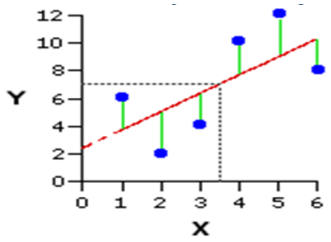
Welk is de r? en de scatterplot?
En de regressielijn?
Fouten bij de voorspelling?
Voorspelling perfect, want Pearson gelijk aan 1

10.1 Enkelvoudige regressielijn

Twee vragen:

- Hoe vinden we de parameters van de regressielijn (Y = a + bX)?
- Hoe goed kunnen we de Y waarden voorspellen op basis van dit model?

- Hoe vinden we deze a en b coëfficiënten?
  - **a:** dit is de uitslag van Y indien X nul bedraagt;
    a = Y – bX (intercept) (toename X, afname Y, afname X, toename Y))
  - **b:** dit is de richtingscoëfficiënt; deze is functie van de r en de verhouding tussen de beide SD. (Pearson +, b +, Pearson -, b -)
    b = r_{YX}. SD_Y/SD_X (slope) (1e = afh, 2e onaf) Deze b geeft aan hoe de Y waarde verandert wanneer de X waarde met één eenheid toeneemt.
- Op grond van deze vergelijking kunnen we voor elke score van X een verwachting voor Y formuleren.
  Y = a + bX
  Y = 2,4 + 1,31X
  bv. X = 3, Y = ?
  Y = 2,4 + 1,31*3
  Y = 6,31
- De regressielijn voldoet aan het criterium van het kleinste kwadraat. D.w.z. dat de gekwadrateerde afwijking van de verwachte uitslag t.o.v. de feitelijke uitslag minimaal is. (best mogelijke tendens weergeven in scatterplot.)



- **Algemene werkformule van de regressielijn:**
  $Y = Y + r_{XY} (X_i – X)*SD_Y/SD_X$

Vergelijking van de ***best*** passende lijn, waarbij de Y waarden zo goed mogelijk geschat kunnen worden op grond van de X waarden. De regressielijn is ***niet*** symmetrisch.

Hoe goed voldoet dit model om de werkelijkheid te voorspellen?

- **Waarde regressielijn:** Hoe goed verklaart het model de werkelijke gegevens?
  - Hoeveel % van de verschillen kan je verklaren = determinatiecoëfficiënt. (relatieve maat)
    - →0.40 = 16% (Pearson) wel verklaren, dus 84 % niet.
    - Wrm boeiend te tekenen → Pearson .70.
    - Proportie verklaarde variantie: r^2_{XY}
      - Deze determinatiecoëfficiënt geeft de gemeenschappelijk variantie weer.
      - Kan berekend worden via het kwadraat van de r. Voorstelling: r^2
    - Proportie niet-verklaarde variantie: $(1- r^2_{XY})$
  - Het verschil tussen de verwachte en de feitelijke score van Y is de schattingsfout.
    - De standaarddeviatie van deze fout is de standaardschattingsfout.
    - Dit komt overeen met de standaarddeviatie van de verschillen tussen de verwachte en feitelijke uitslag. (= absolute maat)
    - Standaarddeviatie van verdeling van fouten
    - $SD_{Y.X} = SD_Y*V1-r^2$

10.2 Standaardschattingsfout

- In PASW wordt deze standaardfout aangeduid middels ‘std. error of the estimate’.

- Deze standaardschattingsfout geeft een indicatie van de (on)nauwkeurigheid van de voorspelling.
- (omgekeerde) relatie met r_{XY}
- **PASW:** afh. boven = dependent, onafh. vanonder = independent. (analyse → regression linear)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,914[a]	,835	,808	4,219

a. Predictors: (Constant), IQ

- **R-square:** meer variabelen, meer naar 1.
- **Predictors:** Pearson-correlatie
- **Std.:** Sd van de fouten
- **R:** determinatiecoëfficiënt.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	541,088	1	541,088	30,402	,001[a]
	Residual	106,787	6	17,798		
	Total	647,875	7			

a. Predictors: (Constant), IQ

b. Dependent Variable: Schooluitsl

- Welk is de waarde van het model voor de populatie?
  - **Sig.:** significantieniveau, als groter dan 0.5, niet van toepassing op model. Des te kleiner, des te beter → significant en van toepassing op de populatie.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16,457	10,654		1,545	,173
	IQ	,517	,094	,914	5,514	,001

a. Dependent Variable: Schooluitsl

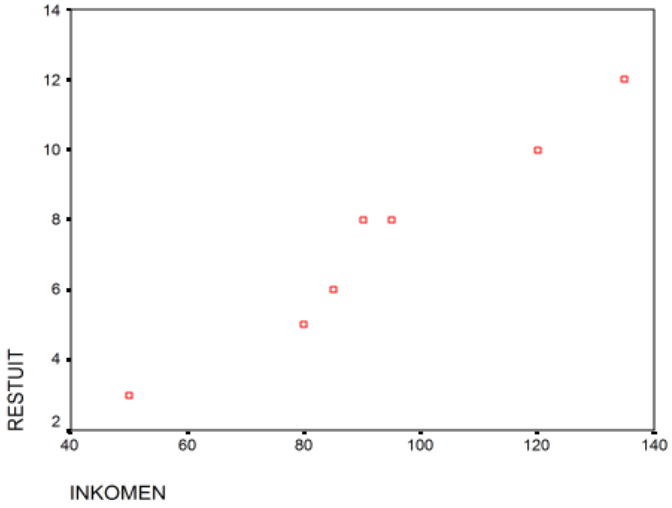
- SV =16,46 + 0,52*IQ
- Z-waarde Y voorspellen door Z-waarde X → Z-waarde X. 0.91

- Uitgaven restaurant en inkomen:

  - Wat is het verband tussen het inkomen en de uitgaven aan restaurantbezoek?
- Hoe zal uitgaven voor restaurantbezoek toenemen in functie van inkomensverandering?

Ppn	uitg. rest(afh.)	inkomen(onaf.)
1	10	120
2	5	80
3	6	85
4	3	50
5	12	135
6	8	90
7	8	95

- Eerste inzicht in de relatie tussen de twee variabelen via de puntenwolk (spreidingsdiagram)
- Bestaat er een rechtlijnig verband tussen beide variabelen? Ja → Pearson = zinvol.



- **Pearson correlatiecoëfficiënt**

Correlations

		RESTUIT	INKOMEN
RESTUIT	Pearson Correlation	1,000	,978**
	Sig. (2-tailed)	,	,000
	N	7	7
INKOMEN	Pearson Correlation	,978**	1,000
	Sig. (2-tailed)	,000	,
	N	7	7

**. Correlation is significant at the 0.01 level (2-tailed).

- = Covariantie gestandardiseerd
- Geeft een aanduiding van de sterkte en de richting van het verband tussen twee variabelen.
- Significantietoets: zie verder inductieve statistiek
- Pearson = 0.98 = hoog verband.

- **Output regressieanalyse**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2,657	1,000		-2,657	,045
	INKOMEN	,108	,010	,978	10,457	,000

a. Dependent Variable: RESTUIT

- 0.978 = Z-waarde inkomen.
- Restuit = -2,66 + 0,11*Inkomen

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,978[a]	,956	,948	,6980

a. Predictors: (Constant), INKOMEN

- R square is de determinatiecoëfficiënt.
- Geeft de verhouding aan tussen de verklaarde variantie en de totale variantie.

- **De regressievergelijking**
  - Restaurantuitgaven = -2,66 + 0,11 * Inkomen
  - Bijvoorbeeld: inkomen is 200, welk is dan de restaurantuitgave? -2,66 + 0,11 * 200 = 19,34.
  - En als X = 0 ?
- Wat is de waarde van deze vergelijking? Dekt het model de werkelijkheid?
  - 96% van de verschillen in de restaurantuitgaven kunnen verklaard worden door de verschillen in inkomen

  - De standaardfout van estimatie bedraagt: 0,70. Dwz dat in 2/3 van de gevallen de fout in de voorspelling kleiner zal zijn dan 0,70

- **Meervoudige regressie:**
  - In dit geval zijn er meerdere X variabelen, op grond waarvan de Y variabele geschat wordt.
  - Niet meer mogelijk via handmatige uitwerking, enkel via PASW.
  - Veel gebruikte procedure om aan te geven hoe diverse onafhankelijke variabelen gezamenlijk een invloed uitoefenen op de afhankelijke variabele. Diverse onafhankelijke variabelen worden t.o.v. mekaar uitgespeeld.
  - De afzonderlijke bètacoëfficiënten bieden een inzicht in het impact van elke onafhankelijke variabele, onder constant houding van de overige variabelen.