Statistiek 2

Samenvatting (cursus + notities)

Nieuw!!!

Online samenvattingen kopen via

www.quickprintershop.be

Like us on Facebook!

www.facebook.com/quickprintershop

Samenvatting Statistiek II

Academiejaar 2014-2015

HOOFDSTUK 1
VAN PROBLEEM NAAR ANALYSE

1.1 Inleiding

Multivariate analyse is een verzamelnaam voor een groep statistische technieken gericht op de analyse van samenhang tussen drie of meer variabelen onderling. De keuze van de multivariate analysetechniek varieert in functie van de onderzoeksvraag.

- Het kan een analyse zijn van een probleemkenmerk, zoals politiek vertrouwen of zittenblijven. In dit geval zijn we op zoek naar factoren of onafhankelijke variabelen die het kenmerk verklaren.
- Het kan een analyse zijn van een probleemrelatie, zoals het verschil in objectieve bestaansonzekerheid naar gewest. Dan stellen we ons de vraag of een verschil in objectieve bestaansonzekerheid verklaard kan worden aan de hand van regionale variantie in socio-economische positie van het gezinshoofd.
- Er kan ook een veelheid van items of uitspraken bevraagd worden in een survey en dan vragen we ons af welke achterliggende opinies of attitudes er gemeten worden (synthese en datareductie).

Het is belangrijk ook op te merken dat het meetniveau van de afhankelijke variabele, het aantal onafhankelijke variabelen, het meetniveau van de onafhankelijke variabelen, het (niet) lineair karakter van de effecten, additieve karakter van effecten, ... leiden tot verschillende **modelspecificaties en analysetechnieken**.

1.2 Notatie

Er zijn vier verschillende types van variabelen die we hier verder gaan bekijken, de vorm van variabele is belangrijk voor verdere modelspecificatie en keuze voor de beste analysetechniek.

- **Continue variabele** van interval- of rationiveau is manifest opgemeten (bv. de leeftijd van respondenten, en kan zowel de rol van afhankelijke als van onafhankelijke variabele opnemen.
- **Nominale of ordinale variabelen met twee categorieën (dichotome variabelen)** zijn ook manifest opgemeten (bv. geslacht) en kan zowel de rol van afhankelijke als van onafhankelijke variabele opnemen.
- **Polytome categorische variabele** is een variabele van nominaal of ordinaal meetniveau met drie of meer categorieën. Deze wordt manifest opgemeten en kan zowel de rol van afhankelijke als onafhankelijke variabele aannemen.
- **Latente variabele** van interval- of rationiveau is niet rechtstreeks opgemeten bij respondenten, bv. een complexe schaal berekend op basis van opgemeten items.

Er zijn ook verschillende types van samenhang die uit de analyse kunnen komen.

- **Symmetrische samenhang** tussen twee kenmerken, er wordt dan geen onderscheid gemaakt tussen onafhankelijke en afhankelijke variabelen.

- **Asymmetrische samenhang** waarbij de onafhankelijke variabele een **lineair** (causaal) effect uitoefent op een afhankelijke variabele. Bij een lineair effect heeft eenzelfde verandering in de onafhankelijke variabele steeds eenzelfde verandering in de afhankelijke variabele tot gevolg.
- **Asymmetrische samenhang** waarbij onafhankelijke variabele een **niet-lineair** effect uitoefent op afhankelijke variabele. Bij een niet-lineair effect wordt de verandering in afhankelijke variabele ten gevolge van eenzelfde verandering in onafhankelijke variabele gradueel groter of kleiner.
- **Interactie-effect** is een asymmetrische samenhang waarbij de combinatie van twee of meer onafhankelijke variabelen een causaal effect uitoefent op de afhankelijke variabele.

1.3 Sociaal-wetenschappelijke probleemstellingen

1.3.1 Inactiviteit en overgewicht

Uit een gezondheidsenquête blijkt dat overgewicht, dit wil zeggen een BMI hoger dan 27, vaker voorkomt bij mensen die niet beroepsactief zijn dan bij werkenden. Is er dan sprake van een causale relatie tussen beroepsactiviteit en overgewicht? Dit is een **bivariate causale structuur** die kijkt naar het effect van beroepsactiviteit op overgewicht.

We kunnen de samenhang tussen inactiviteit en overgewicht niet zonder meer interpreteren als een causale oorzaak-gevolg relatie. Onderzoek geeft namelijk aan dat BMI om verschillende redenen toeneemt met de leeftijd en werkzaamheid ligt lager bij oudere leeftijdsgroepen. Causaliteit kan dus niet afgeleid worden uit de bivariate samenhang omdat de relatie tussen inactiviteit en overgewicht veroorzaakt wordt door associatie van beide kenmerken met leeftijd: ouderen werken minder vaak en hebben een hoger BMI, terwijl jongeren beroepsactief zijn en gekenmerkt worden door een lager BMI, maar beide kenmerken zijn verder mogelijk niet geassocieerd. Dit is wat we noemen **schijnbare causaliteit**.

Dit zal onderzocht worden aan de hand van een analyse van **kruistabellen en elaboratie**. In de configuratie van schijnbare causaliteit verdwijnt samenhang tussen inactiviteit en overgewicht na controle voor leeftijd, dit wil zeggen wanneer er gekeken wordt naar respondenten met dezelfde leeftijd. Het principe van statistische controle neemt bij regressietechnieken de vorm aan van een vergelijking van verschillende **modelspecificaties**.

1.3.2 Regionale verschillen in objectieve bestaansonzekerheid

Uit de budget enquête in 1997 blijkt dat objectieve bestaansonzekerheid (= inkomen onder EU armoedelijn) drie keer frequenter voorkomt in Wallonië dan in Vlaanderen. Is er dan een causaal effect van regio op armoede of kunnen regionale verschillen verklaard worden door regionale variatie in sociaal-economische activiteit van het gezinshoofd? Onderzoek geeft aan dat gezinshoofden in Wallonië vaker inactief of werkloos zijn dan het geval is in Vlaanderen.

Hier hebben we te maken met **indirecte causaliteit**, in de causale relatie van regio naar objectieve bestaansonzekerheid vormt socio-economische activiteit een **intermediaire of tussenliggende variabele**, waardoor er sprake is van indirecte causaliteit.

Dit zal ook onderzocht worden aan de hand van analyse van **kruistabellen en elaboratie**. In de configuratie van indirecte causaliteit verdwijnt samenhang tussen regio en objectieve bestaansonzekerheid na controle voor sociaal-economische activiteit van het gezinshoofd.

Regionale verschillen verdwijnen met andere woorden wanneer er wordt gekeken naar gezinnen met eenzelfde sociaal-economische activiteit. Een kruistabel tussen regio en objectieve bestaansonzekerheid ongeacht sociaal-economische activiteit van gezinshoofd wordt vergeleken met een kruistabel tussen beide kenmerken voor gezinnen met eenzelfde sociaal-economische activiteit.

1.3.3 Leeftijd en subjectieve bestaansonzekerheid

Uit de budget enquête in 1997 blijkt dat er nauwelijks sprake is van samenhang tussen leeftijd van het gezinshoofd (jonger of ouder dan 65 jaar) en subjectieve bestaansonzekerheid (inkomen dat als voldoende wordt gepercipieerd om rond te komen). Nochtans beschikken gepensioneerden over een lager gezinsinkomen en een lagere levensstandaard dan personen op beroepsactieve leeftijd.

Als er aanvankelijk geen samenhang is, maar na controle voor een bijkomende factor wel, dan is er sprake van **suppressie** van een samenhang. Het zwakke bivariate verband tussen leeftijd en subjectieve bestaansonzekerheid is het gevolg van twee tegengestelde effecten die elkaar opheffen: ouderen hebben een lager inkomen dan gezinshoofden jonger dan 65 jaar en bij eenzelfde inkomen zijn ouderen minder vaak bestaansonzeker dan gezinshoofden jonger dan 65 jaar.

In het verband tussen subjectieve bestaansonzekerheid en leeftijd is inkomen een **suppressor variabele**: het negatieve verband tussen beide variabelen komt pas aan het licht na controle voor gezinsinkomen. Elaboratie neemt bij regressietechnieken de vorm aan van een vergelijking van verschillende **modelspecificaties**, namelijk het vergelijken van de regressieparameter voor en na controle.

1.3.4 Beroepsstatus en subjectieve gezondheid

Onderzoek gebaseerd op de sociaal-economische enquête in 2001 en Nationale Databank Mortaliteit toont aan dat België, net als andere Europese landen, wordt gekenmerkt door een uitgesproken socio-economische gradiënt in subjectieve gezondheid en mortaliteit. Gradiënt in subjectieve gezondheid naar beroepsklasse varieert sterk in functie van leeftijd: het gradiënt is beperkt in leeftijdscategorie 18-29 jaar en wordt groter naar oudere leeftijdscategorieën.

Effecten van leeftijd en beroepsklasse op subjectieve gezondheid kunnen niet zonder meer worden opgeteld, het effect van beroepsklasse wordt sterker naarmate leeftijd toeneemt. Er is dus sprake van een gecombineerd effect of **interactie** tussen de effecten van leeftijd en beroepsklasse.

Door middel van een analyse van kruistabellen wordt geïllustreerd hoe de samenhang tussen beroepsklasse en subjectieve gezondheid sterker wordt bij opeenvolgende leeftijdscategorieën, al zijn hier meestal meer gevorderde analysetechnieken voor. Bij regressietechnieken kan de interactieve structuur worden onderzocht door modelspecificatie uit te breiden met producttermen tussen onafhankelijke variabelen onderling.

1.3.5 Politiek vertrouwen

Politiek vertrouwen vormt een basisvoorwaarde voor het functioneren van democratie en politiek bedrijf. Politiek vertrouwen wordt geassocieerd met een hogeren kans op betalen van belastingen, een lagere kans op free-rider gedrag en steun voor herverdelend sociaal beleid. Een steeds conservatiever beleid in de VS sinds jaren 1960 wordt geassocieerd met een daling van politiek vertrouwen. In welke mate wordt politiek vertrouwen beïnvloed door factoren als leeftijd, opleidingsniveau en de perceptie of publieke opinie over het functioneren van overheidsinstellingen?

Politiek vertrouwen wordt verklaard aan de hand van opleiding en perceptie over instellingen, dit is een **convergente causale structuur** en komt aan bod bij meervoudige regressie. Er zijn verschillende modelspecificaties mogelijk (ook niet-lineaire en interactie effecten), die we onderling kunnen vergelijken door te controleren voor bijkomende factoren.

- Bivariate regressie
- Additief model met ongecorreleerde onafhankelijke variabelen
- Additief model met gecorreleerde onafhankelijke variabelen (multicolineariteit)
- Additief model met gecorreleerde onafhankelijke variabelen en niet-lineaire effecten
- Interactiemodel met gecorreleerde onafhankelijke variabelen
- Interactiemodel met gecorreleerde onafhankelijke variabelen en niet-lineaire effecten

1.3.6 Etnische identiteit bij minderheden

De Turkse en Marokkaanse gemeenschappen in België worden gekenmerkt door sterk verschillende motieven tot migratie bij pioniersmigranten in de jaren 1950/60. Turkse migratie is sterker vanuit economische motieven, terwijl Marokkaanse migratie naast economische motieven een ook sterkere culturele en politieke component kent. In welke mate wordt de etnische identiteit benadrukt door Turkse en Marokkaanse mannen in België en hoe varieert dit tussen beide nationaliteitsgroepen, en in functie van migratieleeftijd en vestigingsplaats in België? We gaan met andere woorden de afhankelijke variabele "integratie en assimilatie" onderzoeken aan de hand van onafhankelijke variabelen "nationaliteit", "woonplaats" en "migratieleeftijd". Dit is meervoudige regressie aan de hand van **dummyvariabelen.**

- Meervoudige regressie met een dummy onafhankelijke variabele
- Regressie met polytome onafhankelijke met k categorieën en k-1 dummyvariabelen
- Additief model met ongecorreleerde onafhankelijke variabelen
- Additief model met gecorreleerde onafhankelijke variabelen
- Interactiemodel met gecorreleerde onafhankelijke variabelen
- Interactiemodel met gecorreleerde onafhankelijke variabelen en niet-lineaire effecten

Datareductie = PCA en PFA

ZOEMGROEP 1.1

Bepaal de gepaste multivariate analysetechniek voor.

- Een analyse van inkomen (EUR) in functie van leeftijd in jaren, onderwijsniveau in jaren en anciënniteit in aantal jaren gewerkt – *meervoudige regressie*
- Een analyse van partijkeuze bij gemeenteraadsverkiezingen in functie van geslacht, leeftijd in jaren, opleidingsniveau in vijf klassen en huishoudentype – *multinomial logit*
- De constructie van schalen op basis van 35 items opgemeten – *PCA en PFA*
- Een analyse van werkloosheid in functie van geslacht, leeftijd in jaren, nationaliteit in zes groepen, opleidingsniveau in 5 klassen en arbeidservaring in 8 klassen – *logistische regressie*
- Een analyse van inkomen in euro, in functie van geslacht, leeftijd in jaren, onderwijsniveau in 5 klassen, en anciënniteit in aantal jaren gewerkt – *dummy regressie en variantie analyse*

HOOFDSTUK 2
METEN & MEETNIVEAUS

Bij een dataset in SPSS is het belangrijk dat we de getallen op een legitieme manier gebruiken. In de rijen van een dataset zijn meestal de respondenten gegeven en in de kolommen telkens de verschillende kenmerken die opgemeten werden. Opletten want getallen voor inkomen zijn niet hetzelfde als de getallen voor provincie en opleiding, die laatsten zijn nummers zonder betekenis.

1.1 Wat is meten

Meten is het verdelen van de populatie (P) in equivalentieklassen (bv. alle mensen uit de provincie Antwerpen). Er worden dan kenmerken of schalen gemeten van die equivalentieklassen aan de hand van een variabele. Aan nominale of ordinale variabelen kunnen we een getal toekennen, maar dit zijn arbitraire waarden, zonder intrinsieke betekenis.

1.2 Eigenschappen van variabelen of meetschalen

- **Ordenbaarheid**: de variabele of meetschaal is ordenbaar als wanneer voor elk paar elementen bepaald kan worden of het ene groter/kleiner is dan de andere. Bv. leeftijd, opleidingsniveau, inkomen, ... woonplaats is niet ordenbaar. De ordenbaarheid van de waarden weerspiegelt een bestaande ordening tussen equivalentieklassen van het bestudeerde kenmerk.
- Het bestaan van een **meeteenheid**
- Het bestaan van een **absoluut nulpunt**

Bewerkingen toegelaten voor variabelen van een bepaald meetniveau mogen worden toegepast op variabelen van een hiërarchisch hoger meetniveau, maar niet op variabelen van een lager meetniveau.

1.3 Dummy variabelen

Zijn categorische variabelen (nominaal of ordinaal) met twee categorieën, waarbij gebruik gemaakt wordt van 0/1 codering. Dummyvariabelen worden behandeld als variabelen van interval- of ratiomeetniveau.

ZOEMGROEP 2.1

HOOFDSTUK 3
FREQUENTIEVERDELINGEN EN GRAFISCHE VOORSTELLINGEN

1.1 Inleiding

Bij het aanvatten van een multivariate analyse is het belangrijk om eerst de afzonderlijke variabelen en frequenties in detail te bekijken. Enkele inleidende concepten:

- Steekproefgrootte of effectief N
- Absolute frequenties
- Relatieve frequenties

1.2 Schikking van gegevens op een nominale schaal

De waarden voor een nominale schaal zijn niet geordend. De volgorde waarin de waarden worden opgenomen is willekeurig volgens numerieke of alfabetische volgorde of volgens stijgende of dalende overeenkomende frequenties, bv. gewest. Grafisch kan dit worden weergegeven in een histogram/pictogram/cirkeldiagram waarbij de oppervlakte recht evenredig is met de frequenties.

1.3 Schikking van gegevens op een ordinale schaal

De frequentietabel wordt op dezelfde manier opgesteld als die van nominale variabelen, maar de volgorde van de frequenties in de frequentietabel is bij ordinale variabelen gebaseerd op de ordening van de waarden. We kunnen hierbij ook de cumulatieve frequenties berekenen. Een voorbeeld hiervan is de veiligheidsmonitor "hoe vaak overkomt u een onveiligheidsgevoel?" met als waarden: nooit, zelden, soms, vaak of altijd. Kan grafisch worden voorgesteld aan de hand van een staafdiagram/histogram/... met een gerichte X-as. Ook een cumulatieve frequentiefunctie is mogelijk (trapvorm). De cumulatieve functies zijn niet-strikt monotoon stijgend.

1.4 Schikking van gegevens op interval- of ratioschaal

Hierbij kunnen we een onderscheid maken tussen niet in klassen gegroepeerde gegevens en in klassen gegroepeerde gegevens.

- Niet in klassen gegroepeerde gegevens
  Er is nu sprake van een meeteenheid omdat de verschillen tussen waarden nu een betekenis hebben. Deze verschillen moeten door recht evenredige verschillen worden voorgesteld op de abscis. Omdat de verschillen op de X-as nu een betekenis hebben, krijgen ook oppervlakten onder de functies een betekenis. Dit laat het gebruik toe van frequentieveelhoeken en gebruik van lineaire interpolatie. Deze gegevens kunnen grafisch worden voorgesteld aan de hand van een staafdiagram, een frequentiepolygloon, een histogram of een cumulatieve frequentiefunctie. De X-as heeft nu een meeteenheid, de oppervlakte is evenredig met de absolute of relatieve frequentie in het interval.

- In klassen gegroepeerde gegevens

  Het aantal verschillende waargenomen waarden (n) is vaak te groot. In dat geval is het onoverzichtelijk of onmogelijk de klassen apart te beschouwen. Continue variabelen worden daarom vaak in klassen ingedeeld. Klassengrenzen zijn vaak inhoudelijk ingegeven bv. voor de berekening van afhankelijkheidsratio's worden leeftijdsklassen gehanteerd tot 18 jaar, 18-65 jaar of 65 en ouder. Omdat er een meeteenheid is hebben verschillen tussen waarden nu een betekenis. Deze verschillen moeten door recht evenredige verschillen worden voorgesteld op de X-as.

  Bepaling van de klassen

  1. Bepalen van de variatiebreedte of range: dit geeft het verschil weer tussen de grootste en de kleinste waargenomen waarde.
  2. Aantal klassen *k:* we veronderstellen dat alle waarden in een klasse equivalent zijn, deze hypothese geldt alleen als de verschillen niet te groot zijn. Meestal zijn er tussen de 5-15 klassen.
  3. Klasselengte: klassen van gelijke lengte is aangeraden en in geval van klassen van gelijke lengte, is de klasselengte bij benadering de variatiebreedte gedeeld door het aantal klassen. *V/k*
  4. Klassengrenzen (belangrijk of deze erbij gerekend zijn of niet)
  5. Klassemidden: het bepalen van de klassengrenzen bij

     discrete veranderlijken elke discrete waarde wordt vervangen door het overeenkomstige continue interval bv. 23 = (22.5-23.5)

     continue veranderlijken de exacte klasse komt overeen met de waarnemingsklasse

     Klassemidden wordt berekend als het gemiddelde van exacte klassegrenzen (zowel bij discrete als bij continue veranderlijken)

1.5 Doelstellingen H3

- Altijd frequenties bekijken en missing values niet meenemen in de analyse
- SPSS kan op verschillende manieren grafische voorstellingen geven, afhankelijk van het meetniveau
- Het omgekeerde van H2: hoge meetniveaus mogen geanalyseerd worden met lageren multivariate analysetechnieken (soms wenselijk zoals bij niet-lineair) bv. bij een verschil in belang van migratie dat op jonge leeftijd belangrijk is en afneemt naarmate men ouder is.

HOOFDSTUK 4
UNIVARIATE STATISTISCHE PARAMETERS

Univariate statistische parameters zijn belangrijke kengetallen om een bepaalde verdeling te schetsen. Hier gaat het nog niet over samenhang, maar vooral over centrummaten en maten voor spreiding. Frequentietabellen en grafieken zijn vaak onvoldoende om de informatie vervat in brutowaarnemingen te vatten. Om de gegevens te synthetiseren wordt gebruik gemaakt van kenmerkende waarden of kenwaarden, deze worden **parameters** genoemd. Er zijn typisch drie verschillende soorten parameters

1. Parameters van ligging
2. Parameters van spreiding
3. Parameters van vorm

4.1 Parameters van ligging of positie

Deze parameters laten toe de verdeling op de X-as te situeren en deze moet steeds tussen de kleinste en de grootste waargenomen waarde liggen (en met één van de waarden overeenkomen in geval van nominale variabelen). De keuze voor een parameter (bv. modus, mediaan, gemiddelde, ...) om een verdeling te beschrijven is afhankelijk van het meetniveau van de beschouwde verdeling. Deze parameters laten toe om efficiënt groepen te vergelijken: bv. kijken vrouwen meer televisie dan mannen, verschilt kijkgedrag naar opleidingsniveau, ... Dit is info die je gewoon aan de hand van een tabel niet kan zien.

Centraliteitsparameters of centrummaten

Dit is een *deelverzameling* van de parameters van ligging en geven aan rond welke waarde op X-as de verdeling gecentreerd is en welke waarde representatief is voor de verdeling. De keuze voor de centrummaat is afhankelijk van het meetniveau van een veranderlijke.

- Modus bij **nominale** schaal is de waargenomen waarde met de hoogste frequentie. Bv. de modus voor de veiligheidsmonitor is 2 = zelden. De modus heeft als voordeel dat deze gemakkelijk te bepalen is op basis van de frequentietabel, maar heeft als nadelen dat deze niet noodzakelijk uniek is (er kunnen meerdere modale klassen zijn) Bovendien houdt deze geen rekening met de andere waargenomen waarden.
- Mediaan bij een **ordinale** schaal is de waarde van de variabele die toelaat de waarnemingen in twee gelijke delen op te delen zodat er evenveel waarnemingen kleiner dan of gelijk zijn aan de mediaan als er groter dan of gelijk zijn aan. Bij in klassen gegroepeerde waarnemingen wordt de mediaan bepaald door middel van lineaire interpolatie (homogeniteitshypothese) De mediaan is enkel afhankelijk van de volgorde van de waarnemingen en niet alle waarden worden dus in rekening gebracht (minder gevoelig voor extreme waarden)
- Het **rekenkundig gemiddelde** is gelijk aan de som van de waarnemingen gedeeld door het effectief. De waarden van alle waarnemingen worden gebruikt bij de berekening van het gemiddelde, veruit de meest gebruikte centrummaat voor interval- of ratioveranderlijken. Het gemiddelde laat toe om verschillende groepen efficiënt te vergelijken wat betreft hun score voor een bepaald kenmerk. Regressie analyse zal later worden gebruikt om bv de gemiddelde kijkduur naar leeftijd te beschrijven.

- Het **meetkundig gemiddelde** wordt gebruikt bij logistische regressie en wordt onder meer gebruikt voor de berekening van gemiddelde groeivoeten en in statistische analyse.

ZOEMGROEP 4.1

Parameters van spreiding

Het begrip spreiding kijkt naar verschillen tussen respondenten bv. in inkomen. Een onderzoeksvraag van multivariate analysetechnieken die een vraag is naar spreiding is: in hoeverre kunnen verschillen die worden vastgesteld tussen personen voor mijdingsgedrag verklaard worden door verschillen voor andere kenmerken zoals leeftijd, geslacht, opleidingsniveau,...? Wanneer er vastgesteld wordt dat er verschillen zijn tussen personen wat betreft mijdingsgedrag (spreiding voor het kenmerk mijdingsgedrag), rijst de vraag naar de oorzaken van die verschillen en de verklaring voor die verschillen (zoals leeftijd, geslacht, ...) Spreiding en spreidingsmaten spelen een belangrijke rol in analysetechnieken die gericht zijn op de verklaring van dergelijke verschillen.

Spreidingsmaten geven weer in welke mate eenheden (respondenten, steden, ...) van elkaar verschillen voor een bepaald kenmerk of variabele, dit door weer te geven in hoeverre eenheden geconcentreerd liggen rond de centrummaat (modus, mediaan, gemiddelde), dan wel gespreid liggen over de hele range van een variabele. De keuze voor een spreidingsmaat wordt bepaald door het meetniveau van de veranderlijke.

Regels voor spreidingsmaten: wanneer alle waarnemingen eenzelfde waarde hebben (i.e. geen spreiding of verschillen tussen waarnemingen), dan moet de spreidingsmaat gelijk zijn aan nul. De spreidingsmaat wordt groter naarmate waarnemingen meer gespreid zijn of onderling meer verschillen.

Spreiding bij interval- en ratiovariabelen

- **Variatie** *(s)* of kwadratensom weerspiegelt de som van de gekwadrateerde afwijkingen van het rekenkundig gemiddelde.
- **Variantie** *(s^2)* weerspiegelt de gemiddelde gekwadrateerde afwijking van het gemiddelde. Variantie is het centraal moment van rang 2 rond het gemiddelde. Dit speelt een belangrijke rol in regressie- en variantieanalyse om na te gaan in hoeverre spreiding van een variabele verklaard kan worden door de spreiding van een andere variabele.
- **Standaardafwijking** *(s)* is de vierkantswortel uit de variantie en weerspiegelt de standaard-afwijking rond het rekenkundig gemiddelde.

ZOEMGROEP 4.2

Toepassing: de oppervlakte onder een normale verdeling

De standaardafwijking heeft een belangrijke interpretatie in termen van oppervlakten onder de normale verdeling. In een normale verdeling ligt

- 50% van de waarnemingen 2/3 sa links en 2/3 sa afwijkingen rechts van het gemiddelde
- 68% van de waarnemingen 1 sa links en 1 sa van het gemiddelde
- 95% van de waarnemingen 1.96 sa links en 1.96 sa rechts van het gemiddelde
- 99.7% van de waarnemingen tussen 3s links en 3s rechts van het gemiddelde

Gemiddelde en standaardafwijking laten toe de relatieve plaats te bepalen van waarnemingen op de verdeling van een variabele. Bv. van normaalverdelingen: lengte en gewicht van mensen, afwijkingen van het gemiddelde bij industriële productie.

Een standaardnormale verdeling heeft een gemiddelde van nul en een standaardafwijking van 1. De normale verdeling beschrijft aan de hand van de parameters μ en sigma de klokvormige verdeling van een stochastische variabele X. Aangezien de normale verdeling symmetrisch is rond het gemiddelde, kan aan de hand van de standaardafwijking worden bepaald in welk interval rond μ een bepaalde proportie van de observaties is begrepen. Vooral het 95% betrouwbaarheidsinterval is belangrijk voor significantietesten.

Toepassing: gestandaardiseerde scores

De gestandaardiseerde score van een waarneming (bv. een stad, een respondent) geeft weer hoeveel die waarneming boven of onder het gemiddelde ligt. Een gestandaardiseerde score is geen spreidingsmaat, maar een score die voor elke waarneming wordt berekend om die waarneming te situeren ten opzichte van andere waarnemingen. Dit wordt gebruikt bij de berekening van gestandaardiseerde regressiecoëfficiënten. *Hoeveel standaardafwijkingen zit een respondent boven of onder het gemiddelde?*

ZOEMGROEP 4.3

- **Variatiecoëfficiënt** is de standaardafwijking gedeeld door het gemiddelde. Deze is onafhankelijk van de meeteenheid van een variabele en laat toe om de mate van spreiding bij verschillende variabelen onderling te vergelijken.

Parameters van vorm

Twee frequent bestudeerde kenmerken van vorm van een verdeling zijn de symmetrie en de afplatting.

- De empirische **coëfficiënt van Pearson** S: bij een positief scheve verdeling geldt doorgaans dat de modus kleiner is dan de mediaan, kleiner is dan het gemiddelde.
  Gemiddelde – mediaan / standaardafwijking
  - × Symmetrisch indien S = 0
  - × Positief of linkse indien S > 0
  - × Negatief of rechts indien S < 0
- Een t-verdeling en een normale verdeling zijn symmetrisch en een Chi^2 is asymmetrisch

Afplatting kan voorkomen als platkurtisch, mesokurtisch of leptokurtisch.

HOOFDSTUK 5
INDUCTIEVE STATISTIEK

Inductieve statistiek is enkel relevant bij toevalssteekproeven en is belangrijk in verdere hoofdstukken voor significantietoetsen bij meervoudige regressie.

1.1 Inleiding en basisbegrippen

Sociaal-wetenschappelijk onderzoek is vaak gebaseerd op steekproeven. Er zijn verschillende doelstellingen van inferentiële of inductieve statistiek.

- Veralgemening van steekproefresultaten naar de populatie
- Door het formuleren van kansuitspraken d.m.v. hypothese of betrouwbaarheidsinterval
- Kan enkel bij toevalssteekproeven
- Is niet mogelijk bij non-probability samples want de kans om erin terecht te komen is onbekend, dus er kan geen gewicht aan gegeven worden in de analyse

Er zijn twee verschillende types van kansuitspraken

- Betrouwbaarheidsinterval: stelling dat het gemiddelde inkomen in de populatie waaruit de steekproef getrokken werd met een vooropgestelde zekerheid van 95 procent binnen een bepaald interval ligt.
- Hypothesetoets: toets waarbij de uitspraak die stelt dat het gemiddelde inkomen in Vlaanderen gelijk is aan een bepaald bedrag met een vooropgestelde zekerheid van 99 procent al dan niet wordt verworpen.

Een **steekproefgrootheid** is elk getal dat op basis van een steekproef kan worden berekend. Een schatting van een populatieparameter is een steekproefgrootheid, of het nu om een gemiddelde, een percentage, een correlatie of een regressiecoëfficiënt gaat. Sommige steekproefgrootheden vormen geen schatting van een populatieparameter, maar worden berekend voor een ander doel (bv. teststatistiek Chi² voor samenhang in een kruistabel, F-test in variantieanalyse, ...)

Steekproevenverdeling

Iedere steekproefgrootheid heeft een steekproevenverdeling. Je kan bijvoorbeeld duizend steekproeven trekken van dezelfde grootte (bv. N=10) uit eenzelfde populatie en je berekent telkens de gemiddelde leeftijd.

- Steekproefgemiddelden zullen van elkaar verschillen naargelang toevallig meer ouderen of jongeren werden getrokken.
- Steekproefgemiddelden liggen gespreid rond reële gemiddelde leeftijd in de populatie
- Van steekproefgemiddelden kan een verdeling worden opgesteld, de **steekproevenverdeling.**

Soms kan de steekproevenverdeling empirisch worden bepaald, maar vaak gaat het om een theoretische, hypothetische verdeling die mathematisch kan worden afgeleid.

Twee belangrijke eigenschappen van schatters houden verband met de kenmerken van hun steekproevenverdeling: zuiverheid en variabiliteit