Hoorcollege: Inspecteren van data: verdelingen

Er zijn drie verschillende soorten beschrijvend onderzoek

- Surveys: vragenlijst, interview
- Demografisch: statistisch onderzoek, bijvoorbeeld de statistiek die het CBS bijhoudt
- Epidemiologisch onderzoek: aanwezigheid van ziekten en psychologische aandoeningen

Er zijn drie verschillende soorten surveys

- Cross-sectioneel: een éénmalige doorsnede van de populatie, dus als je bijvoorbeeld een vragenlijst afneemt bij verschillende personen doe je dit éénmalig en daaruit trek je een conclusie.
- Opeenvolgende onafhankelijke steekproeven: je herhaalt de steekproef één of meerdere keren en vergelijkt de resultaten met elkaar. Je moet je wel afvragen of deze steekproeven vergelijkbaar zijn.
- Longitudinaal: je onderzoekt veranderingen over de tijd met één (dezelfde) groep. Dit kan voor problemen zorgen als er mensen uitvallen. Na verloop van tijd kan iemand bijvoorbeeld besluiten zijn deelname aan het onderzoek te beëindigen om wat voor reden dan ook.

Beschrijven en presenteren van data

Uit een onderzoek komen vaak een heleboel cijfers (ruwe data). De ruwe data zijn lastig te interpreteren. Daarom vatten we ze samen. Dit kan numeriek en/of grafisch. Een grafische weergave van een verdeling heet een plot. Dit verschilt van een numerieke verdeling (bijvoorbeeld een tabel). Voor een goede beschrijving van de data moet de data accuraat, beknopt en begrijpelijk zijn. Er is een spanningsveld tussen beknoptheid en accuraatheid. Hoe beknopter iets wordt beschreven hoe minder accuraat het is en andersom.

Hoe beschrijven we een verdeling?

- Algehele patroon
- Vorm: is de vorm van de verdeling symmetrisch of scheef? Hoeveel pieken zijn er te zien in de verdeling? (unimodaal één piek, bimodaal twee pieken en multimodaal meer dan twee pieken)
- Centrale tendentie/locatie: Welke vormen van centrale tendentie kunnen we beschrijven? (gemiddelde, mediaan, modus)
- Spreiding: Is er veel of weinig spreiding?
- Opvallende afwijkingen
- Uitbijters: waarnemingen die in hoge mate afwijken van de rest van de waarnemingen
- Staarten: is de staart dik of dun van een normaalverdeling. Wat betekent dit?

Absolute en relatieve frequenties

Absolute frequenties (f): aantal proefpersonen met een bepaalde score. Bijvoorbeeld bij een de afname van IQ-test bij mensen:

IQ van 120 8 keer

IQ van 122 6 keer

IQ van 135 1 keer

Een nadeel van absolute frequenties is dat ze moeilijk te vergelijken en interpreteren vallen. Is bijvoorbeeld de bovenstaande absolute frequentie IQ van 120 8 keer veel of weinig.

Relatieve frequenties (P): Proportie van een bepaalde score ten opzichte van het totaal (P = f/n). Bijvoorbeeld als 10 mensen een IQ-test maken en 4 daarvan hebben een score van 120 dan is de relatieve frequentie 4/10= 0.4. P = 0.4. Een voordeel van relatieve frequenties is dat je scores makkelijk kunt vergelijken en interpreteren.

Frequentieverdeling: gegroepeerde tabel

Frequentieverdelingen kun je in een gegroepeerde tabel plaatsen. Je moet dan eerst intervallen maken. De vuistregel voor de hoeveelheid intervallen is de wortel van N. Bij 49 groepen heb je dus 7 intervallen. Daarna moet je het bereik (range) bepalen. Range= hoogste waarde – laagste waarde. Als laatste moet je ervoor zorgen dat je gelijke intervalbreedtes hebt.

De intervalbreedte is: $\frac{Range}{Aantal\ intervallen}$. Een interval moet uitputtend & wederzijds exclusief zijn.

Als je de intervallen bij elkaar optelt, dan moet je ook de frequenties van de bij elkaar behorende intervallen bij elkaar optellen.

Frequentieverdeling: cumulatieve tabel

Bij een absolute cumulatieve frequentieverdeling tel je alle voorgaande frequenties bij elkaar op. Je kunt ook de proporties van een variabele cumulatief weergeven. Bijvoorbeeld bij de volgende data:

	Cumulatieve absolute frequentie (f)	Cumulatieve relatieve frequentie (F)
a 6 keer een IQ-score tussen de 70-79	6	0.200
b 8 keer een IQ-score tussen de 80-89	14	0.467
c 12 keer een IQ-score tussen de 90-110	26	0.867
d 2 keer een IQ-score tussen de 121-130	28	0.933
e 2 keer een IQ-score van 130 of hoger	30	1

Toelichting

Cumulatieve absolute frequentie (f): bij a zijn er 6 scores dus de absolute frequentie is gewoon 6. Bij b tel je daar de 8 frequenties bij op, dat maakt 14. Bij c tel je er vervolgens weer 12 bij op, dat maakt 26, etc. 30 is het totaal aantal absolute frequenties. Bij de cumulatieve relatieve frequenties is het principe hetzelfde, alleen hier tel je telkens de proportie van een frequentie ten op zichte van het totaal bij elkaar op. Bijvoorbeeld bij a 6/30= 0.200 -> 20% van de totale scores zijn tussen de 70 en 79. Bij b tel je de proportie 8/30= 0.267 bij de proportie van a op en dat maakt 0.267 + 0.200 = 0.467, etc. Alle cumulatieve relatieve frequenties bij elkaar opgeteld maakt 1 en dat is 100%.

Een cumulatieve frequentieverdeling kun je weergeven met een polygoon.

Maten voor centrale tendentie

De modus is de meest voorkomende waarneming.

Kijk bijvoorbeeld naar de volgende getallenreeks: 2.8.8.8.6.5.4.3.6.6.8.8

8 is hier de modus.

De mediaan ((n+1)/2) is de middelste waarneming. Hiervoor moet je alles eerst op volgorde zetten.

Kijk bijvoorbeeld naar de volgende getallenreeks: 2.4.6.9.12

6 is hier de mediaan

Bij de volgende getallenreeks zie je 2 middelste getallen: 2.4.6.7.9.12

6.5 is dan de mediaan (6+7= 13, 13/2= 6,5)

Het gemiddelde (□yi/n) is alle scores bij elkaar opgeteld delen door het aantal waarnemingen. Kijk bijvoorbeeld naar de volgende getallenreeks: 1 2 3 6.

3 is hier het gemiddelde ((1+2+3+6)/4= 3.

De modus en de mediaan blijven redelijk constant bij een normaalverdeling ongeacht de scores. Het gemiddelde daarentegen is erg gevoelig voor uitbijters.

Maten van spreiding

Range (R): hoogste score- laagste score

Variantie (σ2 of s2): $s^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$

Standaarddeviatie (σ of s): $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$

De standaarddeviatie en variantie zijn alleen geschikt voor spreiding rond het gemiddelde en ze zijn niet bestand tegen uitschieters. Kijk terug naar de samenvatting van college 1 voor de beschrijving van de verschillende letters/tekens van de variantie. De standaarddeviatie is de wortel van de variantie.

Interquartile range (IQR): Q3-Q1

De IQR kan verduidelijkt worden aan de hand van de Five-number summary:

Minimum: laagste score die geen uitbijter is

Q1: 25e percentiel; 25% van de scores is lager dan Q1 en 75% hoger.

Mediaan (Q2): 50e percentiel; 50% van de scores is hoger dan Q2 en 50% lager.

Q3: 75e percentiel: 75% van de scores is lager dan Q3 en 25% hoger.

Maximum: hoogste score die geen uitbijter is

De IQR is Q3 – Q1.

Een vuistregel voor uitbijters is dat als een observatie 1.5 keer de IQR boven de Q3 of 1.5 keer onder de Q! ligt dat het een uitbijter is. Bijvoorbeeld:

Data 3 13 17 19 22 24 25 28 35 39 44 45 83 86 93

Minimum: 3

Maximum: 93

Mediaan: (Q2): 28 ((n+1)/2 = (15+1)2 = 8) het achtste getal in deze reeks is 28.

Vervolgens deel je de data op in twee groepen: de data links naast de mediaan en de data rechts van de mediaan. Dan pas je dezelfde formule toe als bij de mediaan en dan is het antwoord op de data links naast de mediaan Q1 en rechts naast de mediaan Q3. In het geval van deze datareeks: Q1= 19 en Q3 is 45.

IQR = 45 – 19 = 26

Q1 – (26 x 1.5) = -20

Q3 + (26 x 1.5) = 84

Er zijn geen waarnemingen onder de -20. Er zijn 2 waarnemingen boven de 84, namelijk 86 en 93. Dit zijn uitbijters.

Meetniveau ‘s, grafieken en centrale tendenties

Bij een nominaal meetniveau hoort een staafdiagram of een taartdiagram. De bijbehorende centrale tendentie is de modus. Bij een ordinaal meetniveau kun je een staafdiagram, taartdiagram of een Boxplot gebruiken. Hierbij kun je de mediaan of de modus gebruiken als centrale tendentie. Bij een interval meetniveau of hoger kun je een taartdiagram, staafdiagram, boxplot of een histogram gebruiken. De centrale tendenties die mogelijk zijn bij een interval meetniveau of hoger zijn de modus, het gemiddelde en de mediaan.