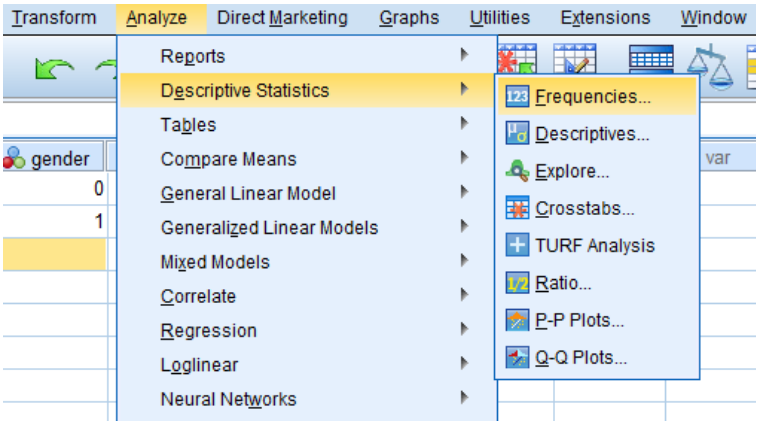
25. Data opschonen

Typefouten

Het is altijd heel belangrijk om je data even door te lopen op bijvoorbeeld typefouten. Je kan dan natuurlijk alle ingevoerde data nog een keer controleren aan de hand van de oorspronkelijke data, maar dit kost erg veel tijd. Een makkelijkere manier is het opvragen van Frequencies. Dit doet je door de volgende stappen te volgen: **Analyze ⟶ Descriptive Statistics ⟶ Frequencies.**



Het screenen en opschonen van de data

Voordat je je data kunt analyseren is het van belang om je databestand te controleren voor errors, mogelijke fouten. Als eerst is het belangrijk om te kijken of je typefouten hebt gemaakt (zie boven). Daarnaast is het essentieel om te onderzoeken of er andere fouten zijn met je data. Je volgt hiervoor de volgende stappen:

- **Stap 1:** Het controleren op errors. Eerst is het noodzakelijk om alle scores na te gaan van alle variabelen. Je onderzoekt dan of er bepaalde scores zijn die buiten de normale range vallen.
- **Stap 2:** Het vinden en controleren van error in het databestand. Vervolgens is het noodzakelijk om uit te zoeken waar de error zich bevindt in het databestand. Deze error dient dan of gecorrigeerd te worden of te worden verwijderd.

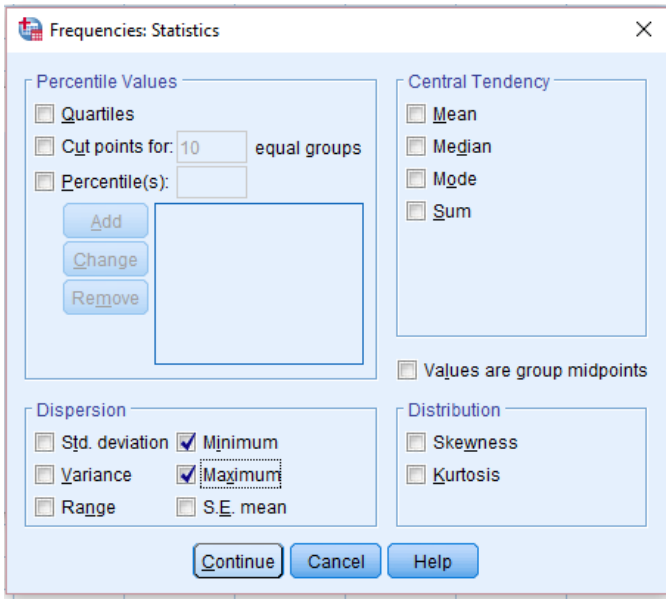
Stap 1: Het controleren op errors

Wanneer je je bestand controleert op errors ga je met name na of er waarden zijn die buiten de normale range van mogelijke scores vallen. Bijvoorbeeld: wanneer variabele ‘geslacht’ gecodeerd is met 0 of 1 (waarbij geldt 0 = man en 1 = vrouw), is het niet mogelijk om scores te vinden anders dan 0 of 1. Scores die een ander getal dan 0 of 1 hebben (bijvoorbeeld 2 of 3) dienen daarom te worden verwijderd of te worden aangepast. Er zijn verschillende manieren om errors te vinden met IBM SPSS. Deze kunnen grofweg worden verdeeld in twee methoden: één voor error bij categorische variabelen en één voor error bij continue variabelen.

Het checken van categorische variabelen

Volg de volgende procedure om error te controleren bij categorische variabelen.

1. Klik op **Analyze** en vervolgens op **Descsriptive Statistics** en dan op **Frequencies**.
2. Kies de variabelen die je wil checken (bijvoorbeeld geslacht). Om een variabele gemakkelijk te vinden kun je je variabelenlijst sorteren op alfabet.
3. Klik op de pijltjestoets (wijzend naar rechts) om de gewenste variabelen te verschuiven naar het variabelenvenster.
4. Klik vervolgens op **Statistics**. Vink **Minimum** en **Maximum** aan in de **Disperson** sectie.
5. Klik vervolgens op **Continue** en dan op **OK** (of op **Paste** om alles op te slaan in de Syntax Editor).



De syntax wordt als volgt gegenereerd:

```
FREQUENCIES VARIABLES=geslacht
  /STATISTICS=MINIMUM MAXIMUM
  /ORDER=ANALYSIS.
```

geslacht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	man	12	40,0	40,0	40,0
	vrouw	17	56,7	56,7	96,7
	2	1	3,3	3,3	100,0
	Total	30	100,0	100,0	

In bovenstaand voorbeeld zie je dat er één error is in het databestand. Er is namelijk één proefpersoon waarbij het geslacht is gecodeerd met cijfer 2 (in plaats van 0 of 1). Kijk daarom bij deze proefpersoon na of er sprake is van een mannelijk geslacht of vrouwelijk geslacht. Verander daarna de data van deze proefpersoon. Het kan ook voorkomen dat er bij een proefpersoon vergeten is om data in te voeren voor de desbetreffende variabele. In de tabel kun je deze vinden bij 'Missing'. In onderstaand voorbeeld is bijvoorbeeld te zien dat bij één proefpersoon de data voor variabele geslacht ontbreekt. Zoek deze proefpersoon op en kijk of je de data kunt corrigeren (zie beneden bij stap 2).

Stap 2: Het vinden en corrigeren van error in het databestand.

Wat te doen wanneer je responsen hebt gevonden die buiten de normale range vallen? Dan is het belangrijk om deze proefpersonen op te sporen. Dit kun je doen door de volgende stappen te ondernemen:

geslacht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	man	12	40,0	41,4	41,4
	vrouw	17	56,7	58,6	100,0
	Total	29	96,7	100,0	
Missing	System	1	3,3		
Total		30	100,0		

- Klik op **Data** en vervolgens kies je voor **Sort Cases**.

- In het dialoogvenster kies je vervolgens de variabele waarvoor je wist dat er sprake was van error (in dit geval dus ‘geslacht’). Klik op de pijltjestoets (wijzend naar rechts) en verplaats de variabele naar het **Sort By** venster. Kies dan uit **ascending** (van laag naar hoog) of **descending** (van hoog naar laag). In ons voorbeeld willen we graag de proefpersoon vinden die bij geslacht antwoordoptie ‘2’ had. We kiezen in dit geval dus voor aflopend (descending).
- Klik dan op **OK.**

Case summaries

Summarize Cases geeft je een tabel met daarin specifieke informatie voor elke proefpersoon. Je volgt de volgende stappen om deze samenvatting te verkrijgen:

1. Klik op **Analyze**, ga naar **Reports** en kies dan voor **Case Summaries**.
2. Kies de variabelen waarin je geïnteresseerd bent (in dit geval geslacht, provincie en leeftijd).
3. Klik op **Statistics** en verwijder **Number of Case** van het **Cell Statistics** venster. Klik dan op **Continue**.
4. Klik op **Options** en verwijder **Subheadings for totals**.
5. Klik op **Continue** en vervolgens op **OK** (of op **Paste** als je de analyse wil opslaan in de Syntax Editor).

   De syntax wordt als volgt gegenereerd:

```
SUMMARIZE
  /TABLES=geslacht provincie leeftijd
  /FORMAT=VALIDLIST NOCASENUM NOTOTAL LIMIT=5
  /TITLE='Case Summaries'
  /MISSING=VARIABLE
  /CELLS=NONE.
```

Case Summaries[a]

	geslacht	provincie	leeftijd
1	2	Noord-Holland	21
2	vrouw	Zuid-Holland	24
3	vrouw	Groningen	23
4	vrouw	Friesland	25
5	vrouw	Drenthe	26

a. Limited to first 5 cases.

In het voorbeeld is alleen een samenvatting gegeven van de eerste vijf proefpersonen. Dit kun je aangeven door onder **Display Cases** bij **Limit cases to first** het aantal te noteren (in dit geval 5).