Slide 1

Initiële data analyse

(Truuks en Flessenhalzen)

Herman Adèr

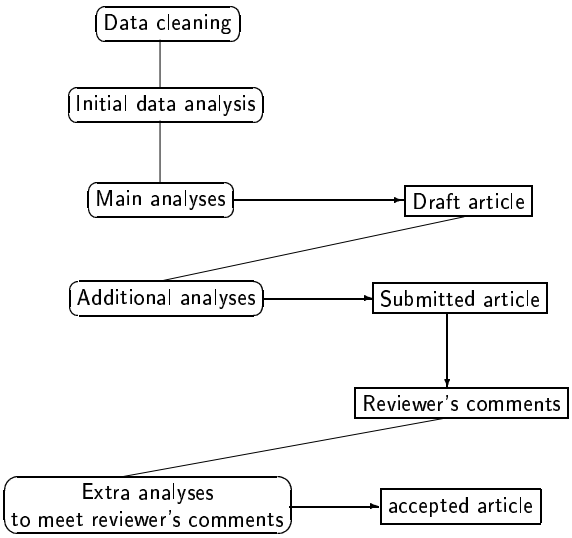
13 Mei, 2003

Slide 2

Overzicht

- Fasen in de data analyse
- Data kwaliteit
- Initiële Data Analyse
- Behoud van informatie
- Ontbrekende waarnemingen
- Meetniveau van de variabelen
- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen
- Principe van bootstrapping
- Transformaties
- Overzicht

Slide 3



Slide 4

Overzicht

- Fasen in de data analyse
- $\Longrightarrow$ *Data kwaliteit*
- Initiële Data Analyse
- Behoud van informatie
- Ontbrekende waarnemingen
- Meetniveau van de variabelen
- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen
- Principe van bootstrapping
- Transformaties
- Overzicht

Slide 5

Typen kwaliteit:

- Methodologische kwaliteit
- Kwaliteit van de rapportage
- Data kwaliteit

Slide 6

Kwaliteitscontrole tijdens het onderzoeks-proces:

- Initiële data analyse
- Systematisch reviewen
- Citation index
- Peer review

Slide 7

Overzicht

- Fasen in de data analyse
- Data kwaliteit

$\Longrightarrow$ *Initiële Data Analyse*

- Behoud van informatie
- Ontbrekende waarnemingen
- Meetniveau van de variabelen
- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen
- Principe van bootstrapping
- Transformaties
- Overzicht

Slide 8

Typische vragen die gedurende de initiële data analyse beantwoord moeten worden:

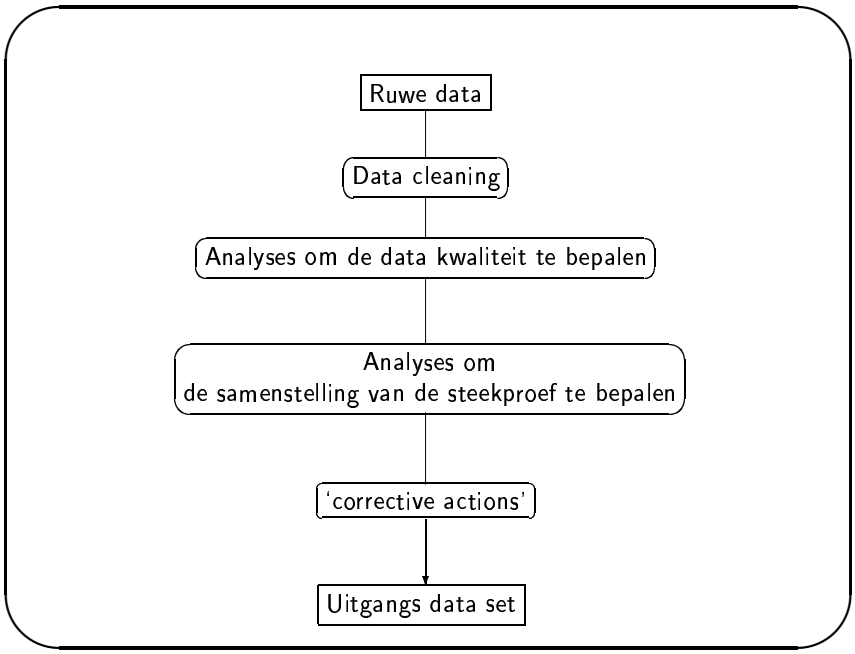
1. Wat is de kwaliteit van de data?
2. Is het onderzoeks-ontwerp geslaagd?
3. Wat is de samenstelling van de steekproef?

Vraag: Wat is het belang van ieder van deze punten voor de rest van het onderzoek?

Commentaar op Slide 8:

1. De *kwaliteit* van de data bepaalt de betrouwbaarheid van de analyse-resultaten
2. Hier moet men denken aan zaken als *randomisatie*, maar in een pilot kan het ook van belang zijn dat alle subgroepen van een populatie vertegenwoordigd zijn.
3. Dit punt is natuurlijk van algemeen belang bij het beschrijven van het onderzoek. Meer in het bijzonder speelt het een rol wanneer men sommige analyses graag in *subgroepen* wil kunnen uitvoeren.

Slide 9



Slide 10

Principe 1 *Gedurende IDA doen we geen analyses gericht op het beantwoorden van de onderzoeksvraagstellingen.*

Vraag: *Waarom niet?*

Slide 11

Waaraan moeten de meest gebruikte statistische analyse technieken voldoen?

Kruistabellen (χ^2-toets): Geen structurele nullen in de cellen.

T-test: De verdeling mag niet te scheef zijn in de groepen.

Multiple lineaire regressie-analyse (en (M)ANOVA, GLM, MLwiN): Normaal verdeelde *residuen.*

Variantie-analyse: Geen lege of slecht gevulde cellen.

(Confirmatieve) factor-analyse: Geen missings, geen 'slecht' verdeelde variabelen.

Cox regressie (Survival analyse): Normaal verdeelde residuen. Censurering onafhankelijk van overlevingstijd.

Vraag: Wat is de relevantie van de bovenstaande voorwaarden voor de initiële data analyse?

VOORBEELD 1 (**Twijfelachtige data waarden.**).

Van een meetinstrument dat per seconde 10 waarden registreert, staat vast dat de uitkomsten integer waarden moeten zijn in de range $-100 - - + 100$. Maar vreemd genoeg wordt in een data set een waarde 222 ontdekt.

Onderzoeker A neemt aan dat tijdens registratie een schrijfarm heeft gehaperd en dat de waarde in de de data set eigenlijk 22 had moeten zijn.

Onderzoeker B neemt liever het zekere voor het onzekere en verandert de waarde in een ontbrekende waarde.

Vraag: *Wat zou uw aanpak zijn?*

In veel gevallen kan men eenvoudig nazoeken (bijvoorbeeld in de patient status) wat de oorspronkelijke waarde had moeten zijn. Maar in het geval dat in Slide 12 wordt beschreven is dat onmogelijk.
In zo'n geval is het belangrijkste dat zowel de oorspronkelijke waarde (222) als de aangepaste waarde bewaard blijft (22 of missing) bewaard blijft.
Dit 'bewaren' kan op verschillende manieren, maar het eenvoudigst is om een extra variabele v' aan de data set (meestal: het SPSS system file) toe te voegen. De oorspronkelijke waarden blijven nu bewaard in de oorspronkelijke variabele v, terwijl bij de nieuwe, extra variabele v' op de plaats waar in v twijfelachtige waarden stonden, door de onderzoeker vervangende waarden zijn ingevuld. Dit vereist weld at verschillen tussen de twee variabelen worden gedocumenteerd, bijvoorbeeld in de value labels van de nieuwe variabele v'.
Het voordeel van de bovenstaande methode is, dat de oorspronkelijke waarden beschikbaar blijven, zodat een eventuele foutieve beslissing over de vervangende waarden kan worden teruggedraaid. Een ander voordeel is dat de aanpassingen traceerbaar blijven voor iedereen die met het data file werkt.

Slide 12

Overzicht

- Fasen in de data analyse
- Data kwaliteit
- Initiële Data Analyse

$\Longrightarrow$ *Behoud van informatie*

- Ontbrekende waarnemingen
- Meetniveau van de variabelen
- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen
- Principe van bootstrapping
- Transformaties
- Overzicht

Slide 13

Voorbeelden van situaties waarin veranderingen in de data worden aangebracht:

- Tijdens Data cleaning
- Samennemen van subgroepen
- In categoriën indelen van (continue) variabelen (leeftijd !)
- Imputeren van ontbrekende waarnemingen
- Vervangen van uitbijters of extreme waarden

Het voorbeeld in Slide 12 is een bijzonder geval van een veel algemener principe dat geldt tijdens alle fasen van de data analyse (zie Slide 15).

Slide 14

Principe 2 (Behoud van informatie.) *Bij de opeenvolgende data manipulaties dient alle informatie uit voorgaande stappen behouden en toegankelijk te blijven.*

Er zijn twee verschillende methoden om dit te verwezelijken: Zoals in het voorbeeld in Slide 12, kan men een nieuwe variabele toevoegen waarin veranderingen worden aangebracht. Een andere methode is om het system file te kopieren en in de *kopie* veranderingen aan te brengen in de oorspronkelijke variabelen. De laatste methode is vooral zinvol als men anders grote hoeveelheden nieuwe variabelen zou moeten aanmaken, ieder met een nieuwe naam, of wanneer de oorspronkelijke variabelen bij de verdere data analyse geen rol zullen spelen (en het dus onzin is om ze de rest van de analyses mee te slepen).

Slide 15

Overzicht

- Fasen in de data analyse
- Data kwaliteit
- Initiële Data Analyse
- Behoud van informatie

$\Longrightarrow$ *Ontbrekende waarnemingen*

- Meetniveau van de variabelen
- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen
- Principe van bootstrapping
- Transformaties
- Overzicht

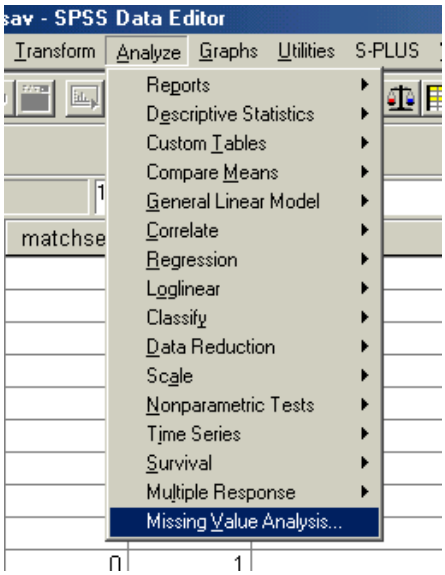
Slide 16

Principe 3 (Ontbrekende waarnemingen) *De ontbrekende waarden moeten worden gecodeerd en het imputeren moet worden gedocumenteerd.*

Vraag: *Wat is imputeren?*

Omdat de noodzaak om ontbrekende waarnemingen te imputeren meestal ligt in de erop volgende analyses (factor-analyse, variantie-analyse), worden de geimputeerde variabelen meestal in een apart file opgeslagen (dat is ook de manier waarop SPSSgeïmputeerde variabelen opslaat)

Slide 17




```
MVA
  cesd1 cesd2 cesd3 cesd4 cesd5 cesd6 cesd7 cesd8 cesd9 cesd10 cesd11 cesd12
  cesd13 cesd14 cesd15 cesd16 cesd17 cesd18 cesd19 cesd20
  /MPATTERN
  /TPATTERN PERCENT=1
  /EM ( TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25
  OUTFILE='C:\a\emgo27\Data-analyse-cursus\Analysis\imputed.sav' ).
```

Slide 18

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent
CESD1	202	.38	.62	1282	86.4
CESD2	202	.22	.53	1282	86.4
CESD3	202	.18	.50	1282	86.4
CESD4	202	2.22	1.04	1282	86.4
CESD5	202	.49	.76	1282	86.4
CESD6	202	.43	.72	1282	86.4
CESD7	202	.72	.87	1282	86.4
CESD8	202	2.16	.99	1282	86.4
CESD9	201	.24	.59	1283	86.5
CESD10	202	.25	.51	1282	86.4
CESD11	202	.63	.82	1282	86.4
CESD12	202	2.30	.95	1282	86.4
CESD13	201	.44	.68	1283	86.5
CESD14	202	.41	.75	1282	86.4
CESD15	202	.24	.58	1282	86.4
CESD16	202	2.35	.94	1282	86.4
CESD17	202	.16	.47	1282	86.4
CESD18	202	.36	.64	1282	86.4
CESD19	202	.27	.57	1282	86.4
CESD20	202	.67	.81	1282	86.4

Slide 19

EM Correlations[a]

	CESD1	CESD2	CESD3	CESD4	CESD5	CESD6	CESD7	CESD8	CESD9	CESD10	CESD11	CESD12
CESD1	1.000											
CESD2	.178	1.000										
CESD3	.344	.169	1.000									
CESD4	-.248	-.067	-.257	1.000								
CESD5	.351	.233	.413	-.361	1.000							
CESD6	.313	.140	.522	-.340	.388	1.000						
CESD7	.338	.168	.439	-.449	.450	.476	1.000					
CESD8	-.309	-.199	-.280	.566	-.381	-.386	-.449	1.000				
CESD9	.279	-.014	.483	-.468	.381	.518	.438	-.418	1.000			
CESD10	.300	.219	.694	-.239	.304	.498	.385	-.336	.484	1.000		
CESD11	.257	.279	.345	-.201	.250	.332	.398	-.369	.290	.352	1.000	
CESD12	-.289	-.163	-.347	.517	-.399	-.362	-.372	.701	-.428	-.311	-.400	1.000
CESD13	.377	.184	.493	-.258	.384	.427	.305	-.331	.297	.394	.224	-.320
CESD14	.404	.196	.581	-.316	.357	.406	.376	-.356	.478	.508	.317	-.418
CESD15	.339	.149	.385	-.117	.218	.242	.240	-.220	.166	.238	.266	-.193
CESD16	-.310	-.118	-.479	.561	-.433	-.400	-.404	.679	-.512	-.437	-.351	.763
CESD17	.283	.174	.519	-.292	.377	.457	.360	-.337	.438	.376	.432	-.371
CESD18	.345	.160	.614	-.349	.397	.616	.414	-.403	.551	.571	.422	-.464
CESD19	.357	.065	.428	-.309	.328	.338	.303	-.252	.298	.338	.169	-.279
CESD20	.322	.206	.518	-.444	.539	.479	.626	-.356	.546	.455	.370	-.386

a. Little's MCAR test: Chisquare = 136.087, df = 58, Prob = .000

Slide 20

Met repeated measurement multilevel analyse (MLwiN) en GEE is het mogelijk een data set waar op bepaalde tijdstippen waarnemingen ontbreken, te analyseren.

Dat kan niet bij repeated measures GLM in SPSS.

GEE: **G**eneralized **E**stimating **E**quations.
GLM: **G**eneralized **L**inear **M**odelling.

Het zelfde soort opmerkingen als voor ontbrekende waarnemingen kan worden gemaakt voor uitbijters. Alleen leveren uitbijters nog een extra methodologisch probleem: *Het is soms moeilijk om precies aan te geven wat een echte uitbijter is en waarom hij in de data voorkomt.* Pas wanneer ze goed geïdentificeerd zijn, kunnen ze als ontbrekende waarnemingen worden behandeld.

Principe 4 (Uitbijters en extreme waarden.) *Uitbijters mogen alleen worden verwijderd, wanneer hun* aanwezigheid *onafhankelijk is van de primaire uitkomst variabele. Extreme waarden mogen nooit worden verwijderd.*

Vraag: *Wat betekent 'verwijderd' in het bovenstaande principe?*
Vraag: *Wat betekent 'onafhankelijk van de primaire uitkomst variabele'?.*
Vraag: *Waarom mogen extreme waarden niet worden verwijderd?*
Een belangrijke vuistregel wordt gegeven in Slide **??**.
Wanneer aan de bovengenoemde onafhankelijkheid niet is voldaan, is het het handigste om een nieuwe variabele te introduceren, bijvoorbeeld: `Uit`, die de aanwezigheid van een uitbijter aangeeft en de uitbijters zelf als missing te behandelen. In de hoofdanalyses wordt `Uit` eerst in het (regressie) model opgenomen. Wanneer de aanwezigheid van uitbijters niet samenhangt met de waarde van de afhankelijke variabele, wordt de, eventueel geïmputeerde orginele variabele in het model opgenomen. Als de aanwezigheid van uitbijters wel invloed heeft, werken we in plaats daarvan met `Uit`.

Slide 21

Overzicht

- Fasen in de data analyse
- Data kwaliteit
- Initiële Data Analyse
- Behoud van informatie
- Ontbrekende waarnemingen

$\Longrightarrow$ *Meetniveau van de variabelen*

- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen
- Principe van bootstrapping
- Transformaties
- Overzicht

Slide 22

Meetniveau van de variabelen:

- Categorisch.
  Voorbeelden: Sexe, Cases/Controls, Onderzoeksgroep, Huisarts/Verplegend personeel;
- Ordinaal.
  Voorbeelden: Hoeveelste kind in het gezin? Niet mee eens enz. Tentamen beoordeling.
- Continue variabelen.
  Voorbeelden: VAS score, BMI, Lengte, Leeftijd, Bloeddruk.

Slide 23

EDUC educ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 primary school	171	11.5	11.9	11.9
	2 LBO	259	17.5	18.0	29.8
	3 MULO, MAVO	277	18.7	19.2	49.0
	4 MBO	200	13.5	13.9	62.9
	5 MMS, HAVO, HBS, VWO	163	11.0	11.3	74.2
	6 HBO	230	15.5	16.0	90.2
	7 Universiteit	88	5.9	6.1	96.3
	8 andere opleiding	54	3.6	3.7	100.0
	Total	1442	97.2	100.0	
Missing	System	42	2.8		
Total		1484	100.0		

Merk op dat de tabel in Slide 25 verraadt dat de aantallen in sommige groepen wel erg klein zijn. er is reden om klassen bij elkaar te nemen. We komen daar later op terug, wanneer we over transformaties praten.

GENDER gender * MARSTAT marstat Crosstabulation

			MARSTAT marstat				
			1 married/cohabit	2 not married	3 widow(er)	4 divorced	Total
GENDER gender	1.00 male	Count	593	120	15	18	746
		% within GENDER gender	79.5%	16.1%	2.0%	2.4%	100.0%
		% within MARSTAT marstat	53.8%	55.0%	15.5%	37.5%	50.9%
		Adjusted Residual	3.9	1.3	-7.2	-1.9	
	2.00 female	Count	509	98	82	30	719
		% within GENDER gender	70.8%	13.6%	11.4%	4.2%	100.0%
		% within MARSTAT marstat	46.2%	45.0%	84.5%	62.5%	49.1%
		Adjusted Residual	-3.9	-1.3	7.2	1.9	
Total		Count	1102	218	97	48	1465
		% within GENDER gender	75.2%	14.9%	6.6%	3.3%	100.0%
		% within MARSTAT marstat	100.0%	100.0%	100.0%	100.0%	100.0%

Uitvoer HILOGLINEAR:

```
Tests of PARTIAL associations.

  Effect Name                DF  Partial Chisq    Prob  Iter


  SEX*MARSTAT*EDUC           21         37.742   .0138     4
  SEX*MARSTAT*HA              3           .000  1.0000     3
  SEX*EDUC*HA                 7           .000  1.0000     2
  MARSTAT*EDUC*HA            21           .000  1.0000     3
  SEX*MARSTAT                 3         43.840   .0000     3
  SEX*EDUC                    7         75.078   .0000     3
  MARSTAT*EDUC               21        108.657   .0000     3
  SEX*HA                      1           .225   .6353     4
  MARSTAT*HA                  3          4.670   .1976     4
  EDUC*HA                     7          3.252   .8607     4
  SEX                         1           .159   .6904     2
  MARSTAT                     3       1689.848   .0000     2
  EDUC                        7        268.660   .0000     2
  HA                          1       1947.867   .0000     2
```

De bijbehorende setup is:

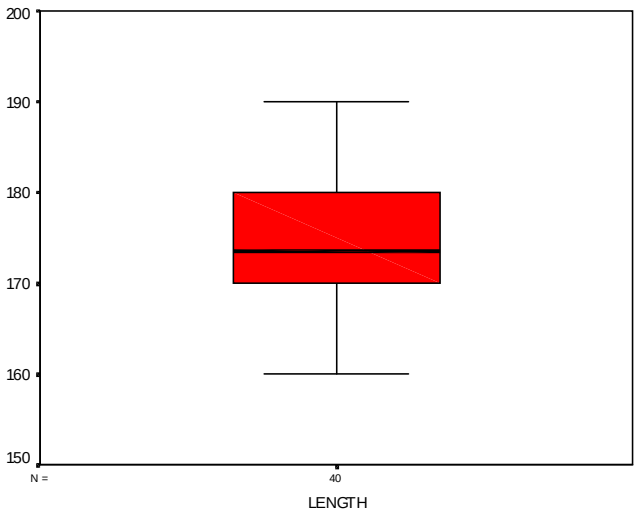
```
HILOGLINEAR
  sex(1 2) marstat(1 4) educ(1 8) ha(0 1)
 /METHOD=BACKWARD
  /CRITERIA MAXSTEPS(10) P(.05) ITERATION(20) DELTA(.5)
  /PRINT=association
  /DESIGN.
```

Hierboven staat de output van het programma HILOGLINEAR in SPSS. Hiermee kunnen alleen categoriale variabelen tegen elkaar worden uitgezet en de meest opmerkelijke kruistabellen worden gevonden, maar er is nog ander programma (LOGLINEAR) waarmee ook de invloed van continue en ordinale confounders kan worden bestudeerd. Het gebruik daarvan hoort meer in de hoofdanalyses thuis.

Descriptives

			Statistic	Std. Error
CESDSUM	Mean		15.1248	.3655
	95% Confidence Interval for Mean	Lower Bound	14.4040	
		Upper Bound	15.8456	
	5% Trimmed Mean		14.7136	
	Median		14.0000	
	Variance		26.990	
	Std. Deviation		5.1952	
	Minimum		4.00	
	Maximum		39.00	
	Range		35.00	
	Interquartile Range		5.0000	
	Skewness		1.523	.171
	Kurtosis		3.641	.341

Slide 24



Slide 25

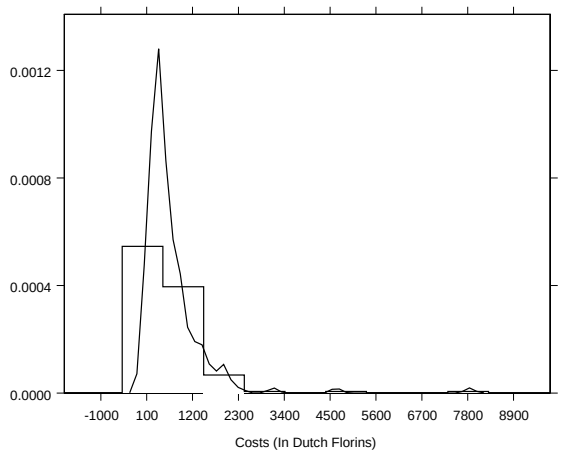
```
LENGTH Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1.00       16 .  0
     7.00       16 .  5555555
    12.00       17 .  000000000002
     7.00       17 .  5555578
     6.00       18 .  000003
     5.00       18 .  55559
     2.00       19 .  00

 Stem width:  10.00
 Each leaf:       1 case(s)
```

Slide 26



Bootstrapping.
De nu volgende vier slides (Slide 30–32) betreffen eigenlijk niet de initiële data analyse, want bootstrapping wordt toegepast in de fase waarin de hoofd-analyses worden uitgevoerd.
De reden dat het onderwerp is opgenomen is dat al diegenen die met met kosten-effectiviteits analyse te maken zullen krijgen, ook met de principes van bootstrapping vertrouwd zullen moeten zijn.

Slide 27

Overzicht

- Fasen in de data analyse
- Data kwaliteit
- Initiële Data Analyse
- Behoud van informatie
- Ontbrekende waarnemingen
- Meetniveau van de variabelen
- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen

$\Longrightarrow$ *Principe van bootstrapping*

- Transformaties
- Overzicht

Slide 28

Slide 29

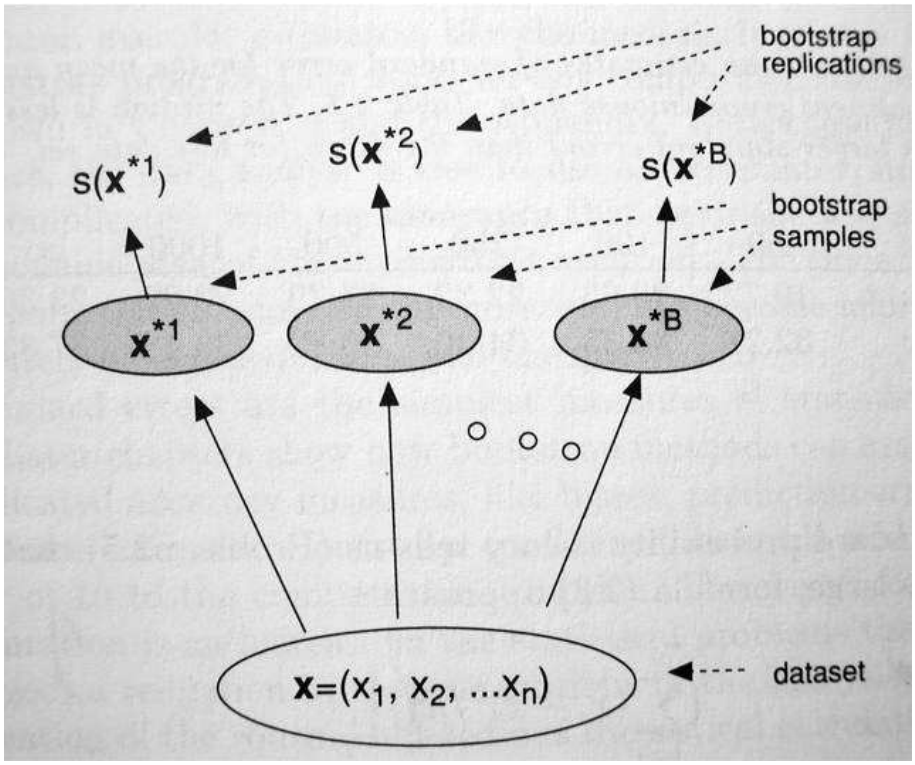
Voorbeeld van bootstrap samples $(n = 5, B = 7)$**.**

$1, 2, 3, 4, 5$

Bootstrap samples:

$$3, 2, 3, 1, 5$$
$$1, 2, 3, 3, 5$$
$$2, 2, 3, 1, 4$$
$$1, 3, 2, 1, 5$$
$$5, 4, 3, 3, 1$$
$$4, 5, 1, 3, 2$$
$$4, 1, 1, 4, 5$$

In Slide 31 wordt een voorbeeld van een zeer kleine steekproef gegeven: het toont wat men zich bij het begrip *bootstrap sample* moet voorstellen. Met ziet goed dat eenzelfde waarneming meerdere malen in eenzelfde sample mag voorkomen.



Slide 30

Situaties waarin bootstrap procedures nuttig kunnen zijn

- De waarschijnlijkheids-verdeling van de statistic die ons interesseert is onbekend of theoretisch te gecompliceerd.
- De steekproef omvang is klein
- Power berekeningen

Het eerste komt vaak voor bij het analyseren van trials waarbij ook kosten-variabelen zijn verzameld (kosten-effectiviteit studies). De kosten zelf zijn vaak niet-normaal verdeeld (zie Slide 28) dus vergelijkingen tussen groepen worden vaak met behulp van bootstrapping gedaan. Daarnaast wordt voor het berekenen van betrouwbaarheids-intervallen van de *kosten-effectiviteits-ratio* ook vaak bootstrapping gebruikt.

Op de laatste mogelijkheid (power berekeningen) gaan we hier niet in.

Slide 31

Overzicht

- Fasen in de data analyse
- Data kwaliteit
- Initiële Data Analyse
- Behoud van informatie
- Ontbrekende waarnemingen
- Meetniveau van de variabelen
- IDA van (i) Categoriale variabelen (ii) Continue variabelen (iii) Kosten variabelen
- Principe van bootstrapping

$\Longrightarrow$ *Transformaties*

- Overzicht

Slide 32

Vraag: *Wanneer zijn transformaties nodig?*

De analyses die tijdens IDA worden gedaan zijn onder andere bedoeld om informatie te krijgen om te kunnen beoordelen of transformaties nodig zijn.

Slide 33

Transformaties:

Missings en Uitbijters: Imputatie.

Categoriale variabelen: Herindelen in kleinere groepen.

Ordinale variabelen: Som (Schaal)-scores.

Continue variabelen: Indelen in Hoog–Middel–Laag.

Continue variabelen: Log-transformatie

Slide 34

Categoriale variable: groepen bij elkaar nemen.

```
compute cateduc=educ.
recode cateduc (1,2=1)(3,4=2)(5=3)(6,7=4)(8=9).
missing values cateduc (9).
value labels cateduc 1 'prim-LBO' 2 'MUL-MAV-MBO'
 3 'MMS-HAVO-VWO' 4 'HBO,Uni' 9 'Anders,missing'.
freq cateduc.
```

De oorspronkelijke frequentie-verdeling is te vinden op Slide 25. Wanneer de setup in Slide 36 wordt uitgevoerd, worden een aantal klassen bij elkaar genomen.

Slide 35

Continue (ordinale) variabelen: somscore berekenen

```
compute cesdsum = means.19(cesd1, cesd2, cesd3,
cesd4, cesd5, cesd6, cesd7, cesd8, cesd9, cesd10,
cesd11, cesd12, cesd13, cesd14, cesd15,
cesd16, cesd17, cesd18, cesd19, cesd20)*20.
examine cesdsum/plot=none.
```

Slide 36

Continue variabelen: in klassen indelen

```
compute catage=age.
recode catage (lo thru 35=1)(35 thru 45=2)
            (45 thru 65=3)(65 thru hi=4).
value labels catage 1 '<36' 2 '36-45'
                    3 '46-65' 4 '>65'.
freq catage.
```

Slide 37

CATAGE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 <36	297	20.0	20.3	20.3
	2.00 36-45	228	15.4	15.6	36.0
	3.00 46-65	620	41.8	42.5	78.4
	4.00 >65	315	21.2	21.6	100.0
	Total	1460	98.4	100.0	
Missing	System	24	1.6		
Total		1484	100.0		

Slide 38

Continue variabelen: logaritme nemen

```
compute lncessum=ln(cesdsum).
examine lncessum/plot=none.
```

Uitvoer **RELIABILITY**:

```
R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)


       N of Cases =       201.0

                                                    N of
Statistics for       Mean    Variance    Std Dev  Variables
      Scale        2.6617    10.4050     3.2257          7

Item Means           Mean     Minimum     Maximum      Range    Max/Min   Variance
                     .3802      .1791      .7114      .5323     3.9722      .0332

Item Variances       Mean     Minimum     Maximum      Range    Max/Min   Variance
                     .4368      .2478      .7563      .5086     3.0526      .0352

Inter-item
Correlations         Mean     Minimum     Maximum      Range    Max/Min   Variance
                     .4232      .2999      .6933      .3934     2.3121      .0087


Item-total Statistics

               Scale          Scale      Corrected
               Mean         Variance       Item-         Squared          Alpha
             if Item        if Item        Total         Multiple        if Item
             Deleted        Deleted    Correlation    Correlation       Deleted

CESD1         2.2935         8.5084        .4446         .2016           .8186
CESD3         2.4826         8.2309        .6744         .5626           .7903
CESD5         2.1791         7.6278        .5286         .2983           .8088
CESD6         2.2338         7.3600        .6401         .4267           .7872
CESD7         1.9502         6.9475        .5892         .3550           .8025
CESD9         2.4179         8.0445        .6046         .3839           .7954
CESD10        2.4129         8.3636        .6073         .5242           .7982


Alpha =   .8238           Standardized item alpha =   .8370
```

Bijbehorende setup:

```
RELIABILITY
  /VARIABLES=cesd1 cesd2 cesd3 cesd4 cesd5 cesd6 cesd7 cesd8 cesd9 cesd10
    /FORMAT=NOLABELS
  /SCALE(ALPHA)=ALL/MODEL=ALPHA
  /STATISTICS=SCALE
  /SUMMARY=TOTAL MEANS VARIANCE CORR.
```

Slide 39

Overzicht van belangrijke punten:

1. Initiële data analyse is voor een belangrijk deel gericht op het nagaan van de kwaliteit van de data.
2. Gedurende IDA doen we geen analyses gericht op het beantwoorden van de onderzoeksvraagstellingen.
3. Bij de opeenvolgende data manipulaties dient alle informatie uit voorgaande stappen behouden en toegankelijk te blijven.
4. In het bijzonder moeten bij *imputatie* van ontbrekende waarden of uitbijters en bij *transformaties* moeten zowel de oorspronkelijke waarden als de nieuwe waarden bewaard blijven.
5. Het meetniveau bepaald welke analyses het meest geschikt zijn tijdens IDA.
6. De slechte verdelingseigenschappen bij kosten-variabelen kunnen vaak met behulp van bootstrapping worden opgevangen.