Simple procedures

Karl B Christensen
http://publicifsv.sund.ku.dk/~kach/SPSS

2. Simple procedures

- transformation
- descriptive procedures
- selection
- sorting data: `split file`

- graphical representation

Example: Lung function in cystic fibrosis patients

Data from O'Neill et.al. (1983)

Table 12.11 Data for 25 patients with cystic fibrosis (O'Neill *et al.*, 1983)

Sub	Age	Sex	Height	Weight	BMP	FEV_1	RV	FRC	TLC	PEmax
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80
7	11	1	139	30.7	89	28	305	179	119	65
8	12	1	150	28.4	69	18	369	198	103	110
9	12	0	146	25.1	67	24	312	194	128	70
10	13	1	155	31.5	68	23	413	225	136	95
11	13	0	156	39.9	89	39	206	142	95	110
12	14	1	153	42.1	90	26	253	191	121	90
13	14	0	160	45.6	93	45	174	139	108	100
14	15	1	158	51.2	93	45	158	124	90	80
15	16	1	160	35.9	66	31	302	133	101	134
16	17	1	153	34.8	70	29	204	118	120	134
17	17	0	174	44.7	70	49	187	104	103	165
18	17	1	176	60.1	92	29	188	129	130	120
19	17	0	171	42.6	69	38	172	130	103	130
20	19	1	156	37.2	72	21	216	119	81	85
21	19	0	174	54.6	86	37	184	118	101	85
22	20	0	178	64.0	86	34	225	148	135	160
23	23	0	180	73.8	97	57	171	108	98	165
24	23	0	175	51.1	71	33	224	131	113	95
25	23	0	179	71.5	95	52	225	127	101	195

http://publicifsv.sund.ku.dk/~kach/SPSS/pemax.sav

http://publicifsv.sund.ku.dk/~kach/SPSS/pemax.txt

http://publicifsv.sund.ku.dk/~kach/SPSS/pemax.xlsx

Definition of new variables

We want to study body mass index

```
DATASET ACTIVATE DataSet2.
COMPUTE BMI=weight/(height/100) ** 2.
EXECUTE.
```

Transformations/Arithmetics

- The usual operators: + - * /
- Raising to a power: **, e.g.. `x**2`
- Square root: `SQRT(x)`
- Logarithms: `LN(x)`, `LG10(x)`

Summary statistics

- Measures of location, centre
  - Average

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$$

    interpreted as the centre of gravity - heavily influenced by outlying observations
  - Median = the middle observation, is not influenced by outlying observations (*robustness*)
  - Variance

$$s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$$

  - Standard deviation $SD = \sqrt{\text{variance}}$ is on the original scale
  - Quantiles (cutpoints dividing distribution into intervals with equal probabilities)
    - median: 50% quantile
    - quartiles: 25%, 50% and 75% quantiles

Summary statistics in SPSS

Syntax

```
MEANS TABLES=pemax BY sex
  /CELLS=MEAN COUNT STDDEV MEDIAN MIN MAX.

FREQUENCIES VARIABLES=pemax
  /NTILES=4
  /STATISTICS=MEAN STDDEV MEDIAN
  /ORDER=ANALYSIS.
```

gives us the output

Report

pemax

sex	Mean	N	Std. Deviation	Median	Minimum	Maximum
1	117,50	14	38,618	100,00	70	195
2	98,45	11	22,827	90,00	65	134
Total	109,12	25	33,437	95,00	65	195

Categorical variables

Means are not the right way to illustrate distributions of categorical variables. Use

```
GET FILE = 'p:\bissau.sav'.
DISPLAY NAMES.
```

to get the `bissau.sav` data set. Tables for `bcg`, `dtp` and `dead`:

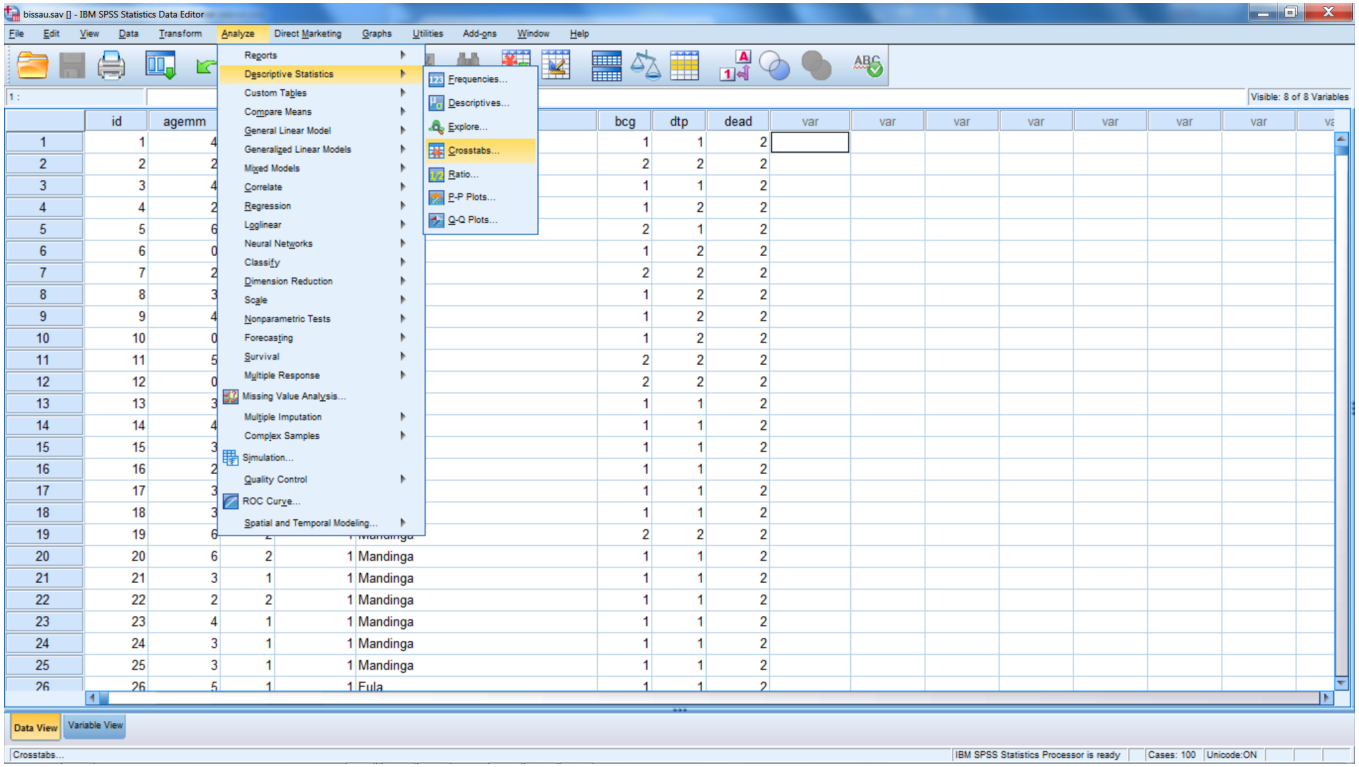
```
FREQUENCIES VARIABLES=bcg dtp dead
  /ORDER=ANALYSIS.

CROSSTABS
  /TABLES=bcg BY dead
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

Note: row percentages are chosen, because these have an interpretation

Categorical variables

Use



remember to click 'Paste'

Categorical variables

Crosstabs

bcg * dead Crosstabulation

			dead		
			1	2	Total
bcg	1	Count	124	3176	3300
		% within bcg	3,8%	96,2%	100,0%
	2	Count	97	1876	1973
		% within bcg	4,9%	95,1%	100,0%
Total		Count	221	5052	5273
		% within bcg	4,2%	95,8%	100,0%

Note: row percentages are chosen, because these have an interpretation

Filtering data

Can select subsets

Obs	age	csex	fev1	pemax	bmi
:	:	m	:	:	:
:	:	m	:	:	:
:	:	m	:	:	:
:	:	f	:	:	:
:	:	f	:	:	:
:	:	f	:	:	:

Obs	age	csex	fev1	pemax	bmi
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:

How to make a smaller data set

Can keep or delete variables. Keep three variables

```
*Set working directory.
cd 'P:\'.
*Open data file.
GET FILE='P:\bissau.sav'.
* Make small data set.
SAVE OUTFILE= 'P:\small.sav'
  /KEEP bcg dtp dead.
```

can also specify which variables we want to keep

```
GET FILE='P:\bissau.sav'.
SAVE OUTFILE='P:\alsosmall.sav'
  /DROP id agemm sex region ethnic.
```

Select subset

```
GET FILE='P:\bissau.sav'.
SELECT IF (agemm <= 3).
FREQUENCIES VARIABLES=dtp dead.
```

P:\bissau.sav

Statistics

		dtp	dead
N	Valid	3489	3489
	Missing	0	0

Frequency Table

dtp

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	916	26,3	26,3	26,3
	2	2573	73,7	73,7	100,0
	Total	3489	100,0	100,0	

dead

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	139	4,0	4,0	4,0
	2	3350	96,0	96,0	100,0
	Total	3489	100,0	100,0	

Sorting data - 'split file'

Use

http://publicifsv.sund.ku.dk/~kach/SPSS/F2_gif1.gif

or

```
SORT CASES  BY sex.
SPLIT FILE SEPARATE BY sex.
```

Now data are sorted by sex and all analyses are stratified until we specify

```
SPLIT FILE OFF.
```

Sorting data - 'split file'

Runs analyses within groups (stratified analyses)

```
GET FILE='P:\pemax.sav'.

SORT CASES  BY sex.
SPLIT FILE SEPARATE BY sex.

FREQUENCIES VARIABLES=pemax
  /FORMAT=NOTABLE
  /NTILES=4
  /STATISTICS=MEDIAN
  /ORDER=ANALYSIS.
```

Output

sex = 1

Statistics[a]

pemax

N	Valid	14
	Missing	0
Median		100,00
Percentiles	25	92,50
	50	100,00
	75	161,25

a. sex = 1

sex = 2

Statistics[a]

pemax

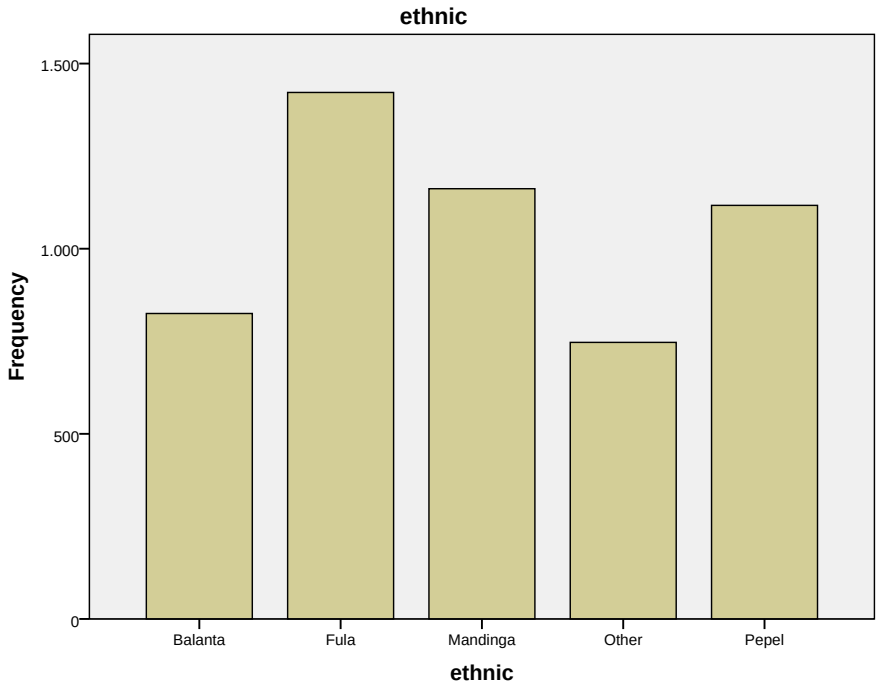
N	Valid	11
	Missing	0
Median		90,00
Percentiles	25	85,00
	50	90,00
	75	120,00

a. sex = 2

Descriptive statistics - bar charts

```
GET FILE='P:\bissau.sav'.

FREQUENCIES ethnic region
/FORMAT NOTABLE
/BARCHART.
```



The Juul data set

Serum IGF-I (Insulin-like Growth Factor) reference data set

Age	N	Source
0-5	44	Circumcision, hernia operation
5-20	833	4 schools in the Copenhagen area
20+	153	Hospital staff

Anders Juul et al., Dep. GR, Rigshosp.

AGE	age
MENARCHE	1st menstrual period occurred (1/2, 2 for yes)
SEXNR	1 for boys, 2 for girls
SIGF1	Serum IGF-I
TANNER	Puberty stage (1-5)
TESTVOL	Testicular volume
WEIGHT	weight

http://publicifsv.sund.ku.dk/~kach/SPSS/juul2.sav

Exercise: Simple procedures

1. Find the data set `juul2.sav` on the homepage and save on your computer.
2. Read the data set into SPSS using syntax. Compute `lsigf1=LN(sigf1)`
3. Calculate median and IQR of `sigf1` for each Tanner group using `split file`.
4. Use 'crosstabs' and bar charts to compare the distribution of the variable Tanner across the two genders.
5. Make a new variable with BMI for each person
6. Describe BMI distribution for each Tanner stage.