Statistics

Section: Data management

Exercises

0 Import data into SPSS

0.1 Import data from "Titanic_learn.xls" into SPSS

Description:

The list includes the names of all 1,310 Passengers (not including crew) known to be aboard the Titanic on April 14, 1912.[1]

Two children were under one year of age:

1. Dean, Miss Elizabeth Gladys, "better known as Millvina, was born on 2 February 1912. She was the daughter of Bertram Frank Dean and Georgette Eva Light Dean. In April, 1912 she was only nine-weeks-old and was, with her parents and elder-brother Bertram, about to emigrate to Wichita, Kansas where her father hoped to open a tobacconist shop."

2. "Goodwin, Sidney Leslie was born on 9th September 1910.
He boarded the Titanic with his parents, Frederick,and Augusta and siblings Lillian, 16; Charles, 14; William, 11; Jessie, 10; and Harold, 9.
The entire family was lost in the sinking."

Exercise

- Import data into SPSS
- Choose suitable measures (scale/ordinal/nominal) to all variables
- Use the shortened codebook below to label variables and values
- Define correct missing data
- Generate a codebook of your result via "Display Data Info"
- Paste the description above and your generated codebook into the Data File Comment[2]
- Save this original file under a suitable name into a suitable directory like (See **Appendix: File Structure**). You should never touch it again. This is your original file. Work only with copies.

Variable Information

Variable	Position	Label	Measurement Level Nominal/Ordinal/Scale
name	1	Name	
class	2	Class	
sex	3	Sex	
way	4	Way of traveling	
age	5	Age	
status	6	Status	

Variable Values

Value		Label
class	1	1. Class
	2	2. Class
	3	3. Class
sex	1	male
	2	female
way	1	single
	2	group
status	1	survived
	2	died

[1] slightly modified and translated from: Bühl A (2006) SPSS 14, Pearson, München
also see http://www.encyclopedia-titanica.org/

[2] Hint: this is a little bit tricky, you have to use MS-WORD as an intermediate reformatter.

0.2 Human development report 1999 (HDR) "HDR99_learn.xls" into SPSS

Description:

The data set contains some of the variables of the Human development report which is released annually from UNDP[3].

Some argue that data show that high divorce rates caused high suicide rates in male.

Exercise

- Import data into SPSS
- Choose suitable measures (scale/ordinal/nominal) to all variables
- Use the shortened codebook below to label variables and values
- Define correct missing data (if any)
- Generate a codebook of your result via "Display Data Info"
- Paste the description above and your generated codebook into the Data File Comment
- Save this original file under a suitable name into a suitable directory like (See **Appendix: File Structure**). You should never touch it again. This is your original file. Work only with copies
- Discuss the above argument

Variable Information

Variable	Position	Label	Measurement Level Nominal/Ordinal/Scale
country	1	Country	
life97	2	Life expectancy at birth 1997	
lit97	3	Adult literacy rate 1997	
cfste97	4	Combined first-sec. and third level gross enrolment ratio 1997	
gdp97	5	Real GDP per capita (PPP$) 1997	
hdi97	6	Human development index 1997	
hdirg97	7	HDI-rank	
indust	8	Industrialized countries Eastern Europe and the CIS(*)	
lifefem	9	Female: Life expectancy at birth 1997	
lifemale	10	Male: Life expectancy at birth 1997	
suimale	11	Male: Suicides per 100,000 1990-95	
suifem	12	Female: Suicides per 100,000 1990-95	
divorces	13	Divorces as % of marriages 1996	
unemplra	14	Total unemployment rate (%) 1997	
lngunfem	15	Female: Incidence of long term unemployment (% of total)	
lngunmal	16	Male: Incidence of long term unemployment (% of total)	
sthdi	17	Stadium of Human Development	

(*) CIS=Commonwealth of Independent States, a modern-day political entity consisting of 11 former Soviet Union Republics

Variable Values

Value		Label
indust	0	No
	1	yes
sthdi	1	High human development
	2	Medium human development
	3	Low human development

[3] Source: Human development report 1999 (UNDP, 1999)

0.3 "Grade Point Average" survey (GPA)[4] (GPA.xls) (STATA or SPSS)

Data are from a "Grade Point Average" survey which took all 78 seventh-grade students in a rural Midwestern school.

Some argue that data show a relationship between the students "self concept" and their academic performance.

id : Identification of students (11 students dropped out)
gpa : **g**rade **p**oint **a**verage
iq : Score on a standard IQ test
gender : 1 = Female 2=Male
concept: Self-concept test score (Piers-Harris Children's Self-Concept test)

Population center values (IQ=100; Self-Concept=60)
gender, gpa and iq were taken from school records.

Exercise

- Import data from the GPA.xls into SPSS or STATA
- Choose suitable measures (scale/ordinal/nominal) to all variables (SPSS)
- Define correct missing data
- Generate a codebook of your result via "Display Data Info"
- Paste the description above and your generated codebook into the Data File Comment
- Consider statistical descriptive and or inferential procedures
- Discuss the result

[4] Moore (2006), p.39 (Data provided by Darlene Gordon, Purdue University)

0.4 "High School and Beyond" survey (HSB)[5] (HSB_learn.xls) (STATA or SPSS)

Data are a random sample from the " High School and Beyond " survey conducted 1980 in the USA Principal Investigator: National Center for Education Statistics this data collection contains information from the first wave of the longitudinal study of American youth conducted by the National Opinion Research Center on behalf of the National Center for Education Statistics (NCES).

Data were collected from 58,270 high school students (28,240 seniors and 30,030 sophomores) and 1,015 secondary schools in the spring of 1980.

Data here contain just a few variables:

Variable	Label
id	
female	
race	
ses	Socioeconomic status
schtyp	type of school
prog	type of program
read	reading score
write	writing score
math	math score
science	science score
socst	social studies score

Variable	Value	Label
female		male
	1	female
race	1	hispanic
	2	asian
	3	african-amer
	4	white
ses	1	low
	2	middle
	3	high
schtyp	1	public
	2	private
prog	1	general
	2	academic
	3	vocation

Socio-Economic Status (SES)

"Definition
The socio-economic status is characterized by the economic, social and physical environments in which individuals live and work, as well as demographic and genetic factors. Measures for SES may include: Income or Income Adequacy, Education, Occupation, or Employment.
…
Social scientists have shown continued interest in SES even though there has never been complete consensus on precisely what it represents (Liberatos et al. 1988, McLoyd 1997)."
Robert H. Bradley , Robert F. Corwyn, Annual Review of Psychology, 2002

[5] Data source: Philip B. Ender < **http://www.gseis.ucla.edu/courses/ed230bc1/stata.htm >**

0.5 Import data from the General Social Survey (GSS93_learn.xls) into SPSS

Description:

The GSS has been ongoing since1972.
Year by year, careful multi-stage sampling has produced a representative sample of around 1,500 English-speaking people over the age of 17 (⩾18) within the continental United States.
The GSS consists of more than 1,000 questions.

Exercise

- Import data from the General Social Survey (GSS93.xls) into SPSS
- Choose suitable measures (scale/ordinal/nominal) to all variables
- Use the shortened codebook in **Appendix: GSS93** to label variables and values
- Define correct missing data
- Generate a codebook of your result via "Display Data Info"
- Paste the description above and your generated codebook into the Data File Comment
- Save this original file under a suitable name into a suitable directory like (See **Appendix: File Structure**). You should never touch it again. This is your original file. Work only with copies.

Hint:
Use copy and paste to transfer value labels from one variable to another

See **Appendix: Example for a code book**
how a appropriate description of your own codebook could look like

See http://webapp.icpsr.umich.edu/GSS/
about coding instructions and other important topics

Appendix: shortened Codebook for GSS93[6]

Variable Information

Variable	Position	Label	Measurement Level Nominal/Ordinal/Scale	Missing Values
id	1	Respondent ID Number		
wrkstat	2	Labor Force Status		0, 9
marital	3	Marital Status		9
agewed	4	Age When First Married		0, 98, 99
sibs	5	Number of Brothers and Sisters		98, 99
childs	6	Number of Children		9
age	7	Age of Respondent		0, 98, 99
birthmo	8	Month in Which R Was Born		0, 98, 99
zodiac	9	Respondents Astrological Sign		0, 98, 99
educ	10	Highest Year of School Completed		97, 98, 99
degree	11	RS Highest Degree		7, 8, 9
padeg	12	Father's Highest Degree		7, 8, 9
madeg	13	Mother's Highest Degree		7, 8, 9
sex	14	Respondent's Sex		
race	15	Race of Respondent		
income91	16	Total Family Income		0, 98, 99
rincom91	17	Respondent's Income		0, 98, 99
region	18	Region of Interview		0
xnorcsiz	19	Expanded N.O.R.C. Size Code		0
size	20	Size of Place in 1000s		-1
partyid	21	Political Party Affiliation		8, 9
vote92	22	Voting in 1992 Election		0, 8, 9
polviews	23	Think of Self as Liberal or Conservative		0, 8, 9
cappun	24	Favor or Oppose Death Penalty for Murder		0, 8, 9
gunlaw	25	Favor or Oppose Gun Permits		0, 8, 9
grass	26	Should Marijuana Be Made Legal		0, 8, 9
relig	27	Religious Preference		8, 9
life	28	Is Life Exciting or Dull		0, 8, 9
chldidel	29	Ideal Number of Children		*, 9
pillok	30	Birth Control to Teenagers 14-16		0, 8, 9
sexeduc	31	Sex Education in Public Schools		0, 8, 9
spanking	32	Favor Spanking to Discipline Child		0, 8, 9
letdie1	33	Allow Incurable Patients to Die		0, 8, 9
news	34	How Often Does R Read Newspaper		0, 8, 9
tvhours	35	Hours Per Day Watching TV		-1, 98, 99
bigband	36	Bigband Music		0, 8, 9
blugrass	37	Bluegrass Music		0, 8, 9
country	38	Country Western Music		0, 8, 9
blues	39	Blues or R & B Music		0, 8, 9
musicals	40	Broadway Musicals		0, 8, 9
classicl	41	Classical Music		0, 8, 9
folk	42	Folk Music		0, 8, 9
jazz	43	Jazz Music		0, 8, 9
opera	44	Opera		0, 8, 9
rap	45	Rap Music		0, 8, 9
hvymetal	46	Heavy Metal Music		0, 8, 9
attsprts	47	Attended Sports Event in Last Yr		0, 8, 9
visitart	48	Visited Art Museum or Gallery in Last Yr		0, 8, 9
tvshows	49	How Often R Watches TV Drama or Sitcoms		0, 8, 9
tvnews	50	How Often R Watches TV News		0, 8, 9
tvpbs	51	How Often R Watches Public TV Shows		0, 8, 9
scitest4	52	Humans Evolved From Animals		0, 8, 9
partners	53	How Many Sex Partners R Had in Last Year		-1, 98, 99
sexfreq	54	Frequency of Sex During Last Year		*, 8, 9
dwelown	55	Homeowner or Renter		0, 8, 9
sei	56	Respondent Socioeconomic Index		.0, 99.8, 99.9
cohort	57	Year of Birth		0, 9999
income4	58	Total Family Income		
degree2	59	College Degree		7, 8, 9
agecat4	60	Age Categories		
politics	61	Political Outlook		
region4	62	Region		
married	63	Married ?		

Variables in the working file

[6] Data: slightly relabeled from SPSS 13's data set GSS93.sav which is a subset of 1,500 observations from the GSS in 1993

Value		Label
wrkstat	0[a]	NAP
	1	Working fulltime
	2	Working part-time
	3	Temp not working
	4	Unempl., laid off
	5	Retired
	6	School
	7	Keeping house
	8	Other
	9[a]	NA
marital	1	married
	2	widowed
	3	divorced
	4	separated
	5	never married
	9[a]	NA
agewed	0[a]	nap
	98[a]	dk
	99[a]	na
sibs	98[a]	dk
	99[a]	na
childs	8	Eight or More
	9[a]	NA
age	98[a]	DK
	99[a]	NA
birthmo	0[a]	NAP
	1	January
	2	February
	3	March
	4	April
	5	May
	6	June
	7	July
	8	August
	9	September
	10	October
	11	November
	12	December
	98[a]	DK
	99[a]	NA
zodiac	0[a]	NAP
	1	Aries
	2	Taurus
	3	Gemini
	4	Cancer
	5	Leo
	6	Virgo
	7	Libra
	8	Scorpio
	9	Sagittarius
	10	Capricorn
	11	Aquarius
	12	Pisces
	98[a]	DK
	99[a]	NA
educ	97[a]	NAP
	98[a]	DK
	99[a]	NA
degree	0	Less than HS
	1	High school
	2	Junior college
	3	Bachelor
	4	Graduate
	7[a]	NAP
	8[a]	DK
	9[a]	NA
padeg	0	LT High School
	1	High School
	2	Junior College
	3	Bachelor
	4	Graduate
	7[a]	NAP
	8[a]	DK
	9[a]	NA

Value		Label
madeg	0	LT High School
	1	High School
	2	Junior College
	3	Bachelor
	4	Graduate
	7[a]	NAP
	8[a]	DK
	9[a]	NA
sex	1	Male
	2	Female
race	1	white
	2	black
	3	other
income91	0[a]	NAP
	1	LT $1000
	2	$1000-2999
	3	$3000-3999
	4	$4000-4999
	5	$5000-5999
	6	$6000-6999
	7	$7000-7999
	8	$8000-9999
	9	$10000-12499
	10	$12500-14999
	11	$15000-17499
	12	$17500-19999
	13	$20000-22499
	14	$22500-24999
	15	$25000-29999
	16	$30000-34999
	17	$35000-39999
	18	$40000-49999
	19	$50000-59999
	20	$60000-74999
	21	$75000+
	22	Refused
	98[a]	DK
	99[a]	NA
rincom91	0[a]	NAP
	1	LT $1000
	2	$1000-2999
	3	$3000-3999
	4	$4000-4999
	5	$5000-5999
	6	$6000-6999
	7	$7000-7999
	8	$8000-9999
	9	$10000-12499
	10	$12500-14999
	11	$15000-17499
	12	$17500-19999
	13	$20000-22499
	14	$22500-24999
	15	$25000-29999
	16	$30000-34999
	17	$35000-39999
	18	$40000-49999
	19	$50000-59999
	20	$60000-74999
	21	$75000+
	22	Refused
	98[a]	DK
	99[a]	NA
region	0[a]	Not Assigned
	1	New England
	2	Middle Atlantic
	3	E. Nor Central
	4	W. Nor Central
	5	South Atlantic
	6	E. South Central
	7	W. South Central
	8	Mountain
	9	Pacific

[a] missing value
NAP stands for “Not applicable”
DK stands for “Don’t Know”
NA is for “Not Answered”

Appendix: File structure

Let's say today is November 30, 2006.

Last time when you modified your data was November 28, 2006

So you

1. create a new directory 20061130
2. copy the data file GSS93_20061128.sav from 20061128 to 20061130
3. rename the copied data file to GSS93_20061130.sav and start working

Your work directory structure on November 30, 2006 might look like this:

After finishing your work you save your work into a zip file GSS93_20061130.zip

Backup data regularly.

Appendix: Example for a codebook

Code book

http://www.owlnet.rice.edu/~poli502/Codebook.txt

The following codebook lists all variables included in both data files. Each variable is listed with an abbreviated variable name (in capital letters), the exact wording of the question used in the GSS survey, and the response categories with their corresponding codes. Both the variable names and the codes for each response category are included in the SPSS data files.

ABANY

Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason?

1	Yes
2	No
8	Don't know
9	No answer
0	Not applicable

EDUC

Respondent's education

00	No formal schooling
01	1st grade
02	2nd grade
03	3rd grade
04	4th grade
05	5th grade
06	6th grade
07	7th grade
08	8th grade
09	9th grade
10	10th grade
11	11th grade
12	12th grade
13	1 year in college
14	2 years
15	3 years
16	4 years
17	5 years
18	6 years
19	7 years
20	8 years
98	Don't know
99	No answer

…

EQWLTH

Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Others think the government should not concern itself with reducing this income difference between the rich and the poor. What score between 1 and 7 comes closest to the way you feel?

1	Government should do something to reduce income differences between rich and poor.
7	Government should not concern itself with income differences.
8	Don't know
9	No answer
0	Not applicable

…

HRS1

If working full or part time, how many hours did you work last week, at all jobs?

0	0-9 hours
1	10-19 hours
2	20-29 hours
3	30-39 hours
4	40-49 hours
5	50-59 hours
6	60-69 hours
7	70-79 hours
8	80 or more hours
9	No answer, Don't know, Not applicable

…

INCOME

In which of these groups did your total family income, from all sources, fall last year before taxes, that is? Just tell me the letter.

01	Under $1,000
02	$ 1,000 to 2,999
03	$ 3,000 to 3,999
04	$ 4,000 to 4,999
05	$ 5,000 to 5,999
06	$ 6,000 to 6,999
07	$ 7,000 to 7,999
08	$ 8,000 to 9,999
09	$10,000 to 14,999
10	$15,000 to 19,999
11	$20,000 to 24,999
12	$25,000 or over
13	Refused
98	Don't know
99	No answer
0	Not applicable

…

LIBATH

There are always some people whose ideas are considered bad or dangerous by other people. For instance, somebody who is against all churches and religion . . . If some people in your community suggested that a book he wrote against churches and religion should be taken out of your public library, would you favor removing this book, or not?

1	Favor
2	Not favor
8	Don't know
9	No answer
0	Not applicable