



Fundamentals of SPSS

Participant Workbook

American Samoa
Community Cancer Network

April-May 2008

Sara Krosch, MA



Contents

Introduction	ii
Session 1: Data Collection: Surveys	1
Session 2: Introduction to SPSS – Defining and Entering Data	13
Week 1 Task: Defining and Entering Data	23
Session 3: Descriptive Statistics and Assessing Normality	24
Session 4: Confidence Intervals and Recoding Variables	37
Week 2 Task: Working with Original Data	43
Session 5: Inferential Statistics and Correlations/Associations	44
Session 6: One-sample and paired t-tests	51
Session 7: Independent (two-sample) t-tests and One-way Analysis of Variance (ANOVA)	56
Week 3 Task: Working with Original Data	63
Statistics Overview	64
Statistics Terms	65

Dealing with opened-ended responses in SPSS 174-181
“Chapter 20: Multiple response analysis and multiple dichotomy analysis”
excerpt from *SPSS: Analysis Without Anguish (Version 15.0 for Windows) Student
Version* by Coakes, Steed and Price.

sample data sets for this section available at
<http://www.johnwiley.com.au/highered/spssv15/student-res/index.html>

Introduction

This Participant Workbook was created to accompany a workshop-style training in the fundamentals of data collection and statistical analysis using SPSS 16.0.1 software for health professionals in American Samoa, coordinated by the American Samoa Community Cancer network. Each session may take up to 2 hours to complete. Participants gain hands-on experience through working examples and weekly tasks that are to be completed outside of workshop sessions. The workshop culminates with participant presentations of work-related data that has been analyzed using SPSS.

This workbook guides participants through the very basics of defining and entering data, generating descriptive statistics and conducting basic statistical analysis to test hypotheses using SPSS. No prior knowledge of statistics is assumed so fundamental statistical terms and concepts are covered. No experience using SPSS is assumed either, however participants are expected to have experience gathering data and entering it into Excel spreadsheets.

The overall **goal** of the training utilizing this workbook is to build the capacity of indigenous cancer researchers and health professionals to gather and analyze local data in support of scientifically rigorous inquiry.

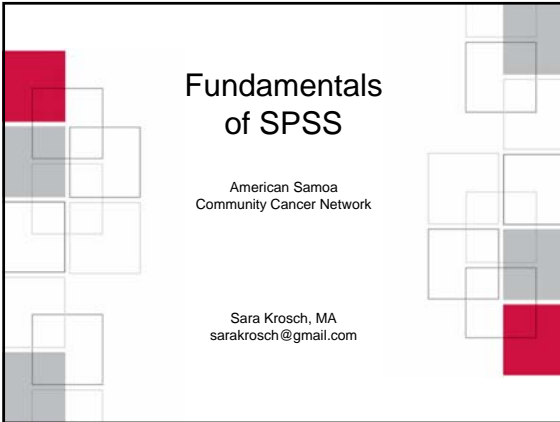
The specific **objectives** of this workbook are to enable workshop participants to:

- Develop and execute random sampling plans to gather local data via surveys
- Understand basic statistics concepts and terms as they apply to SPSS
- Define and enter data in SPSS
- Generate descriptive statistics and interpret output in SPSS
- Test for normality and make necessary transformations to achieve normal distributions of sample data in SPSS
- Conduct inferential statistical tests (hypothesis tests) and interpret output in SPSS
- Perform both descriptive and inferential statistics functions on original, local data and share results

This workbook is accompanied by a disk containing sample data sets used in working examples for each session as well as the booklet (.pdf) SPSS Survey Tips. Some of the content for sessions was adapted from *SPSS: Analysis Without Anguish (Version 15.0 for Windows) Student Version* by Coakes, Steed and Price (2008) available from John Wiley Publishers. This text serves as an excellent step-by-step guide for further reading.

This workbook was compiled and created by Sara Krosch for the American Samoa Community Cancer Network in March 2008. Any reproduction, distribution and/or use of this workbook is freely granted for educational purposes only. Reproduction, distribution and/or use of this workbook for profit is strictly forbidden unless by consent of the author.

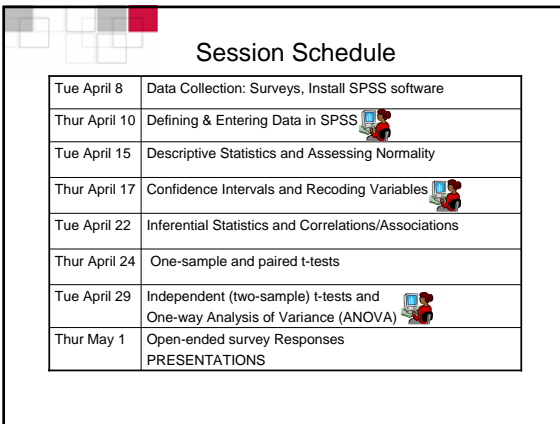
The author can be reached at sarakrosch@gmail.com






Fundamentals of SPSS

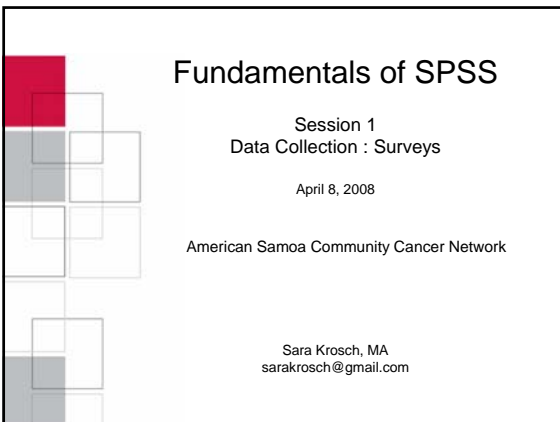
American Samoa
Community Cancer Network

Sara Krosch, MA
sarakrosch@gmail.com



Session Schedule

Tue April 8	Data Collection: Surveys, Install SPSS software
Thur April 10	Defining & Entering Data in SPSS 
Tue April 15	Descriptive Statistics and Assessing Normality
Thur April 17	Confidence Intervals and Recoding Variables 
Tue April 22	Inferential Statistics and Correlations/Associations
Thur April 24	One-sample and paired t-tests
Tue April 29	Independent (two-sample) t-tests and One-way Analysis of Variance (ANOVA) 
Thur May 1	Open-ended survey Responses PRESENTATIONS




Fundamentals of SPSS

Session 1
Data Collection : Surveys

April 8, 2008

American Samoa Community Cancer Network

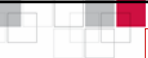
Sara Krosch, MA
sarakrosch@gmail.com



Session 1 - Data Collection: Surveys

Data is only as good
as the instrument and
methods used to gather it.

Garbage in...garbage out.




Session 1 - Data Collection: Surveys

Step 1: Define the Area of Interest
Step 2: Define the Population
Step 3: Decide on a Sampling Plan
Step 4: Choose Data Collection Methods
Step 5: Design your Instrument (survey)
Step 6: Pilot Test and Refine Instrument

}
TODAY

Step 7: Gather Data
Step 8: Analyze Data
Step 9: Report Results
Step 10: ACT on Data


SPSS



Session 1 - Data Collection: Surveys

Step 1: Define the Area of Interest

- What is the health issue you want to study?
- Is this the first time this topic has been studied?
- Do you need general or specific data?
- Is the topic sensitive?
- How will the data fit into the program?

Examples


Session 1 - Data Collection: Surveys

Step 1: Define the Population
total group of interest

American Samoans

American Samoans on Tutuila Island

American Samoan males on Tutuila Island

American Samoan males who smoke on Tutuila Island

American Samoan males ages 20-40 who smoke on Tutuila Island

Session 1 - Data Collection: Surveys

Step 1: Decide on a Sampling Plan

Census vs. Sample

Sample: **representative** subset of the population

Sampling Frame: list from which sample is drawn

Population: American Samoans
Sample: **characteristics?**

Population: American Samoan males ages 20-40 who smoke on Tutuila Island
Sample: **characteristics?**

Session 1 - Data Collection: Surveys


```

graph TD
    A[Sampling Plans] --> B[Probability Sampling  
(Random Sampling)]
    A --> C[Non-Probability Sampling  
(Non-Random Sampling)]
    B --> D[• Simple  
• Systematic  
• Stratified  
• Cluster (Area)]
    C --> E[• Convenience  
• Snowballing/Networking  
• Purposive]
  
```

Random Sampling: selection process that makes choosing any particular subject as likely or probable as the next

Session 1 - Data Collection: Surveys

Random Sampling: selection process that makes choosing any particular subject as likely or probable as the next



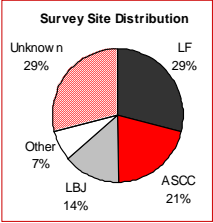
Why is random sampling important?

- Captures random variation in population
- Limits bias
- Can make more confident statements about the population

Session 1 - Data Collection: Surveys

Sampling Plans

Survey Site Distribution



Non-Probability Sampling (Non-Random Sampling)


- Convenience
- Snowballing/Networking
- Purposive (focus groups)

Session 1 - Data Collection: Surveys

Sampling Plans

Probability Sampling (Random Sampling)

- Simple
- Systematic
- Stratified
- Cluster (Area)



4

Session 1 - Data Collection: Surveys

Simple Random Sampling

All patients

Patient sample

- 1) Sampling Frame: all patients in past 12 months (N=1,000)
- 2) Decide on sample number (n=100 or 10%)
- 3) Pick out 100 names out of a hat OR Use a computer program (Excel) to randomly select 100

+ Easiest
- Could miss population variation

Using Excel to choose a Simple Random Sample

- 1) copy and paste the list of client names into a column in an EXCEL spreadsheet.
- 2) In the column right next to it paste (paste special> formulas) the function =RAND()
- 3) Sort both columns – the list of names and the random number – by the random numbers. This rearranges the list in random order from the lowest to the highest random number.
- 4) Take the first hundred names in this sorted list.

Session 1 - Data Collection: Surveys

Systematic Sampling

- 1) Sampling Frame: all patients in past 12 months (N=1,000)
List the population in random order for the characteristic of interest (gender, status)
- 2) Number all the cases
- 3) Decide on the sample number (n=100 or 10%)
- 4) Pick a number between 1-5 (3)
- 5) Start with the 3rd case, and choose every 10th case

+ easy, more random than simple random sampling, good for large lists
-Homogenous, does not focus on sub-group variability

Session 1 - Data Collection: Surveys

Stratified Random Sampling

All patients 1,000

100

50 30 20

Random sample of each strata in representative proportions

- 1) Sampling Frame: all patients in past 12 months (N=1,000)
- 2) Decide on sample number (n=100 or 10%)
- 3) Divide the population into non-overlapping groups (strata)
- 4) Do a simple random sample from each strata with equal or representative proportions

+ Represent more traits of the population (key sub-groups)
- Requires more detailed information

Session 1 - Data Collection: Surveys

Cluster/ Multi-stage (Area) Sampling

60 Villages divided into 7 groups

Each red village chosen

50

Random sample within red villages

- 1) Divide the population into Clusters (geographic boundaries)
- 2) Do a simple random sample of clusters
- 3) Do a simple random sample from within chosen clusters

+ Covers large area, saves time and \$
- Requires coordination

Session 1 - Data Collection: Surveys

Sample Size

- Minimum 50-100 (if representative)
- Maximum 1,000 – 1,500
- Amount of error willing to tolerate

For example, if an error of five percent (±5 percent) is acceptable, the formula calculates the required sample size as:

$$N = \frac{1}{.05^2} = \frac{1}{.0025} = 400$$

Session 1 - Data Collection: Surveys

Sampling Plans

**Probability Sampling
(Random Sampling)**

- Simple
- Stratified
- Cluster (Area)
- Systematic

**Non-Probability Sampling
(Non-Random Sampling)**

- Convenience
- Snowballing/Networking
- Purposive

Which have you used?

Session 1 - Data Collection: Surveys

Step 4: Choose a Data Collection Method


- Observations
- Interviews
- Focus Groups
- Surveys/Questionnaires
- Experiments

Session 1 - Data Collection: Surveys

Step 4: Choose a Data Collection Method

- Observations
- Interviews
- Focus Groups
- Surveys/Questionnaires**
- Experiments

- Data from a large number of people
- Low budget
- Do not need in-depth information
- Data collection standardized
- Want to comment about the population based on a sample

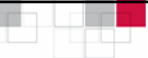


Session 1 - Data Collection: Surveys

Step 5: Design your Instrument (survey)

Things to consider...

- Goal of survey → type
- Method: self- or interviewer-administered
- Question and Answer Format
- Threats to Validity and Reliability




Session 1 - Data Collection: Surveys

What is the goal of your survey?

- Pilot
- Baseline
- Follow-up (reliability)

Common Types of Surveys

- KAB/KAP (knowledge, attitude, behaviors/practices)
- QOL (Quality of Life)
- Satisfaction




Session 1 - Data Collection: Surveys

Will your survey be self-administered or interviewer-administered?

SELF	INTERVIEWER
+ Less time and money	- Takes more time and money
+ Convenient	+ Allows for clarification
- May represent views beyond the respondent	- Must train interviewers
- Usually more missing data	+ Usually more in-depth data gathered

Depends on the topic and the population

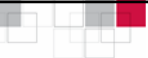


Session 1 - Data Collection: Surveys

Tips on writing questions


- short, < 25 words
- simple language
- use specific language
- avoid “double-barreled” questions
- avoid “loaded” questions
- avoid “leading” questions

Threats to **Validity**
 (are we measuring
 what we think)



Session 1 - Data Collection: Surveys

Examples:



Are you satisfied with the amount and kind of lung cancer prevention information you received?


Does having unprotected sex put someone at risk for HIV/AIDS?

Do you think the hospital is doing a good job of providing cancer screening services?

How do you feel the government is handling the diabetes and obesity crisis in American Samoa?


Do you usually go to the doctor when you are sick?

Do you agree or disagree that the school lunch menu should not include soda and ice cream?



Session 1 - Data Collection: Surveys

Examples:



Double-barreled Are you satisfied with the **amount and kind** of lung cancer prevention information you received?


Leading Does having **unprotected** sex put someone at **risk** for HIV/AIDS?

Leading & Vague Do you think the hospital is doing a **good job** of providing cancer screening services?

Double-Barreled & Loaded How do you feel the government is handling the **diabetes and obesity crisis** in American Samoa?

Vague & Leading Do you **usually** go to the doctor when you are sick?

Confusing Do you **agree or disagree** that the school lunch menu **should not** include soda and ice cream?



Session 1 - Data Collection: Surveys

Answer Formats:

Close-ended

- yes/no
- list of choices
- likert scale

1
Very Satisfied

2
Satisfied

3
Dissatisfied

4
Very Dissatisfied

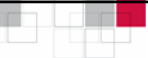
1
Strongly Agree

2
Agree

3
Neither Agree nor Disagree

4
Disagree

5
Strongly Disagree



Session 1 - Data Collection: Surveys


Answer Formats:

Open-ended

- More depth and variety of data (attitudes, preferences, experiences)

Which ASCCN programs have you participated in?
 (white space not blanks)


What services would you like to see included in our family programs?
 (Put at end of survey)



Session 1 - Data Collection: Surveys

Answer Dilemmas...

- "Don't know" → only knowledge questions not opinion
- "Can't say" or "N/A" → data reflects real experiences, but can avoid answering
- Providing a list of choices vs open-ended
- "all of the above"
- "Other _____"



Session 1 - Data Collection: Surveys

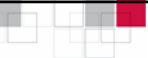
Step 6: Pilot Test and make Refinements

Why pilot test?

- test for understandability
- get extra feedback on format or possible answers
- saves time and money

Pilot test with who?

- co-workers, experts (content/face validity)
- small representative sample (20% of future sample)



Session 1 - Data Collection: Surveys

Step 5: Pilot Test and Make Refinements


Use feedback from pilot test to make necessary changes to question and answer wording


Learn from the experience of administering the survey

Decide whether incentives are needed
tangible vs intangible incentives

"Help our organization better serve people with disabilities in the community."

"Your feedback will help shape future legislation."






Session 1 - Data Collection: Surveys

Pilot Test "Community Violence Survey"

Original survey from Guam with no changes made

- 1) Offer advice on
 - Wording of questions
 - Layout of survey
 - Organization of questions
 - Answer choices
 - Construct validity: do all questions pertain to the topic
- 2) What **sampling plan** would you use and why?
Describe the characteristics of your **population** and **sample**.
- 3) What analysis would you like to do with this data?
What hypothesis do you have?



Session 1 - Data Collection: Surveys

Next Session:
Introduction to SPSS - Defining and Entering Data

Action Items:
1) Install SPSS 16.0.1 software
2) Bring Excel spreadsheet of data to be analyzed

Using Excel to choose a Simple Random Sample

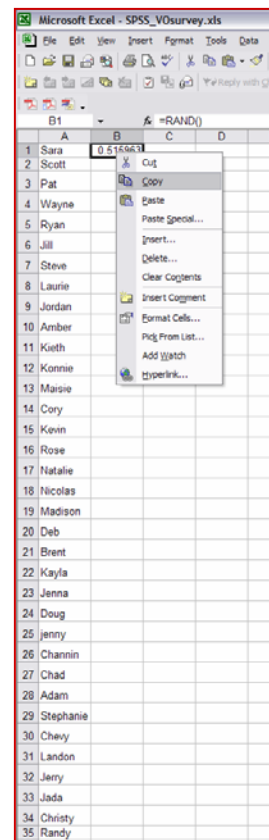
- 1) copy and paste the list of client names into a column in an EXCEL spreadsheet.
- 2) In the column right next to it paste (paste special> formulas) the function =RAND()
This is EXCEL's way of putting a random number between 0 and 1 in the cells.
- 3) Sort both columns -- the list of names and the random number -- by the random numbers. This rearranges the list in random order from the lowest to the highest random number.
- 4) Take the first hundred names in this sorted list.



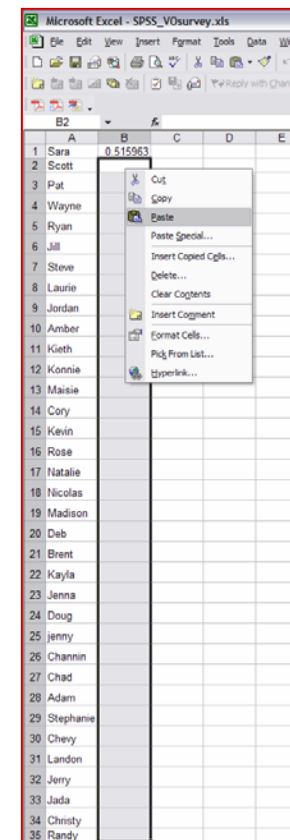
	A	B	C
1	Sara		
2	Scott		
3	Pat		
4	Wayne		
5	Ryan		
6	Jill		
7	Steve		
8	Laurie		
9	Jordan		
10	Amber		
11	Kieth		
12	Konnie		
13	Maisie		
14	Cory		
15	Kevin		
16	Rose		
17	Natalie		
18	Nicolas		
19	Madison		
20	Deb		
21	Brent		
22	Kayla		
23	Jenna		
24	Doug		
25	Jenny		
26	Channin		
27	Chad		
28	Adam		
29	Stephanie		
30	Chevy		
31	Landon		
32	Jerry		
33	Jada		
34	Christy		
35	Randy		



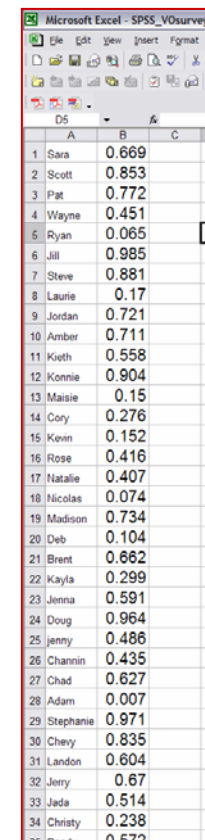
	A	B	C
1	Sara	=RAND()	
2	Scott		
3	Pat		
4	Wayne		
5	Ryan		
6	Jill		
7	Steve		
8	Laurie		
9	Jordan		
10	Amber		
11	Kieth		
12	Konnie		
13	Maisie		
14	Cory		
15	Kevin		
16	Rose		
17	Natalie		
18	Nicolas		
19	Madison		
20	Deb		
21	Brent		
22	Kayla		
23	Jenna		
24	Doug		
25	Jenny		
26	Channin		
27	Chad		
28	Adam		
29	Stephanie		
30	Chevy		
31	Landon		
32	Jerry		
33	Jada		
34	Christy		
35	Randy		



	A	B	C	D	E
1	Sara	0.515963			
2	Scott				
3	Pat				
4	Wayne				
5	Ryan				
6	Jill				
7	Steve				
8	Laurie				
9	Jordan				
10	Amber				
11	Kieth				
12	Konnie				
13	Maisie				
14	Cory				
15	Kevin				
16	Rose				
17	Natalie				
18	Nicolas				
19	Madison				
20	Deb				
21	Brent				
22	Kayla				
23	Jenna				
24	Doug				
25	Jenny				
26	Channin				
27	Chad				
28	Adam				
29	Stephanie				
30	Chevy				
31	Landon				
32	Jerry				
33	Jada				
34	Christy				
35	Randy				



	A	B	C	D	E
1	Sara	0.515963			
2	Scott				
3	Pat				
4	Wayne				
5	Ryan				
6	Jill				
7	Steve				
8	Laurie				
9	Jordan				
10	Amber				
11	Kieth				
12	Konnie				
13	Maisie				
14	Cory				
15	Kevin				
16	Rose				
17	Natalie				
18	Nicolas				
19	Madison				
20	Deb				
21	Brent				
22	Kayla				
23	Jenna				
24	Doug				
25	Jenny				
26	Channin				
27	Chad				
28	Adam				
29	Stephanie				
30	Chevy				
31	Landon				
32	Jerry				
33	Jada				
34	Christy				
35	Randy				



	A	B	C
1	Sara	0.669	
2	Scott	0.853	
3	Pat	0.772	
4	Wayne	0.451	
5	Ryan	0.065	
6	Jill	0.985	
7	Steve	0.881	
8	Laurie	0.17	
9	Jordan	0.721	
10	Amber	0.711	
11	Kieth	0.558	
12	Konnie	0.904	
13	Maisie	0.15	
14	Cory	0.276	
15	Kevin	0.152	
16	Rose	0.416	
17	Natalie	0.407	
18	Nicolas	0.074	
19	Madison	0.734	
20	Deb	0.104	
21	Brent	0.662	
22	Kayla	0.299	
23	Jenna	0.591	
24	Doug	0.964	
25	Jenny	0.486	
26	Channin	0.435	
27	Chad	0.627	
28	Adam	0.007	
29	Stephanie	0.971	
30	Chevy	0.835	
31	Landon	0.604	
32	Jerry	0.67	
33	Jada	0.514	
34	Christy	0.238	
35	Randy	0.572	

Session 2: Introduction to SPSS – Defining and Entering Data

Participants will be able to:

- Open and close the SPSS program
- Open an already existing data file
- Import an Excel data file
- Define and recode variables
- Enter new data
- Save data
- Handle missing data
- Practice with working examples

Key Terms:

case
continuous variable (numerical)
categorical variable
(string: binary, ordinal, nominal)

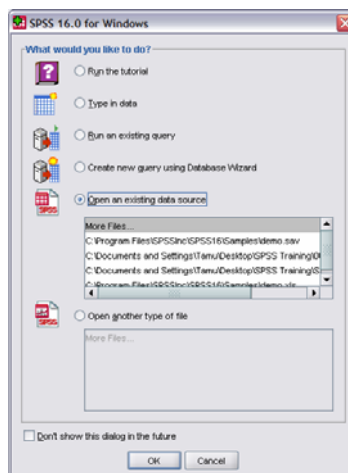
1 Opening SPSS on your computer

Start > All Programs> SPSS Inc. > SPSS 16.0

2 Opening an already existing data file

SPSS data files have the file extension .sav

When you start a new SPSS session and you know the pre-existing data set you want to work with from the initial window choose: >Open and existing data source



OR If SPSS is already open from the Menu Bar choose:

File > Open > Data > Program Files > SPSS Inc> SPSS 16 > locate the folder containing the data file > select the .sav file > Open



OR the > open file button



Working Example 1 : Open the Sample File *demo.sav*

File > Open > Data > Program Files > SPSS Inc> SPSS 16 > Samples > demo.sav > Open

3 Reading a Data File

A data file is displayed in the Data Editor. SPSS data files are made up of rows and columns. Each row is a **case** (an individual respondent to a survey). Each column is a **variable** (question on a survey). Variables can be **categorical** (binary, ordinal, nominal) or **continuous** (numbers).

	age	marital	address	income	inccat	car	carcat	ed	employ	retire	empcat	jobsat
1	55	Married	12	72.00	\$50 - \$74	36.20	Luxury	Did not co...	23	No	More than 15	Highly sati...
2	56	Unmarried	29	153.00	\$75+	76.90	Luxury	Did not co...	35	No	More than 15	Somewhat ...
3	28	Married	9	28.00	\$25 - \$49	13.70	Economy	Some colle...	4	No	Less than 5	Neutral F...
4	24	Married	4	26.00	\$25 - \$49	12.50	Economy	College de...	0	No	Less than 5	Highly diss...
5	25	Unmarried	2	23.00	Under \$25	11.30	Economy	High schoo...	5	No	5 to 15	Somewhat ...
6	45	Married	9	76.00	\$75+	37.20	Luxury	Some colle...	13	No	5 to 15	Somewhat ...
7	42	Unmarried	19	40.00	\$25 - \$49	19.80	Standard	Some colle...	10	No	5 to 15	Somewhat ...
8	35	Unmarried	15	57.00	\$50 - \$74	28.20	Standard	High schoo...	1	No	Less than 5	Highly diss...
9	46	Unmarried	26	24.00	Under \$25	12.20	Economy	Did not co...	11	No	5 to 15	Highly sati...
10	34	Married	0	89.00	\$75+	46.10	Luxury	Some colle...	12	No	5 to 15	Somewhat ...
11	55	Married	17	72.00	\$50 - \$74	35.50	Luxury	Some colle...	2	No	Less than 5	Neutral F...
12	28	Unmarried	3	24.00	Under \$25	11.80	Economy	College de...	4	No	Less than 5	Highly sati...
13	31	Married	9	40.00	\$25 - \$49	21.30	Standard	College de...	0	No	Less than 5	Somewhat ...
14	42	Unmarried	8	137.00	\$75+	68.90	Luxury	Some colle...	3	No	Less than 5	Highly diss...
15	35	Unmarried	8	70.00	\$50 - \$74	34.10	Luxury	Some colle...	9	No	5 to 15	Somewhat ...
16	52	Married	24	159.00	\$75+	78.90	Luxury	College de...	16	No	More than 15	Highly sati...
17	21	Married	1	37.00	\$25 - \$49	18.60	Standard	Some colle...	0	No	Less than 5	Highly diss...
18	32	Unmarried	0	28.00	\$25 - \$49	13.70	Economy	Did not co...	2	No	Less than 5	Somewhat ...

Hover over a column title for a more descriptive name.

OR from the Menu Bar choose: View > Value Labels

OR the > Value Label Button



There are two tabs at the bottom of the Data Editor: Data View and Variable View



Working Example 2 : Using the Data View and Variable View answer the following about *demo.sav*

- How many people were surveyed?
- How many questions (variables) were in the survey?
- What did question 14 ask in the survey?

4 Saving files and ending a session

To save a file choose: File > Save as > type the file name and make sure it is a .sav file > select the location of the file > Save

To end a session without saving choose: File > Exit
(You will be prompted to save the contents of each window)

5 Importing from Excel (.xls files)

Data can be imported from Excel. (To import Access or Text Files see detailed instructions in the Tutorial.) The Excel column name will become the variable name. Names will be converted to have no spaces.

To import an Excel spreadsheet choose: File> Open> Data > Samples Folder> Select .xls as the file type to view > SPSS_VOsurvey.xls > Open > Check read variable names from first row or choose the worksheet or range > Continue



Working Example 3: The results of a 2006 community survey of violence in American Samoa were entered into an Excel spreadsheet. Locate the file SPSS_VOsurvey.xls. Open it in SPSS and save it as *06ViolenceSurvey.sav* using the directions above.

6 Entering new data directly

You can enter data directly into SPSS. Data is entered in Data View. Variables are defined in Variable View.

In the Data Editor, Data View rows represent cases (observations) and the columns represent variables. In Variable View each row is a variable and each column is the attribute associated with that variable.

Before entering new data in the Data View it is best to define your variables in Variable View. The main types of variables are **continuous** (numbers, currency, dates) and **categorical** (strings or text).

6.1 Defining variables in Variable View

Imagine you conducted a follow-up study to the *06ViolenceSurvey.sav* file. The survey contained questions: gender; age; Have you witnessed family violence? You have completed the second survey and now you want to enter the data directly into SPSS.

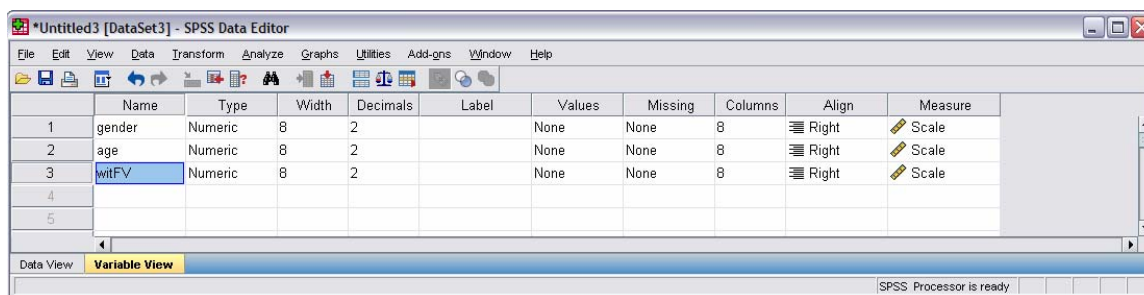


Working Example 4:

First, open a new data sheet File > New > Data

Go to Variable View to enter and define the variables

Under the variable column **Name** enter each variable: 1 gender, 2 age, 3 witFV



Notice that variables are automatically given a numeric data type. age is a continuous numeric variable while gender and witFV are categorical string variables.



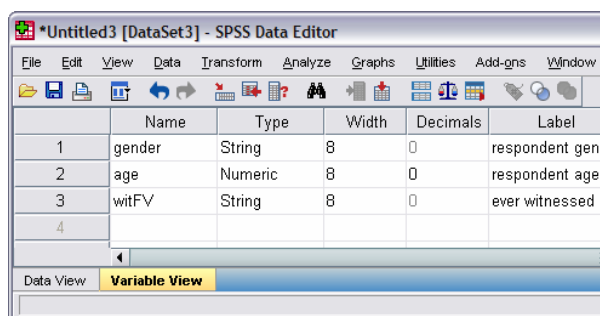
Working Example 5: Define gender and witFV as string variables.

Click in the **Type** column for gender and click the [...]. Choose string and characters: 8 > OK Repeat this for witFV.

None of the variables have decimals. Change this by clicking on the **Decimals** column and change the number to 0 for each variable.

The column **Label** allows you to enter a longer more descriptive variable name. This is especially helpful when there are many variables or similar variables. Enter the following by typing in the Label column for each variable:

Variable Name	Variable Label
gender	respondents gender
age	respondents age in years
witFV	ever witnessed family violence



Defining **Value Labels** is useful for statistical reports and charts. A value label assigns a number to a categorical string label.

Value Label gender

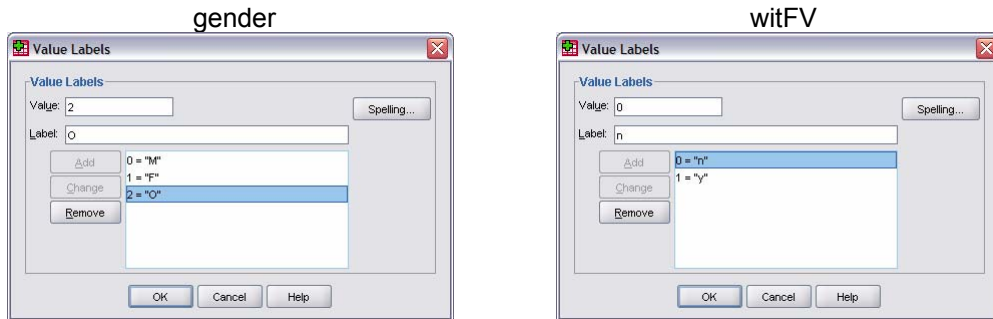
Survey Options	Data Name	Type: String
male	M	0
female	F	1
other	O	2

Value Label witFV

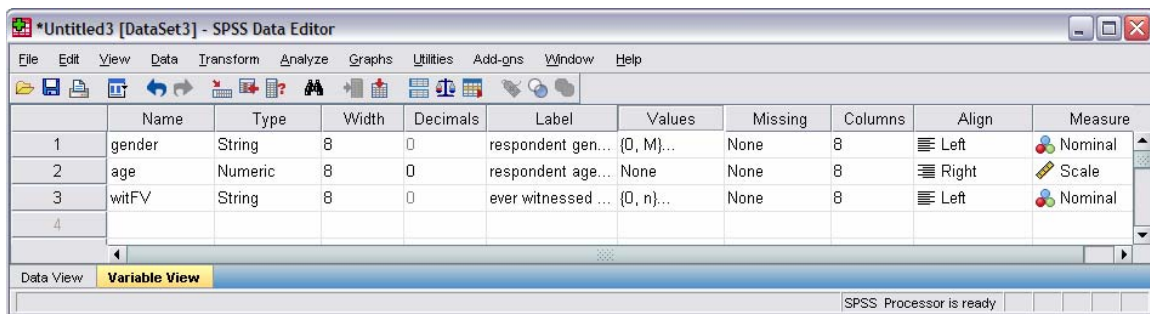
Survey Options	Data Name	Type: String
yes	y	1
no	n	0



Working Example 6: Click on the Values column for gender and [...]. Enter the value (number) you want to assign to the first gender label (M) and then Add. When you have entered each choice for gender click OK. Do the same for witFV.



Now that your variables have been defined entering the survey data will be quicker and you are less likely to make errors.



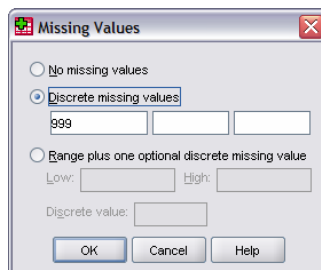
6.2 Missing data

It is rare to obtain complete data sets. Survey respondents can skip a question or not answer it correctly. If you do not filter missing data your analysis can have inaccurate results. You may also want to analyze the type of respondents who did not answer the question or the questions frequently skipped.

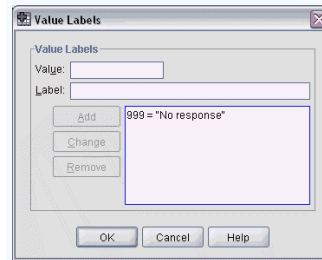
When you have missing data you can leave the cell blank or assign a missing code value (usually 999 or the variable mean). Missing codes must be of the same data type as the data they represent (missing numeric data must also have numeric missing value codes). Missing codes cannot appear in the data set.



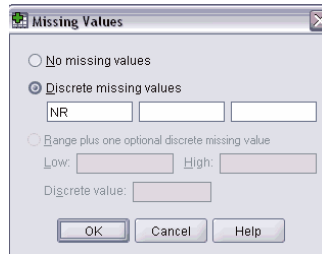
Working Example 7 : in the data set below one respondent's age is missing. To create a missing value click in the Missing Column for age > Discrete missing values > 999 > OK



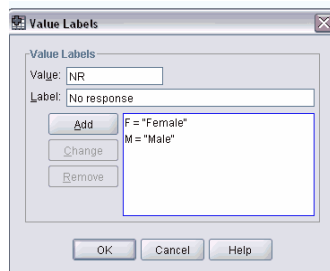
Now that a missing data value has been added, a label can be applied to that value.



Missing String data (gender and witFV)



Now you can add a label for your missing string data



6.3 Entering data

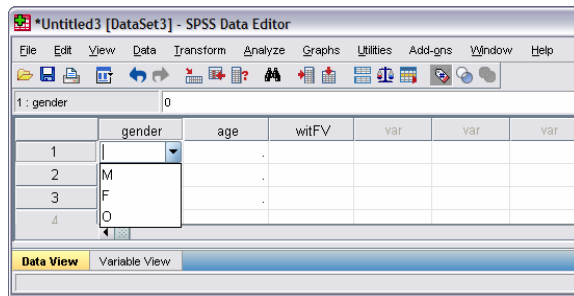


Working Example 8 : Go to Data View to begin entering your survey data found below.

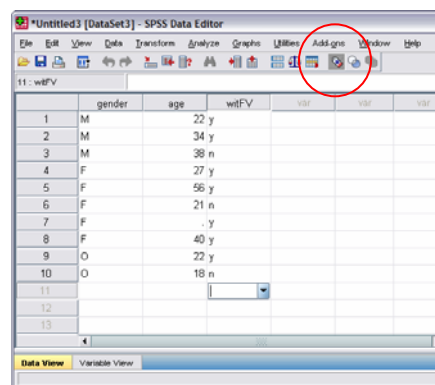
Follow-up Survey Data

case	gender	age	witVO
1	M	22	y
2	M	34	y
3	M	38	n
4	F	27	y
5	F	56	y
6	F	21	n
7	F	.	y
8	F	40	y
9	O	22	y
10	O	18	n

Notice the choices you defined for each variable will appear in a drop down box for you to choose. You can also type the data.



When you are finished entering the Data you can see what values are associated with your labels by clicking View in the menu bar and choosing value labels or click on the Value Label button.



Tip: Inserting and deleting cases and variables

Often data will need to be inserted in an existing data file. You can add or delete cases (rows) and variables (columns) in Data View.

To insert a new case between existing cases, select any cell in the row **below** the position where you want to insert the new case. From the menus toolbar select Data > insert case

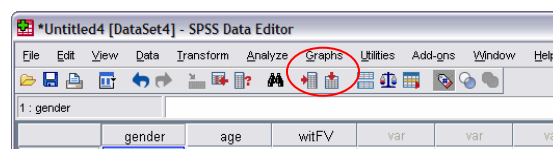
OR click the Insert Case tool

OR right click on the number above the row you want to insert a new case and choose insert case.

To insert a new variable between existing variables, select any cell to the right of the position where you want to insert the new variable. From the menus toolbar select Data > insert variable

OR click the Insert variable tool

OR right click on the variable to the right of where you want to insert a new variable and choose insert variable.



To delete an entire case or variable right click on the number of the case or the name of the variable and choose clear.



Tip: Copying and pasting variable attributes

After you have defined a variable's attributes you can copy these attributes and apply them to other variables. For example, several of your survey questions may have yes/no or a likert scale format (strongly agree, agree, no opinion, disagree, and strongly disagree).

In Variable View, add your new variable (name, label). Click on the values cell for a previously entered variable whose attributes you want to copy. From the Menus bar choose: Edit > copy
Click on the value cell for the new variable. From the Menus bar choose: Edit > paste

Now the defined values from the earlier entered variable are now applied to the new variable.

To copy all attributes of one variable to a new variable click the number of the row of the variable you want to copy. From the Menus bar choose: Edit > copy
Click on the row number for the new empty row. From the Menus bar choose: Edit > paste

All attributes from the first variable are now applied to the new variable.

	Name	Type	Width	Decimals	Label	
1	age	Numeric	8	0	Respondent's Age	(9)
2	marital	Numeric	8	0	Marital Status	(0)
3	income	Dollar	12	0	Household Income	No
4	sex	String	8	0	Gender	(F)
5	agedwed	Numeric	8	2	Age Married	(9)
6	VAR00001	Numeric	8	2		(9)
7	VAR00002	Numeric	8	2		(9)
8	VAR00003	Numeric	8	2		(9)

	Name	Type	Width	Decimals	Label	
1	age	Numeric	8	0	Respondent's Age	(9)
2	marital	Numeric	8	0	Marital Status	(0)
3	income	Dollar	12	0	Household Income	No
4	sex	String	8	0	Gender	(F)
5	agedwed	Numeric	8	2	Age Married	(9)
6	VAR00001	Numeric	8	2		(9)
7	VAR00002	Numeric	8	2		(9)
8	VAR00003	Numeric	8	2		(9)
9	VAR00004	Numeric	8	0	Marital Status	(0)



Note on copy/paste: Copying and pasting selected data cells in Data View only copies the data values with no variable attribute definitions. Copying an entire variable in Data View by selecting the name at the top of the column will also past that variables defined attributes.

	age	marital	address	income	inccat	car
1	55	1	12	72.00	3.00	36.20
2	56	0	29	153.00	4.00	76.90
3	28	1	9	28.00	2.00	13.70
4	24	1	4	26.00	2.00	12.60
5	24	0	7	23.00	1.00	11.30

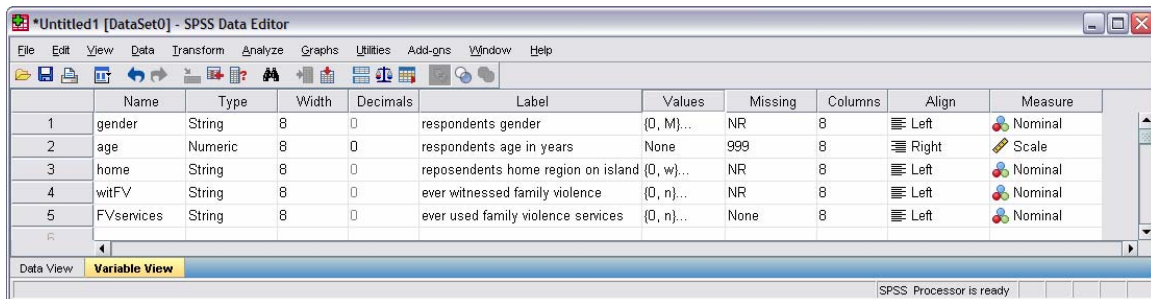
	Name	Type	Width	Decimals	Label	Val
1	age	Numeric	4	0	Age in years	None
2	marital	Numeric	4	0	Marital status	(0, Unm
3	address	Numeric	4	0	Years at cure	None
4	income	Numeric	8	2	Household inc	None
5	inccat	Numeric	8	2	Income catego	(1.00, L



Working Example 9: Imagine more data from your follow-up survey has just arrived. You now have data for 2 more questions: home region on the island (home); Have you used family violence services (FVservices)?

Enter the following variable information. Insert a column so that the variable *home* comes between *age* and *witFV*. Since both *home* and *FVservices* are categorical string variables, a variable value can be assigned to each. And since *FVservices* uses the same value codes, variable attributes from *witFV* can be copied and pasted.

variable number	variable name	variable type	variable label	variable value
3	home	string	respondents home region on island	0= west 1= central 2= east
5	FVservices	string	ever used family violence services	1= yes 0=no



Now enter your new data for *home* and *FVservices* in Data View. **Save this data set as *Vsurvey.sav***

Additional Follow-up Survey Data

case	home	FVservices
1	w	y
2	c	y
3	c	.
4	c	y
5	e	n
6	c	n
7	w	n
8	c	y
9	e	y
10	c	n

*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

12: FVservices Visible: 5 of 5 Variables

	gender	age	home	witFV	FVservices	VS
1	M	22 w	y	y		
2	M	34 c	y	y		
3	M	38 c	n		No Respon...	
4	F	27 c	y	y		
5	F	56 e	y	n		
6	F	21 c	n	n		
7	F	999 w	y	n		
8	F	40 c	y	y		
9	O	22 e	y	y		
10	O	18 c	n	n		

Data View Variable View

SPSS Processor is ready

*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

12: FVservices Visible: 5 of 5 Variables

	gender	age	home	witFV	FVservices	VS
1	0	22 0	1	1		
2	0	34 1	1	1		
3	0	38 1	0		NR	
4	1	27 1	1	1		
5	1	56 2	1	0		
6	1	21 1	0	0		
7	1	999 0	1	0		
8	1	40 1	1	1		
9	2	22 2	1	1		
10	2	18 1	0	0		

Data View Variable View

Value Labels

SPSS Processor is ready



Week 1 Task: Defining and Entering Data

Please complete before Session 3 on Tuesday April 15.

You have been given the following data to define and enter in SPSS.

- Define a variable name for each variable, a variable label and value labels (for categorical variables)
- NOTE- all variables can be considered “numeric” type as long as the data entered fits this type
- Enter the data for each variable
- Check that the data has been entered correctly
- Save the data as **diabetes.sav**

Description of diabetes.sav data set

Variable Number	Variable Name	Description
1	DIABETES	Diagnosis of diabetes (1= yes, 0=no)
2	AGE	Age in years
3	SEX	0=male, 1=female
4	WEIGHT	Weight in Kgs
5	HEIGHT	Height in cms
6	CIGS	cigarette consumption, average number per day
7	ALCOHOL	alcohol consumption, average number standard drinks per week

	DIABETES	AGE	SEX	WEIGHT	HEIGHT	CIGS	ALCOHOL
1	1	45	0	97.1	200	12	12
2	1	43	1	55.5	160	0	14
3	1	67	0	85.2	175	15	21
4	1	54	1	67.0	163	20	7
5	1	47	1	74.0	170	12	14
6	1	42	0	88.0	183	9	0
7	1	38	0	90.5	190	25	30
8	1	66	1	60.0	166	30	0
9	1	64	1	68.8	163	0	20
10	1	55	1	70.0	165	6	7
11	0	43	0	85.8	191	0	35
12	0	42	0	84.0	184	0	14
13	0	36	0	79.5	175	10	7
14	0	48	1	55.4	164	15	0
15	0	51	0	86.1	180	0	15
16	0	52	1	56.0	163	0	15
17	0	61	0	70.2	168	18	0
18	0	44	1	70.0	166	0	5
19	0	40	0	75.4	175	10	8
20	0	35	1	67.7	165	20	0

Session 3: Descriptive Statistics and Assessing Normality

Participants will be able to:

- Distinguish categorical and continuous variables
- Create frequency tables and bar charts for categorical variables
- Create histograms, box plots and scatterplots for continuous data
- Assess normality for continuous variables
- Make normality transformations

Key Terms:

categorical variable
continuous variable
frequency
descriptive statistics
mean, median, mode
histogram/box plot
scatterplot
normal distribution curve

1 Types of Variables

Different types of statistical analysis are done on different types of variables. Variables fall into two categories: categorical or continuous.

Categorical Variables: data with a limited number of distinct values or categories

Also be called qualitative data (Examples: gender, marital status, diagnosis)

Categorical variables can be string (text), or numeric where number codes represent categories (0= unmarried 1= married)

There are three types of categorical variables:

- Binary- only two possible answers (yes no)
- Nominal – a name or category with no implied order (type of cancer: breast, cervical, lung, prostate)
- Ordinal- name of category in a meaningful order (high, medium, low) but the distance between variables cannot be calculated

Continuous Variables: data with an infinite range of values and a distinct distance between each value. Also call quantitative or scale data.

(Example: \$72,195 is higher than \$52,398 and the distance between the two values is \$19,797)



Working Example 1: In the data set you define and entered, **diabetes.sav**, which variables are categorical and which are continuous?

DIABETES	continuous	categorical
AGE	continuous	categorical
SEX	continuous	categorical
WEIGHT	continuous	categorical
HEIGHT	continuous	categorical
CIGS	continuous	categorical
ALCOHOL	continuous	categorical

2 Descriptive Statistics

Descriptive statistics are used to summarize and describe the data gathered. Descriptive statistics are useful in making basic observations about the data such as the number of males and females, the age range and the mean (average) number of cigarettes smoked. Other statistics such as standard deviation give more information about the distribution of each variable.

2.1 Descriptive Statistics for Categorical Variables

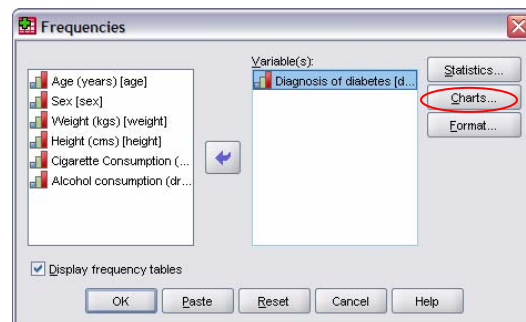
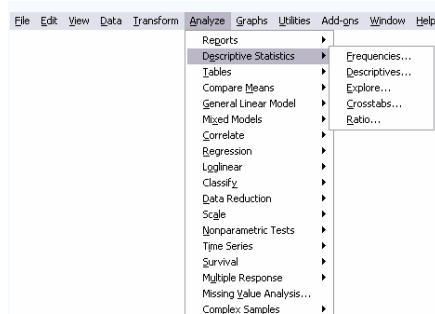
Categorical variables are summed up in terms of their **frequency** or number/percent. A Frequency Distribution shows how often a value was present (how many males and females). Frequencies can be shown in tables of counts and percents or in bar charts or pie charts.



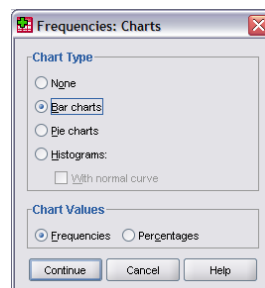
Working Example 2: The Frequency procedure produces frequency tables that display both the count and percent of cases for each variable. Open the data set **diabetes.sav**

Analyze > Descriptive Statistics > Frequencies

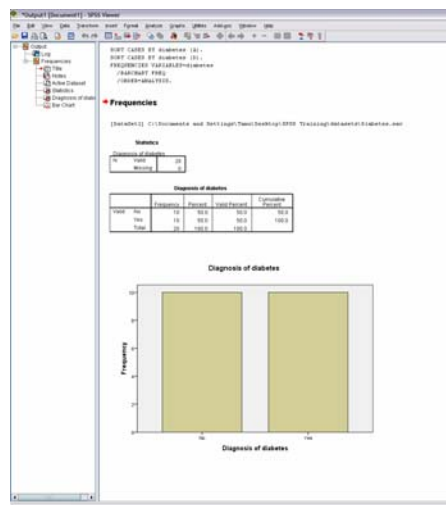
Chose the categorical variable > > check Display frequency tables



Click on Charts > bar chart > frequencies > continue > OK



A frequency table and bar chart appear in the Output window.



Other important statistics for categorical variables include the mode and the median. The **mode** is the category with the greatest number of cases (most often). For ordinal data the **median**, the value at which half of the cases fall above and half below (middle) is useful when there are many categories.

2.2 Descriptive Statistics for Continuous Variables

Most statistical analysis deals with continuous variables because this data has the most variability. There are several measures to continuous variables.

Measures of Central Tendency include the **mean** (average) and **median** (middle value).

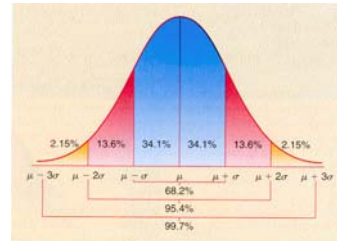
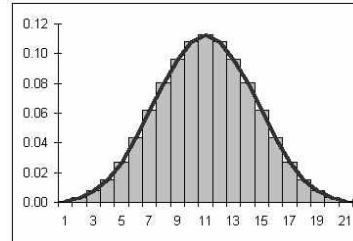
Measures of Dispersion show the amount of variation or the spread of the data. These include **variance**, **standard deviation** and minimum and maximum values.

- **Variance** - The variance is how much individuals score differ (or deviate) from the mean. Variance is based on squared deviations of scores about the mean.
- **Standard deviation** - The standard deviation is a measure of variability expressed in the same units as the data. In a normal distribution, 68% of the scores are fall ± 1 standard deviation from the mean, 95% are ± 2 standard deviations from the mean and nearly all scores, 99%, are with in ± 3 standard deviations of the mean.

Measures of Central Tendency and Dispersion show if a variable has a normal distribution or if the data is not normal (skewed- clumped to the right or left). A **Normal distribution** has

- a symmetrical bell-shaped curve
- only one peak
- no outliers (extreme values)
- and the mean and median are the same.

Normal Distribution Curve (continuous variables)



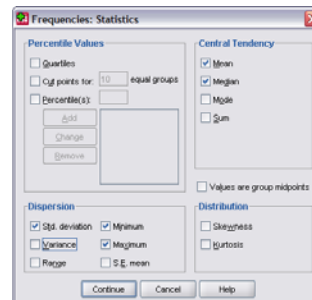
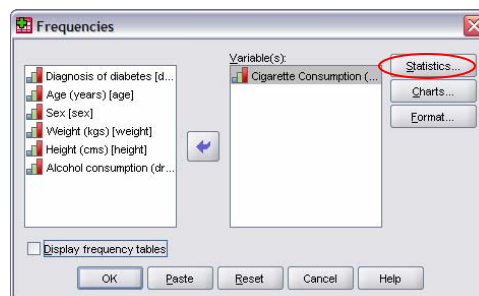
Continuous variables are displayed in **histograms**. Histograms are similar to bar charts but their bars touch indicating the continuous nature of the data. Each bar represents a range of values and the height of the bar is determined by the mean of each range.



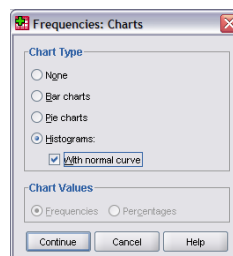
Working Example 3: The Frequency procedure produces tables and graphs (histograms and box plots) showing the mean, median and distribution of a continuous variable.

Using diabetes.sav Analyze > Descriptive Statistics > Frequencies > select the continuous variable > *Cigarette consumption* > (uncheck display frequency tables)

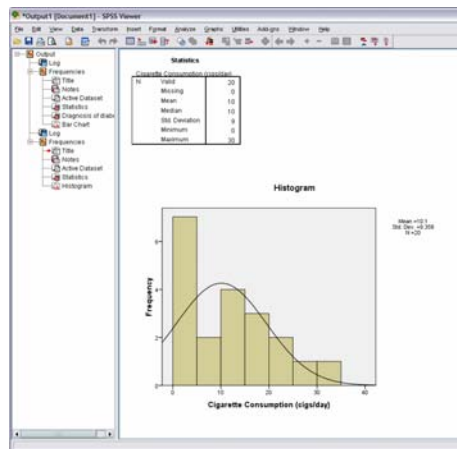
Click on Statistics > check mean, median, standard deviation, minimum and maximum > Continue



Click Charts> check histogram> with normal curve> Continue > OK



A frequency statistics table and histogram is displayed in the Output viewer.



Output Analysis Questions for *Cigarette Consumption*:

Does the curve appear to be normal or skewed?

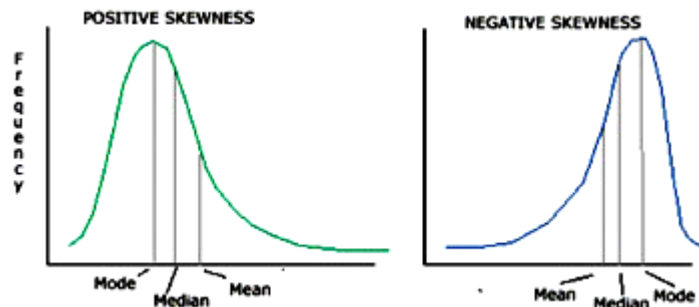
Are the mean and the median the same?

What are the minimum and maximum responses?

What is the standard deviation?

IF this were a normal distribution 68% of respondents smoke between _____ and _____ cigarettes per day?

The extreme values found in the data set diabetes.sav shifted the mean to the right so that the histogram is positively **skewed**. Notice how extreme values have no effect on the median.



3 Assessing Normality

Most statistical analysis assumes that continuous variables are normally distributed. If distributions are not normal or skewed they can be transformed before further analysis. Normality is only assessed for continuous variables.

There are several graphs that can be generated to determine if a continuous variable has a normal distribution. They include:

- Histograms
- Boxplots
- Scatterplots
 - Normal (QQ) Probability plots

→ Detrended Normal QQ Probability plots

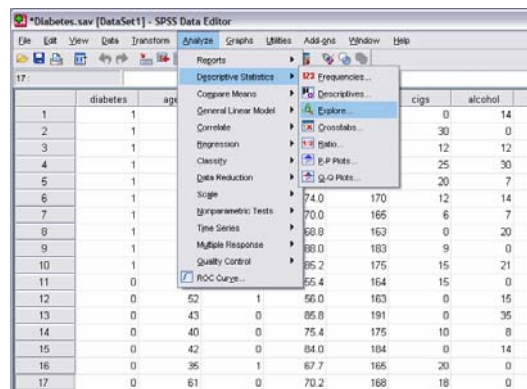
Some statistics can also be used to check for normality:

- skewness
- kurtosis
- Kolmogorov-Smirnov and Shapiro-Wilks

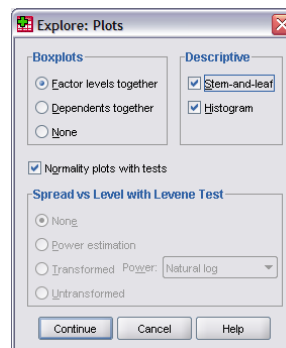
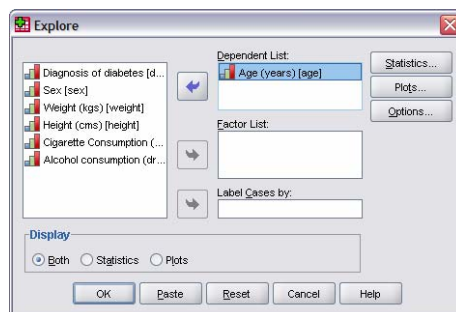
3.1 Normality Graphs



Working Example 4: To obtain graphs and statistics used in determining if a continuous variable is normally distributed use the Explore procedure.
Analyze > Descriptive Statistics > Explore >

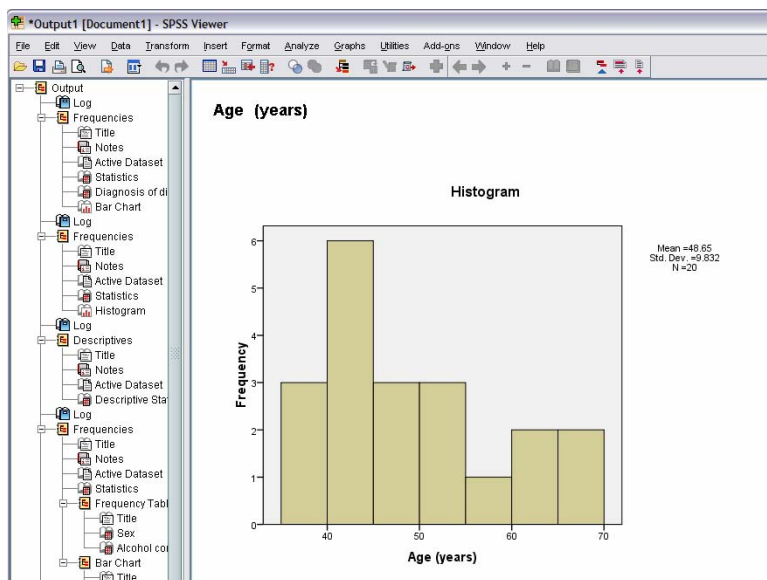


Using the data set diabetes.sav select the variable AGE > move it to the Factor List > Display Both > Click Plots > check histograms, Normality plots with tests > factor levels together under Boxplots > Continue > OK



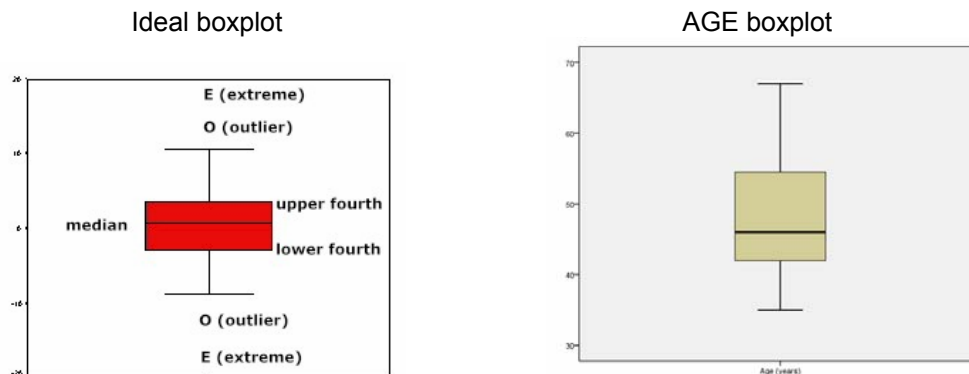
All tables and graphs are displayed in the Output Window.

Let us first look at the **histogram** for the variable AGE.



The Y axis shows frequency of cases. The x-axis values are the midpoints of the value ranges (each bar covers a range of 5 years). Compared to the ideal normal distribution curve, this histogram's shape is positively skewed.

The **Boxplot** shows the summary statistics for AGE.



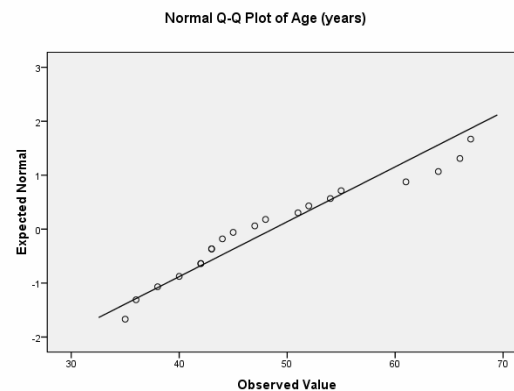
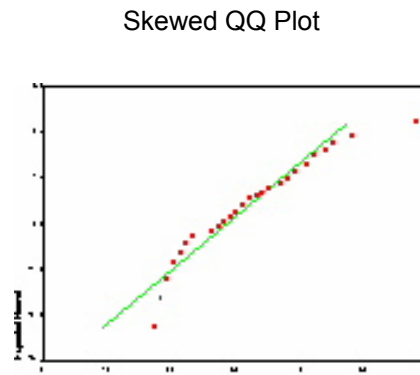
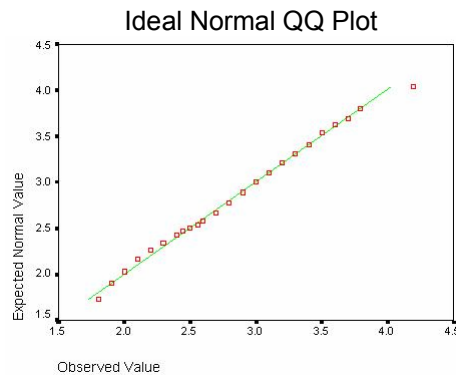
The boxplot shows the median, the spread of the data, the 25th and 75th percentiles and any outliers. The lower boundary of the box is the 25th percentile and the upper boundary is the 75th percentile. The median is the horizontal line in the box. The smallest and largest observed values are shown in the length of the whiskers. Any outliers are shown with a (o).

If the distribution is normal

- The median line will be centered in the box
- the whiskers will be of equal length
- and no outliers are present

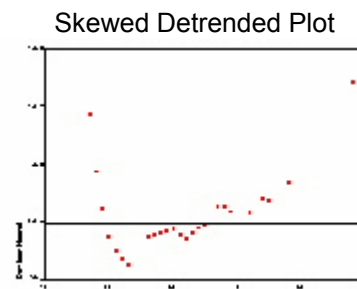
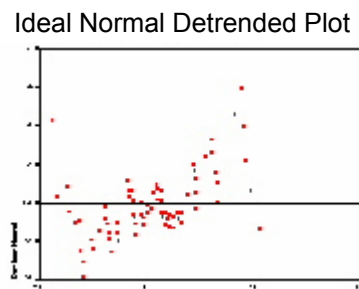
As with the histogram the boxplot shows that AGE is slightly skewed.

The Normal QQ Plot is constructed by plotting each observation by its z-score (how many standard deviations it is away from the mean). In the **Normal Probability (QQ) Plot** cases will follow a straight line along a diagonal if the distribution is normal. Systematic departures from this line show lack of normality.



The Normality plot for AGE looks slightly skewed.

The **Detrended Normal QQ Plot** plots deviations from the Normal or the deviations from the diagonal line in the Normal QQ plot. The points should show a random pattern distributed around the zero line, with no apparent clustering of points (points do not follow lines).



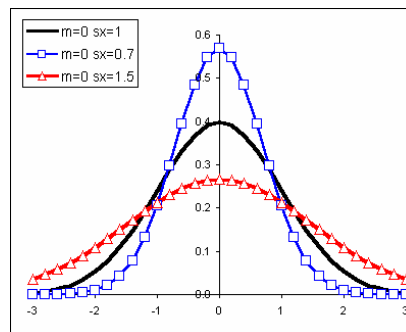


AGE values are scattered and do not appear to be aligning, but some values are far from the zero line.

3.2 Normality Statistics

Skewness and **kurtosis** refer to the shapes of the distribution. Both skewness and kurtosis values are zero (0) if the distribution is perfectly normal. Positive values for skew indicate a positive skew. Negative values for skew indicate negative skewness.

Positive values for kurtosis indicate a distribution that is sharply peaked (**leptokurtic**). Negative values for kurtosis indicate a distribution that is flatter (**platykurtic**).



Descriptives				Statistic	Std. Error
Age (years)	Mean			48.65	2.198
	95% Confidence Interval for Mean	Lower Bound		44.05	
		Upper Bound		53.25	
	5% Trimmed Mean			48.39	
	Median			46.00	
	Variance			96.661	
	Std. Deviation			9.832	
	Minimum			35	
	Maximum			67	
	Range			32	
	Interquartile Range			13	
	Skewness			.605	.512
	Kurtosis			-.673	.992

AGE has a positive skew and a flatter than normal kurtosis. We also know that mean and median are equal in normal distributions. Age mean is 48.65 while the median is 46.0, which are nearly equal.

The **Kolmogorov-Smirnov** and **Shapiro-Wilk** statistics correspond to the QQ and detrended probability plots. If the significance level (.Sig) is greater than .05 then we assume the data fits the normal distribution. Shapiro-Wilk should be calculated if the sample size is less than 100.

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age (years)	.145	20	.200*	.929	20	.145

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

These statistics tell us that AGE is normally distributed.

CONCLUSION: Combining our analysis of graphs and statistics for normality, we can conclude that although AGE is not perfectly distributed it is not extremely skewed. A transformation of the data is not necessary. The variable has a near-normal distribution.



Working Example 5: Continue working with the data set diabetes.sav

- Obtain descriptive statistics and histograms for all continuous variables
- Decide if each variable has a normal distribution or is greatly skewed

Variable Name	Mean	Median	Standard Deviation	Distribution Curve (norm, pos, neg)
AGE	48.65	46.00	9.632	Normal

4 Variable Transformation

Variables rarely conform to the classic normal curve. When skewness and kurtosis are extreme, transformation is an option. The decision to transform variables depends on the severity of their departure from the norm.

A variable can be transformed 4 different ways:

- Natural logarithm (ln) – most common and good to reduce effect of outliers
- Square root (\sqrt{x})
- Reciprocal ($1/x$)
- Square (x^2)

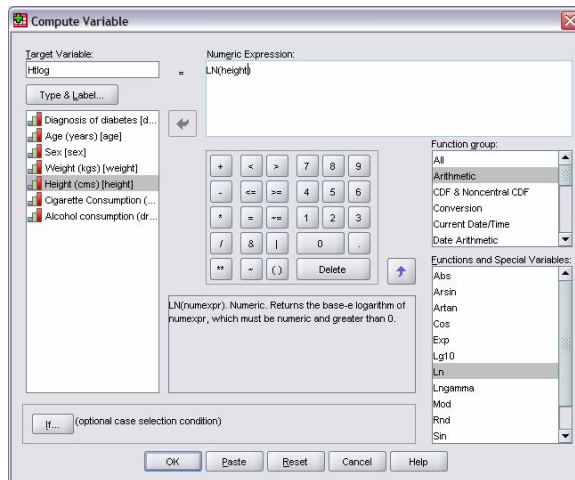
If a variable has a non-normal distribution, try transforming it via each of these 4 ways. The transformed data can then be assessed for normality. The transformed variable that is closest to normal should be used for further analysis.

In **Working Example 5** you discovered that some variables in diabetes.sav are not normally distributed. Choose one variables that is significantly skewed to perform a series of transformations.

All transformations are done using the **Compute** command. The Instructions below describe how to transform the non-normal variable HEIGHT.

4.1 Logarithmic Transformation (Ln)

Transform > Compute > in the Target variable box type an appropriate variable name (HTlog) > from the Function group select *arithmetic* > from Functions and Special variables box select *Ln* (short of Log) and press ▲ > select the variable (HEIGHT) and press ► > OK

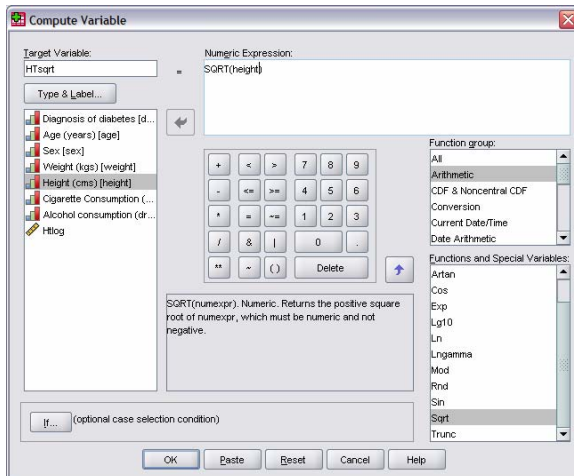


Now the new variable HTlog has been created and tests for normality can be performed.

	diabetes	age	sex	weight	height	cigs	alcohol	HTlog
1	1	43	1	55.5	160	0	14	5.08
2	1	66	1	60.0	166	30	0	5.11
3	1	45	0	97.1	200	12	12	5.30
4	1	38	0	90.5	190	25	30	5.25
5	1	54	1	67.0	163	20	7	5.09
6	1	47	1	74.0	170	12	14	5.14
7	1	55	1	70.0	165	6	7	5.11
8	1	64	1	68.8	163	0	20	5.09
9	1	42	0	88.0	183	9	0	5.21
10	1	67	0	85.2	175	15	21	5.16
11	0	48	1	55.4	164	15	0	5.10
12	0	52	1	56.0	163	0	15	5.09
13	0	43	0	85.8	191	0	35	5.25
14	0	40	0	75.4	175	10	8	5.16
15	0	42	0	84.0	184	0	14	5.21
16	0	35	1	67.7	165	20	0	5.11
17	0	61	0	70.2	168	18	0	5.12
18	0	44	1	70.0	168	0	6	5.11

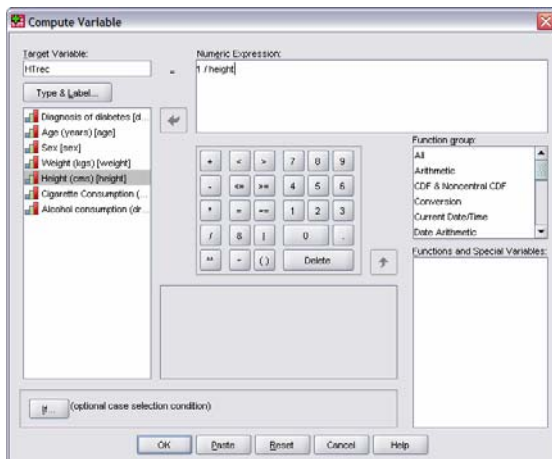
4.2 Square Root Transformation ($\sqrt{}$)

Transform > Compute > in the Target variable box type an appropriate variable name (HTsqrt) > from the Function group select *arithmetic* > from Functions and Special variables box select *sqrt* and press ▲ > select the variable (HEIGHT) and press the ► > OK



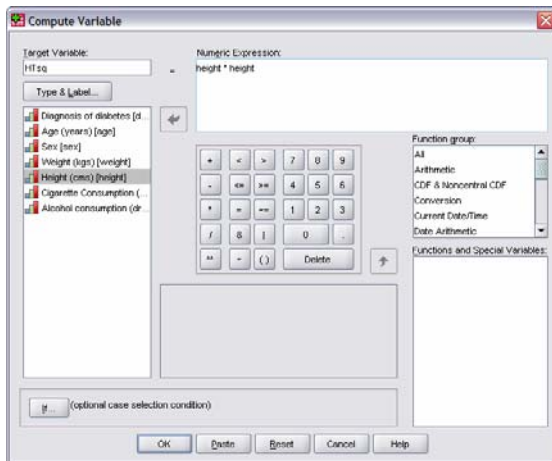
4.3 Reciprocal Transformation ($1/x$)

Transform > Compute > in the Target variable box type an appropriate variable name (HTrec) > using the keypad, type 1/ in the Numeric Expression box > select the variable (HEIGHT) and press ► > OK



4.3 Square Transformation (x/x)

Transform > Compute > in the Target variable box type an appropriate variable name (Htrsq) > select the variable (HEIGHT) and press ► > using the keypad, type * in the Numeric Expression box > select the variable (HEIGHT) and press ► > OK



Now the new variables HTsqrt, HTrec and HTsq have been created. Tests for normality can be performed on each of these new variables. Which ever variable version of HT__ is most normal should replace the original non-normal variable HEIGHT in future statistical analysis.



Working Example 6: Follow the instructions above to figure out which transformation of the HEIGHT variable produces a distribution that is closest to normal.

Session 4: Confidence Intervals and Recoding Variables

Participants will be able to:

- Obtain and explain confidence intervals for continuous variables
- Recode variables

Key Terms:

random error
systematic error
confidence interval

1 Types of Error

When reporting the results of statistical analysis some level of error is always expected. There are two types of error: random error and systematic error.

Random error occurs naturally in a sample when the cases chosen for observation have a mean value above or below the norm (a skewed curve). To correct for natural random error repeated samples of the same size or a larger sample size can be observed.

Systemic error usually occurs as a result of poor sampling techniques or when a confounding factor is not controlled for in analysis. For example, if you conducted a survey by distributing your questionnaire in front of the Post Office in Pago Pago on Good Friday your sample is likely to be unrepresentative of the population because it can only contain people who:

- are in Pago Pago at the time
- who go to the Post Office
- who are not attending church

2 Confidence Intervals

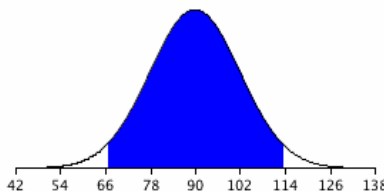
Researchers should always try to have minimal random and systematic error in order to arrive at more normal distributions. When a distribution of a variable is normal we can be confident in reporting our results.

Confidence intervals are used to indicate the reliability of findings from a sample. When you compute a confidence interval, you compute the mean of a sample in order to estimate the mean of the population. For example, a CI can be used to describe how reliable survey results are. Confidence intervals are often stated at the 95% level. "We are 95% confident that population mean weight of people with diabetes is between _____ and _____ kilograms."

A survey result with a small CI is more reliable than a result with a large CI.

95% of scores fall within the blue with a confidence interval of 66-114.

"We are 95% confident that the population mean will fall between 66Kg and 114Kg."

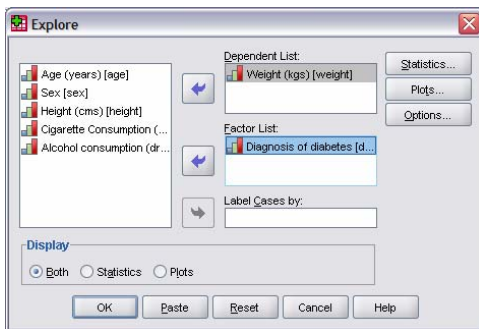


Open the data set diabetes.sav. We will obtain 90%, 95% and 99% confidence intervals for the population mean weight of diabetics and non-diabetics.

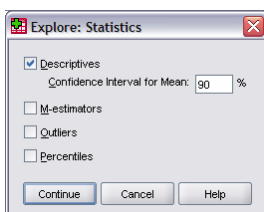
Dependent Variable – the effect, changes in response to the independent variable

Independent Variable (factor)- causes a change, sometimes deliberately manipulated

Analyze > Descriptive Statistics > Explore > Dependent List WEIGHT > Factor List DIAGNOSIS >



Statistics > 90% > Continue > OK



Descriptives					Statistic	Std. Error
Diagnosis of diabetes						
Weight (kgs)	No	Mean			73.010	3.5717
		90% Confidence Interval for Mean	Lower Bound		66.463	
			Upper Bound		79.557	
		5% Trimmed Mean			73.261	
		Median			72.800	
		Variance			127.572	
		Std. Deviation			11.2948	
		Minimum			55.4	
		Maximum			86.1	
		Range			30.7	
		Interquartile Range			19.7	
	Yes	Skewness			-.447	.687
		Kurtosis			-.942	1.334
		Mean			75.610	4.3864
		90% Confidence Interval for Mean	Lower Bound		67.569	
			Upper Bound		83.651	



Working Example 1: Obtain 95% and 99% CI to complete the table.

	mean weight	90% CI	95% CI	99%CI
Diabetics	75.610Kg	67.569 - 83.651		
Non-diabetics	73.010Kg	66.463 - 79.557		

What do you notice about the length of the intervals when the confidence level (%) changes?



Working Example 2: Obtain a 95% CI for population mean cigarette consumption for diabetics and non-diabetics

	mean cigs	95% CI
Diabetics		
Non-diabetics		

Express the results: "We are 95% confident that _____"

3 Recoding Data

Sometimes you may want to recode variable data to:

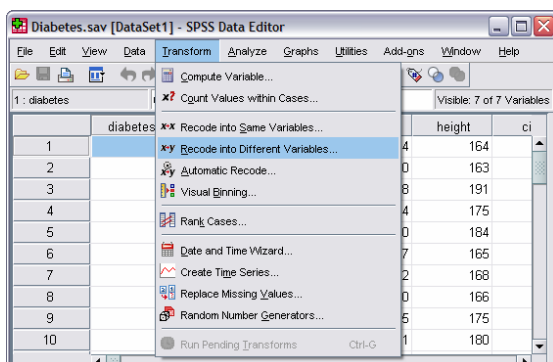
- turn a continuous variable into a categorical variable (example: individual ages into age range categories)
- combine several response categories into a single category (education level → educated)
- create a new variable that is the computed difference between two existing variables (first WEIGHT – second WEIGHT)
- recode negatively worded scale items (strongly agree --- strongly disagree)
- replace missing values and bring outlying cases into the distribution

3.1 Create categories for a continuous variable

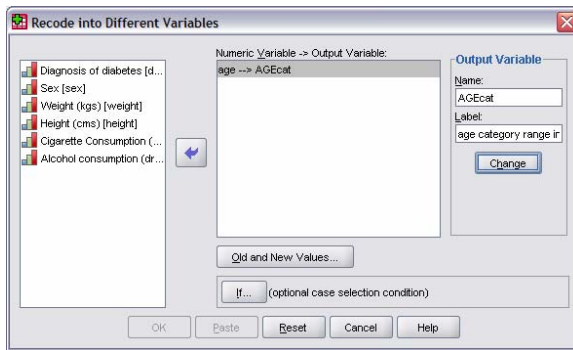
You are working with the data set diabetes.sav and you would like to create age categories from the continuous variable AGE. Since there are few cases you will use a *median split* to create two new age categories.

You can figure out median age by Analyze > Descriptive Statistics > Frequencies > AGE . Statistics > Median> The median (middle) age of respondents = _____

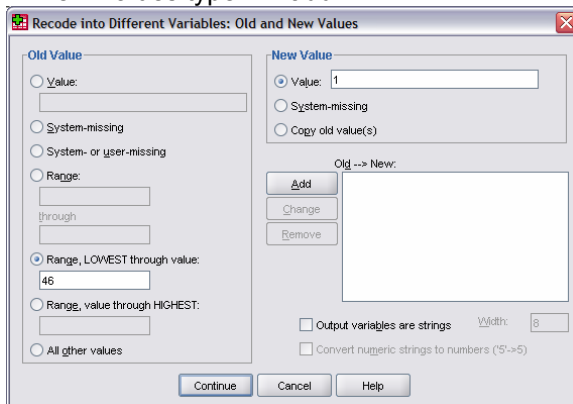
Transform > **Recode Into Different Variables** (this will retain the original data)



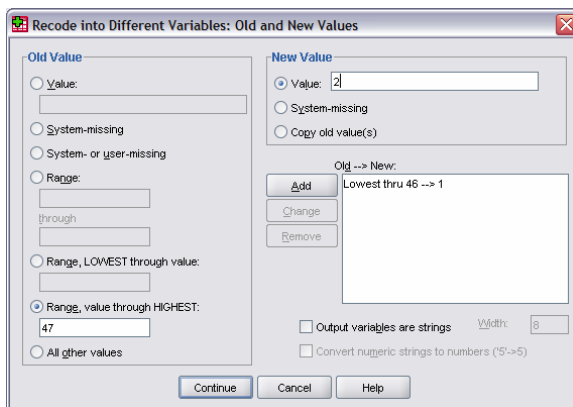
Select the variable AGE > type a new variable name (AGEcat) and variable label > Change



Click on Old and New values > click the third Rang button and type the median age for AGE > In New Values type 1 > add >



Click on the third range button > type the median +1 > new Value 2 > Add > Continue > OK



Now the variable AGEcat has been added to the data set with two possible values
1 = 0 - 46 years and 2 = 47+ years
This variable can now be used for analysis as a categorical variable.

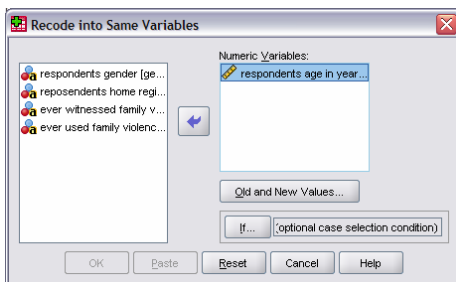
	age	sex	weight	height	cigs	alcohol	AGEcat	
1	48	1	55.4	164	15	0	2.00	
2	52	1	56.0	163	0	15	2.00	
3	43	0	85.8	191	0	35	1.00	
4	40	0	75.4	175	10	8	1.00	
5	42	0	84.0	184	0	14	1.00	
6	35	1	67.7	165	20	0	1.00	
7	61	0	70.2	168	18	0	2.00	
8	44	1	70.0	166	0	5	1.00	
9	36	0	79.5	175	10	7	1.00	
10	51	0	86.1	180	0	15	2.00	
11	43	1	55.5	160	0	14	1.00	
12	66	1	80.0	166	30	0	2.00	
13	45	0	97.1	200	12	12	1.00	
14	38	0	90.5	190	25	30	1.00	
15	54	1	67.0	163	20	7	2.00	
16	47	1	74.0	170	12	14	2.00	



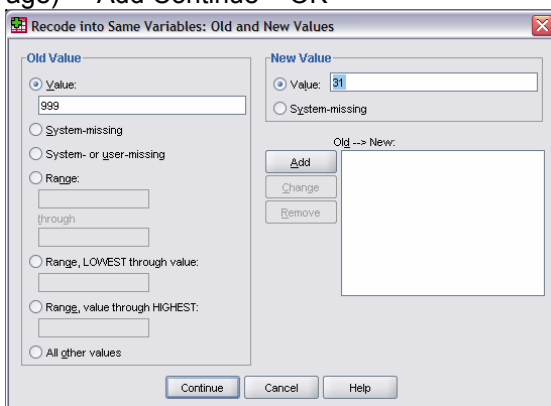
Working Example 3: In Session 2 you saved the dataset Vsurvey.sav. Open this dataset. Vsurvey contains one missing age for a respondent. You can replace this missing value with *mean substitution*.

First find the mean age in Vsurvey.sav. (31)

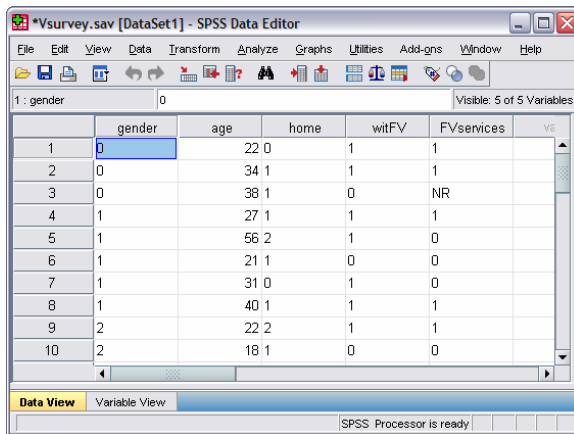
Transform > Recode into Same Variables > add AGE >



Click on Old and New Values > the old value for missing data was 999 > new Value 31 (mean age) > Add Continue > OK



(*If a missing data value had not been previously assigned select System/user-missing button.)



	gender	age	home	witFV	FVservices	VS
1	0	22 0	1	1		
2	0	34 1	1	1		
3	0	38 1	0	NR		
4	1	27 1	1	1		
5	1	56 2	1	0		
6	1	21 1	0	0		
7	1	31 0	1	0		
8	1	40 1	1	1		
9	2	22 2	1	1		
10	2	18 1	0	0		

Now the missing age value (999) has been replaced with the mean age.



Working Example 4: Recode a variable in diabetes.sav

Use the CIGS variable to create a new variable SMOKER which identifies smokers and non-smokers.

If CIGS = 0 then define SMOKER = 1 (smokes 0 cigarettes per day)

If CIGS ≥ 1 then define SMOKER = 2 (smokes at least 1 cigarette a day)

Check the recoding by comparing the values of CIGS with those of SMOKER

Label the values of the new variable as smoker and non-smoker.

Determine the number of smokers and non-smokers in the sample by obtaining a frequency distribution (bar chart) of the new variable SMOKER



Working Example 5: Recode a variable in diabetes.sav.

Use the ALCOHOL variable to create a new variable DRINKER which divides individuals into alcohol consumption level categories: non-drinkers, occasional drinkers and drinkers.

If ALCOHOL = 0 then define DRINKER = 1 (does not drink alcohol)

If ALCOHOL ≥ 1 and ≤ 7 then define DRINKER = 2 (drinks 1-7 standard drinks on average per week)

If ALCOHOL > 7 then define DRINKER = 3 (drinks 8 or more standard drinks on average per week)

- Check the recoding by comparing the values of ALCOHOL with those of the new variable DRINKER.
- Label the values of the new variable.
- Determine the number of cases in the sample in each DRINKER category.

Save your data as diabetes.sav



Week 2 Task: Working with Original Data

Please complete before Session 5 on Tuesday April 22.

Locate a dataset relevant to your work to complete the following steps. The dataset can be raw data (surveys) or data already entered into an Excel spreadsheet. If you have no such data either ask a co-worker for data or inform your facilitator and she will provide you with a generic Public Health data set. The data should have at least 20 cases if possible.

Using the skills and concepts outlined in Sessions 2-4 complete the following steps with your original data set.

- Define all variables
- Enter case data
- Check that the data has been entered correctly
- Generate Descriptive Statistics
 - frequencies for categorical variables.
 - histograms, box plots and scatterplots for continuous variables
- Assess normality for continuous variables
- Make normality transformations for non-normal continuous variables
- Recode at least one continuous variable into a categorical variable
- Create and test at least one correlation hypothesis for categorical variables (crosstab) or continuous variables (scatterplot)
- **Save the data your data set and your output.**

At the end of this training workshop (May 1, 2008), you will be asked to present some of your data analysis to the other participants.

Session 5: Inferential Statistics and Correlations/Associations

Participants will be able to:

- Understand the difference between Descriptive and Inferential Statistics
- Determine if there is a relationship between two categorical variables
- Determine the strength of relationships between two continuous variables

Key Terms:

null hypothesis
alternative hypothesis
correlation
assumptions
p-value

[1 Descriptive Statistics](#)

Sessions 2-4 focused on **Descriptive Statistics**--statistical methods used to describe and summarize information obtained from the sample. Because it is usually impossible to survey or observe the entire population of interest, data is gathered from a representative sample of that population. If the sample data has a normal distribution, or the data was transformed to achieve a normal distribution, further analysis can be done.

[2 Inferential Statistics](#)

After you have generated Descriptive Statistics about each of your variables you can begin to conduct **Inferential Statistics**—using the results of samples (means of variables) to test hypotheses. Inferential statistics compare the sample mean to a hypothesized value.

A hypothesis is a statement you seek to either prove or disprove with evidence from the sample data. There are two types of hypotheses for each inferential statistical test:

The **null hypothesis** is always tested and then accepted or rejected

Ho: the population parameter (mean) = specific value

The **alternative hypothesis** says

Ha: the population parameter (mean) \neq specific value

There are usually assumptions that must be met before conducting inferential statistics tests, most importantly:

- the data comes from a random sample of the population
- the data comes from a normal distribution

When conducting inferential statistics always

1) check for assumptions first, if assumptions are not met proceeding with statistical tests will most likely end in inaccurate results that cannot be generalized to the population

2) define your null (Ho) and alternative (Ha) hypotheses by writing each out before testing the null

[3 Correlation](#)

One inferential statistics test is **correlation** which looks for a relationship between two variables.

The Pearson coefficient describes the relationship between two continuous variables. Correlation between two categorical variables is discovered through crosstabulation tables (crosstabs).

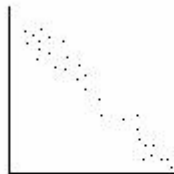
[3.1 Testing correlation between two continuous variables](#)

When looking for a correlation (linear association) between two continuous variables, scatterplots are generated. If the data is normally distributed all data points will fall between +1 and -1.

If all points line up along +1 they are positively correlated (as one value increases the other value increases). If all points line up along -1 they are negatively correlated (if one value goes up the other value goes down). If all points line up along the 0 line or they are randomly scattered there is no correlation between the variables.

Scatter Diagram Correlation

Strong Negative Correlation



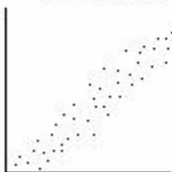
Weak Negative Correlation



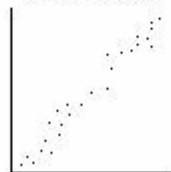
No Correlation



Weak Positive Correlation



Strong Positive Correlation



Hypotheses for continuous variable correlation test

Ho: no association between variables

Ha: there is an association between variables

There are 3 correlation **assumptions for continuous variables**:

- 1) data comes from a random sample and from a normal distribution
- 2) no missing data cases for the variables tested
- 3) the relationship between the variables must be linear

The test statistic of interest is Pearson's correlation coefficient r which will always be less than .05

If the Pearson's value is between

0 - 0.25 there is **little or no** association between the variables

0.25 - 0.5 there is a **fair** degree of association between the variables

0.5 - 0.75 there is a **moderate to strong** association between the variables

>0.75 there is a **very strong** association between the variables



Working Example 1: You want to see if there is an association between some of the variables in the diabetes.sav dataset. You suspect that the variables WEIGHT and ALCOHOL are associated, and probably positively associated.

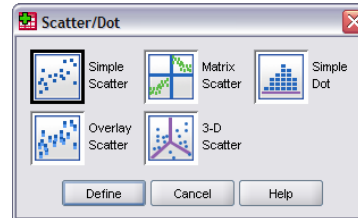
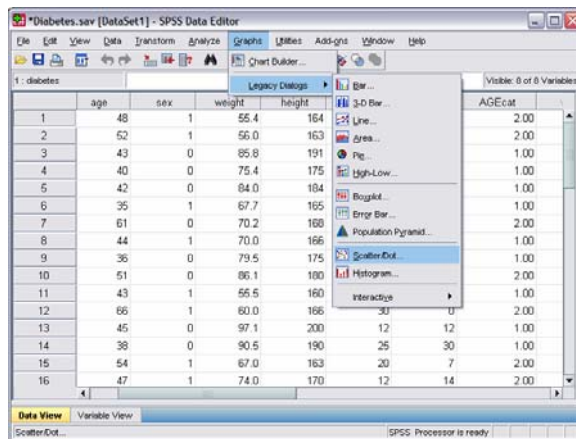
First check that WEIGHT and HEIGHT fulfill the assumptions.

Next write out your null and alternative hypotheses.

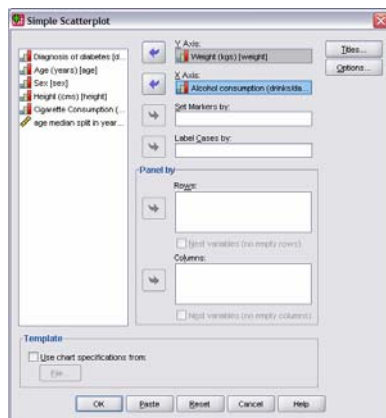
Ho: _____

Ha: _____

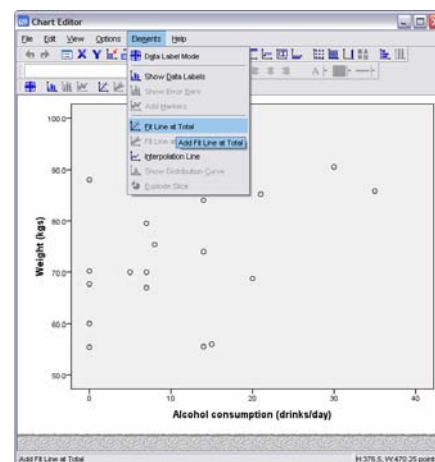
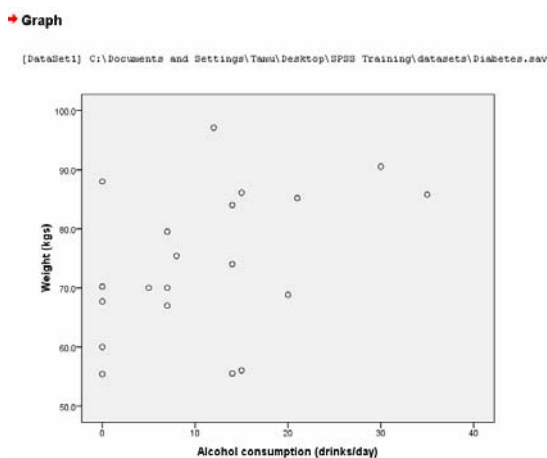
To test the null hypothesis obtain a scatterplot
Graphs > Legacy Dialogs > Simple Scatter > Define

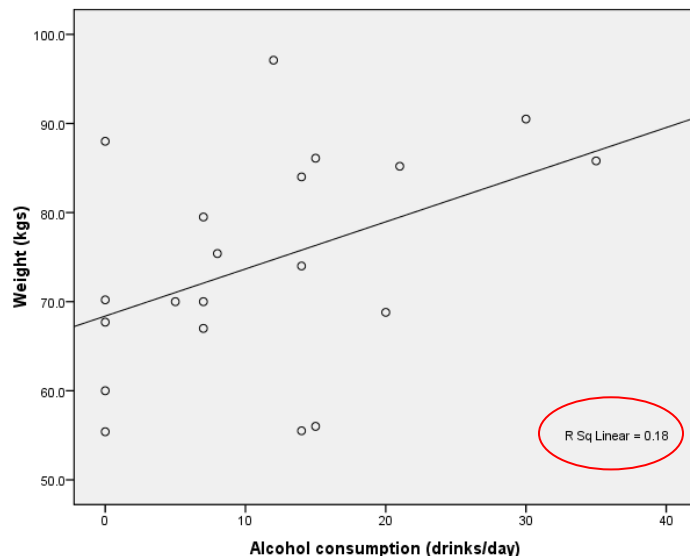


Select the variables WEIGHT and ALCOHOL > OK



The scatterplot that appear can be better interpreted if you add a fit line to the graph. Double click on the scatterplot > Elements > Fit line at Total



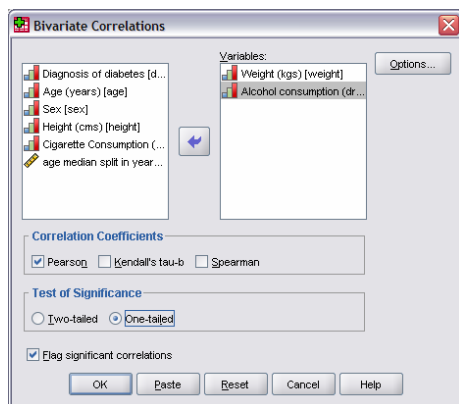


Even without the fit line the points do not seem to have a strong linear correlation. Pearson's R statistic shows the variability in the dependent variable (y axis) that can be explained by the independent variable (x axis). The closer R is to 1 the better the data fit the line.

$R = 0.18 \rightarrow 0 - 0.25$ there is **little or no** association between the variables

To test our correlation H_0 further we can generate Pearson table.

Analyze > Correlate > Bivariate > add the variables WEIGHT and ALCOHOL > check Pearson > one-tailed > flag significant correlations > OK



Correlations			
		Weight (kgs)	Alcohol consumption (drinks/day)
Weight (kgs)	Pearson Correlation	1.000	.424*
	Sig. (1-tailed)		.031
	N	20.000	20
Alcohol consumption (drinks/day)	Pearson Correlation	.424*	1.000
	Sig. (1-tailed)	.031	
	N	20	20.000

*. Correlation is significant at the 0.05 level (1-tailed).

$R = .424 \rightarrow$ there is a **fair** degree of association between the variables

Based on the analysis of the scatterplot and the bivariate correlation table, do we accept or reject the Null Hypothesis (H_0)? _____

3.1 Testing correlation between two categorical variables

A crosstabulation table (crosstab) and a chi-square test is used to show independence or relatedness between two categorical variables.

Crosstabs display the relationship between two or more categorical (nominal or ordinal) variables. The size of the table is determined by the number of distinct values for each variable with each cell in the table representing a unique combination of values. Numerous statistical tests are available to determine whether there is a relationship between the variables in the table. The most common test is a Chi-square.

The **hypotheses** for a Chi-square test for correlation are

Ho: there is no association between the two categorical variables

Ha: There is an association between the categorical variables

There are 4 **assumptions** that must be verified before conducting chi-square tests:

- 1) data comes from a random sample and from a normal distribution
- 2) cells are mutually exclusive (no case is counted twice)
- 3) cells are exhaustive (all cases are included)
- 4) when the number of cells is less than 10 and when the total sample size is small, all expected frequencies are 5 and above

Chi-square tests are useful in proving whether or not there is any association between categorical variables based on examining the p-value.

The **p-value** is the probability of the observed value when the null hypothesis (Ho) is true. Because most statistical tests are run at a 95% confidence interval the p-value is set at 5% (.05). So, if the p-value is less than .05 the observations from the data are significant. A p-value under .05 means that the data can only happen when the null hypothesis is false. A p-value greater than .05 shows that the data results could happen often when the Ho is true.

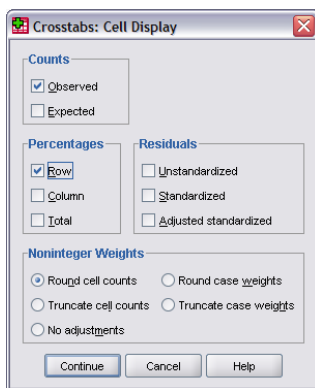
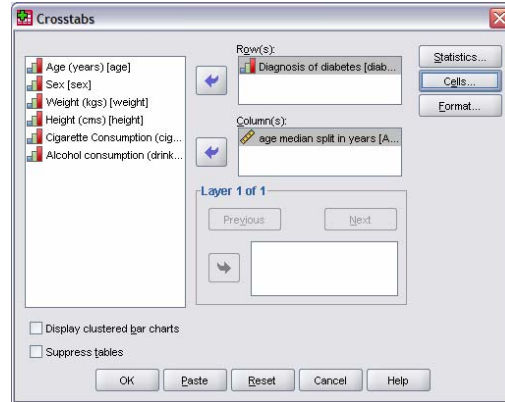
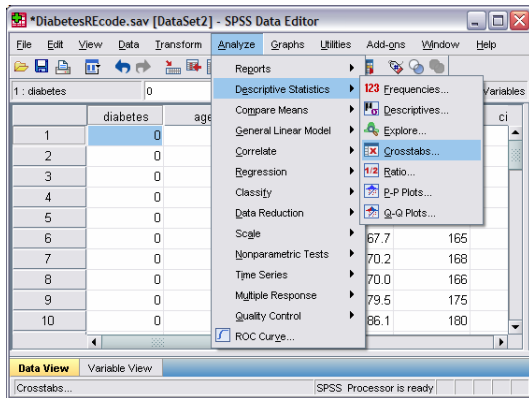
- If the p-value is **less than .05 we should reject the Ho** and say there IS an association between the two variables
- If the **p-value is greater than .05 we should accept the Ho** and say there is NOT a significant association between the two variables



Working Example 2: Generate a Crosstab

Let's examine the variables DIABETES and AGEcat (the categorical variable we created from the continuous variable AGE) to see if there is an association between them.

Analyze > Descriptive Statistics > Crosstabs > select DIABETES as row > select AGEcat as column > cells > Under Percentages click Row > Continue > OK



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Diagnosis of diabetes * age median split in years	20	100.0%	0	0%	20	100.0%

Diagnosis of diabetes * age median split in years Crosstabulation

		age median split in years			Total
		up thru 46	47 or older		
Diagnosis of diabetes	No	Count	6	4	10
		% within Diagnosis of diabetes	60.0%	40.0%	100.0%
	Yes	Count	4	6	10
		% within Diagnosis of diabetes	40.0%	60.0%	100.0%
Total		Count	10	10	20
		% within Diagnosis of diabetes	50.0%	50.0%	100.0%

3.1.1. Reading a Crosstab

Crosstabs show counts and percents.

How many diabetics are 46 years old or younger?

How many non-diabetics are 47 years or older?

Does there appear to be an obvious relationship between age and diagnosis of diabetes?

3.1.2 Significance Testing for Crosstabs – Chi-square

You may think there is a relationship between the variables but could this be just a random variation (naturally occurring) in your sample? The most common test to determine if the relationship between the variables is a true correlation or just random chance is a Chi-square.

The Pearson chi-square statistic tests the hypothesis that the row and column variables are independent (unrelated) tests the strength of the correlation. The value of the statistic is not very important. Instead, look at the Significance value (sig.) or p-value.

The lower the p-value, the less likely the two variables are independent (unrelated) the higher the p-value the more likely the two variables are related.

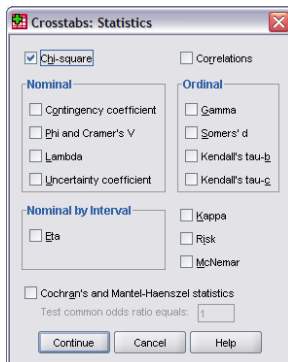
Remember our hypotheses

Ho: there is no association between the two categorical variables

Ha: There is an association between the categorical variables



Working Example 3: Generate a Chi-square
Analyze > Descriptive Statistics > Crosstabs > Statistics> Chi-square > Continue > OK



Diagnosis of diabetes * age median split in years Crosstabulation

			age median split in years		
			up thru 46	47 or older	Total
Diagnosis of diabetes	No	Count	6	4	10
		% within Diagnosis of diabetes	60.0%	40.0%	100.0%
	Yes	Count	4	6	10
		% within Diagnosis of diabetes	40.0%	60.0%	100.0%
	Total	Count	10	10	20
		% within Diagnosis of diabetes	50.0%	50.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.800 ^a	1	.371		
Continuity Correction ^b	.200	1	.655		
Likelihood Ratio	.805	1	.369		
Fisher's Exact Test				.656	.328
Linear-by-Linear Association	.760	1	.383		
N of Valid Cases	20				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.00.

b. Computed only for a 2x2 table

The Pearson Chi-square p-value is greater than .05

Do we accept or reject the null hypothesis? _____



Working Example 4
Obtain a Crosstab for DIABETES and SEX.

Then generate a Chi-square for these variables.

What conclusion can you make about the null hypothesis? (accept or reject) Why?

Session 6: One-sample and paired t-tests

Participants will be able to:

- Check for t-test assumptions
- Write hypotheses for t-tests
- Conduct a one-sample t-test
- Conduct a two sample (repeated measures) t-test

Key Terms:

null hypothesis (H_0)
alternative hypothesis (H_a)
assumptions
p-value
confidence interval
t-statistic

1 Inferential Statistics

T-tests are used in inferential statistics. They are statistical tests used to make statements about the population based on the characteristics of a sample.

2 Types of t-tests

A t-test is used to determine whether there is a significant difference between two sets of scores. There are three main types of t-tests:

- One-sample
- Repeated-measures
- Independent groups

T-tests make hypotheses about population means

Each statistical test has certain assumptions that must be met before analysis. If the assumptions are violated the test will not yield accurate results and generalizations about the population may be false.

The **assumptions** for all t-tests are:

- Tests are conducted on continuous variables.
- Scores were gathered from a random sample of the population.
- Scores are normally distributed.
- Scores are independent (data do not overlap, observations are not shared)

2.1 One-sample t-test

The one-sample t-test is used when you have data from a single sample of participants and you want to know if the mean of the population (μ) from which the sample is drawn is the same as a hypothesized mean.

The **hypotheses** for a one-sample t-test are

$H_0: \mu = \text{specific value}$

$H_a: \mu \neq \text{specific value}$

We **accept the H_0** when the p-value is less than .05 and if 0 is not in the 95% confidence interval range (between lower and upper). If the p-value is less than .05 the population mean is significantly different than the hypothesized mean. If 0 is not in the confidence interval we cannot be confident that our hypothesized mean will also be the population mean.

Remember to check assumptions and define your hypotheses before conducting the statistical test.

We will use the data set **pulse.sav** to practice t-tests. The *pulse* data was collected from a random selection of 92 youth who participated in a simple experiment. Each participant's resting heart rate pulse was recorded. Then each person flipped a coin. If the coin came up heads they

ran in place for one minute and then their pulse was recorded again. If the coin came up tails the participant did not run, waited one minute and had their second pulse recorded.

Pulse.sav contains the following:

Variable	Description
Pulse1	first pulse rate in beats per minute
Pulse2	second pulse rate in beats per minute
Ran	1= ran 2= did not run
Smokes	1= smokes regularly 2 = does not smoke regularly
Sex	1 = male 2= female
Height	height in inches
Weight	weight in pounds
Activity	usual level of physical activity 1= slight 2 = moderate 3 = a lot

Locate and open **pulse.sav**



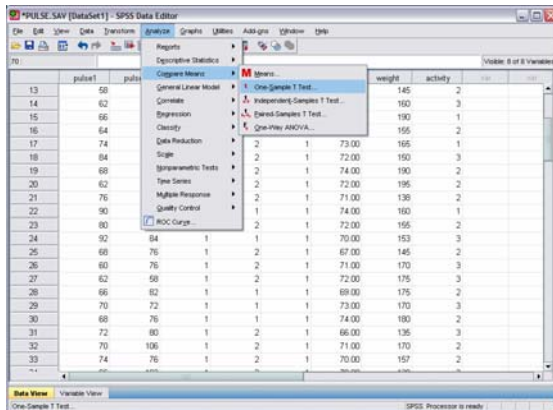
Working Example 1: Population mean weight one-sample t-test

Conduct a one-sample t-test with the following hypotheses

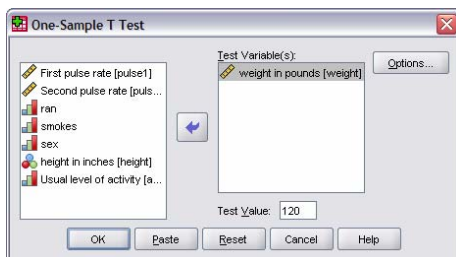
$H_0: \mu \text{ weight} = 120 \text{ pounds}$

$H_a: \mu \text{ weight} \neq 120 \text{ pounds}$

Analyze > Compare means > One-Sample t-test



Select the variable Weight > type the test value (120) > OK



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
weight in pounds	92	145.15	23.739	2.475

One-Sample Test

	Test Value = 120					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
weight in pounds	10.162	91	.000	25.152	20.24	30.07

Do we accept or reject the $H_0: \mu \text{ weight} = 120$ pounds?

Why?

What do our sample data tell us about the population?

2.1.1 What is a t statistic?

The t statistic is a measure of how extreme a statistical estimate is. You compute this statistic by subtracting the hypothesized value from the statistical estimate (μ) and then dividing by the estimated standard error. In many, but not all situations, the hypothesized value would be zero.

The hypothesized value is reasonable when the t-statistic is close to zero. The hypothesized value is not large enough when the t-statistic is large positive. The hypothesized value is too large when the t-statistic is large negative.

What does the t-statistic tell us about our sample?



Working Example 2: Continue working with the Pulse.sav data set. Test the following hypotheses:

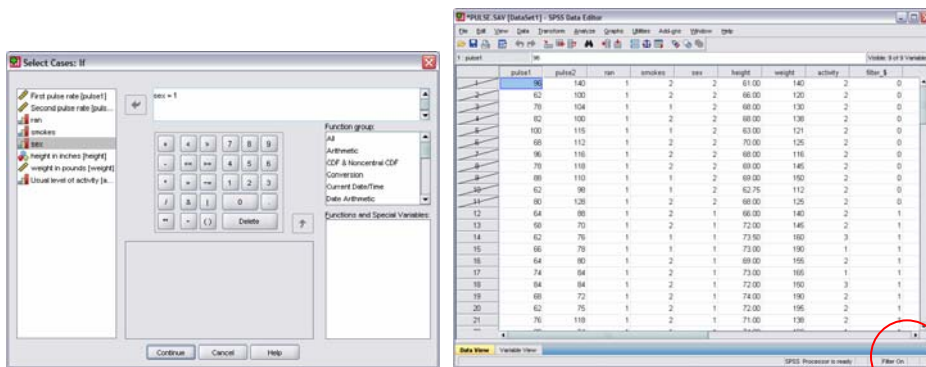
The mean population height for males is 75 inches.

The mean population weight for females is 120 pounds.



Tip: To test the first hypothesis for males you must select the males using the **Select Cases Procedure**.

Data > Select Cases > If condition is satisfied > If > select sex > type = 1 > Continue > OK



Now SPSS will only analyze male cases (where sex = 1). Notice the Filter on at the bottom of the screen. After you have tested the male hypothesis, change your "If" statement to Sex = 2 to select the female cases. When you are finished analyzing the female cases, restore the male students by clicking on *All* cases in the Select cases box or simply "reset".

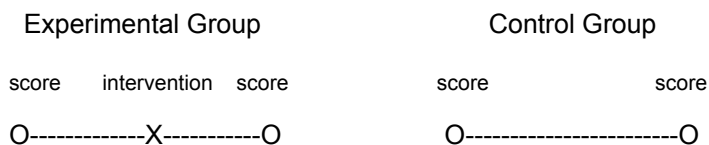
What did you find?

Ho: μ height for males = 75 inches. Accept or Reject

Ho: μ weight for females = 120 inches. Accept or Reject

2.2 Paired t-test (repeated-measures)

A paired or repeated-measures t-test is used when you have pre-test and post-test data from only one group of participants.



If the same randomly sampled population is tested twice over time and there is a difference in their means, that difference can be attributed to the independent variable or the treatment effect (exposure to the intervention), as long as confounding factors are controlled for.

The paired t-test has the **same assumptions as the one-sample t-test**. However, if your sample size is under 30 the difference between the scores for each participant (pre- and post-) should be normally distributed. This assumption is tested by testing the normality of each variable (score 1 and score 2) which will allow you to assume that the difference scores are also normally distributed.

The **hypotheses** for the paired t-test are:

Ho: $\mu_1 = \mu_2$ (no change) Ha: $\mu_1 \neq \mu_2$ (is change)

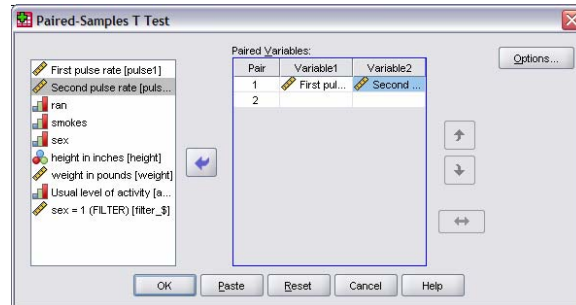
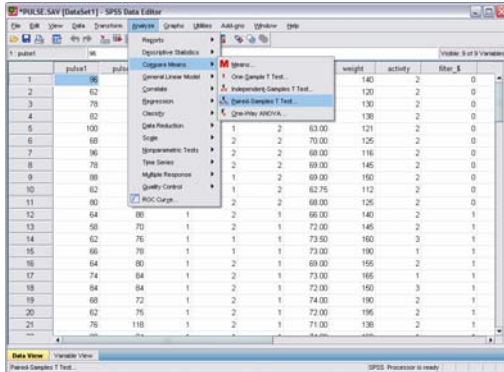
We **accept the Ho** when the p-value is less than .05 and if 0 is not in the 95% confidence interval range (between lower and upper).



Working Example 3: Conduct a paired t-test on pulse.sav data. Test the hypothesis:

$H_0: \mu \text{ pulse1} = \mu \text{ pulse 2}$ for all participants

Analyze > Compare means > Paired sample t-test > select Pulse 1 and Pulse 2 > OK



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	First pulse rate	72.87	92	11.009	1.148
	Second pulse rate	80.00	92	17.094	1.782

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	First pulse rate & Second pulse rate	92	.616	.000

Paired Samples Test

		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	First pulse rate - Second pulse rate	-7.130	13.471	1.404	-9.920	-4.341	-5.077	91	.000

What did you find?

$H_0: \mu \text{ pulse1} = \mu \text{ pulse 2}$ for all participants

Accept or Reject

Why?



Working Example 4: Test the following hypothesis:

$H_0: \mu \text{ pulse1} = \mu \text{ pulse 2}$ for all female runners

What did you find?

Session 7: Independent (two-sample) t-tests and One-way Analysis of Variance (ANOVA)

Participants will be able to:

- Check for t-test assumptions
- Write hypotheses for t-tests
- Conduct an independent group, two-sample t-test
- Conduct a one-way analysis of variance (ANOVA) with post hoc analysis

Key Terms:

null hypothesis (Ho)
alternative hypothesis (Ha)
variance (between-groups, within-groups)
p-value
normality
confidence interval
Levene's test
Tukey's test

1 Inferential Statistics

T-tests are used in inferential statistics. They are statistical tests used to make statements about the population based on the characteristics of a sample.

2 Independent groups (two-sample) t-tests

The independent groups (two-sample) t-test is used when the participants in one condition are different from the participants in another condition. A two-sample t-test is conducted on random samples of the same population (females and males, smokers and non-smokers).

The means of the two groups are compared to determine whether the difference between means for the two sets of scores is significant.

The Hypotheses for an independent two sample t-test are:

Ho: $\mu_1 = \mu_2$ Ha: $\mu_1 \neq \mu_2$

Two-sample t-test **assumptions** are:

- Tests are conducted on continuous variables.
- Scores were gathered from a random sample of the population.
- Scores are normally distributed.
- Scores are independent (participants should only appear in one group and these groups should be unrelated)

Additionally the two-sample t-test has the assumption:

- Homogeneity of variance- the groups should come from populations with equal variances

Variance - a measure based on the deviations of individual scores from the mean.



Tip: To test for homogeneity of variance use the **Levene's test for equality of variances**.

The **hypotheses for Levene's test** of variance are:

Ho: no significant difference between the variances of the group scores

Ha: there is a significant difference between the variances of the group scores

If the Levene's test result is not significant (p-value is greater than .05) we accept the null hypothesis (Ho). If this test statistic is significant (a p-value less than .05) you reject the null hypothesis and a Post Hoc Comparison is done.

2.1 Conducting a two-sample t-test

Using the pulse.sav data set you want to find out if the population mean weight is the same for males and females?

Write your hypotheses:

Ho: _____

Ha: _____

2.1.1 Check for normality

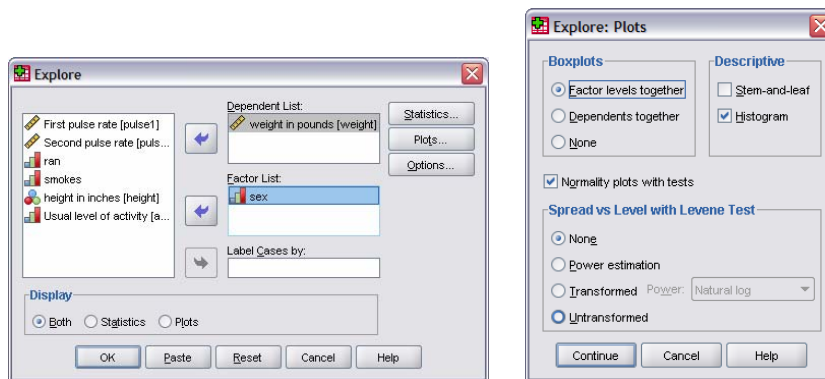
To satisfy the normality assumption and because we are working with two independent groups it is good to first check the normality of each set of scores separately.



Working Example 1: Check male/female WEIGHT for normality

Data > select cases > If condition is satisfied > If > sex = 1 > Continue > OK

Analyze > Descriptive Statistics > Explore > add WEIGHT to the Dependent list > add SEX to the Factor List > plots > histogram > check Normality plots with test > Continue > OK



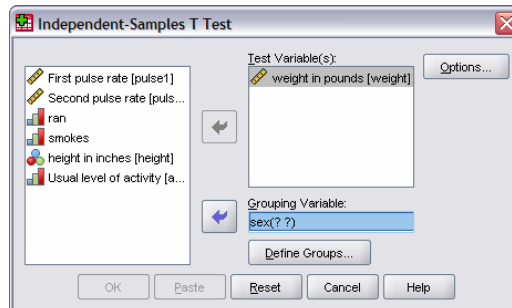
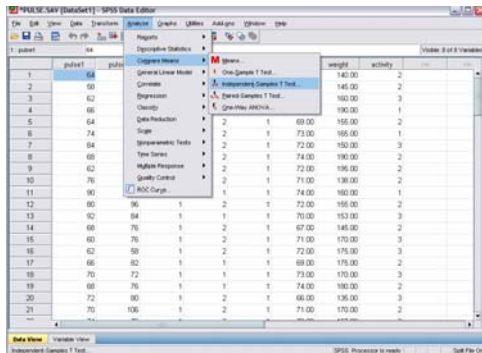
Are Male WEIGHT and Female WEIGHT normally distributed variables?

If not, what steps should you take next?

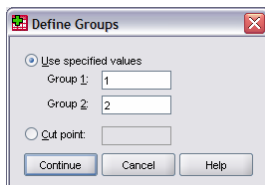
2.1.2 Conduct and independent groups (two-sample) t-test



Working Example 2: Analyze > Compare means > Independent samples t-test > add the variable WEIGHT to the Test variable > add SEX to the Grouping variable



Click Define groups > type the lowest value for the variable (1 = male, 2 = female) > Continue > OK



First we must check Levene's test for equal variance.

Group Statistics				
	sex	N	Mean	Std. Deviation
weight in pounds	Male	57	158.2632	18.63611
	Female	35	123.8000	13.37205

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
weight in pounds	Equal variances assumed	3.251	.075	9.529	90	.000	34.46316	3.61672	27.27791 41.64841
	Equal variances not assumed			10.297	87.713	.000	34.46316	3.34693	27.81153 41.11479

Levene's p-value is greater than .05. Do we accept or reject the Ho?

Ho: no significant difference between the variances of the group scores



If Levene's assumption is not met the two groups **cannot** be compared using a independent samples t-test. Then Levene's test becomes a meaningful result showing that population group variances are different.



If the groups have equal variance we then look at the p-value in the "Equal variances assumed" row. If the p-value is less than .05 we reject the Ho. If 0 is not included in the Confidence interval range we reject the Ho.

What do we conclude about the independent sample t-test hypotheses?
Ho: $\mu_1 = \mu_2$ accept or reject

3 One-way, between groups Analysis of Variance (ANOVA) with post hoc comparisons

The independent two sample t-test tests the hypothesis that two population means are equal. When you want to compare the means of more than two groups or levels of an independent variable a one-way analysis of variance (ANOVA) can be done.

The ANOVA test looks at two types of variance: **between groups variance** and within groups variance. Between groups variance measures the effect of the independent variable (treatment effect). **Within groups variance** is how much variance exists between individual cases in a group (determined by normality and Levene's tests).

ANOVA looks at the **F-ratio** which is a ratio of the two types of variance. A significant F-ratio (less than .05) indicates that the population means are probably not equal and the Ho is rejected. But we go on to analyze where the significant differences in population means lie through post hoc analysis.

Post hoc analysis involves hunting through the data for any significance or doing a set of comparisons. The post hoc test we will use is **Tukey's honestly significant difference (HSD)**.

The **hypotheses** for ANOVA are:

Ho: all population means are equal ($\mu_1 = \mu_2 = \mu_3 \dots$)

Ha: not all population means are equal

The ANOVA **assumptions** are:

- Random samples representative of the population in a normal distribution
- Observations are independent (participants should only appear in one group and these groups should be unrelated)
- Population variance is equal (proved with a Levene's test)

3.1 Check for normality

All population samples should have a normal distribution. Each sample should first be checked for normality. If a sample is non-normal a transformation can be performed.

3.2 Check for Homogeneity of variance (Levene's test)

ANOVA tests also assume that each group should have similar variances. Levene's test determines if variances are equal or not.

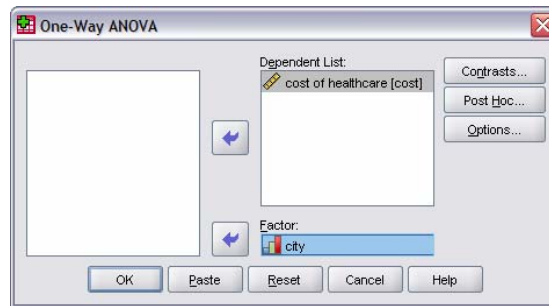
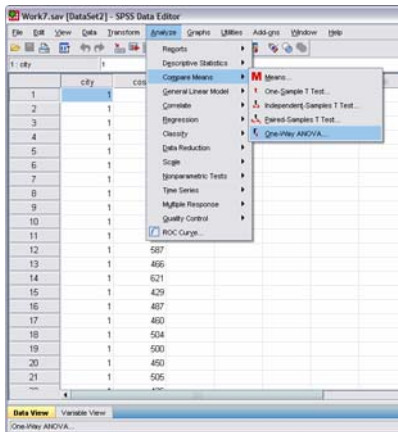
3.3 Conduct an ANOVA with post hoc analysis

The data set Work7.sav contains information on monthly household healthcare spending in four Australian cities. Random samples of 25 2-person households from each city were asked to record their healthcare spending over 6 months. (This is an independent groups design because each group were in different cities. If the same subjects were in all the same conditions a repeated measures test could be done).

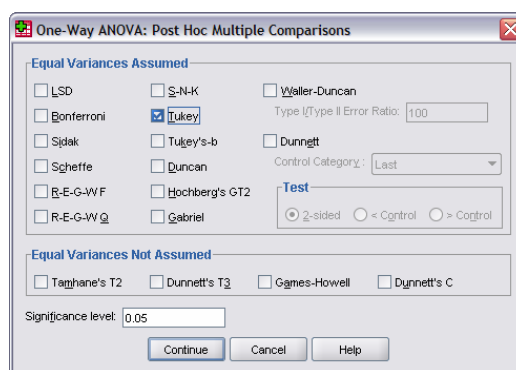
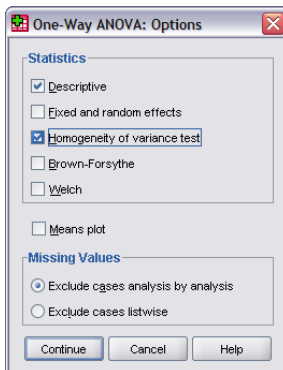


Working Example 3: First open the dataset Work7.sav

Analyze > Compare means > One-way ANOVA > add cost of healthcare to the Dependent list > add city to the factor list>



Click Options > Check Descriptive and Homogeneity of variance test > continue > Click Post Hoc > check the box for Tukey > Continue > OK



Descriptives

cost of healthcare									
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean			Minimum	Maximum
					Lower Bound	Upper Bound			
Adelaide	25	497.28	56.628	11.326	473.91	520.65		383	621
Hobart	25	515.84	56.529	11.306	492.51	539.17		397	647
Melbourne	25	531.20	63.976	12.795	504.79	557.61		397	677
Perth	25	555.12	72.576	14.515	525.16	585.08		429	739
Total	100	524.86	65.385	6.539	511.89	537.83		383	739

Test of Homogeneity of Variances

cost of healthcare			
Levene Statistic	df1	df2	Sig.
.817	3	96	.488

To interpret the output first look at Levene's test.

What can we conclude about equal variance? Accept or reject Ho?

Ho: no significant difference between the variances of the group scores

ANOVA

cost of healthcare					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	44947.000	3	14982.333	3.802	.013
Within Groups	378299.040	96	3940.615		
Total	423246.040	99			

Next look at the F-statistic p-value. If it is less than .05 we reject the ANOVA hypothesis

Ho: all population means are equal ($\mu_1 = \mu_2 = \mu_3 \dots$)

If we reject the Ho then we can look at the Post Hoc tests (Tukey) to determine where the significant difference in population means occurs.

Between which cities is there a significant difference in monthly healthcare spending?

Post Hoc Tests

Multiple Comparisons

cost of healthcare Tukey HSD						
(i) city	(j) city	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Adelaide	Hobart	-18.560	17.755	.723	-64.98	27.86
	Melbourne	-33.920	17.755	.231	-80.34	12.50
	Perth	-57.840 [*]	17.755	.008	-104.26	-11.42
Hobart	Adelaide	18.560	17.755	.723	-27.86	64.98
	Melbourne	-15.360	17.755	.823	-61.78	31.06
	Perth	-39.280	17.755	.127	-85.70	7.14
Melbourne	Adelaide	33.920	17.755	.231	-12.50	80.34
	Hobart	15.360	17.755	.823	-31.06	61.78
	Perth	-23.920	17.755	.535	-70.34	22.50
Perth	Adelaide	57.840 [*]	17.755	.008	11.42	104.26
	Hobart	39.280	17.755	.127	-7.14	85.70
	Melbourne	23.920	17.755	.535	-22.50	70.34

*. The mean difference is significant at the 0.05 level.



Working Example 4: Open the data set fevlect.sav. It contains data from a study that looked at the lung function in respiratory patients treated in outpatient departments of three different hospitals.

- Which variable contains the respiratory measurements? _____
- Which variable identifies the three hospitals? _____

Look for any evidence that, on average, the lung function of respiratory patients differs between the three hospitals.

Ho: _____

Ha: _____

Run an ANOVA with post hoc comparison test.

- What is the Levene's p-value? _____
- What is the F p-value? _____
- Do you accept or reject the Ho ? _____
- What are the mean FEV1 levels observed in the three hospitals?

Hospital 1 _____ Hospital 2 _____ Hospital 3 _____

- Do these means appear to be different? _____
- Is a post hoc test required? _____



Week 3 Task: Working with Original Data

Please complete before Session 8 on Thursday May 1.

Continue working with your original data set.

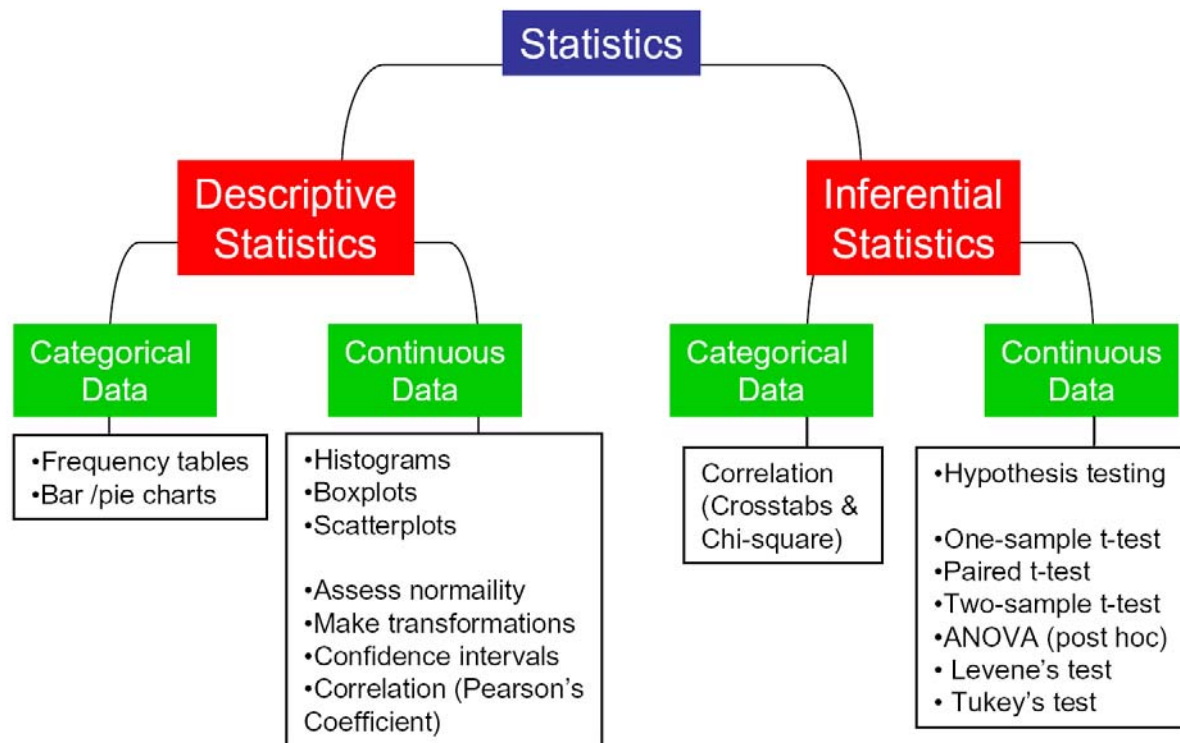
Using the skills and concepts outlined in Sessions 5-7 complete at least one of the following with your original data set.

- Create and test at least one correlation hypothesis for categorical variables (crosstab) or continuous variables (scatterplot)
- Conduct at least one t-test (one sample, paired sample, two-sample)
- Conduct an ANOVA with post hoc comparison test

Save the data your data set and your output.

At the net session (May 1, 2008), you will be asked to present some of your data analysis to the other participants.

Statistics Overview



Statistics Terms

- **Statistics** - a set of concepts, rules, and procedures that help us to:
 - **organize** numerical information in the form of tables, graphs, and charts;
 - **understand** statistical techniques underlying decisions that affect our lives and well-being; and
 - **make** informed decisions.
- **Data** - facts, observations, and information that come from investigations.
 - **Continuous data** sometimes called quantitative data -- the result of using some instrument to measure something
 - **Categorical data** also referred to as frequency or qualitative data. Things are grouped according to some common property(ies) and the number of members of the group are recorded (e.g., males/females, vehicle type).
- **Variable** - property of an object or event that can take on different values. For example, college major is a variable that takes on values like mathematics, computer science, English, psychology, etc.
 - **Discrete Variable** - a variable with a limited number of values (e.g., gender (male/female), college class (freshman/sophomore/junior/senior).
 - **Continuous Variable** - a variable that can take on many different values, in theory, *any* value between the lowest and highest points on the measurement scale.
 - **Independent Variable** - a variable that is manipulated, measured, or selected by the researcher as an antecedent condition to an observed behavior. In a hypothesized cause-and-effect relationship, the independent variable is the cause and the dependent variable is the outcome or effect.
 - **Dependent Variable** - a variable that is not under the experimenter's control -- the data. It is the variable that is observed and measured in response to the independent variable.
 - **Qualitative Variable** - a variable based on categorical data.
 - **Quantitative Variable** - a variable based on continuous data.
- **Graphs** - visual display of data used to present frequency distributions so that the shape of the distribution can easily be seen.
 - **Bar graph** - a form of graph that uses bars separated by an arbitrary amount of space to represent how often elements within a category occur. The higher the bar, the higher the frequency of occurrence. The underlying measurement scale is discrete (nominal or ordinal-scale data), not continuous.
 - **Histogram** - a form of a bar graph used with interval or ratio-scaled data. Unlike the bar graph, bars in a histogram touch with the width of the bars defined by the upper and lower limits of the interval. The measurement scale is continuous, so the lower limit of any one interval is also the upper limit of the previous interval.
 - **Boxplot** - a graphical representation of dispersions and extreme scores. Represented in this graphic are minimum, maximum, and quartile scores in the form of a box with "whiskers." The box includes the range of scores falling into the middle 50% of the distribution (**Inter Quartile Range** = 75th percentile - 25th percentile) and the whiskers are lines extended to the

- minimum and maximum scores in the distribution or to mathematically defined ($\pm 1.5 \times \text{IQR}$) upper and lower fences.
- **Scatterplot** - a form of graph that presents information from a bivariate distribution. In a scatterplot, each subject in an experimental study is represented by a single point in two-dimensional space. The underlying scale of measurement for both variables is continuous (measurement data). This graph is the most useful techniques for gaining insight into the relationship between two variables.
 - **Measures of Center** - Plotting data in a frequency distribution shows the general shape of the distribution and gives a general sense of how the numbers are bunched. Several statistics can be used to represent the "center" of the distribution. These statistics are commonly referred to as measures of **central tendency**.
 - **Mode** - The mode of a distribution is simply defined as the most frequent or common score in the distribution. The mode is the point or value of X that corresponds to the highest point on the distribution. If the highest frequency is shared by more than one value, the distribution is said to be **multimodal**. It is not uncommon to see distributions that are bimodal reflecting peaks in scoring at two different points in the distribution.
 - **Median** - The median is the score that divides the distribution into halves; half of the scores are above the median and half are below it when the data are arranged in numerical order. The median is also referred to as the score at the **50th percentile** in the distribution. The **median location** of N numbers can be found by the formula $(N + 1) / 2$. When N is an odd number, the formula yields a integer that represents the value in a numerically ordered distribution corresponding to the median location. (For example, in the distribution of numbers (3 1 5 4 9 9 8) the median location is $(7 + 1) / 2 = 4$. When applied to the ordered distribution (1 3 4 5 8 9 9), the value 5 is the median, three scores are above 5 and three are below 5. If there were only 6 values (1 3 4 5 8 9), the median location is $(6 + 1) / 2 = 3.5$. In this case the median is half-way between the 3rd and 4th scores (4 and 5) or 4.5.
 - **Mean** - The mean is the most common measure of central tendency and the one that can be mathematically manipulated. It is defined as the average of a distribution is equal to the $\sum X / N$. Simply, the mean is computed by summing all the scores in the distribution ($\sum X$) and dividing that sum by the total number of scores (N). The mean is the balance point in a distribution such that if you subtract each value in the distribution from the mean and sum all of these **deviation scores**, the result will be zero.
 - **Measures of Spread** - Although the average value in a distribution is informative about how scores are centered in the distribution, the mean, median, and mode lack context for interpreting those statistics. Measures of **variability** provide information about the degree to which individual scores are clustered about or deviate from the average value in a distribution.
 - **Range** - The simplest measure of variability to compute and understand is the range. The range is the difference between the highest and lowest score in a distribution. Although it is easy to compute, it is not often used as the sole measure of variability due to its instability. Because it is

based solely on the most extreme scores in the distribution and does not fully reflect the pattern of variation within a distribution, the range is a very limited measure of variability.

- **Interquartile Range (IQR)** - Provides a measure of the spread of the middle 50% of the scores. The IQR is defined as the 75th percentile - the 25th percentile. The interquartile range plays an important role in the graphical method known as the **boxplot**. The advantage of using the IQR is that it is easy to compute and extreme scores in the distribution have much less impact but its strength is also a weakness in that it suffers as a measure of variability because it discards too much data. Researchers want to study variability while eliminating scores that are likely to be accidents. The boxplot allows for this distinction and is an important tool for exploring data.
- **Variance** - The variance is a measure based on the deviations of individual scores from the mean. As noted in the definition of the mean, however, simply summing the deviations will result in a value of 0. To get around this problem the variance is based on squared deviations of scores about the mean. When the deviations are squared, the rank order and relative distance of scores in the distribution is preserved while negative values are eliminated. Then to control for the number of subjects in the distribution, the sum of the squared deviations, $\sum(X - \bar{X})^2$, is divided by N (population) or by $N - 1$ (sample). The result is the average of the sum of the squared deviations and it is called the variance.
- **Standard deviation** - The standard deviation (s or σ) is defined as the **positive square root** of the variance. The variance is a measure in squared units and has little meaning with respect to the data. Thus, the standard deviation is a measure of variability expressed in the same units as the data. The standard deviation is very much like a mean or an "average" of these deviations. In a normal (symmetric and mound-shaped) distribution, about two-thirds of the scores fall between $+1$ and -1 standard deviations from the mean and the standard deviation is approximately $1/4$ of the range in small samples ($N < 30$) and $1/5$ to $1/6$ of the range in large samples ($N > 100$).
- **Measures of Shape** - For distributions summarizing data from continuous measurement scales, statistics can be used to describe how the distribution rises and drops.
 - **Symmetric** - Distributions that have the same shape on both sides of the center are called symmetric. A symmetric distribution with only one peak is referred to as a **normal distribution**.
 - **Skewness** - Refers to the degree of asymmetry in a distribution. Asymmetry often reflects extreme scores in a distribution.
 - **Positively skewed** - A distribution is positively skewed when it has a tail extending out to the right (larger numbers). When a distribution is positively skewed, the mean is greater than the median reflecting the fact that the mean is sensitive to each score in the distribution and is subject to large shifts when the sample is small and contains extreme scores.
 - **Negatively skewed** - A negatively skewed distribution has an extended tail pointing to the left (smaller numbers) and reflects

bunching of numbers in the upper part of the distribution with fewer scores at the lower end of the measurement scale.

- **Kurtosis** - Like skewness, kurtosis has a specific mathematical definition, but generally it refers to how scores are concentrated in the center of the distribution, the upper and lower tails (ends), and the shoulders (between the center and tails) of a distribution.
 - **Mesokurtic** - A **normal distribution** is called mesokurtic. The tails of a mesokurtic distribution are neither too thin or too thick, and there are neither too many or too few scores in the center of the distribution.
 - **Platykurtic** - Starting with a mesokurtic distribution and moving scores from both the center and tails into the shoulders, the distribution flattens out and is referred to as platykurtic.
 - **Leptokurtic** - If you move scores from shoulders of a mesokurtic distribution into the center and tails of a distribution, the result is a peaked distribution with thick tails. This shape is referred to as leptokurtic.