PARTICIPANT WORKBOOK

Managing Data

Created: 2013

Managing Data. Atlanta, GA: Centers for Disease Control and Prevention (CDC), 2013.

Managing Data

Introduction

Learning Objectives

At the end of the training, you will be able to:

1. Create a data dictionary that includes, at a minimum:
    a. Variable names
    b. Variable descriptions or labels
    c. Variable types
    d. Response options and allowable values

2. Clean the data
    a. Identify errors, including duplications, missing data, miscodes, and outliers
    b. Use statistical software to identify and correct errors

Estimated Completion Time

The workbook should take between 6 and 7 hours to complete.

Target Audience

The workbook is designed for FETP fellows who specialize in NCDs; however, you can also complete the module if you are working in infectious disease.

Pre-work and Prerequisites

Before participating in this training module, you must complete training in:

- Basic epidemiology and surveillance;
- Statistical software program your FETP is using (e.g., SPSS, Epi Info).

About this Workbook and the Activity Workbook

You will read information about creating a data dictionary and cleaning data in the **Participant Workbook.** To practice the skills and knowledge learned you will complete three exercises. To apply what you have learned you will refer to the **Activity Workbook** and create a data dictionary and clean data for an NCD study in your country.

ICON GLOSSARY

The following icons are used in this workbook:

Image Type	Image Meaning
	**Stop** – a point at which you should consult a mentor or wait for the facilitator to provide locally relevant information about the topic
	**Activity**- an activity or exercise that you should complete
	**Light bulb** – key idea to note and remember or supplemental information

ACKNOWLEDGEMENTS

Many thanks to the following people from the Centers for Disease Control and Prevention (CDC) who contributed to this module:

- Fleetwood Loustalot, PhD, FNP, Andrea Neiman, MPH, PhD (Division for Heart Disease and Stroke Prevention), and Edward Gregg, PhD (Division of Diabetes Translation), for creating the hypertension case study.

- Indu Ahluwalia, (Senior Scientist, Division of Reproductive Health, National Centers for Chronic Disease Prevention and Health Promotion), and Richard Dicker, MD, MPH, from the Centers for Global Health, Division of Global Health Protection, for their subject matter expertise and for reviewing the training module.

Some of the content of this module was taken from a training manual developed by CDC's Division of Epidemiology and Surveillance Capacity Development: *Advanced Management and Analysis of Data Using Epi Info for Windows: Risk Factors for Sexually Transmitted Infections in Kuwadzana, Zimbabwe; 2006.*

Section 1: Overview of Data Management

DATA MANAGEMENT PRACTICES

In the *Creating an Analysis Plan* module you learned how to develop an analysis plan -- creating blank templates, or table shells, to use during data analysis. Before you begin data analysis, there are two additional tasks to complete, which you will learn in this module:

- Creating a data dictionary
- Cleaning the data

DATA DICTIONARY

If you are analyzing data that you did not collect, you must first familiarize yourself with the dataset. You will create a data dictionary, also called a codebook, to understand the meaning of the collected data. It should describe how the data are arranged in the computer file and what the various numbers and letters mean.

Whether or not you collected the data yourself, you should always use the data dictionary during data analysis so that the meaning of a variable and the coding used will never be in question. The data dictionary is the place where you will look up which codes correspond to each possible response.

CLEANING THE DATA

Every dataset contains some errors. Cleaning data is the process you will use to identify inaccurate, incomplete, or improbable data, and then correct it when possible. Data cleaning is a two-step process that includes **detection** and **correction.**

Section 2: Data Dictionary

COMPONENTS OF A DATA DICTIONARY

A data dictionary should include, at a minimum:

- Variable names;
- Variable descriptions or labels;
- Variable types; and
- Response options and codes used to represent the response options.

Some data dictionaries also include the column from the questionnaire where the variable can be found.

Variable Name

Identify each variable from the survey or questionnaire and give it a name. Use the name to identify variables in the database and during the analysis.

Ideal features of a **name**:

- Easily identifies the question on the data collection form (if one is used) or type of information collected
- Begins with a letter
- Cannot end with a period
- Can have special symbols or characters
- Should be short, with a maximum length of 64 characters
- Limits the use of symbols

Variable Description or Label

Provide a description (or label) of the variable that explains the variable name.

Variable (Field) Type

Indicate the type of variable (field). Most common types are:

- Numeric
- Text/alpha
- Date

Response Options (or Values)

Some variables use actual values. Other variables use codes that can be text or numeric. Identify the accepted response options or values for each variable. For example, it is common to use numeric coding for nominal response options, such as marital status (see example below) or codes “1”

for “yes”, “0” or “2” for “no”, and “9” or “99”for “unknown”, “don’t know”, or “missing”.

For open-ended text fields, where there are a large number of possible responses or characters of text such as the country of birth, you can indicate that the response can contain up to a certain number of characters.

The following is a basic example of the first few lines of a data dictionary. Note that the first three examples are actual values and the fourth example is coded. Often, we code nominative responses and use the actual value for numerical responses. Also notice that for the variable DOB, which represents the patient’s date of birth, a date field type is used; any date between the first of January, 1900, and the 30th of November, 2009 (the last date of the study), could be a valid value for that variable. If you had certain age restrictions for your study, you could alter the allowable values so that only dates that match those restrictions could be valid values.

Variable Name	Description	Field type	Response Options[1]
IDnum	ID number	Numeric	0001 – 9999
DOB	Date of birth	Date (dd/mm/yyyy)	01/01/1900 – 30/11/2009
CntryBth	Country of birth	Text	{up to 60 characters of text}
MarStat	Marital status	Numeric	1 = Single 2 = Married 3 = Divorced 4 = Widowed 99 = Missing

Some variables in the data dictionary are derived variables. These variables are created from other variables to enable more detailed analysis

[1] *Some data dictionaries also include a fifth column for **measures** to record the type of variables (scale, nominal, or ordinal).*

of the data. For example, a BMI variable can be created using the weight and height variables.

Stop

Let the facilitator or mentor know you are ready for the group discussion.

KEY POINTS TO REMEMBER

Use the space below to record any key points from the facilitator-led discussion:

Activity

Practice Exercise #1 (Estimated Time: 30 minutes)

Background:

For this exercise you will work individually, in pairs, or in a small group to create a data dictionary based on the information provided in Figure 1, a handout provided by your facilitator, and the case study and questionnaire from the *Creating an Analysis Plan* module.

Instructions:

1. Read the information in Figure 1.
2. Review the handout (questionnaire and blood pressure and height/weight measurements).
3. Use the table on the following page to create a data dictionary for **questions # 13 – 18**. (The first 12 lines have already been entered.)

Note: Selections included in the 'Measure' column in the data dictionary below are reflective of choices available in SPSS (i.e., Nominal, Ordinal, and Scale). Additional columns may be used if the data dictionary is created in a different format.

Ask your facilitator to review your work.

Figure 1: Hypertension case study

The Panel Members from the Ministry of Health have posed three questions that need to be addressed. A recent survey was conducted in your country that may provide the data to support your responses. After reviewing the information about the survey (e.g., sample, methodology), you need to review the questions in the survey that could be used for analysis. Common demographic and descriptive questions and measurements are frequently included in surveys. In addition, questions and measurements about the primary outcome of interest (i.e., hypertension) would need to be reviewed.

#	Variable Name	Type	Label	Value	Measure
1	SEQN	Numeric	Identification number		Scale
2	AGEY	Numeric	Age in years	555 = Missing 777 = Refused 999 = Don't Know	Scale
3	SEX	Numeric	Gender	1 = Female 2 = Male 6 = Missing 77 = Refused 99 = Don't Know	Nominal
4	EDU	Numeric	Educational level	1 = Less than High School 2 = High School Graduate 3 = Some College 4 = College Graduate 5 = <25 years of age 6 = Missing 77 = Refused 99 = Don't Know	Nominal
5	ETH5C	Numeric	Racial/ethnic group	1 = non-Hispanic white	Nominal

				2 = non-Hispanic black 3 = Hispanic 4 = Other 6 = Missing 77 = Refused 99 = Don't Know	
6	HCP	Numeric	Current health care provider	1 = Yes, one 2 = Yes, more than one 3 = No 6 = Missing 77 = Don't Know 99 = Refused	Nominal
7	BPTOLD1	Numeric	Ever told you had high blood pressure	1 = Yes 2 = No 6 = Missing 77 = Refused 99 = Don't Know	Nominal
8	BPTOLD2	Numeric	Told you had high blood pressure on 2+ occasions	1 = Yes 2 = No 6 = Missing 77 = Refused 99 = Don't Know	Nominal
9	BPMED	Numeric	Taking blood pressure medication	1 = Yes 2 = No 6 = Missing	Nominal

				77 = Refused 99 = Don't Know	
10	BPSYS	Numeric	Systolic blood pressure	6666 = Missing 9999 = Refused	Scale
11	BPDIA	Numeric	Diastolic blood pressure	6666 = Missing 9999 = Refused	Scale
12	BPHI	Numeric	BP >=140/90	1 = Yes 2 = No 6 = Missing	Nominal
13					
14					
15					
16					
17					
18					

Activity

Take out the activity workbook and complete skill assessment #1.
Then continue reading in the participant workbook.

Section 3: Cleaning Data

Overview

Few datasets are free of errors and missing values. It is important to review the dataset to identify errors before beginning analysis. When you find errors, correct them in the dataset and document the changes made. Failure to correct errors may result in false analysis results and invalid conclusions. Even after you clean the data you may find additional errors during analysis. You will also correct (and document) those errors.

When you clean data, look at the distribution (frequency of values) for each variable to:

- Assess for accurate and consistent data entered
- Check for completeness of data or missing values
- Determine whether to create or collapse data categories

If more than one person has collected or entered the data you should familiarize yourself with these aspects of the data before analyzing the data.

Documenting Errors and Changes to the Dataset

A key principle of data management is to write down everything, such as:

- Changes to the dataset
- Decisions about how to assess certain fields

This documentation will ensure that you make consistent decisions and will provide a reference for those who may have questions about your analysis

Tip

It is difficult to remember to return to a problem, so fix an error as soon as it occurs (or you become aware of it).

This can be particularly important when you are entering data from the field.

Use a table such as the one below to document changes made to the dataset for missing, miscoded, and out-of-range values:

Variable name	Format	Problem	Record IDs affected and resolution

The following is an example of how you can use this table during data cleaning:

Variable name	Format	Problem	Record IDs affected and resolution
DOB	dd/mm/yyyy	Missing data	0012 – not collected; action: none 0075-unclear writing; action: left blank 0103-not entered; action: entered from questionnaire
DOB	dd/mm/yyyy	Improbable value	0024-month=15;data was mis-entered; action: corrected in database

Use a table such as the one below to document changes made to the dataset for duplicate records:

Number	Primary ID	Secondary ID	Problem	Record IDs affected and resolution
1	SSN		duplicate	3125 - removed
2	Participant ID number		duplicate	3241 – removed
3	Missing	Social Security Number (SSN)	duplicate	3278 - removed

Tip

Because these edits permanently change the dataset it is important to make a "working copy" of the original dataset. Make changes to the working copy only. If you make any mistakes, you can access the original database and start over.

Common Sources and Types of Errors in Epidemiologic Data

Some of the most common errors occur during the data collection phase. Other sources of error are measurement errors, improper functioning of measurement equipment, and interviewer mistakes (often due to inadequate training). The respondents may also cause errors if they provide the incorrect response. This can occur if they incorrectly read or interpret a question, or if they intentionally provide a false answer. Errors can also occur after the data have been collected, most often by data entry mistakes or coding errors.

Common types of data errors are entering duplicate information, miscoding, assignment of missing values, and inclusion of out-of-range values.

Detecting and Correcting Duplicate Records

Duplicate records can occur for many reasons:

- Data entry errors in which the same case is accidentally entered more than once.
- Multiple cases share a common primary ID value but have different secondary ID values, such as family members who live in the same house.
- Multiple cases represent the same case but with different values for variables other than those that identify the case. For example, the same person or company makes multiple purchases of different products or at different times.

Duplications can arise in several ways; most often it happens when the same person's information is entered into the same database more than once. Less often a person might accidentally be enrolled in the study twice, or they may be asked to complete the same interview or questionnaire twice.

Good record keeping and study organization will usually prevent these types of errors from occurring. Sometimes duplications can arise when multiple databases are merged together into one.

To check for duplicate records:

1. Identify how many records are in the database. Use your statistical software to check the **record count.**
2. Determine if the number of records matches the number of questionnaires.
3. If the number of records is more than the number of questionnaires, run a **frequency** listing to look for multiple records with the same identifying information (such as ID number or name).
4. If there are two records with the same ID number or name, select the records and examine them to determine if they are identical (a duplicate record) or whether an ID number or name was entered incorrectly.

Deleting Duplicate Records

When you have identified the correct record, remove the inaccurate or incomplete duplicate record from the dataset so that it will not be a problem in your analysis.

Stop

Let the facilitator or mentor know you are ready for the group discussion.

KEY POINTS TO REMEMBER

Use the space below to record any key points from the facilitator-led discussion:

Activity

Practice Exercise #2 (Estimated time: 30 minutes)

Background:

For this exercise you will practice checking for duplicate records for the hypertension case study.

Instructions:

1. Check for and identify duplicate records.
2. Correct any errors and document in the table below.
3. Let your facilitator know when you have completed the exercise.

Figure 2:

A cross-sectional survey was conducted during 2009–2010 among adults aged 18 years and older (N = 993). The survey was conducted in all provinces (or regions) in Country X. The purpose of the survey was to provide estimates of the current health of the adults in the country as well as health conditions in the country. Assessments of high blood pressure were included in the survey. Data were collected by trained interviewers in the homes of participants using paper and pencil questionnaires and measurements. Responses were checked for completion and entered into a database manually. The dataset needs to be reviewed for errors.

Use the table below to document the problem and resolution:

Number	Primary ID	Secondary ID[2]	Problem	Record IDs affected and resolution

DETECTING AND CORRECTING MISSING, MISCODED, AND OUT-OF RANGE VALUES

Very few datasets are 100% complete or accurate. Usually there are a few missing, miscoded or out-of-range values. Since several people can collect and record survey data, errors can occur during the data collection stage, in recoding forms, or during data entry. Values can be coded incorrectly or data from one variable column can be mistakenly entered under an adjacent column. It can be easy to detect errors in data entry or column shifts if you have the completed survey forms available during data cleaning.

Line List

ID	state	sex	age	Education	Marital Status	Employment	Mo_Income	Children	Hlth_
75	Virginia	female	39	Primary	Married	Homemaker	<20,000	3	
81	Virginia	male	23	Secondary	Single	Refused	<20,000	0	
91	Virginia	female	82	None	Widowed	Homemaker	20,000-60,000	0	
96	Iowa	male	34	Primary	Married	Student	20,000-60,000	5	
103	Nebraska	male	34	Primary	Single	Civil Servant	20,000-60,000	1	
104	Nebraska	male	28	Secondary	Married	Civil Servant	20,000-60,000	2	
106	Virginia			52	Secondary	Married	Student		
108	Nebraska	female	21	University	Single	Retired	<20,000	2	
110	Nebraska	female	29	Secondary	Married	Homemaker	20,000-60,000	6	

Sometimes missing data occurs randomly and sometimes it occurs in patterns. For example, you may find that many records are missing data on

[2] *Many datasets have secondary IDs. This dataset does not.*

the same variable. When there are patterns to missing data this may provide clues to why it is missing. (Please see example on the following page.)

A large amount of missing data for a single subject may indicate that the interview or questionnaire was terminated early. Or it might mean that the person was lost to follow-up. Sometimes data are intentionally missing, such as a skip pattern on a questionnaire, to avoid asking subjects irrelevant questions.

Out-of-range values, known as **outliers**, are values that fall outside the range of values of the majority of responses. You will most often detect outliers by examining descriptive statistics for each variable. This includes the minimum and maximum values, measures of central tendency, and frequency distributions, which are represented graphically by histograms.

The frequency distribution of a continuous variable such as age might include a few values that seem quite low (young) or high (old). You can sort the variable (ascending or descending) to determine whether the values are accurate (true “outliers”) or the result of miscoding.

For example, a survey of reproductive age women may have respondents who were less than 10 years of age and over 65 years of age; both could be outliers because most women in the survey would be less than 45 years old. The survey should not have included anyone above age 45.

Example of Patterns to missing data (missing high blood pressure information for persons interviewed by Interviewer #3)

ID	Interviewer #	state	sex	age	Education	Marital Status	Employment	Mo_Income	Children	Hlth_stat	HBP	Diabetes	Hrt_Dz	Asthma	SMK_CIG
173	2	Virginia	female	70	None	Married	Homemaker	20,000-60,000	0	3	Yes	No	No	No	No
6	2	Iowa	female	35	Primary	Married	Homemaker	20,000-60,000	0	3	No	No	No	No	No
2274	2	Connecticut	female	23	University	Widowed	Retired	20,000-60,000	5	3	No	No	No	No	No
1574	2	North Dakota	female	20	University	Single	Retired	20,000-60,000	4	3	No	No	No	No	No
3166	2	Virginia	female	41	University	Single	Civil Servant	>100,000	0	2	No	No	No	No	No
2000	2	North Dakota	female	46	University	Married	Homemaker	20,000-60,000	5	2	No	No	No	No	No
286	2	Alaska	female	20	Secondary	Single	Refused	20,000-60,000	2	3	No	No	No	No	Yes
719	2	North Dakota	female	23	University	Single	Civil Servant	Don't Know	2	1	No	No	No	No	No
1630	2	North Dakota	female	45	Primary	Married	Homemaker	Don't Know	1	3	No	No	Yes	No	No
130	2	Nebraska	male	26	Primary	Married	Civil Servant	<20,000	3	1	No	No	No	No	No
70	2	Virginia	female	22	University	Single	Civil Servant	20,000-60,000	0	1	No	No	No	No	No
667	2	North Dakota	female	37	University	Married	Civil Servant	20,000-60,000	2	3	No	No	No	No	No
2680	2	Virginia	female	45	Secondary	Married	Homemaker	<20,000	6	3	Yes	No	No	No	No
3255	2	Virginia	female	40	Vocational	Married	Homemaker	20,000-60,000	6	3	No	No	No	No	No
2579	2	Virginia	female	56	Primary	Married	Homemaker	<20,000	0	3	Yes	No	No	No	No
2619	3	Virginia	female	70	None	Married	Homemaker	<20,000	0	3	.	No	No	No	No
3115	3	Virginia	female	53	Primary	Married	Homemaker	20,000-60,000	1	4	.	No	No	No	No
149	3	Virginia	female	39	Secondary	Single	Homemaker	<20,000	0	2	.	No	No	No	No
2341	3	Connecticut	female	40	None	Married	Homemaker	20,000-60,000	5	4	.	No	No	No	Yes
2304	3	Connecticut	female	46	Primary	Married	Homemaker	20,000-60,000	3	3	.	No	No	No	No
2260	3	Connecticut	female	29	Primary	Married	Homemaker	Refused	7	1	.	No	No	No	No
62	3	Iowa	female	24	Secondary	Married	Homemaker	20,000-60,000	1	2	.	No	No	No	No
2441	3	Oklahoma	female	35	Primary	Married	Homemaker	20,000-60,000	5	2	.	No	No	No	No
2435	3	Oklahoma	female	26	None	Married	Homemaker	20,000-60,000	0	2	.	No	No	No	No
1192	3	Tennessee	female	29	Secondary	Married	Homemaker	20,000-60,000	3	2	.	No	No	No	Yes
1199	3	Tennessee	female	30	Secondary	Married	Homemaker	65,000-99,000	0	2	.	No	No	No	No
1323	3	Tennessee	female	50	University	Married	Homemaker	20,000-60,000	4	3	.	No	No	No	No
1392	3	Tennessee	female	24	Primary	Married	Homemaker	20,000-60,000	0	1	.	No	No	No	No
1409	3	Tennessee	female	20	Secondary	Married	Homemaker	Refused	1	1	.	No	No	No	No
3496	3	North Dakota	female	34	Vocational	Married	Homemaker	20,000-60,000	4	2	.	No	No	No	No
989	4	North Dakota	female	24	University	Single	Homemaker	20,000-60,000	1	2	No	No	No	No	No
1803	4	North Dakota	female	33	Secondary	Married	Homemaker	Refused	5	3	No	Yes	No	Yes	No
1922	4	North Dakota	female	45	Primary	Married	Homemaker	Don't Know	2	4	No	No	No	No	No
760	4	North Dakota	female	33	Vocational	Married	Homemaker	65,000-99,000	3	2	No	No	No	No	No
608	4	North Dakota	female	27	Secondary	Married	Homemaker	20,000-60,000	1	1	No	No	No	No	No
572	4	North Dakota	female	20	Secondary	Single	Homemaker	65,000-99,000	3	1	No	No	No	No	No

FREQUENCIES

In addition to using a graph (or histogram) to quickly detect errors, you can examine the data in *each variable* by conducting a **frequency distribution**. The statistical software program you use will have a frequency command that allows you to select *all* variables. Review the individual variables and look for values that are out-of-range or inconsistent with other data in the record or where data are missing.

Identifying Records with Missing Values

The best way to ensure the statistical software displays missing values in your frequency distribution is to select a command to show missing values. Alternatively, you can identify variables for which you expect to have a certain number of responses and those for which you do not expect to see missing values. For example, you might expect responses to all the basic demographic questions from each person interviewed in the study. If you interviewed 2,242 people, you would expect 2,242 responses to each demographic question.

In the table below, when you look at the frequency for *sex,* you see the following results:

Sex	**Frequency**	**Percent**	**Cum Percent**	
Female	1070	47.8%	47.8%	
Male	1170	52.2%	100%	
Total	2240	100.0%	100.0%	

Note that there are only 2,240 responses rather than 2,242. This means that two records had missing values on gender.

Correcting Missing Data

As a first step, find the questionnaires that are missing a value for this variable and determine if the data were missing from the questionnaire or were not entered. Use the appropriate software commands to select all records with the variable name (e.g., *Sex*) equal to “Missing”; identify the records to review. Return to the questionnaires and determine if the correct information is available.

If missing data are *not* caused by errors in data entry, correct the error by contacting the study participant(s); however, this is often not possible.

There are other approaches to dealing with missing data as described below.

Handling Missing Values in Analysis

Some investigators use complete case analysis when the amount of missing data is small (less than 10% for each variable). This method deletes records with missing data from the analysis so that the analysis dataset include only records with complete data.

Other investigators only use the data that are available. Records with missing values for just a few variables are not included in the analyses involving those variables; they are included in analyses of variables for which values are available. For example, if a value for “sex” is missing from a record, you may choose to exclude that record for analysis of the sex variable.

Another method is **not** to analyze the variables that have a large amount of missing data. Of course, if the variable is important for your analysis, and is related to your hypothesis, then do not use this method!

For datasets with larger amounts of missing data, (more than 10% for a variable), you can use **imputation** techniques. Imputation is a method for assigning values for missing data by making statistical inferences from similar records with known values. Consult a statistician before undertaking imputation.

Identifying Records with Miscoded Values

You can avoid miscoded values if you use a data entry screen with allowable values for text variables or range checks for numeric variables. By reducing the opportunity to enter data incorrectly you will reduce the need to check for miscoded values.

Correcting Miscodes

The simplest way to correct miscodes is to first look at the original data source from the subject in question (such as the questionnaire) to determine the true value. Then make the correction in the database. **Remember to write down the changes made.**

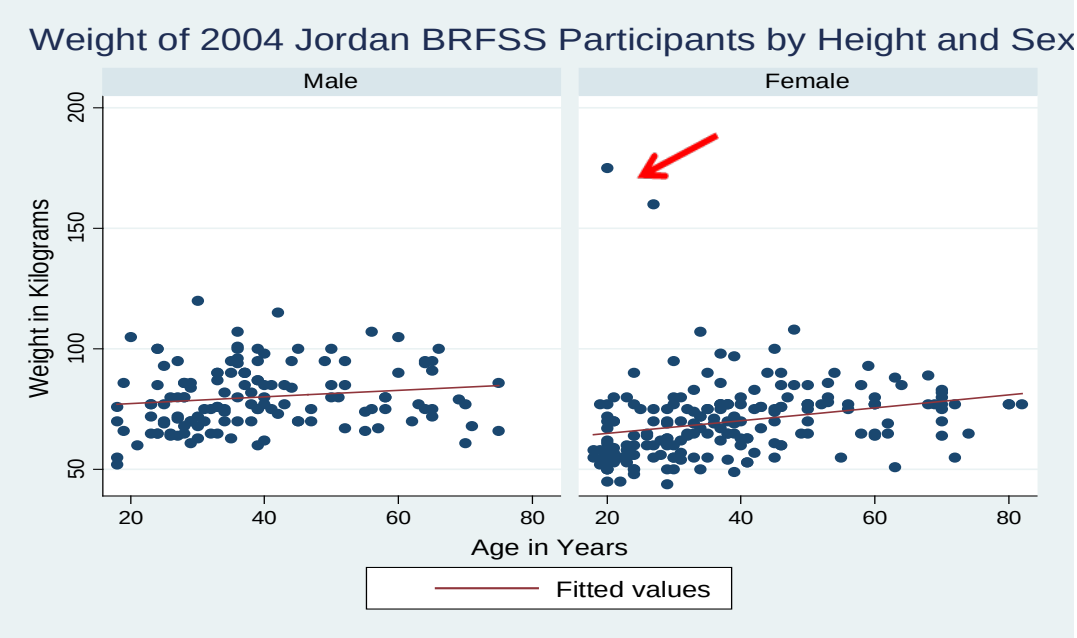
If you do not have access to the original data source, recontact the subject to confirm that the information you have is correct (or incorrect);

unfortunately this option is often not feasible. Another common approach is to recode the value as missing ("999") and deal with it the same way you managed other missing data.

Identifying Records with Out-of-Range Values

Some variables may contain values that seem out-of-range compared to the responses from the other participants in the study. These are often numerical values that may have been incorrectly coded. When you run frequencies on the variables you should notice these out-of-range values or *outliers*; however, some data errors only appear when you compare two variables. Making a **scatterplot** illustrates the value of one variable on the X axis and the value of the other variable on the Y axis. The points that stray from the bulk of the scatterplot points represent the *outliers.* Many statistical software programs have this functionality.

The scatterplot below shows the weights of the Jordan BRFSS participants by age and sex. There would appear to be two extreme weight values among the women (shown below).



Handling Outliers

To determine whether an outlier is a true outlier or an error (e.g., data entry error or miscode), look at the original data source from the subject in question (such as the questionnaire). If it is an error, make the correction in the database. **Remember to write down the changes made.**

When you have an outlier that you cannot resolve by looking at the questionnaire, decide whether to verify the data or leave it as entered. This will depend on the effort required, the importance of that particular variable, and the overall size of the dataset. For a very small dataset and a key variable, it is probably worth the effort to get it right. In another circumstance, having a "missing" value for such a variable may be acceptable.

Tip

It is never okay to change a value just because it does not seem valid.

LOGIC CHECKS

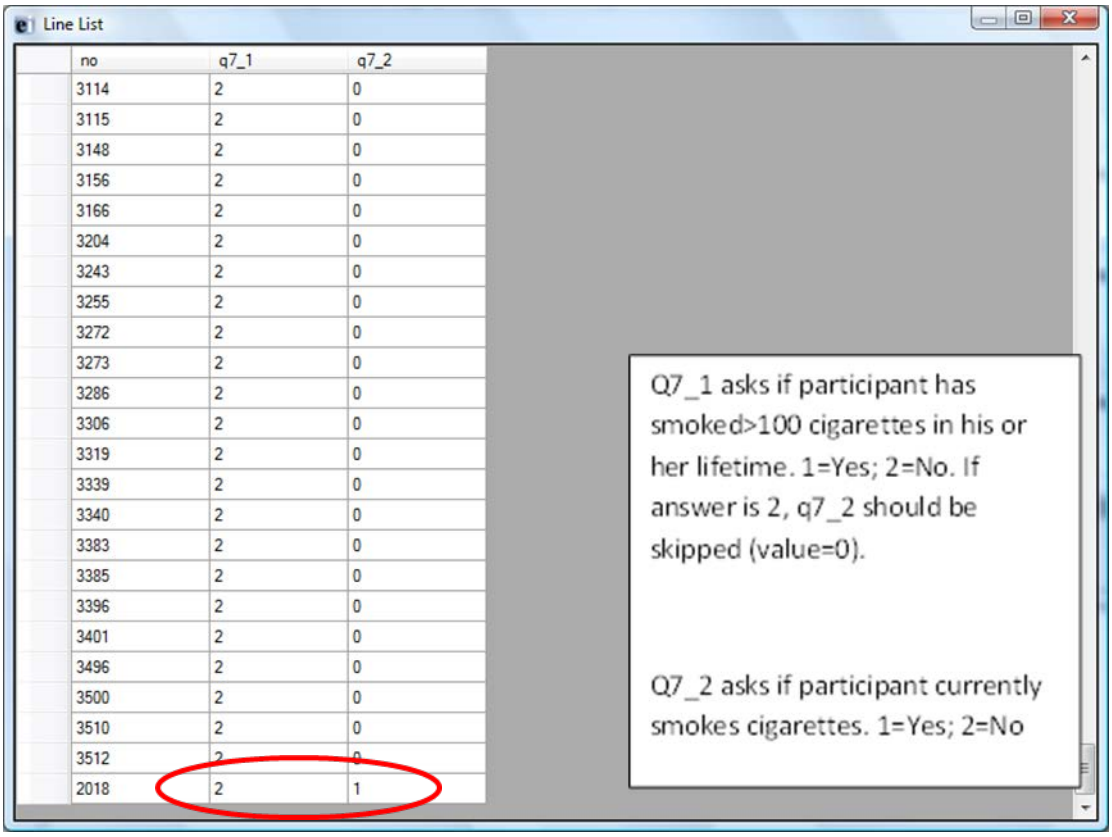
A logic check is when you compare responses of two different variables to determine if they are logical. One type of logic check looks for impossibilities (usually a typo or data misentry). An example of this is a date of discharge for a given hospital stay that is earlier than the date of admission for the same stay. Similarly, we often compare a calculated age based on date of birth to stated age in years.

Another type of logic check is looking for inconsistencies, such as comparing the hysterectomy (or prostate cancer) variable with the gender variable. For example, if a question were asked about a diagnosis of prostate cancer and the reply is marked 'yes', this would not be compatible with “sex = F”. You cannot solve this without doing additional investigation. Maybe the participant was female and the code for prostate cancer is incorrect. Or maybe the participant did have prostate cancer but the sex is miscoded.

A third type is of logic check is ensuring that skip patterns have been followed. For example, a respondent answers "never smoked cigarettes," then answers that she started smoking in 2004.

Some statistical software programs can incorporate **logic checks** so that improbable values are flagged for the investigator to examine.

For example, refer to Question 7.1 in the Jordan BRFSS questionnaire below. It asks if the participant has smoked at least 100 cigarettes in his or her lifetime. The answer choices are '1' for 'Yes' and '2' for 'No'. The next question (and several more that follow) should only be answered by respondents who answered 'Yes' to question 7.1. When you perform a logic check, you find that one respondent (# 2018) answered 'No' to being a smoker (question 7.1=2) and 'Yes' to currently smokes cigarettes (question 7.2=1). You would need to investigate this inconsistency by checking questionnaire **#2018.**



Stop

Let the facilitator or mentor know you are ready for the group discussion.

KEY POINTS TO REMEMBER

Use the space below to record any key points from the facilitator-led discussion:

Activity

Practice Exercise #3 (Estimated time: 45 minutes)

Background:

For this exercise you will detect and correct errors for the hypertension case study dataset.

Instructions:

1. Check for missing data of basic demographic data, such as age and sex, by running **frequencies**.
2. Assuming you cannot contact the study participants, describe how you will resolve the missing data in the table below (next to #1).
3. Check for miscodes by running **frequencies** and **logic checks**.
4. Correct the miscodes by referring to the questionnaire.
5. Document in the table below. (*Number 2 has already been filled in.*)
6. Make a scatterplot to identify any out-of-range values (outliers).
7. Correct the outliers by referring to the questionnaire.
8. Document in the table below.

Number	Variable name	Format	Problem	Record IDs affected and resolution
1	SEX	Numeric	Miscoded (response = 3)	51929 – changed to missing
2				
3				

4				
5				
6				
7				
8				
9				

Activity

Take out the activity workbook and complete skill assessment #2.

Resources

For more information on topics found in this workbook:

Centers for Disease Control and Prevention, Division of Epidemiology and Surveillance Capacity Development. *Advanced Management and Analysis of Data Using Epi Info for Windows: Risk Factors for Sexually Transmitted Infections in Kuwadzana, Zimbabwe; 2006.*

Tulane University: Practical Analysis of Nutritional Data. Available at http://www.tulane.edu/~panda2/Analysis2/datclean/dataclean.htm#1.