TDU

INTRODUCTION TO DATA MANAGEMENT FOR APPLIED PUBLIC HEALTH

Intersession Activity:

Public Health Agency of Canada
Agence de la santé publique du Canada

Canada

Contents

SCENARIO

It is the beginning of June 2014. You are a field epidemiologist who has been mobilised to the small fictional island jurisdiction Kataan. It has been nearly one year since a devastating heat dome affected this region, when over a seven-day period; daily temperatures soared to temperatures above 40˚C. This weather event was unprecedented, Kataan being a temperate, coastal nation.

The occurrence of such extreme weather events are more probable in the future because of climate change. In order to be prepared for such events in the future, public health authorities want to know more about individuals who experienced heat-related injures and deaths to gain a clear picture of vulnerabilities. You are here to support this jurisdiction in acquiring and analysing data to provide useful information to support preparedness initiatives. It is your first day and you have received a rapidly assembled line list of 35 individuals whose deaths were determined by the local coroner's office to be due to heat. Your immediate priority is to provide and interpret a descriptive epidemiological analysis of these data, develop initial caveats for interpretation, and start developing a list of potential recommendations for data quality.

ACTIVITY INSTRUCTIONS

GENERAL CONSIDERATIONS:

Note: as you will likely be working through and editing this dataset in Excel, there are no built-in trackers for recording the steps you have taken to clean and analyse the dataset. As such, you will have to make sure you do this yourself!

1. Start by creating a new tab, and naming it "Data Cleaning"

   - This is where you will keep a record of all the steps you took to take your dataset from messy and confusing, to beautiful and easy-to understand!

   While working through cleaning and preparing your data for analysis, you may find yourself changing variable names or creating new variables. It is always a good idea to keep a description of each variable to make sure that someone else reading your work can follow it easily enough as well!

   Update the "Data Dictionary" with any changes you make to the variable names. The "Data Dictionary" tab contains a list of all variable names and their description.

Additional notes:

CLEANING THE DATASET

Note: there may be several ways of cleaning any of the following variables. Making reasonable assumptions, omissions, or interpretations are all legitimate methods of cleaning data under various circumstances. The important part is that you document any changes you make in sufficient detail so that someone else can follow along and replicate (or reverse) any of the changes that have been made should additional information become available.

Variable names

Look at the variable names included in the dataset: most are abbreviated versions of longer names. However, a few variable names remain written in their long format or contain special characters or spaces. You may wish to replace these variable names by creating abbreviated versions.

- Note that you should update these in your data dictionary – and include a brief explanation of what these terms were originally and their meaning.

Additional notes:

Duplicates

You may have noticed in the previous step that PHN 16194 has been duplicated in your dataset. Before correcting or changing the dataset, it is best to have a look to compare the duplicate entry across all variables. I.e., is this a true duplicated entry, or is there an error in the PHN record? Are there any other duplicates in the dataset? How would you check?

Are there any other duplicated entries in the dataset? Record any findings or decisions that you make.

Additional notes:

Sex vs. gender

We will not spend a lot of time discussing the differences between sex and gender data, but we will note that there certainly can be differences, and it is important to consider what data have been submitted to you, and how to account for it properly. For a more thorough discussion and reflection on sex and gender-based analysis – we recommend the TDU's course on Applied Learning on LGBTQ2S+ Epidemiology (ALLE).

For now, take a look through the 'sex' variable in our dataset and make any changes required, ensuring you record any changes and assumptions in the data cleaning tab.

Additional notes:

Spelling errors and typos

There appear to be several typos present throughout the dataset. Scan through the variables for any spelling errors, mix-ups, aberrant white-space and special characters, fixing them as you go and recording any changes in your data-cleaning tab.

Additional notes:

Date formats

Dates can be one of the trickiest and most frustrating parts of data cleaning. We need to ensure that dates are entered correctly and in a consistent format in order to be able to calculate ranges and ages for individuals entered into our line list. Take special note of date formatting in Excel as this can be particularly challenging when moving data from Excel to other programs. Note: In Excel, you can change the number formatting option using the Number button under the Home tab. You can also change stubborn cells manually.

Additional notes:

Create new variable(s) for interpretation and analysis

It is often useful to create new variables from the existing cleaned ones in order to facilitate analysis and interpretation. We suggest adding a new variable to the form to capture each individual's age at time of death.

Additional notes:

ANALYSIS AND SUMMARY

Now that your dataset is cleaned and you are happy with all the entries and variable, it's time for the fun part: analysis! There are many different ways that you could analyse this dataset. Try some of the following:

- Convert the excel line list into a nicely-formatted table
- Summarize the data using descriptive statistics (e.g., person, place, time). If you're comfortable doing so, try making pivot tables to create these summary statistics more easily.

Please see the Excel 101 resource (provided in the course materials folder) if you need a refresher on how to create a pivot table.

Additional notes:

SUMMARY, INTERPRETATION AND CAVEATS

Interpretation

(To be completed individually during the intersession activity)

1) Create a nice summary table that includes descriptive statistics (e.g., person/place/time; mean, median, range) for your cleaned dataset (see template below).

2) In writing, summarize the data in the context of the scenario provided. Think about person/place/time when coming up with your description. (Word limit = 150 words).

(To be completed in small groups, after returning to the virtual classroom)

3) Describe any caveats and limitations of your analysis (e.g., what assumptions did you make when cleaning the data? Are there any data you omitted? Are there any gaps in the dataset you were provided?) How can these caveats and limitations affect your interpretation of the situation? (e.g., What else would you need to know to provide a more complete picture of the scenario?)

4) Based on your interpretation, and given the caveats and limitations discussed above – provide a list of next steps and recommendations for improving data quality, and facilitating interpretation. What would be your recommendations for ensuring that you or your team have all the information required to make well-informed decisions?

EPI-SUMMARY TEMPLATE

Case count:	**Region**	**Sex = 1**	**Sex = 2**
	Region 1	n=##	n=##
	Region 2	n=##	n=##
	Region 3	n=##	n=##
	Region 4	n=##	n=##
	Region 5	n=##	n=##
	Total	**n=##**	**n=##**

Age		**Sex = 1**	**Sex = 2**
	Mean	n=##	n=##
	Median	n=##	n=##
	Range	##-##	##-##

Location of death		**Sex = 1**	**Sex = 2**
	Home	n=##	n=##
	Nursing Home	n=##	n=##
	LTCF*	n=##	n=##

(* Long-term care facility)

Brief Epidemiological Summary (150 words max):

Caveats and Limitations (point-form): (Note: these will be discussed together in groups)

Recommendations to improve data quality and completeness (point-form): (Note: these will be discussed together in groups)