TDU

INTRODUCTION TO DATA MANAGEMENT FOR APPLIED PUBLIC HEALTH

Intercession Activity: Answer Key

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

Canada

Contents

Scenario 3

Activity Instructions 3

General Considerations: 3

Cleaning the dataset 4

Variable names 4

Duplicates 4

Sex vs. gender 5

Spelling errors and typos 5

Date formats 6

Create new variable(s) for interpretation and analysis 7

Analysis and Summary 8

Summary, Interpretation and Caveats 9

Epi-Summary Template 10

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:

There are a few discrepancies in the naming of variables, however. We have ‘dob’ representing ‘date of birth’ and ‘dod’ representing ‘date of death’ but ‘location_of_death’ is spelled out long-form.

1. Change ‘location of death’ to ‘lod’ for consistency.

Keeping in line with removing special characters from variable names, let’s go ahead and update ‘health_region’ and ‘health_authority’.

2. Change ‘health_region’ to ‘region’.
3. Change ‘health_authority’ to ‘authority’.

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

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:

4. PHN 16194 does look like a fully duplicated variable. Delete this record from the dataset, and record the omission in your data-cleaning tab.

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), or consulting the PHAC SGBA+ Toolkit (note: this is only available on the PHAC intranet).

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:

Multiple entries are included under the variable 'sex'. It looks like the majority of these are coded as a 1 or a 2, however, we also have rare instances of '3', 'F', 'Other' and 'Male'. Update the entries in this variable to improve consistency and accuracy.

5. Let us go ahead and make the assumption that 1 = Female; 2 = Male, and 3 = Other. Note these assumptions in your data cleaning tab, then change 'F' to 1 for PHN 10224; 'Other' to 3 for PHN 1662; and 'male' to 2 for PHN 16194.

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:

6. Fix typos and spelling errors in all of the text-based variables.

- ‘Municipality’:
    - Change ‘Andria’ to ‘Andoria’ in PHN 19873
    - Change ‘bejor’ to ‘Bejor’ in PHN 19735
- ‘Region’
    - PHN 18773 – region mix-up. ‘Bejor’ should be part of the ‘Great Barrier’ health region, not ‘Neutral Zone’.
    - PHN 19107 – region mix-up. ‘Kronos’ should be part of the ‘Neutral Zone’ health region, not ‘Great Barrier’.
    - PHN 10518 – The health region includes an underscore that isn’t present in the other examples. Remove underscore in favour of a space.
- ‘Health authority’
    - PHN 18773 – ‘Neutral Zone’ should be in the ‘Beta’ health authority
    - PHN 10981 – ‘Nekrit Expanse’ should be in the ‘Delta’ health authority
    - PHN 19107 – Provided you’ve also fixed the health region for this entry, the health authority should be ‘Beta’
    - PHN 11986 – Health authority should be ‘Beta’, not ‘Bta’
- ‘lod’
    - Several entries for Home have been entered inconsistently. Change the following entries to ‘Home’ while being mindful of removing unnecessary white space.
        - PHN 19107
        - PHN 19735
        - PHN 13412
        - PHN 11761
        - PHN 18773
        - PHN 13531
        - PHN 20725

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:

7. Fix issues with dates in both the ‘dob’ variable and the ‘dod’ variable.

- dob
  - Apply the date format to this field by highlighting and selecting ‘Short Date’ from either the Excel number-format drop-down in the ribbon of the Home tab, or by highlighting the entries, clicking Ctrl-1 on your keyboard, and selecting Date from the number formatting menu.
- dod
  - There are several issues with the entries in the ‘dod’ variable. Many of the dates do not make sense for the scenario provided. The heat-dome event occurred in 2013 so all the dod entries should reflect the same year. Update the year in the following entries to 2013.
    - PHN 11986
    - PHN 14437
    - PHN 16662
    - PHN 16243
    - PHN 2563
    - PHN 5435
  - Several dates are in the wrong format. Correct the formats to ISO standard for the following entries:
    - PHN 2563
    - PHN 16662
    - PHN 5435
    - PHN 14437

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:

8. Calculate the age of each individual by subtracting their date of birth (‘dob’) from their date of death (‘dod’), then dividing the result by 365.25 and rounding down to the nearest year. Save this variable as ‘Age’.

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:

9. Convert the linelist into a nicely-formatted table
10. If comfortable doing so, construct a pivot table to summarize the data using descriptive statistics (e.g., person, place, time).
11. Summarize the dataset using the template provided below.

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
	Badlands	n=1	n=3
	Delphic Expanse	n=2	n=0
	Great Barrier	n=12	n=6
	Nekrit Expanse	n=4	n=0
	Neutral Zone	n=1	n=6
	Total	**n=20**	**n=15**

Age		Sex = 1	Sex = 2
	Mean	76	77
	Median	75	71
	Range	64-98	57-103

Location of death		Sex = 1	Sex = 2
	Home	n=18	n=14
	Nursing Home	n=1	n=0
	LTCF*	n=0	n=1
	Outside	n=1	n=0

(* Long-term care facility)

Brief Epidemiological Summary (150 words max):

[To debrief together in small groups]

Caveats and Limitations (point-form):

[To debrief together in small groups]

Recommendations to improve data quality and completeness (point-form):

[To debrief together in small groups]