Lab 01A: Week 2

Contents

The **specific aims** of this lab are:

- build experience in R, RStudio and generating R Markdown documents
- improve your statistical literacy
- generate discussion and provide an opportunity to practice *statistical thinking* and *communicating statistical concepts*
- practice chi-squared tests for categorical data

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO2 Extract and combine data from multiple data resources.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO8 Create a reproducible report to communicate outcomes using a programming language.

1 Quick quiz

1 Quick quiz

See the Sway section of Ed.

2 Group exercise

As a group (in breakout rooms on Zoom) discuss and brainstorm the following:

1. Select one of the data sets from below:

- what questions could you ask of the data?
- draft some visualisations you could use to answer these questions (you don't have to actually create them, just think about what you could possibly create)
- are there any properties of the data which might confound the question or make it difficult to answer?
- which of the questions you brainstormed were the hardest to answer visually?

2. Present the outcomes of your discussion to the rest of the class (your tutor might share a GoogleDoc for you to add your comments to).

2.1 Australian road fatalities

Australian road fatalities since 1989. Data sourced from the Australian Roads and Deaths Database (Bureau of Infrastructure Transport and Regional Economics 2018). You can download the data to June 2021 from here.

```
library("tidyverse") # loads readr, dplyr, ggplot2, ...
# If you download the data file first you can read it in using the
# readxl package if it's in your current working directory:
fdata = readxl::read_excel("ardd_fatalities_jun2021.xlsx", sheet = 2, skip = 4, na =
        c("","-9"), guess_max = 1e6)
```

```
glimpse(fdata)
```

```
Rows: 52,572
Columns: 23
$ `Crash ID`                      <dbl> 20214003, 20215045, 20215003…
$ State                           <chr> "SA", "WA", "WA", "NSW", "Ql…
$ Month                           <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6…
$ Year                            <dbl> 2021, 2021, 2021, 2021, 2021…
$ Dayweek                         <chr> "Sunday", "Tuesday", "Sunday…
$ Time                            <dttm> 1899-12-31 00:00:00, 1899-1…
$ `Crash Type`                    <chr> "Single", "Single", "Single"…
$ `Bus Involvement`               <chr> "No", "No", "No", "No", "No"…
$ `Heavy Rigid Truck Involvement` <chr> "No", "No", "No", "No", "No"…
$ `Articulated Truck Involvement` <chr> "No", "No", "No", "No", "No"…
$ `Speed Limit`                   <chr> "110", "110", "110", "100", …
$ `Road User`                     <chr> "Passenger", "Driver", "Driv…
$ Gender                          <chr> "Male", "Male", "Male", "Mal…
$ Age                             <dbl> 24, 32, 27, 23, 21, 19, 19, …
$ `National Remoteness Areas`     <chr> "Remote Australia", NA, NA, …
$ `SA4 Name 2016`                 <chr> "South Australia - South Eas…
$ `National LGA Name 2017`        <chr> "Kangaroo Island (DC)", NA, …
$ `National Road Type`            <chr> "Sub-arterial Road", NA, NA,…
$ `Christmas Period`              <chr> "No", "No", "No", "No", "No"…
$ `Easter Period`                 <chr> "No", "No", "No", "No", "No"…
$ `Age Group`                     <chr> "17_to_25", "26_to_39", "26_…
$ `Day of week`                   <chr> "Weekend", "Weekday", "Weeke…
$ `Time of day`                   <chr> "Night", "Night", "Night", "…
```

Open the above Excel file and try to work out what each of the parameters are doing, e.g. `sheet = 2`, `skip = 4`, `na = c("","-9")` and `guess_max = 1e6`. The help for the `read_excel()` may also be usefuk `?read_excel`. Also the data dictionary linked from the Australian Roads and Deaths Database homepage. Check what happens if you don't include these parameters, do you get warning messages?

2.2 Cereal

This data is taken from the The Data and Story Library. Missing values are identified as those with a value of `-1`.

```r
path = "https://github.com/DATA2002/data/raw/master/Cereal.csv"
cereal = readr::read_csv(path, na = "-1")
glimpse(cereal)
```

```
Rows: 77
Columns: 16
$ name     <chr> "100%_Bran", "100%_Natural_Bran", "All-Bran", "All-…
$ mfr      <chr> "N", "Q", "K", "K", "R", "G", "K", "G", "R", "P", "…
$ type     <chr> "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "…
$ calories <dbl> 70, 120, 70, 50, 110, 110, 110, 130, 90, 90, 120, 1…
$ protein  <dbl> 4, 3, 4, 4, 2, 2, 2, 3, 2, 3, 1, 6, 1, 3, 1, 2, 2, …
$ fat      <dbl> 1, 5, 1, 0, 2, 2, 0, 2, 1, 0, 2, 2, 3, 2, 1, 0, 0, …
$ sodium   <dbl> 130, 15, 260, 140, 200, 180, 125, 210, 200, 210, 22…
$ fiber    <dbl> 10.0, 2.0, 9.0, 14.0, 1.0, 1.5, 1.0, 2.0, 4.0, 5.0,…
$ carbo    <dbl> 5.0, 8.0, 7.0, 8.0, 14.0, 10.5, 11.0, 18.0, 15.0, 1…
$ sugars   <dbl> 6, 8, 5, 0, 8, 10, 14, 8, 6, 5, 12, 1, 9, 7, 13, 3,…
$ potass   <dbl> 280, 135, 320, 330, NA, 70, 30, 100, 125, 190, 35, …
$ vitamins <dbl> 25, 0, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, …
$ shelf    <dbl> 3, 3, 3, 3, 3, 1, 2, 3, 1, 3, 2, 1, 2, 3, 2, 1, 1, …
$ weight   <dbl> 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.33, 1.0…
$ cups     <dbl> 0.33, 1.00, 0.33, 0.50, 0.75, 0.75, 1.00, 0.75, 0.6…
$ rating   <dbl> 68.40297, 33.98368, 59.42551, 93.70491, 34.38484, 2…
```

3 Projects in RStudio

One of the hallmarks of good science is reproducibility. R Markdown documents help this goal, but it's not so helpful if you can't find the file you just created, or the data set it refers to. It's vital to have an appropriate folder structure on your computer to keep your various analyses in. The structure below is a suggestion:

```
DATA2002/
  |- Labs/
    |- Lab00/
      |- Lab00.rmd
      |- Lab00.html
    |- Lab1a/
      |- Lab1a.rproj
      |- Lab1a.rmd
      |- Lab1a.html
      |- data/
        |- FILE_NAME.csv
  |- Assignment/
    |- Assignment.rproj
    |- Assignment.rmd
    |- Assignment.html
```

Some key elements:

- There are `.rproj` files in some folders, these store information about "RStudio projects." `File > New Project` will let you create a new project. I recommend creating a new project for each lab and for each module report.
- Once you're in a project, the working directory is wherever that `.rproj` file is stored and you can refer to files relative to that working directory.
- You can switch between projects using the top right drop down menu in the RStudio interface. You can also have multiple projects open at the same time - for example you could have the `Lab1a.rproj` project open in one RStudio window and the `Assignment.rproj` open in another window. When you do this the two instances of RStudio don't know about each other, i.e. an object available in one RStudio window is not accessible in the other RStudio window.

You can find out more about R projects here.

4 Exercises

4.1 Tablet devices

Tablet devices are an increasingly important component of the global electronics market. According to a market intelligence research company, the use of tablet devices can be classified into the following user segments.

User Segment	2012 percentages	Current survey frequency
Business-Professional	69%	102
Goverment	21%	32
Education	7%	12
Home	3%	4
Total	100%	150

Do the data provide sufficient evidence to indicate that the figures obtained in the current survey agree with the percentages in 2012?

Some R code to help with the calculations:

```
y_i = c(102, 32, 12, 4)
p_i = c(0.69, 0.21, 0.07, 0.03)
n = sum(y_i)
e_i = n * p_i
```

4.2 Smoking rates

A study of patients with insulin-dependent diabetes was conducted to investigate the effects of cigarette smoking on renal and retinal complications. Before examining the results of the study, a researcher expects that the proportions of four different subgroups are as follow:

Subgroup	Proportion
Nonsmokers	0.50
Current Smokers	0.20
Tobacco Chewers	0.10
Ex-smokers	0.20

Of 100 randomly selected patients, there are 44 nonsmokers, 24 current smokers, 13 tobacco chewers and 19 ex-smokers. Should the researcher revise his estimates? Use 0.01 as the level of significance.

```
y_i = c(  ,  ,  ,  )
p_i = c(0.5, 0.2, 0.1, 0.2)
n = sum(  )
```

```
e_i = n * p_i
```

5 Australian road fatalities

Answer the following questions about the Australian road fatalities data.

1. How are missing values recorded, and why might they occur?

2. How many fatalities occurred since 1989? How many fatal crashes have there been since 1989?

3. What is the most common hour of the day for a fatal crash?

4. What is the most common day of the week for a fatal crash?

5. What is the most common month for a fatal crash?

6. Are fatal crashes uniformly distributed across the months of the year? Filter the data down to one year (e.g. 2019) to do this test. You should write out a full hypothesis test and make an appropriate conclusion.

Step 1: Create a new R Project (e.g. call it Lab01). Step 2: download the fatalities to June 2021 Excel file and save it into the R project folder. Step 3: Open a new R Markdown file.

To get things moving, here's some code that imports the data and makes all the required edits at once. Your tutor will help explain the steps.

```
# fatalities data
fdata = readxl::read_excel("ardd_fatalities_jun2021.xlsx",
                           sheet = 2,
                           skip = 4,
                           na = c("","-9"),
                           guess_max = 1e6) %>%
  janitor::clean_names()

# crash data
cdata = fdata %>%
  select(-road_user, -gender, -age, -age_group) %>%
  distinct() %>%
  group_by(crash_id) %>%
  slice(1) %>%
  ungroup() %>%
  mutate(hour = lubridate::hour(time))
```

You might want to investigate the **lubridate** package which provides a range of functions for working with dates and times.

As you're working through this question, spend some time making the output in your compiled HTML file "presentable," i.e. remove any spurious output (messages or warnings) using the chunk options, turn on code folding, include sufficient commentary as text such that you don't need to read the code to know what is going on, if you generate plots, edit the axis labels so that they are meaningful (i.e. not just the raw variable name).

6 For after the lab

6.1 Recap

R functions used:

- `chisq.test()` for performing a chi-square test
- `readr::read_csv()` for importing data from a CSV file
- `dplyr::glimpse()` glimpse a data frame
- `janitor::clean_names()` for cleaning the column names
- `dplyr::select()` select or remove columns from a data frame
- `dplyr::distinct()` keep only unique rows in a data frame
- `dplyr::group_by()` group a data frame by one or more variables and `dplyr::ungroup()` undoes the grouping operation
- `dplyr::slice()` slice rows out of a data frame
- `dplyr::mutate()` create or overwrite columns in a data frame
- `lubridate::hm()` takes a character vector that looks like `"hh:mm"` and converts it into a time vector
- `lubridate::hour()` takes a time vector and extracts just the hour component

6.2 Pollution

To deal with the water pollution problem, three proposals are suggested:

1. Remove the industrial plant
2. Relocate the industrial plant to the river mouth; and
3. Build a sewage plant.

Thirty government officials are interviewed and their opinions are given below. Test, at the 5% level of significance, the null hypothesis that there is no preference among the three proposals. To facilitate the working, you may complete the following table:

Proposal	Observed y_i	Expected e_i	$(y_i - e_i)$	$\frac{(y_i-e_i)^2}{e_i}$
Remove the plant	6			
Relocate the plant to river mouth	9			
Build a sewage plant	15			
Total	30			$t_0 =$

```
y_i = c(6, 9, 15)
p_i = c(  ,   ,   )
n = sum(  )
e_i = n * p_i
```

7 Data dictionary

7.1 Cereal

Nutritional information for breakfast cereals ($n = 77$)

- `name` Name of cereal
- `mfr` Manufacturer
    - 1 = American Home Foods
    - 2 = General Mills
    - 3 = Kellogg's
    - 4 = Nabisco
    - 5 = Post
    - 6 = Quaker Oats
    - 7 = Ralston Purina
- `type` Type of cereal
    - 1 = cold
    - 2 = hot
- `calories` Calories per serving
- `protein` Protein grams
- `fat` Fat grams

- `sodium` Sodium millimeters
- `fiber` Fiber
- `carbo` Carbohydrates
- `sugar` Sugar
- `Potass` Potassium
- `vitamin` Vitamins
- `shelf` Shelf position in store
    - 1 = bottom
    - 2 = middle
    - 3 = top
- `weight` Weight (grams)
- `cups` Cups in serving
- `rating` Taste rating

References

Bureau of Infrastructure Transport and Regional Economics. 2018. "Australian Road Deaths Database." https://bitre.gov.au/statistics/safety/fatal_road_crash_database.aspx.