Week 2 RHMI Exercise Answers
Andrew Perfors

Day 2

1. What is your favourite colour?
2. How tall are you in cm?
3. Rank the following from best to worst: (a) bunnies; (b) bears; (c) doggies

Measurement

1. For Bunny's survey question about colour, what is the construct, the measure, and the observation?

   The construct is the psychological concept she is trying to get at, which presumably is something like the internal feeling of having a particular favourite colour. The measure is the question she asked her participants, i.e., survey question #1. And the observation is what they said.

2. For each of Bunny's survey questions, say what kind of variable it yields (e.g., nominal, etc).

   #1 yields nominal, categorical discrete data.
   #2 yields numeric, non-categorical, continuous data (presuming one permits fractions)
   #3 yields numeric, non-categorical, ordinal, discrete data.

3. How do you think reliability of these survey questions might differ between themselves? Consider each of the kinds of reliability, and say how Bunny might go about measuring them.

   Test-retest: this is if you'd get the same answer when asking the exact same questions at different times. #1 and #3 might have the lowest of this, as long as you asked the questions soon after each other, if people don't have strong opinions and forget what they said the first time. #2 would probably be very high if you tested and retested close in time but very low if the test and retest were separated by years, because the participants might have grown!

   Inter-rater: this is if different people asking the exact same question would get the same answers. This is probably pretty high for the first two questions but might be lower depending on whether a bunny, a bear, or a doggie is asking!

   Internal consistency: this is if you'd get the same answers if you tried to measure in slightly different ways. So variants on #1 might include "of all of the colours, what do you like best?" And variants of #3 might put the three in different orders in the question. One would hope that these changes wouldn't change the answers but people are complicated; you can't be sure without trying!

Markdown

1. Make the following changes to gladlysurvey_modified.Rmd:

   Change the title to "My first R Markdown document" and the name to yours.

```
---
title: "My first Markdown Document"
author: "Andy"
date: "07/03/2022"
---
```

2. Add a line of text that says “Here is my new stuff! I *love* it.” and then follow it by a chunk that creates a vector called hobbies that contains three of your favourite hobbies. Then use the `print()` function to print hobbies.

````
18
19  Here is my new stuff. I *love* it!
20
21  ```{r hobbies}
22  hobbies <- c("eating","sleeping","reading")
23  print(hobbies)
24  ```
25
````

3. After doing #2, see if you can figure out the difference between the code chunk arguments `echo`, `include`, and `eval`, as well as what other arguments there are. (Hint: you can do this by experimenting with it yourself and also googling! Google is your friend).

> Probably the easiest way to answer this was to google. I did so and found lots of great references, including this handy one:
> https://rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf
>
> Here’s the relevant info from that document:

option	default value	description
Code evaluation		
child	NULL	A character vector of filenames. Knitr will knit the files and place them into the main document.
code	NULL	Set to R code. Knitr will replace the code in the chunk with the code in the code option.
engine	'R'	Knitr will evaluate the chunk in the named language, e.g. **engine = 'python'**. Run **names(knitr::knit_engines$get())** to see supported languages.
eval	TRUE	If **FALSE**, knitr will not run the code in the code chunk.
include	TRUE	If **FALSE**, knitr will run the chunk but not include the chunk in the final document.
purl	TRUE	If **FALSE**, knitr will not include the chunk when running **purl()** to extract the source code.
Results		
collapse	FALSE	If **TRUE**, knitr will collapse all the source and output blocks created by the chunk into a single block.
echo	TRUE	If **FALSE**, knitr will not display the code in the code chunk above it’s results in the final document.
results	'markup'	If **'hide'**, knitr will not display the code’s results in the final document. If **'hold'**, knitr will delay displaying all output pieces until the end of the chunk. If **'asis'**, knitr will pass through results without reformatting them (useful if results return raw HTML, etc.)
error	TRUE	If **FALSE**, knitr will not display any error messages generated by the code.
message	TRUE	If **FALSE**, knitr will not display any messages generated by the code.
warning	TRUE	If **FALSE**, knitr will not display any warning messages generated by the code.

> Note: you might think it’s kind of silly for me to expect you to google things for this class — if it matters, why don’t I tell you? I do this for a very good reason: there is no possible way I can teach you everything you ever need to know in R — it’s a huge language and is changing all of the time — which means that one of the main things I want to teach you is how to find things out for yourself. This is, in fact, half of the job description of professional statisticians and programmers.

4. Make a copy of your datafile called `gladlysurvey3.csv` and put it in a folder called `mydata`. See if you can figure out how to load it from `gladlysurvey.Rmd` (Hint: you'll have to change the arguments to your `here()` function) and make sure to actually create the folder first!

   a. First you want to create a folder called `mydata` on your machine. You do this outside RStudio, go to this folder (with the Rproj and so forth in it) where it is on your machine and create a folder inside it called mydata (the normal way you make new folders on your machine).

   b. Then you want to make a copy of `gladlysurvey2.csv` called `gladlysurvey3.csv` and put it in `mydata`. You can do this either not through RStudio (i.e., the normal way you make copies of files and put them in folders on your machine) or you can do it through RStudio. To do it through RStudio, use the following commands:

   ```
   loc <- here("mydata/gladlystudy3.csv")
   write.csv(gdata, loc)
   ```

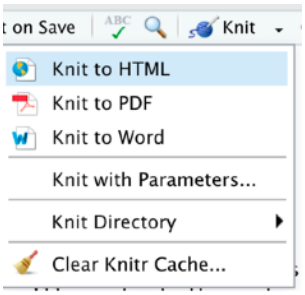
   c. To read it from the markdown, you need to type, with `loc` defined as in (b) above.

   ```
   gdata3 <- read_csv(file=loc)
   ```

   This will read it and put it in a variable called `gdata3`.

5. Knit your markdown file to html instead of word.

   This is easy: just choose "html" from the Knit options menu:

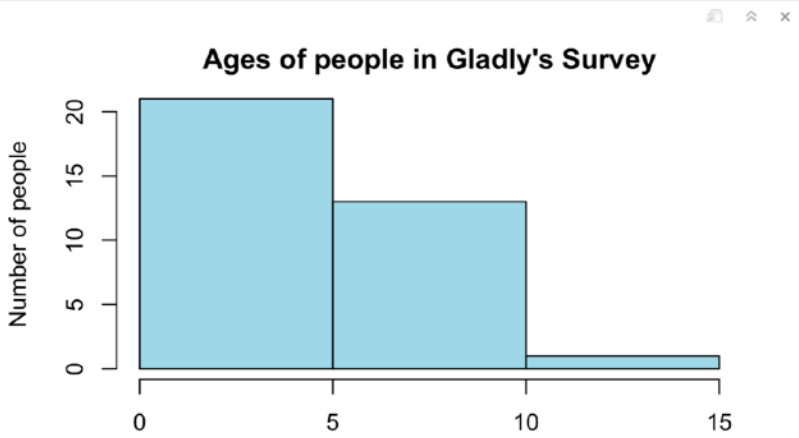


You can also try knitting to html. For many people knitting to pdf won't work without downloading and installing some extra software on their computer (not R packages). So we will never require you to knit to pdf in this subject.

Descriptive statistics

1. Change your histogram of ages to look like the one on the right. (Hint: you might need to google around to find names of colours to use).

```
```{r gdhistogrammod}
hist(x=gdata$age,main="Ages of people in Gladly's Survey",xlab="Years
Old",ylab="Number of people",col="lightblue",breaks=3,xlim=c(0,15))
```
```



2. Without using `summary()`, calculate the mean, standard deviation, and median of the variables `carrot`, `cake`, and `mud`. (Hint: you'll need to use the `na.rm` argument in your functions). Where you can, check your answers against the ones shown by `summary()`.

```
> mean(gdata$carrot,na.rm=TRUE)
[1] 7.911765
> sd(gdata$carrot,na.rm=TRUE)
[1] 1.524897
> median(gdata$carrot,na.rm=TRUE)
[1] 8
> mean(gdata$cake,na.rm=TRUE)
[1] 8.529412
> sd(gdata$cake,na.rm=TRUE)
[1] 1.186676
> median(gdata$cake,na.rm=TRUE)
[1] 9
> mean(gdata$mud,na.rm=TRUE)
[1] 1.470588
> sd(gdata$mud,na.rm=TRUE)
[1] 0.5632855
> median(gdata$mud,na.rm=TRUE)
[1] 1
```

We needed to specify `na.rm=TRUE` because our data had `NAs` and if we didn't, it would have returned `NA`, like here:

```
> mean(gdata$carrot)
[1] NA
```

3. Calculate the 10th and 90th percentile for `age`.

   Here we need to use the `quantile()` function:

```
> quantile(gdata$age,c(0.1,0.9))
10% 90%
  3   8
```

4. How would you interpret the responses to the questions about eating mud, carrots, and cake?

   This was open ended, and what I wanted you to do was to try to figure out what they meant about the participants. A few things you might have noticed was that the answers to the questions about carrots and cake were all quite high — with medians of 8 and 9 on a scale from 1 to 10, and high means too. This might mean that the participants just didn't understand the question (like how to rate things on a scale of 1 to 10) but the answers to the mud question were low, as one would expect since nobody likes eating mud! This suggests either that all of the participants just really like all food, or they're really hungry a lot of the time, or something else. We'll investigate further in the next classes.