Week 2 Computer Lab Exercises (2nd version 29 July 2018)

Put your group member names and ID numbers here

Due no later than 5pm Tuesday of Week 3 (7 August 2018)

Modifications from first release of this Lab

The first version of this document was posted on Moodle on Saturday night 29 July 2018. That version was, unfortunately, not the final version. The modified questions, with a description of the changes made, are listed below:

- Part II, Question 1c. The words "difference in the" has been deleted.
- Part II, Question 3e. This whole question has been deleted, as it is essentially the same as Question 4a.
- Part II, Question 4a. There is no change to the question, however an additional comment has been inserted here to alert students to the potential issue that some of their permuted samples may not result in a well-defined (i.e. finite or non-missing) difference in sample proportions. If so, some adjustments to the code provided in **Workshop1_GenderStudy.Rmd** file may be required.

Instructions

This assignment is a group assignment. All group members must belong to the same tutorial session. Your tutors will finalise the group membership during the Week 2 tutorial. If you miss that tutorial, be sure to contact your tutor so you can join a group. Groups will normally be comprised of 3 or 4 people.

Only one submission for each group is required, but all group member names and ID numbers must be stated in the "YAML" header at the top of each and every submitted file. (If you don't know what the "YAML" header is, you can look at the "rmarkdown2.0" Cheat Sheet available in the "Some handy R Cheat Sheets" folder under the Computing Resources on the unit Moodle page!)

Follow instructions provided for each Part below. Write your answers in the corresponding RMarkdown file (where required) so that it compiles to produce the desired results. Insert your response to (or code chunk for) each question part in the space immediately following the relevant question part.

Remember that you are part of a GROUP, and that everyone in your group needs to understand what is going on with these exercises! Be sure to talk to each other, and check that everyone understands not only what the "answers" are, but also how they were obtained. If you are not sure, ask another group to explain what they understand.

Once all tasks are completed, you will need to upload **two (2)** separate files to the Week 2 Lab 2018 Assignment link on Moodle for your group. These files will have names such as:

1. `GroupName_Lab2.Rmd`
2. `GroupName_Lab2.html`

Do not forget to include all group member names and ID numbers in the YAML header of every RMarkdown file, and ensure that they appear in every rendered version (e.g. HTML, DOC, DOCX or PDF) of the file that you submit.

Note that although the Week 2 Computer Lab assignment is not due until Week 3, you should review the questions here prior to attending your Week 2 tutorial session. This will give you an opportunity to ask questions where clarification is required, and also will prepare you for discussions and activities that will likely occur during the Week 2 tutorials.

Due to the need to finalise groups during the tutorial sessions, the Moodle link to upload the Week 2 Computer Lab files will not appear until the weekend of 4th August 2018.

Assessment marks

The final mark for this assessment task will be based on
i. Completion of all tasks outlined in this file (50%);
ii. Completeness and clarity of responses (30%); and
iii. The ability of the submitted .Rmd files to compile immediately without error (reproducibility) (20%).

Background

Read Chapter 1 of **IntroStat with Randomization and Simulation**, by OpenIntro Statistics, by Diez, Barr and Cetinkaya-Rundel [https://www.openintro.org/stat/textbook.php?stat_book=os]. Take note of Section 1.8 (p 50-54) where the Gender discrimination case study is reviewed.

Students are reminded of the file “Some useful notes on R” that is available under the Workshops Section of the unit Moodle site, as well as the various resources under the “Computing Resources” Section. These materials should assist you in completing these Computer Lab Exercises.

Another resource that students may find useful is a website (Using R Markdown for Class Reports, by Cosma Shalizi)[http://www.stat.cmu.edu/~cshalizi/rmarkdown/]. In particular, this page (and others it points to) explains the use of LaTeX to generate mathematical symbols when compiling RMarkdown.

PART I

Question 1

Download the **Workshop1_GenderStudy.Rmd** RMarkdown file from Moodle. Render the RMarkdown file (without modification) in RStudio to produce the output in `html` format. Read through the `html` so you are familiar with what each what each code chunk produces. Once you have finished, answer each of the questions below.

a. What is a *tibble*? (A simple one sentence answer is sufficient.)

   Answer:

b. What are the main components required for a data frame to be in "tidy data format"?

   Answer:

c. In the `RStudio` *Console* window, type in

   vignette("dplyr")

   and read through the vignette called **Introduction to dplyr** that will appear in the `RStudio` *Help* window. (You may wish to work through the exercises there if you choose - they are there to help you understand what each "verb" function in the `dplyr` package does - but this is not part of the current exercise).

   Find the nine **Single table verbs** listed in the dplyr vignette, and list all nine of them in your answer below.

   Answer:

d. Which of these dyplr "verb" functions is used most often in the **Workshop1_GenderStudy.Rmd** file? Explain what type of action this most frequently used dplyr verb does.

   (Hint: You might find it helpful to use the `RStudio` search function (Edit/Find...) to be sure you find all instances of each verb.)

   Answer:

Question 2

Review the available `R` help information about the `R` function **sample()**. Then complete the following questions:

a. Write a code chunk using the `sample` function to permute the elements of a vector defined by `x = seq(1, 100, 2)`. You can insert your code chunk below, before the question asked in part b.

Insert code chunk

b. Explain why it is important to use the `set.seed()` function in a program at some point before implementing the `sample()` function.

   Answer:

PART II

Is yawning contagious? An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't anyone yawning near them (control). The following table shows the results of this experiment, where a "Yes" response indicates the subject yawned and a "No" response indicates the subject did not yawn.

Group	No	Yes	Total
control	12	4	16
treatment	24	10	34

Question 1

a. Overall, how many subjects participated in the experiment?

Answer:

b. Insert a code chunk below that executes simple mathematical calculations to obtain the sample proportion of subjects who yawned for each of the treatment and control groups. Also include in your code chunk a calculation for the difference in proportions between the treatment and control groups.

Insert code chunk

c. *(This question has been modified from the original version.)* Replicate the RMarkdown code below to produce the table from part b. with an additional column that reports the sample proportions calculated in your code chunk from part b.

Replicate here:

Question 2

The null hypothesis for the experiment is given by the expression:

$$H_0: p_{control} - p_{treatment} = 0.$$

a. Write a sentence in English equivalent to the stated null hypothesis above.

Answer:

b. Which of the following alternative hypotheses is most relevant to the MythBusters yawning study?

$$i. H_1: p_{treatment} - p_{control} \neq 0$$

$$ii. H_1: p_{treatment} - p_{control} < 0$$

$$iii. H_1: p_{treatment} - p_{control} > 0$$

Answer:

c. Provide a justification for your answer to part b.

Answer:

Question 3

a. Insert a code chunk that uses the `tibble` package to produce a dataframe containing a collection of responses (one for each subject) from the Mythbusters yawning experiment. The dataframe should contain a single response variable (in a column) for each subject (in a row), as well as another variable (in another column) that indicates whether the subject was in the `treatment` or `control` group.

Insert code chunk

b. Insert a code chunk that uses verb functions from the `dplyr` package to manipulate the `tibble` dataframe produced in part a. to create a contingency table for the MythBusters yawning experiment data. (The result should look much like the table produced in RMarkdown, shown above Question 1.)

Insert code chunk

c. Insert a code chunk and use functions from the `dplyr` package to add a column to your contingency table produced in part b. that also contains the sample proportion of subjects who yawned in each group.

Insert code chunk

d. Insert a code chunk to calculate the difference in sample proportions between the treatment and control groups from the contingency table you produced in part c.

Insert code chunk

e. *(This question has been deleted from the original version.)*

Question 4

a. Insert a code chunk containing a function you have created that will

i. Input the tibble you produced in Question 3 part a,

ii. Permute the outcome of the yawning variable (i.e. mix up treatment and control labels amongst the different subjects), and

iii. Return the difference between the proportions of the resulting treatment and control groups.

(This note has been added to the original version.) **Note: Students should take care to ensure that all permuted samples results in a well-defined (i.e. finite or non-missing) difference in sample proportions.**

Insert code chunk

b. Insert a code chunk that will run your function (using an appropriate loop) 10000 times, saving the result.

Insert code chunk

c. Insert a code chunk that makes a dotplot (or a histogram) of the results from part b., and add a vertical line on the plot that represents the difference between the treatment and control group proportions from the original data. (Note, you may need to modify the value of the `bandwidth` option for the graph to fit nicely in the document.

Insert code chunk

d. Insert a code chunk that compute the proportion of times that the permuted data yields a difference larger than the difference of the actual data.

Insert code chunk

e. Explain why the numerical value you reported in part e. may be considered as the (permutation) p-value for testing the null hypothesis.
Answer:

f. Based on your p-value, what is your decision about the null hypothesis (i.e. reject, or fail to reject, the null hypothesis)?

Answer:

g. Write a full sentence stating your conclusion, indicating the presumed significance level of your test.

Answer:

h. Finally, based on these experimental results how would you answer “Is yawning contagious?”

Answer: