

BIOL/STATS 2244

Lab 3: *Analysis and Conclusion*

Objectives

Lab Assignment 3 provides an opportunity to experience the *Analysis* and *Conclusion* stages of the Scientific Inquiry Framework (“PPDAC”), in addition to applying concepts from earlier stages. Specifically, this includes:

- i. applying your understanding of vocabulary and concepts associated with analyzing data;
- ii. identifying an appropriate model for a specific research question;
- iii. checking the conditions for the model underlying the relevant statistical inference procedures;
- iv. interpreting the results of statistical inference procedures, and,
- v. reporting conclusions in appropriate scientific formats.

To achieve these objectives, you will need to draw on content from Online Modules 1 (Working with Data in R), 2 (Visualizing and Summarizing Data in R), 3 (Inference on the centre of a distribution) and 4 (Inference on One Proportion) and from (related) course material.

Background for Assignment

In the first Lab Assignment, you were introduced to a little background information that led/explained the logic behind the following **Research Objective** for 2244 labs:

Characterise the therapeutic impact of aromatherapy foot massage.

In Lab Assignment 1, you were tasked with *Planning* a sampling and study design to collect data to address some research questions related to this Objective. In Lab Assignment 2, you were introduced to the sampling and study design for a research study related to this objective, and provided a datafile from that study. You should re-acquaint yourself with the sampling and study design that generated the data, and the Data Description file you were provided. **You will be working with the same data for Lab Assignment 3.**

Instructions for Assignment

To help you approach this Assignment in a logical/organized fashion, you are encouraged to follow these steps (in the order presented).

1. Refresh your memory (as needed) about the data we are working with, as described in the Data Description.
2. Read through the remainder of this Assignment file, including the “Reminders/Tips for Success”, the Assignment Questions, and the comments related to Marking Rubrics so that you know what you are being asked to do.
3. Open the “Answer template” file. Use this file to type/enter your answers to the Assignment Questions; it is set up with the proper headings for this Assignment; you just need to input your answers (use whatever space you need to do so). Refer back the Assignment Guidelines and Format file (with respect to formatting of R code, graphs, etc. in assignments) that was provided alongside this Assignment.

4. **Import** the data file provided, “labdata.csv” into R or R Studio or R Studio Cloud. We also recommend that you **attach()** the datafile as a dataframe called *labdata*. If you refer to “*labdata*” in your R code at any point in your assignment, we will assume that it is the labadata.csv file that you were provided. This way, if you choose to save the data file under a different name that we don’t recognize (i.e. other than *labdata.csv*), you won’t be penalized if you always refer to it in code as *labdata*.
5. It might be useful to check the structure of the data using a function like *str()* to ensure it imported properly and/or is consistent with the quality check information provided in the Data Description file. This would also be an opportunity to ensure that R has correctly identified the types of variables that exist; for instance, R may have mistakenly identified the “heartrate” variable as a factor, when it should be numeric or integer. You can “correct” this by using the following lines of code:

```
heartrate_num<-as.numeric(as.character(heartrate))
```

Then, just be sure in any subsequent code that you want to use heart rate data, you refer to this new vector, “heartrate_num”. Of course, it would be appropriate to include this line of code in your Assignment answers, since you would be using a new ‘object’ that doesn’t originally exist in the labdata.csv file.
6. Answer the Assignment Questions (below).

Reminders/Tips for Success

1. **Make interpreting your R code easy and ensure that it is functional.** The best approach is to generate an R script file for each numbered Question that requires use of R in any way. Put everything in that R script file that takes you from attaching the dataset to completing the question. Annotate your more complicated lines of code with #comments (as demonstrated in the first in-lab session and in the example provided in the Assignment Guidelines and Format file). Then, simply copy/paste the contents of that script file into your answer (and include the Output if applicable) when asked for R code. We should then be able to copy/paste what you’ve included, and the code should run without problems!
2. **We only know names/variables that are part of the original *labdata.csv* file.** So, when we are looking through/grading your R code, if we see a name for a variable or datafile that is NOT part of the original *labdata.csv* file (e.g. you’ve renamed a variable or created a new one), you must provide information on what that R object represents. As discussed during the second in-lab session, this is best achieved by simply providing the line(s) of R code (plus #comments!) used to create that new object. Failing to provide such new R object definitions is one way to not be awarded full marks for your R code.
3. **Your answers should be written *specifically* for the research study/context (in terms of variables, sample, units, measurements, etc.) with which we are dealing.** For example, it’s insufficient to talk about the “response variable”; we should be talking about the actual name of that variable, e.g. concentration of testosterone (using an example from our in-class case study on heel height and male behavior). This idea of using the context or “language” of the Problem has repeatedly been illustrated in more recent video lectures. Being specific means using the *context* of the research; a sentence like, “the distribution of *student heights* in my sample of 2244 students needs to be symmetric” is explicit about *what* needs to be symmetric AND uses the *vocabulary* of the study.
4. **Demonstrate your understanding of course content through application, not definition.** Questions which asks you to discuss something with reference to particular course concepts (e.g. sampling variability) requires an *application* of those course concepts to the current scenario/situation. That means that simply providing the definition of those concepts is not going to result in points awarded

for the Question. Your answers should demonstrate you understand that concept and why it applies (or doesn't!) or is useful in the particular context.

5. **Use what we have been doing in lecture (in class or video lectures) as cues to understand the questions.** Nearly everything you are asked to do in this Assignment is somehow illustrated/discussed during a lecture. Your first line for trying to understand what a question is asking is to go back to your lecture notes/videos.
6. **Show us what you know, completely.** Most inference procedures have more than one 'condition' that must be met for the underlying model to be valid. If we violate *any* of the conditions, we shouldn't use the procedure. In these situations, it may seem 'redundant' to continue to check the other conditions of a model. Remember that these Assignments are assessing your understanding and application of course material. Occasionally, we violate conditions for a model (it's bound to happen with real data!). Be sure to assess ALL conditions of a model completely, regardless of whether one or more is violated. Show us what you know!
7. **If you get stuck with R, at least tell us what you wanted to do.** We recognize this is your first course that involves using R (and for many of you, any kind of programming language). Some of these Assignments questions will be tough, others should be quite accessible with some careful thought and application of what you are learning in the Online R Modules. If you get stuck and run out of time to get help to "unstuck", don't leave an answer blank. Tell us what you were trying to do, show us the code you were trying to use, or what functions and types of arguments you think would be relevant. That is, walk us through your thought process. It likely won't be worth full marks, but some part marks may be obtainable.
8. **How to write 'symbols' in a document.** For some questions, you may need to use symbols to represent specific values. If you use a word processing software like Word (i.e. .docx) to create your assignment file, you will likely find most of the symbols you would need can be inserted either from the Insert/Symbol menu, or, by using the Equation Editor (also part of the Insert menu). Alternatively, if you aren't sure how to get mathematical symbols in your word processing software, you can use the following "phonetic" symbols; when we see these "words", we will interpret them as the corresponding symbols. In all cases, feel free to use subscripts liberally to help communicate with these symbols.

symbol	"phonetic" version
σ	sigma
μ	mu
\bar{x}	x-bar
\hat{p}	p-hat

For this assignment, if you wish/need to compute a confidence interval, use a confidence level of **85%**.

In addition, you are always welcome—at the end of a given question—to provide **a short commentary** justifying/explaining any choices you made for which variables, subsets, etc. or reasoning you used to answer a question. Help us understand your thought process when working with our data.

Assignment Questions

***Note: Lab learning opportunities versus other situations**

One of the key learning outcomes in 2244 is knowledge of and evaluation of the conditions/model necessary to make different inference procedures valid (i.e. to verify that the underlying model fits). On quizzes, tests, and exams, we adhere **strictly** to only *conducting* analyses if the model truly fits/is valid. However, often in teaching lab settings—**and is the case for this Assignment**—, we need to conduct analyses even though we may fail to meet some/all of the necessary conditions, so that we have the opportunity to practice (and, for your instructor to assess your knowledge/use) with statistical software.

In future courses, research, and careers, we should always check the fit of your selected model (a step that we explicitly take in these labs). If the model is not a good fit, carefully consider investigating alternative, more appropriate models and analyses (this might include consulting a statistician, transforming data, or exploring non-parametric alternatives).

Question 1.

Note: this will be a tough question. If you get stuck, consider moving on to the other questions and then returning to it later.

The dataset we are working with is in the *long format* (as opposed to *wide*). This format is useful for many reasons (and is the preferred format for R), but can make it challenging in repeated measures studies like the one we are working with, when we want to compare “before” and “after” measurements on the same individual, or an individual’s response under two conditions. This question asks you to do just that.

Each participant’s level of anxiety was recorded using the State-Trait Anxiety Index at each of the three timepoints in the study. One of the questions the researchers for the study asked is whether aromatherapy foot massage has a positive impact on anxiety. To address this question, we would need to see whether the anxiety scores observed improved from baseline to after the intervention. The researchers defined an improvement as any change in anxiety score from baseline to after the intervention that was at least **3** points (e.g. moving from a score of 65 to 62 would be considered an improvement).

Use R to create a new vector (i.e. variable) called *change* that characterizes whether each study participant experienced an improvement or not in their anxiety score from baseline to after the intervention.

Report your complete **R code** (with #comments to help us understand the steps in your method) that is used to create this new vector, *change*. As well, use R to generate a table that summarizes the distribution of the *change* variable. Copy and paste the **table** into your answer.

There will be many ways to get R to make this new vector. You have freedom on how you achieve it. Just be sure that your code—as you report in your answer—runs as written (i.e. we may copy/paste it into R and see what happens). And provide annotations to help us understand what you have done.

Note that Online R Module 1 does provide sufficient knowledge to answer this question.

Question 2.

For all parts of this question, write your answers in the language of the Problem.

a) Consider the research question, *what is the mean heartrate for adults who have been using aromatherapy foot massage for an extended period of time?*

Which of the following **inference procedures** would be most appropriate to address this research question? (simply copy/paste your selection as your answer)

- t confidence interval for the population mean
- large sample confidence interval for the population proportion
- large sample test for the population proportion
- t test for the population mean

b) Is it valid to use the inference procedure you selected in *part a* for our data? To answer this question,

- write each **condition** required for the confidence interval to be valid using the language of the Problem
- specify explicitly whether the condition is **valid** or not (i.e. yes or no?);
- provide a **justification** for your decision on its validity (i.e. explain how you came to your conclusion about whether the condition is valid or not...remember, show us you understand!).
- include any **R code and associated Output** you used as part of your justification (if applicable).

Question 3.

Regardless of your answer to *Question 2* (i.e. whether you determined the model fit our data or not), use R to conduct the inference procedure you identified for *Question 2*.

a) Restate the **inference procedure** that you chose in Question 2. Simply copy/paste your answer from Question 2 part a.

b) Provide the complete **R code and associated Output** you used to conduct the inference procedure. It should be clear from the code you report where any values/vectors you use as arguments come from (i.e. we should be able to figure out what you did based on your code).

c) Provide a proper scientific **conclusion** for the results of your analysis.

Question 4.

Consider the following research question, *Does engaging in aromatherapy foot massage have a positive therapeutic impact on blood pressure in healthy adults?* This question could be addressed using a hypothesis test based on our dataset.

Describe what you would do with our dataset to address this research question. Your **description** should include mention of which all variables in the dataset you would need to work with, any sub-setting or data transformation you would need to complete, and specifically what type of comparisons (if any) you would be doing/focused on. Wherever possible, provide a brief explanation of why you are using those variables/subsets/comparisons (i.e. justify/explain your approach). Essentially, this question is asking you to “plan out” an analysis approach for this research question, based on our particular data set. Note that there is NO requirement for use of R (or even referencing R functions) for this question. For reference, a moderate length paragraph (or a series of numbered “steps” with explanation if you prefer) should be quite sufficient to address this question.

Marking Rubric

Like Assignment 2, most of the questions in Assignment 3 have correct vs. incorrect answers and/or approaches. Consequently, the marking scheme for evaluating your answers to certain questions may often have a single 'right' answer/approach for which we are looking. However, how we use R to explore, summarize, and analyse our data can, to some degree, vary in technique. That is, there may be more than one way to ask R to complete a particular 'task'.

So, what does this mean for a student trying to understand expectations when completing this Assignment? In addition to the "Reminders/Tips for Success" provided on pages 2-3 of this file, consider the following general criteria for different types of questions/marking; these criteria will likely play a heavy role in evaluating the answers submitted for the Assignment:

Criteria for R code and output

- ✓ Selection of data, variables, and subsets is relevant to the question or task;
- ✓ Choice of R functions is relevant and appropriate (demonstrating an understanding of the analysis being conducted) for the question, task, and/or type of data being summarized/analyzed;
- ✓ Reported R code for any numbered question is complete and would function (i.e. reproduce the output included/described in the answer) if it were copied/pasted into R, and run on the *labdata.csv* data (assuming we had first successfully imported that data and saved it as a dataframe called *labdata*).
- ✓ Reported R code uses brief *#comments* to help interpret the purpose of more complex commands/functions

Criteria for 'other' questions (i.e. identifying, describing, explaining, discussing, etc.)

- ✓ *Knowledge*: use and application of relevant statistical vocabulary/concepts demonstrates an accurate understanding of those concepts; that is, the vocabulary/concepts are *used/applied* in a manner that is consistent with the definition/understanding. The use moves beyond simply defining the concepts, but actually applies them to the situation. This criterion also connects to whether an answer is consistent with any expectations/guidelines communicated in course content (e.g. lectures).
- ✓ *Connections/Justification*: Answer demonstrates (through explanation and/or description, where appropriate) the relevance or relationship of choices made/vocabulary used to the question(s) or situation. That is, it's clear WHY you have made the choices you did and these choices make logical sense.
- ✓ *Completeness*: Answer provides sufficient detail (whether in written answers or visual content) that another, knowledgeable individual can understand and/or recognize the application of the concepts, without ambiguity or doubt. This also refers to whether the answer has addressed all aspects of the question.
- ✓ *Communication*: Answer uses clear and concise language, and thoughtful organization of ideas to facilitate readability and understanding. That is, we do not have to re-read your answer multiple times to try to understand what you are saying.

Other comments

Remember that these Assignments are meant to assess your understanding of course content. So, many of the questions relate directly to content covered/discussed in (video) lecture, as part of the assigned Independent Study material, and/or from the Online R Modules. Reviewing these 'resources' while working on the Assignment may be quite beneficial/informative!