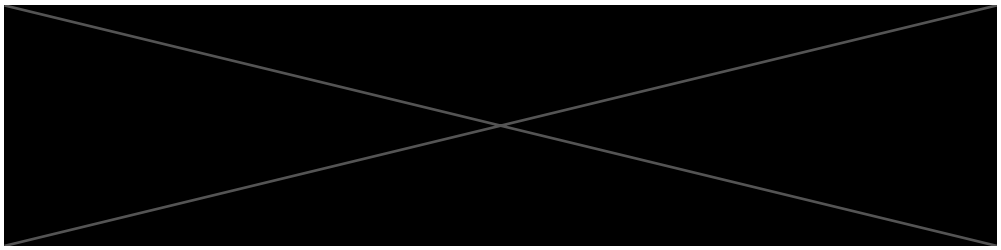


Assignment 3: *Analysis*

Objectives

This Assignment is designed to demonstrate your current mastery of the following Learning Outcomes:

i. Analyse data using inference procedures to address a research question
    a. Evaluate model diagnostics for common parametric inference procedures
ii. Use statistical software to explore, summarize, analyse, interpret, and communicate data;
    a. Use R to conduct common parametric inference procedures, including model diagnostics.
    b. Use R markdown to produce reproducible analyses and reports.
iii. Communicate statistical concepts, analyses, and arguments in an accurate and scholarly manner.
    a. Use conventional and transparent formats for reporting results of statistical analyses in written/graphical form.

To achieve these objectives, students will need to draw on course material from the topics primarily from our Lecture Topic and Lab about hypothesis tests. However, you may also have to draw on ideas and techniques from earlier topics, including Lab 3: Reproducible files using R markdown.

How this Assignment 'works'

This Assignment is the last of three Assignments in the course; it continues our progression through the phases of the PPDAC Framework. The major focus of this Assignment is the Analysis stage of PPDAC.

In the first Assignment, you were introduced to some 'Research Background' related to flexibility and screentime in children. We will continue to work with this background (you can refer back to it if necessary for this assignment). In the second Assignment, you were provided a dataset that is related to the Research Background; we will continue to use this dataset for this Assignment. Recall that the data is in a .csv file named assign.csv, and **you have access to the research article that describes the way the data were collected** (i.e. in the Materials and Methods section of the article).

As was the case with Assignment 2, **you will not use every variable and/or every data point in the datafile for this Assignment!**

Being successful on this Assignment

Remember, the Assignment is evaluating you on three things:

- Your ability to use R in an R markdown file to conduct an inference procedure
- Your ability to identify and evaluate model conditions for an inference procedure
- Your ability to use conventional symbolism and structures in the context of an inference procedure

The knowledge to support the application of the inference procedure and evaluation of model conditions was part of *Topic 9: Understanding hypothesis tests + t procedures*; the skills to use R were exemplified in earlier labs, and in the context of t procedures, *Lab 5: t procedures for the mean in R*; look back at what we did in the Lecture Topic and the Lab to support you for this Assignment. Your best approach would be the following (you don't need to 'answer' these suggestions in your assignment; this is just a guide to how to approach the assignment):

1. Think about the Research Question (given below, in the Assignment Questions section) and the dataset. Which variables in the dataset(s) will you use to answer the question? Use the article's **Materials & Methods** sections to understand what the variables/columns in the data represent and the way the variables were measured (i.e. to understand types of variables).
2. "Analyse" the Research Question, alongside the variables from the data like we have been doing throughout the course: How many variables are you working with? What type of variables are they (again, use the information in the article's Materials & Methods to help you understand the way the variables were measured, and then determine if they are quantitative or categorical)? What is your analysis goal: estimating a value for a population parameter or assessing evidence for a claim about a population parameter? What parameter is involved?
3. Review the Lecture Topic content related to the identified inference procedure; apply the concepts to the specific data and Research Question we are working with.
4. Use the code examples in the relevant Lab to do the work with R.

Dataset check

Here's the R output from *str(assign)* so you can confirm the data imported correctly into R. I have NOT corrected any mistakes R made at identifying variable type.

```
> str(assign)
'data.frame':   394 obs. of  14 variables:
 $ id           : chr  "L1" "L2" "L3" "L4" ...
 $ sex          : chr  "female" "female" "female" "male" ...
 $ pain         : chr  "without" "without" "without" "without" ...
 $ age          : int  9 9 9 9 10 11 9 9 9 9 ...
 $ height       : num  128 132 135 136 138 ...
 $ weight       : num  25.9 29.2 26.2 28.7 28 37.1 24.7 22.4 27.1 32.7 ...
 $ BMI          : num  15.8 16.7 14.3 15.6 14.7 17.2 16.4 14.3 16.6 17.5 ...
 $ physical     : num  5 12 12 7 20 6 11 3 0 0 ...
 $ screen       : num  1 1 1 2 0.5 2 1 1 3 0.5 ...
 $ sleep        : num  9 8 9 9 8 9 9 8 9 8.5 ...
 $ visual_analog: num  0 0 0 0 0 0 0 0 0 0 ...
 $ distance     : num  11.2 4.7 14.7 3.9 14.6 5.9 14 12 15.3 16.8 ...
 $ strength     : num  35.3 22 34.3 39 32.5 43.5 20 23 25.8 20.5 ...
 $ tilt         : num  13 17.5 16.3 12.7 18.6 9.4 20.5 21.4 13.8 19.9 ...
```

[redacted]

Assignment Questions

There is some concern that when school-age children spend time using screened devices (e.g. smartphones, tablets, TVs, computers, etc.), they are not as active as they should be. With lower physical activity, childhood obesity becomes a concern. Research has suggested that poor health in childhood can be a precursor to poor health as an adult. One simple measure of physical health that can be calculated with very little information is the *body mass index* (BMI)[1]. BMI is computed as:

$$BMI = \frac{weight\ (kg)}{(height\ (m))^2}$$

The value of BMI for an individual is compared against a 'normal' range to evaluate the individual's health. The 'normal' range of BMI for school age children is between 15.2 and 18.0 kg/m^2 (with a midpoint of **17.6 kg/m^2**). Arguably, if the use of screens leads to poorer health, then we might expect mean BMI to be higher for children whose mean daily screentime is above the average (which has been reported as **6 h/day**). Consequently, we can address the following **Research Question:**

Do school-age children who have a typical daily screentime of more than 6 h have a body mass index above the 'normal' midpoint for school-age children?

You will be conducting a null hypothesis test, specifically, **the t test for the mean**, to assess evidence for the claim, **the mean BMI for school age children who use screens more than 6 h/day is higher than midpoint BMI of school-age children,** to answer that Research Question.

Question 1.
What are the appropriate statistical hypotheses (null and alternative) that should be used to conduct the hypothesis test to answer the Research Question described above? Write *both* hypotheses using appropriate conventional symbolic format and write the *null* hypothesis in sentence format such that it makes it clear what is the interpretation of the symbolic format. Provide a brief explanation (1-2 sentences) justifying your null and alternative hypotheses.

Question 2.
What are the model conditions that must be met for the t hypothesis test to evaluate whether the mean BMI for school age children who watch more than 6 h/day is higher than the midpoint BMI to be valid? Discuss whether each is met for our data, generating and interpreting appropriate R output if/as necessary.

Question 3.
Use R to conduct the t-test, consistent with the statistical hypotheses you stated in Question 1. You do NOT need to interpret the results.

---

[1] BMI is not generally considered a valid measure of physical health anymore; someone who has a great deal of muscle may be considered "obese" based on BMI. We're just using it for the purpose of this assignment because it's easy!

Notes:

- You will need to use the LaTeX abilities of R markdown to write the symbolic (i.e. equation) format for Question 1; this was covered in Lab 3. **BE VERY CAREFUL:** you cannot copy/paste symbols from a Word file (or other word processing file) into an R markdown file. Doing so will cause knitting problems. You've been warned.
- If you use any R code/output to answer a question, be sure both the code and output are showing in your knitted file.
- This is not meant to be a long assignment. Question 1 is a couple sentences plus the two symbolic/equation forms of the hypotheses. Question 2 will be—perhaps—about 4-8 sentences, perhaps with some R code and output if relevant (just make sure you are complete). Question 3 will be code and output.

Grading for Assignment 3

This Assignment is less 'flexible' that the previous two Assignments. Where the previous Assignments had opportunity for more than one appropriate approach, this Assignment really has some 'correct' answers. Still, your Assignment will be graded using the same general grading scheme as the previous two Assignments.

General Overview of Grading

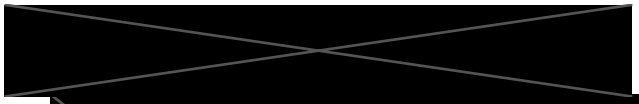
Your answers to the Assignment Questions will be graded based on a 4-level rubric (**on page 8**), which focuses on your mastery of several course-level learning outcomes. This is just like in Assignments 1 and 2. When the graded Assignment is returned to you, you will receive **three categories** that indicate which level on the rubric your submission received for the three Learning Outcomes being evaluated. As an example, you might receive:

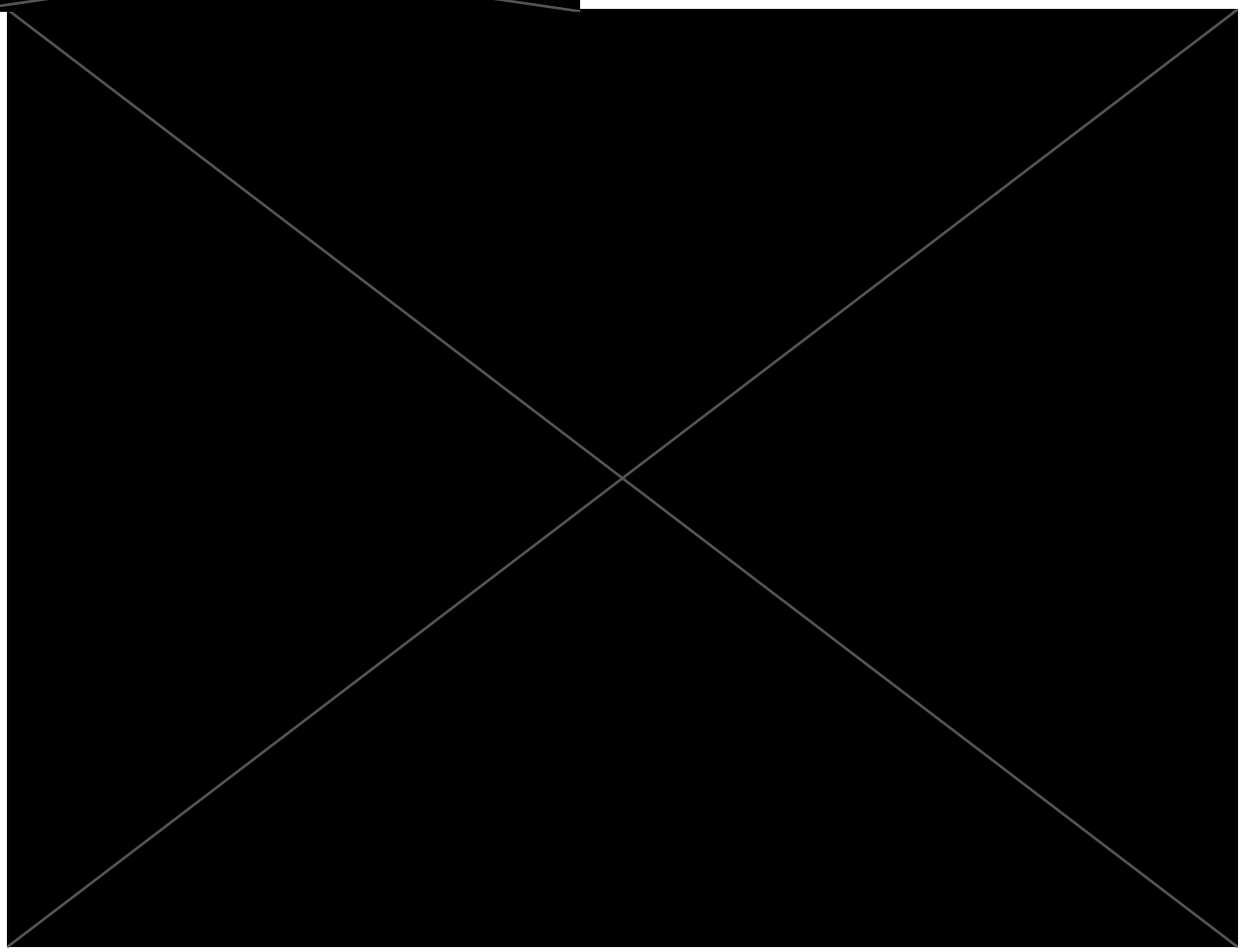
- ***Proficiency*** *for "Conventional formats"*, indicating that your submission demonstrated *proficiency* for the learning outcome, *Use R markdown to produce reproducible analyses and reports PLUS Use conventional and transparent formats for reporting results of statistical analyses in written/graphical form,* and,
- ***Mastery*** *for "Evaluate model"*, indicating your submission demonstrated *mastery* of the learning outcomes, "*Evaluate model diagnostics for common parametric inference procedures"*, and,
- ***Approaching Proficiency*** *for "R for inference"*, indicating your submission demonstrated *approaching proficiency* of the learning outcome, "*Use R to conduct common parametric inference procedures, including model diagnostics".*

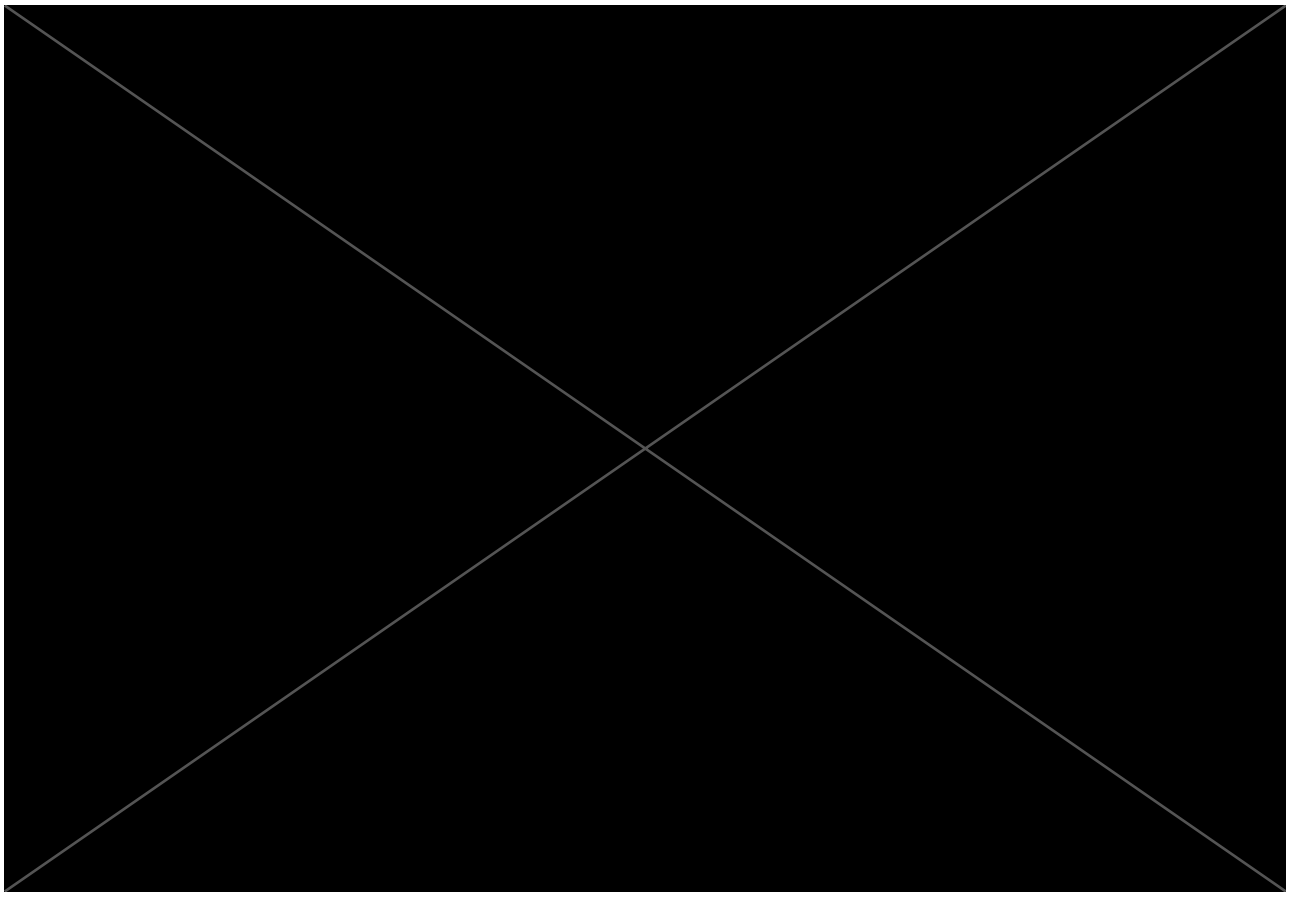
You will be able to see these three levels and some feedback about them through Gradescope when the graded Assignment is returned. These levels will also be communicated in OWL Brightspace Grades entry for each learning outcome with Assignment 3. To be clear, there will be **NO** numeric mark for the Assignment (e.g. like 75% or 8/10, etc.); the three assigned levels are the only outcome of the evaluation.

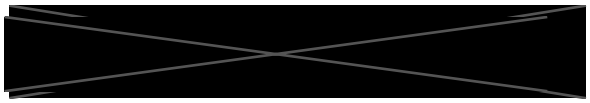
Essential Requirements and Late/Accommodated Assignments

- Failure to submit the Assignment at all will result in a '0' for the Assignment unless Academic Consideration from the Academic Counseling office of your home Faculty is received.









Format of Assignment 3

Structure of your Assignment

We use the same formatting 'rules' as for Assignment 2.

Please follow these rules:

- ✓ ALL aspects of your Assignment (written text, code, and output) are **created in a single R markdown file** which you will knit; **this is part of the grading rubric**.
  - Knit to a .PDF file.
  - If knitting to .PDF directly doesn't work, knit to Word (.docx), and then save the Word file to a .PDF file for submission.
  - If knitting to .PDF or to Word then saving as a PDF doesn't work, knit to HTML and then saving that as a .PDF; knitting to HTML is the least preferred because it doesn't automatically make page breaks (so it is hard to see/grade)

- ✓ name the files whatever you want, but it's a good idea to keep the filename fairly simple. I also find it helpful if you put your name or userID (e.g. jsmith123) in the file name, but this is not strictly required.

- ✓ **ALL of your code needs to show in your knitted .PDF file**. This includes any packages you have loaded, your code to import of the datafile, etc. Be aware; if you see "include = FALSE" at the start of an R chunk, switch it to "include = TRUE". The best approach is to create a new R markdown file, delete all the default 'stuff' that comes in it (except for the title, name, date, etc.). Then, create your own new R chunks as needed for code. I have provided an 'example' R markdown file that you could start with, should you want.

- ✓ **Make sure that long lines of code are still visible in your knitted PDF file**—we need to see your full code so we can evaluate it properly. The file "Tips using R to create a graph" has a suggestion to handle this.

- ✓ Any written text (beyond short #comments that clarify code) should be in the 'text' section of the R markdown file. **Be sure to label different question parts** (i.e. Question 1, 2, etc.).

- ✓ **Do not 'print out' the dataset or variables in your knitted file** (i.e. don't make it so that the raw data shows in your RMD file). Essentially, your resulting file for submission will typically only be a few pages long (there is no strict maximum because it depends on how efficient you are with writing R code; efficiency is NOT being graded!).
  - If you used View(), remove that from your R Markdown file before you knit it
  - When you use read.csv() or read.table() to import the datafile in the R markdown file, it should simultaneously assign the data to a dataframe.
    i.e. a3 <- read.csv(file = "assign.csv")

- ✓ **Do NOT include the Assignment *questions* in your answer file**; just have your answer text/code/output in your file. Your Assignment will be submitted to Turnitin; having the Assignment question text in your submission will inflate the observed textual similarity.

- ✓ In Questions that involve written responses, **do your best to use proper grammar and spelling**. Problems with sentence structure and grammar that significantly detract from the readability of your answer (and therefore, our understanding of what you are saying) may impact our evaluation of your answer.

- ✓ **Complete your Assignment independently**. This Assignment (and the other Assignments in the course) are *individual* Assignments. Evidence of inappropriate collaboration in your Assignment answer will be dealt with according to the University's procedures for issues relating to Academic Integrity; the 2244 policy on the use of AI technology also applies. Your Assignment will be submitted to Turnitin for analysis of textual similarity to other sources.

Submission of your Assignment files

There are TWO places that you must upload your assignment files:

1. The R markdown file (.rmd) and knitted file saved to PDF (.pdf) get uploaded directly to OWL Brightspace

2. The knitted PDF file only gets uploaded to Gradescope.

This is the same approach as we used in Assignments 1 and 2; it's just that you upload both the .RMD and .PDF to Brightspace directly.

Biol/Stat 2244B – Assignment 3 – W24 INC

Learning outcome	Level M (Mastery)	Level P (Proficiency)	Level A (Approaching Proficiency)	Level N (Not Met)
Use R markdown to produce reproducible analyses and reports PLUS Use conventional and transparent formats for reporting results of statistical analyses in written/graphical form.	LaTeX and/or R markdown formatting is used to successfully generate mathematical symbols in statistical hypotheses. Symbols used accurately reflect convention for test being conducted. Phrasing (symbolic and sentence format) accurately reflects the Research Question, claim, and structure of data being analysed (i.e. type of variables, structure of comparison groups). Sentence and symbolic format of null hypothesis are consistent.	Attempt made at using LaTeX and/or R markdown formatting to generate mathematical symbols in statistical hypotheses but one or two of the following occur: • Inconsistency between sentence and symbolic format of null hypothesis, but one is accurate • Attempt at symbols clearly accurate (i.e. appropriate by convention) but symbols don't render properly • Minor misunderstanding in general structure of null or alternative hypotheses • explanation for hypotheses lacking.	Attempt made at using LaTeX and/or R markdown formatting to generate mathematical symbols in statistical hypotheses, but one or more of the following occur: • More than two errors from Level P. • Symbols are not an accurate reflection of conventional symbols relevant to the test being conducted. • Evident misunderstanding of the general structure of null and/or alternative hypotheses • One of the requested hypotheses missing (e.g. no sentence format, or no alternative, or no null)	No attempt at using LaTeX and/or R markdown formatting to generate symbols for the statistical hypotheses AND/OR evident misunderstanding of statistical hypotheses in general.
Use R to conduct common parametric inference procedures, including model diagnostics.	All data transformation, model diagnostics, and analyses are conducted in R. Application of R functions for diagnostics and test accurately reflect statistical hypotheses. All aspects of assignment are completed in R markdown file, which is successfully knit. All R code is successful in producing relevant output. All R code and resulting output is visible in submitted PDF (from knitting).	All data transformation, diagnostics, and analyses are conducted in R, but 1 or 2 of the following occur: • RMD file is not successfully knit but all code would function otherwise • Some code and/or output is not showing in knitted file but is otherwise accurate • Minor mismatch between statistical hypotheses stated and application of hypothesis test with data	R is used to perform model diagnostics and analyses but any of the following occur: • Some data transformation is not conducted in R • Some components of code are not/would not be successful at producing relevant output • Significant mismatch between statistical hypotheses and application of the test with data which suggests a misunderstanding of how to conduct the test	R is not used to analyse data or perform diagnostics, or, attempt in R does not involve relevant functions/techniques (i.e. attempt does not demonstrate an understanding of using R for the purposes required).
Evaluate model diagnostics for common parametric inference procedures.	Model conditions are accurately identified. Discussion reflects an accurate understanding of the conditions, and the appropriate methods for assessing each condition, as taught in 2244 Lecture Topic. Use of R to evaluate model condition(s)—where appropriate—is accurate.	Model conditions are accurately identified. Discussion demonstrates an accurate understanding of the conditions and appropriate methods for assessing each condition, but minor errors in application or interpretation related to our data occur.	Attempt at identifying model conditions suggests some awareness of conditions, even if one or more of the following occur: • Discussion/evaluation suggests misunderstanding of appropriate methods of evaluating condition(s) • Significant misunderstanding of how conditions apply to our data	No attempt at identifying and/or evaluating conditions, or, attempt demonstrates a (near) complete lack of understanding of the relevant model conditions.