STAT 260 Summer 2024: R Assignment 3

Due: Tuesday August 6 BEFORE 6pm (PT) on Crowdmark.

Introduction to R: Before attempting this assignment, read and work through the **Introduction to R Assignment 3** file posted on Crowdmark. This file contains a list of all the R commands needed to complete this assignment.

Submission: Follow the same submission instructions as in R Assignment 1. Make sure to copy/paste your R code and the output into a document file, and then convert it into PDF.

For each question requiring a calculation, include both your R code and its output in your document.

Upload your files for submission to the assignment on Crowdmark, before Tuesday August 6 at 6:00pm (PT). A late penalty of 5% per hour will be applied.

Note: For each of the following, carry out your calculations **only** using R or RStudio. For each part, include a copy of your R code used and the output of your code. (i.e. Copy and paste the relevant pieces into a Word document (or other word processing document).)

⋆⋆ **Bonus** ⋆⋆ This assignment will be graded out of 28 marks (the equivalent of completing Parts 1 - 3). If you choose to complete the optional Part 4, it will be counted as bonus marks. The bonus marks will count only towards your overall R-assignment grade in the course - so if you lost some marks on R-Assignments 1 and 2, here's your chance to make some back.

Note: R can complete basic arithmetic operations such as addition (+), subtraction (-), multiplication (*), division (/), and square root (sqrt()). It should be noted that parameter values can be entered as arithmetic expressions. For example, we could enter an arithmetic expression using multiplication when writing the value of lambda in the desired Poisson function. As such, ***you must complete all of the following calculations using only R*** - no external calculator should be used.

Background: For Parts 1-3 of this R assignment, we'll show how we can use R to quickly and accurate compute confidence intervals and complete hypothesis tests for a single population mean (μ) and a population proportion (p), as well as two independent population means ($\mu_1 - \mu_2$).

Part 1 **[9 marks]** A video game developer claims that their flagship game can run at 120 FPS (frames per second) on the YBox game console, but consumer reviews cast doubt upon this value. A gaming magazine tests the developer's claim by running a series of 28 play tests of the game on the YBox, and finds the following fame rates (in FPS):

97	135	114	111	119	118	114	114	97	135
96	113	128	115	117	115	103	106	131	123
129	99	119	113	93	116	115	104		

(a) Store the data in an appropriately named vector (you do not need to copy/paste code for this step). Is there evidence to suggest that the mean frame rate of the game on the YBox is less than 120 FPS? Conduct the appropriate hypothesis test using R:

- Using the correct notation, define the parameter(s) of interest in this study.
- State the null and alternative hypotheses in terms of the parameter(s).
- Copy and paste the appropriate R command and output for the hypothesis test into your document.

(b) Clearly state the value of the test statistic, the degrees of freedom, and the p-value you found from part (a). What is the strength of the evidence against H_0?

(c) Write a sentence stating your conclusion of the hypothesis test you just performed. (i.e. write a plain language conclusion regarding the game's frame rate on the YBox that a non-statistician could understand.)

(d) Calculate a 99% confidence interval (i.e. a two-sided confidence interval) for the mean frame rate of the game on the YBox. Copy and paste the appropriate R command and output. In an additional line, clearly state the values you get for the confidence interval from your output.

(e) Referring to your confidence interval in (d) (and not to parts (a)-(c)), is it reasonable to assume that the mean frame rate is 110 FPS? Explain in a short sentence.

Part 2 [9 marks] In a certain municipality, a city counsel follows an informal rule that they will only approve large events like multi-day festivals if more than 60% of the local population approve of the event. This year, two possible such events have been proposed: Springfest and Autumnfest.

- ▷ For Springfest, city workers survey 116 voters, and find that 78 are in favour of the event.
- ▷ For Autumnfest, city workers survey 182 voters, and find that 110 are in favour of the event.

(a) Is there evidence to suggest that the proportion of voters in favour of Springfest exceeds 60%? Conduct the appropriate hypothesis test:

- Using the correct notation, define the parameter(s) of interest in this study.
- State the null and alternative hypotheses in terms of the parameter(s).
- Copy and paste the appropriate R command and output for the hypothesis test into your document.

(b) Clearly state the value of the test statistic and the p-value you found from part (a). Using the significance level $\alpha = 0.05$, should we reject the null hypothesis?

(c) Write a sentence stating your conclusion of the hypothesis test you just performed. (i.e. Write a plain language conclusion regarding Springfest in plain language so that a non-statistician could understand the conclusion of the test.)

(d) Calculate a 90% confidence interval (i.e. a two-sided confidence interval) for the true proportion of municipal voters in favour of Autumnfest. Copy and paste the appropriate R command and output into your document. In an additional line, clearly state the values you get for the confidence interval from your output.

(e) According to your confidence interval in part (d), is it reasonable to assume that only 50% of voters are in favour of Expansion B? Explain using your confidence interval.

Part 3 **[10 marks]** An study from 2010 compared the daily number of cigarettes consumed by smokers from England and Scotland. A random sample of that data for 29 Scottish smokers and 48 English smokers is given below:

Daily Cigarettes Consumed by Scottish Smokers:

5	5	10	12	20	10	2	15	15	20
20	15	15	15	20	16	18	20	15	25
60	5	10	15	10	10	10	20	3	

Daily Cigarettes Consumed by English Smokers:

4	24	19	2	15	5	16	10	24	26
21	28	7	22	20	27	20	27	11	1
29	27	19	10	17	23	13	39	6	24
11	16	20	18	16	46	2	18	20	6
13	12	12	16	17	12	16	13		

(a) Store the data for the samples in two appropriately named vectors (we recommend using the `scan()` function to avoid transcription errors). Calculate the sample standard deviations of the two groups. Copy and paste your commands and output for the standard deviations into your document. (You don't need to copy/paste the commands for creating the vectors).

(b) Determine if you can assume the variances are equal between the two groups. Complete the necessary calculation in R, and then copy and paste the ratio of the standard deviations into your document. What do you assume about the variances?

(c) Test the research hypothesis that the mean number of cigarettes consumed daily by is not the same for English and Scottish smokers.

- Using the correct notation, define the parameter(s) of interest in this study.
- State the null and alternative hypotheses in terms of the parameter(s).
- Copy and paste the appropriate R command and output for the hypothesis test into your document.

(d) Clearly state the observed value of the test statistics and the p-value you found from part (c). At significance level $\alpha = 0.10$, should we reject H_0?

(e) Write a plain language sentence stating your conclusion of the hypothesis test you just performed; that is, write your conclusion regarding the difference in mean daily cigarettes consumed by Scottish and English smokers.

(f) Calculate a 80% confidence interval (i.e. a two-sided confidence interval) for the mean difference (English-Scottish) in daily cigarettes consumed by English and Scottish smokers. Copy and paste the appropriate R command and output. In an additional line, clearly state the values you get for the confidence interval from your output.

⋆ The following is an optional bonus question ⋆

Part 4 (OPTIONAL) [10 marks]

Background: *In this question, we will simulate 500 repetitions of an experiment for random variables $X_1, X_2, \ldots, X_{500}$ that follows an exponential distribution. Then, we will compare the histograms and the means for the X_i (for $i = 1, 2, \ldots, 500$) versus the distribution of the sampling mean $\overline{X}$. You may need to refer back to your* Intro to R-Assignment 2 *PDF for a refresher on simulating data.*

Scenario: The lifespan of a certain variety of wireless headset is known to be exponentially distributed, with a mean lifespan of 6.2 years. Suppose that a large tech company purchases 500 of the headsets. We will use R to simulate the lifespans of the 500 headsets.

(a) In R, simulate the 500 headset lifespans, and save that data in a vector named `simHeadset`. Copy and paste only your R code into the document (do not paste in the 500 outcomes). Create a histogram for the simulated 500 lifespans in `simHeadset` (don't forget to include an appropriate title and labels for the axes). Then, use R to determine the **mean** of the sample in `simHeadset`.

(b) *Carefully read the* ***Making Vectors with for-loops*** *section of the* ***Introduction to R Assignment 3*** *before attempting this.*

In R, **initialize** a vector with 300 numeric entries named `simMean`. Using a **for-loop**, fill the entries of `simMean` so that each entry is the **sample mean** of 500 randomly generated headset lifespans. Thus, the first entry in `simMean` will be the **mean** lifespan of the first 500 simulated heatsets, the second entry will be the **mean** lifespan of the second 500 simulated headsets, etc, up to the 300th entry.

Copy and paste only your R code into the document (do not paste the outputs).

Note: each of the 300 entries in ***`simMean`*** *contain a sample mean that was found by generating 500 headset lifespans. Thus, in total (in the background), you will have generated* $300 \times 500 = 150,000$ *headset lifespans.*

Create a histogram for the simulated 300 sample mean lifespans in `simMean` (don't forget to include an appropriate title and labels for the axes). Then, in R, determine the **mean of the sample** in `simMean`.

(c) In a brief sentence, compare the shape (including symmetry/asymmetry) of the histograms from (a) and (b). Do they resemble other distributions? Are they similar to one another? Why or why not? (reference a specific result/theorem from lecture).

(d) In a brief sentence, compare the means found in (a) and (b). Are they similar? Why / why not would you expect them to be so? (reference a specific result/theorem from lecture).