

Stat 21 Homework 1

replace_this_with_your_name

Due: Monday, Jan 31 by midnight

This assignment is due on to be submitted on Gradescope by **midnight** on **Monday, Jan 31**. Use this file as the template for your submission. Do not delete anything from this template unless you are prompted to do so (e.g. where to write your name above, where to write your solutions or code below). Make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `knitr` **before** you attempt to knit this document.

Your completed assignment should be submitted as a single **PDF** using the link under Week 2 titled “Submit HW 1 to Gradescope”. You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and the requested answer/output. You are allowed to work with your classmates on this homework assignment but you are expected to write up your own solutions. Every answer must be supported by a written statement unless otherwise specified.

Part I: Non-coding problems

Problem 1

Elena is selling a TV at a cash auction and also intends to buy a toaster oven in the auction. If X represents the profit for selling the TV and Y represents the cost of the toaster oven, write an equation that represents the net change in Elena’s cash.

Solution Problem 1: [Write your solution here]

Problem 2

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.

		Health Status					
Health Coverage	No	Excellent	Very good	Good	Fair	Poor	Total
	Yes	459	727	854	385	99	2542
Total		4198	6245	4821	1634	578	17476
		4657	6972	5675	2019	677	20000

- (a) If we draw one individual at random, what is the probability that the respondent has excellent health and doesn’t have health coverage?
- (b) If we draw one individual at random, what is the probability that the respondent has excellent health or doesn’t have health coverage?

Solution Problem 2:

- (a) [Write your solution to part (a) here]
- (b) [Write your solution to part (b) here]

Problem 3

Below is a stem-and-leaf plot for the profits (as percent of sales) for 29 different corporations in the US. The stems are split so that each stem represents a span of 5%. Thus the smallest observation is a loss of 9% and the largest observation is a gain of 25%. As another example, -0|3 is interpreted as a a loss of 3%.

```
-0|9 9
-0|1 2 3 4
0|1 1 1 1 2 3 4 4 4
0|5 5 5 5 6 7 9
1|0 0 1 1 3
1|
2|2
2|5
```

- (a) Find the minimum, lower 25% quantile, median, lower 75% quantile, and the maximum of these profits. (These values are collectively referred to as a “5-number summary” of the data.) If you do these calculations by hand, you must attach a picture of your work showing every step to your final homework document.
- (b) Calculate the mean, variance, and standard deviation of these profits. (You may use R for these calculations in the space below or you may use another calculator. Regardless, make sure you show your work. If you do these calculations by hand, you must attach a picture of your work showing every step to your final homework document.)
- (c) Describe the distribution of profits for these corporations in words. Remark on things like symmetry and modality.

Solution Problem 3:

- (a) [Write your solution to part (a) here.]
- (b) [Write your solution to part (b) here.]

Replace this line with code if you wish to use R to calculate your answers for part (b). Otherwise, .

- (c) [Write your solution to part (c) here.]

Problem 4

The Central Limit Theorem essentially states that: “The mean of a random sample of data has a sampling distribution whose shape can be approximated by a Normal model and that the larger the sample is, the better the approximation will be.” What does the term *sampling distribution* refer to? (You may want to do a quick internet search for this term to help inform your answer.) Respond in no more than 5 sentences.

Solution Problem 4: [Write your solution here]

Problem 5

In a large class of introductory Statistics students, the professor has each person toss a coin 16 time and calculate the proportion of each person’s tosses that were heads. The students then report their results, and the professor plots a histogram of these several proportions.

- (a) What shape would you expect this histogram to be? Why?
- (b) Where do you expect the histogram to be centered?

- (c) How much variability would you expect among these proportions?
- (d) Explain why a Normal model should **not** be used here.

Solution Problem 5:

- (a) [Write your solution to part (a) here.]
- (b) [Write your solution to part (b) here.]
- (c) [Write your solution to part (c) here.]
- (d) [Write your solution to part (d) here.]

Problem 6

Census data for a certain country shows that 19% of the adult residents are Latinx. Suppose 72 people are called for jury duty and only 9 of them are Latinx. Does this apparent under-representation of Latinx jurors call into question the fairness of the jury selection system. Explain your answer with statistical reasoning. (You do not need to evaluate a test or calculate an interval.)

Solution Problem 6: [Write your solution here.]

Part II: R coding problems

Problem 7

Review this [cheat-sheet](#) made by a former student on an introduction to R.

Solution Problem 7: [Confirm you have read through this document here. (completion)]

Problem 8

A company with a fleet of 150 cars found that the emissions systems of 7 out of the 22 they tested failed to meet pollution control guidelines. Is this strong evidence that more than 20% of the fleet might be out of compliance? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.

Solution Problem 8: [Write your solution here.]

Uncomment this line and put any r-code you used for your solution here

Problem 9

It is widely believed that regular mammogram screening may detect breast cancer early, resulting in fewer deaths from that disease. One study that investigated this issue over a period of 18 years was published during the 1970s. Among 30,565 people with breast tissue who had never had mammograms, 196 died of breast cancer, while only 153 of 30,131 who had undergone screening died of breast cancer.

Do these results suggest that mammograms may be an effective screening tool to reduce breast cancer deaths? Use appropriate statistical methods to support your answer.

Solution Problem 9: [Write your solution here.]

Uncomment this line and put any r-code you used for your solution here

Problem 10

In July of 2004, the Gallup Poll asked 1005 US adults if they actively try to avoid carbohydrates in their diet. That number increased to 27% from 20% in a similar 2002 poll of the same sample size. Is this what

statisticians would call a “statistically significant” increase? Use either a difference in proportions test or CI to justify your answer.

Solution Problem 10: [Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Optional exercise

If you would like some extra practice getting used to coding basics in R, install the R package called *swirl* in RStudio by navigating to ‘Tools -> Install Packages’ and setting “Install From” to “Repository (CRAN)”, typing in “swirl” under “Packages”, and checking the box “install all dependencies”. Next, call this package into your working library by typing `library("swirl")` in the R console window. Follow the prompts that appear in the console. Select the course option “1: R Programming: The basics of programming in R” and then type in the course option “1”. Complete the following lessons:

- 1: Basic Building Blocks
- 2: Workspace and Files
- 3: Logic

Once you have completed the above lessons you can exit the tutorial by typing `bye()` into the R console.