Stat 21 Homework 1

Solutions

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: X - Y

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					
		Excellent	Very good	Good	Fair	Poor	Total
Health	No	459	727	854	385	99	2542
Coverage	Yes	4198	6245	4821	1634	578	17476
	Total	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: [note there is a typo in the table above, calculations are done based on this typo]

(a) probability of intersection events: 4459/20000

(b) probability of union: 4657/20000 + 2542/20000 - 4459/20000

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: [note there was a typo in the class worksheet associated with this problem importing the data, an extra data point of 1.0 was accidentally included; either version of the calculations below get full credit]

```
stem_data <- c(-0.09, -0.09, -0.01, -0.02, -0.03, -0.04,
               0.01, 0.01, 0.01, 0.01, 0.02, 0.03, 0.04, 0.04, 0.04,
               0.05, 0.05, 0.05, 0.05, 0.06, 0.07, 0.09,
               0.10, 0.10, 0.11, 0.11, 0.13, 0.22, 0.25)
summary(stem_data)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.09000  0.01000  0.04000  0.04724  0.09000  0.25000
```

```
var(stem_data)
```

```
## [1] 0.005699261
```

```
sd(stem_data)
```

```
## [1] 0.07549345
```

```
stem_data_typo <- c(-0.09, -0.09, -0.01, -0.02, -0.03, -0.04,
               0.01, 0.01, 0.01, 0.01, 0.02, 0.03, 0.04, 0.04, 0.04,
               0.05, 0.05, 0.05, 0.05, 0.06, 0.07, 0.09,
               0.10, 0.10, 0.11, 0.11, 1.0, 0.13, 0.22, 0.25)
summary(stem_data_typo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## -0.0900  0.0100  0.0450  0.0790  0.0975  1.0000
```

```
var(stem_data_typo)
```

```
## [1] 0.03576103
```

```
sd(stem_data_typo)
```

```
## [1] 0.1891059
```

(a) must describe how to find quantiles generally... put data in order, etc.

(b) may use R or write out first few terms of sums for these values

(c) unimodal, roughly symmetric, slight right skew

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: sampling distribution describes how much some function of the data (or a random sample) may vary from one random sample to the next; have normal sampling distributions for proportions and means.

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) histogram will take values between 0 and 1 but with empty spaces, large gaps.

(b) centered around $p = 0.5$

(c) variance of $p(1-p) = 0.25$

(d) the sample size is $n = 16$, even though experiment is being repeated as many number of times as the experiment is being conducted

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: answers may vary but should describe valid statistical reasoning such as determining if observation is due to random chance or sampling variability.

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: there is a problem with the necessary assumptions for this test. the sample size is 22 which represents over 10% of the size of the population implying that the assumption of an independent sample is violated. also, there are only 7 observed "successes" and to use the CLT we would need at least 10 observed successes and at least 10 observed failures each. since there is no way to remedy *both* of these issues at once, a different procedure should be used besides the one-sample z-test. for the sake of completion however, the code for calculating the p-value of the one-sample z-test is presented below. *Note that these p-values are very different (although we'd expect them to be the same). this is because R's default in the prop.test() function makes an adjustment - called a continuity correction - for the fact that the number of observed successes is too small. You don't need to worry about this and either of the above methods are valid for our purposes.*

```
prop.test(x=7, n=22, p=0.2, alternative="two.sided")$p.val
```

```
## [1] 0.2630106
```

```
## or equivalantly could do the calculations by hand:
n = 22
phat = 7/n
2*pnorm((phat - 0.2)/sqrt(0.2*0.8/n), mean=0, sd=1, lower.tail = FALSE)
```

```
## [1] 0.1658066
```

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: Code below tests the hypothesis that $H_0 : p_1 - p_2 = 0$ vs $H_A : p_1 - p_2 > 0$ where p_1 = probability of death from breast cancer among people without mammograms and p_2 = probability of death from breast cancer among people with mammogram screening

```
prop.test(x=c(196, 153), n=c(30565, 30131), alternative="greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(196, 153) out of c(30565, 30131)
## X-squared = 4.4977, df = 1, p-value = 0.01697
## alternative hypothesis: greater
```

```
## 95 percent confidence interval:
##  0.0002930113 1.0000000000
## sample estimates:
##      prop 1      prop 2
## 0.006412563 0.005077827
```

```
### or equivalently, could do the following calculations "by hand" (the small difference is due to roun
n1 = 30565
phat1 = 196/n1
n2 = 30131
phat2 = 153/n2
phat_pooled = (196+153)/(30565+30131)
pnorm((phat1 - phat2)/sqrt((phat_pooled)*(1-phat_pooled)*((1/n1) + (1/n2))), mean=0, sd=1, lower.tail=F
```

```
## [1] 0.01483482
```

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: the code below considers $p_1 - p_2$ where p_1 = proportion of adults avoiding carbs in 2004 and p_2 = proportion of adults avoiding carbs in 2002. For a hypothesis test, the p-value should be for a one-sided test (0.00013); a 95% CI would be [0.0678, 0.072]

```
prop.test(x=c(1005*.27, 1005*.20), n=c(1005, 1005), alternative="greater")$p.val
```

```
## [1] 0.0001320198
```

```
## or find a two-sided CI with
prop.test(x=c(1005*.27, 1005*.20), n=c(1005, 1005), conf.level = 0.05)$conf.int
```

```
## [1] 0.06782295 0.07217705
## attr(,"conf.level")
## [1] 0.05
```

Optional exercize

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