INFX 573: Problem Set 4 - Statistical Theory

Shuyang Wu

Due: Tuesday, November 1, 2016

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset4.Rmd` file from Canvas. Open `problemset4.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset4.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps4.Rmd`, knit a PDF and submit the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
```

Problem 1: Triathlon Times

In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the Men, Ages 30 - 34 group while Mary competed in the Women, Ages 25 - 29 group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups.

Here is some information on the performance of their groups:

- The finishing times of the Men, Ages 30 - 34 group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the Women, Ages 25 - 29 group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

(a) Write down the short-hand for these two normal distributions.

N1(=4313, =583), N2(=5261, =807) ##### (b) What are the Z scores for Leo's and Mary's finishing times? What do these Z scores tell you?

```
# Calculations
(4948 - 4313)/583
```

```
## [1] 1.089194
```

```
(5513 - 5261)/807
```

```
## [1] 0.3122677
```

```
1-0.8643
```

```
## [1] 0.1357
```

```
1-0.6176
```

```
## [1] 0.3824
```

Z1=(x−)/ = (4948 - 4313)/583 = 1.089194 Z2=(x−)/ = (5513 - 5261)/807 = 0.3122677 Z scores can be used to identify which observations are more unusual than others. Leo's score is 1.089 standard deviation above the mean, whereas Mary's score is 0.312 standard deviation above the mean, thus Leo's score is more unusual. ##### (c) Did Leo or Mary rank better in their respective groups? Explain your reasoning. Mary rank better in her respective group because her score is closer to the mean than leo's score and their z scores are both positive which means the time they took to finish the races are both above the mean. And in races, time is shorter the better. Mary's z-score is less above the mean comparing to Leo's, thus Mary ranked better. ##### (d) What percent of the triathletes did Leo finish faster than in his group? 1-0.8643 = 0.1357 x 100% = 13.57% ##### (e) What percent of the triathletes did Mary finish faster than in her group? 1-0.6176 = 0.3824 x 100% = 38.24% ##### (f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning. If the distribution is not normal the Z statistic will not be normally distributed. So the percentiles of Z will not be standard normal. Thus we can't use z score to find out how likely the observations were to occur or where they each rank in their respective groups. We can't calculate the percent below or above the observation either. Thus, my answer for (b) would be the same cause I can calculate the z scores, but they don't mean anything nor can I answer the other questions.

Problem 2: Sampling with and without Replacement

In the following situations assume that half of the specified population is male and the other half is female.

```
# Calculations
5/10 * 5/10
```

```
## [1] 0.25
```

```
5/10 * 4/9
```

```
## [1] 0.2222222
```

```
5000/10000 * 5000/10000
```

```
## [1] 0.25
```

```
5000/10000 * 4999/9999
```

```
## [1] 0.249975
```

(a) Suppose you're sampling from a room with 10 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement?

Sampling with replacement: P1 = 5/10 x 5/10 = 1/4 = 0.25 Sampling without replacement: P2 = 5/10 x 4/9 = 2/9 = 0.22 ##### (b) Now suppose you're sampling from a stadium with 10,000 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement? Sampling with replacement: P1 = 5,000/10,000 x 5,000/10,000 = 1/4 = 0.25 Sampling without replacement: P2 = 5,000/10,000 x 4,999/9,999 = 0.249975 ##### (c) We often treat individuals who are sampled from a large population as independent. Using your findings from parts (a) and (b), explain whether or not this assumption is reasonable. This assumption is reasonable because as the population grows larger, the difference of sampling with and without replacement becomes minimal, like in (b) where P2 is 0.249975 which is really close to 0.25. Thus treating individuals as independent does not affect the significance of testing when sampling from a large enough population.

Problem 3: Sample Means

You are given the following hypotheses: $H_0 : \mu = 34$, $H_A : \mu > 34$. We know that the sample standard deviation is 10 and the sample size is 65. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

```
# calculations
10/sqrt(65)
```

```
## [1] 1.240347
```

```
(1.65*1.24)+34
```

```
## [1] 36.046
```

When P value = 0.05, z score is equal to 1.65 from the probability table. The standard deviation of the sampling distribution of the mean is 10/sqrt(65) = 1.24, plug into the p-value formula, sample mean when p value equals to 0.05 is (1.65*1.24)+34 = 36.05.