IMT 573: Problem Set 4 - Statistical Theory

Naga Soundari Balamurugan

Due: Tuesday, October 30, 2018 at 11:59AM

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. If you are using more than just a standard function that you found from another source, please credit the source in the comments. For example:
4. Collaboration on problem sets is acceptable, and even encouraged, but students must turn in an individual write-up in their own words and their 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` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_ps4.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

Setup:

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

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

Problem 1: Triathlon Times

In triathlons, it is common for racers to be placed into age and gender groups. Fred and Catarina both completed the Hermosa Beach Triathlon, where Fred competed in the Men, Ages 30 - 34 group while Catarina competed in the Women, Ages 25 - 29 group. Fred completed the race in 1:22:28 (4948 seconds), while Catarina completed the race in 1:31:53 (5513 seconds). 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.

Men Triathlete's group: x ~ N(4313, 583) Women Triathlete's group: x ~ N(5261, 807)

(b) What are the Z scores for Freds and Catarinas finishing times? What do these Z scores tell you?

```r
#Fred's Z-score
zscore_Fred <- (4948 - 4313) / 583
zscore_Fred
```

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

```r
#Catarina's Z-Score
zscore_Catarina <- (5513 - 5261) / 807
zscore_Catarina
```

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

The Z-Score of any observation gives us the number of standard deviations it falls above/below the mean. Thus, seeing the positive z-scores, we can know that both Fred and Catarina did not perform better in their corresponding groups (since smaller the time, better the performance, the curve is to be read in reverse). Fred is 1.089 standard deviation below the mean and Catarina is 0.31 standard deviation below the mean.

(c) Did Fred or Catarina rank better in their respective groups? Explain your reasoning.

```r
pnorm_Fred <- pnorm(4948, 4313, 583)
pnorm_Fred
```

```
## [1] 0.8619658
```

```r
pnorm_Catarina <- pnorm(5513, 5261, 807)
pnorm_Catarina
```

```
## [1] 0.6225814
```

No, Both Fred and Catarina do not rank better in their respective groups. Based on the z-scores we calculated above, the p-value of an athlete running as fast or faster than Fred did is .86. Similarly, the p-value of an athlete jumping as fast or faster than Catarina did is .62. Between the both of them, Catarina performed better in her group.

(d) What percent of the triathletes did Fred finish faster than in his group?

Fred finished faster than 14%(1 - 0.86) of the triathletes in his group.

(e) What percent of the triathletes did Catarina finish faster than in her group?

Catarina finished faster thar 38%(1 - 0.62) of the triathletes in her group.

(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

Yes, the answers might change if the distributions are not nearly normal. Assume if the distribution is skewed to the left, then the mean might be smaller and the standard deviation also changes which in turn would impact the z-scores and percentile it falls into. It could be just the opposite case for a right skewed distribution with a higher mean.

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.

(a) Suppose you are 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?

```
sampWithRep <- (5/10) * (5/10)
sampWithRep
```

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

```
sampWithoutRep <- (5/10) * (4/9)
sampWithoutRep
```

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

> Sampling with replacement: Since half of the population is female, the probability would be 5/10 for selecting the first female and again the same as the selected female is replaced. Thus the probability of selecting two females in a row with replacement is **0.25**.
>
> Sampling without replacement: The probability of selecting a female in the first pick would be 5/10. As we aren't going to replace the one we have already picked, the total population would be 1 less as like the female count and the probability would be 4/9. Thus the probability of selecting two females in a row without replacement is **0.22222**

(b) Now suppose you are 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?

```
sampWithReplace <- (5000/10000) * (5000/10000)
sampWithReplace
```

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

```
sampWithoutReplace <- (5000/10000) * (4999/9999)
sampWithoutReplace
```

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

> Sampling with replacement: Since half of the population is female, the probability would be 5000/10000 for selecting the first female and again the same(5000/10000) as the selected female is replaced. Thus the probability of selecting two females in a row with replacement is **0.25**.
>
> Sampling without replacement: The probability of selecting a female in the first pick would be 5000/10000. As we aren't going to replace the one we have already picked, the total population would be 1 less as like the female count and the probability would be 4999/9999. Thus the probability of selecting two females in a row without replacement is **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.

> Yes, we treat the individuals sampled from the large population as independent as the sample size is comparatively very similar to the entire population. As from the part (b), we can see that the both are probability with and without replacement are around 0.25 which is not the case in part (a). Thus it makes a very little difference if the sampling is done with or without replacement for a larger 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.

> As the sample mean to be found for p-value equal to 0.05, from the normal probability table, we know that the z-score is 1.65. The formula that connects the z-score with sample mean is z = (xi - x)/se where x is the mean, z is the z-score, xi is the sample mean and se is the standard error. We have all the values to be substituted in this formula except standadard error.
>
> Standard error could be calculated from se = standard deviation/square root of n (from the slide 34 of week 4a).

```r
se <- 10/sqrt(65)

sampleMean <- (1.65 * se) + 34
sampleMean
```

```
## [1] 36.04657
```

> Thus the **sample mean is 36.04** for the p-value 0.05.