INFX 573: Problem Set 4 - Statistical Theory

Amol Surve

Due: Tuesday, November 1, 2016

Collaborators: Abhishek Gupta, Shreya Jain

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.

Normal distribution

Leo: N(mu=4313, std deviation=583)

Mary N(mu=5261, std deviation=807)

(b) What are the Z scores for Leo's and Mary's finishing times? What do these Z scores tell you?

Z Score can be calculated by finding the difference between the value and the mean and then dividing it by standard deviation.

```
leo<-4948 #Leo's time

mary<-5513 #Mary's time

men<-4313 #Men mean

women<-5261 #Women's mean

sd1<-583 #Men's standard deviation

sd2<-807 #Women's standard deviation

zleo<-(leo-men)/sd1 #Calculating z score for Leo

zmary<-(mary-women)/sd2 #Calculating z score for Mary

zleo #Printing z score for Leo
```

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

```
zmary #Printing z score for Mary
```

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

Leo has a Z score of 1.08 i.e. Leo's time is 1.08 standard deviations away from the mean.

Mary has a Z score of 0.31 i.e. Mary's time is 0.31 standard deviations away from the mean.

Since Mary is having lesser z score which means less time, it is what makes the decision in the race. Hence, Mary performed better compared to Leo w.r.t age and gender group.

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

In the race, performance is inversely proportional to time. Thus, less finishing time indicates better performance. Mary and Leo are both 0.31 and 1.08 standard deviations below the mean but Mary's performance is better compared to Leo since she is closer to the mean in her age and gender group.

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

```
LeoPercentage<-(1-pnorm(zleo))*100 #Calculating percentage of atheles Leo finished faster than

LeoPercentage #Print the percentage
```

```
## [1] 13.80342
```

Leo finished faster than 13.8% of the atheletes in the group.

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

```
MaryPercentage<-(1-pnorm(zmary))*100 #Calculating percentage of atheles Mary finished faster than

MaryPercentage #Print the percentage
```

```
## [1] 37.74186
```

Mary finished faster than 37.74% of the atheletes in the group.

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

Not having normal distributions in our data would affect the z scores and it also will not be normally distributed. It doesn't accurately tell whether the standard deviations are accurately distributed above or below the mean and hence, that would affect the performance measures of Lep and Mary w.r.t. thier groups. Hence, the answers to parts (b) and (e) would have become different due to change in the newly calculated z score which is not normally distributed.

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 youre 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?

When we are sampling with replacement, the number of possible sammples of size k from a population of size n is n^k. The order matters as we want them in a row. As we calculate it, we get the probability as 0.25

```
Fout<-5^2 #favorable outcomes

Tout<-10^2 #total outcomes

Probability<-Fout/Tout #Probability

Probability #Printing it
```

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

Without Replacement, population reduces by 1 which will be (5/10)*(4/9)

```r
ProbNew<-(5/10)*(4/9) #probability

ProbNew #Printing it
```

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

The probabilty is 0.22

(b) Now suppose youre 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?

When we are sampling with replacement, the number of possible sammples of size k from a population of size n is n^k. The order matters as they are in a row. As we calculate it, we get the result at 0.25

```r
w<-5000^2 #favorable outcomes

wt<-10000^2 #total outcomes

Probw<-w/wt #Probability

Probw #Print it
```

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

Without Replacement, population reduces by 1 which will be (5000/10000)*(4999/9999)

```r
ProbWR<-(5000/10000)*(4999/9999) #Calculating the probability

ProbWR #Printing the result
```

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

The probabilty is 0.249

(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.

For the population of 10 people - The difference between the probability of samppling with and without replacement is 0.03.

For the population of 10000 people - The difference between the probability of samppling with and without replacement is 0.001.

Hence, as the population increases, the probability of sampling becomes less dependent of previous sampling. Hence, according to me, we can make an assumption that individuals sampled from a large population can be treated independently.

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.

Null Hypothesis- H_0, mu=34

Alternate Hypothesis- H_A. mu>34

The sample size- 65.

Z test will be appropriate to us since population is not available.

For a p-value of 0.05, the confidence interval is 95%. The Z statistic value using qnoem() function is 1.644.

The mean calculations:

```
sMean=34 #Assigning sample mean

SD=10 #Assigning standard deviation

n=65 #Assigning sample size

z=qnorm(0.95) #Z score based on 95% confidence interval

SampleMean=(z*SD)/sqrt(n)+sMean #Calculating sample mean

SampleMean
```

```
## [1] 36.04019
```

We get an answer of 36.04. So we can conclude that for a sample mean of 36.04 we will get a p-value of 0.05.