INFX 573: Problem Set 4 - Statistical Theory

Pierre Augustamar

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.

(a) N(μ = 4313, sd = 583) for Leo
(b) N(μ = 5261, sd = 807) for Mary

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

```
# Load standard libraries
 ZscoreLeo = (4948 - 4313) / 583
 ZscoreMary = (5513 - 5261) / 807

ZscoreLeo
```

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

```
ZscoreMary
```

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

Response

A z-score measures how many standard deviations a data point is away from the mean. To tha end, we can conclude that Mary is 0.31 standard deviation above the mean. In addition, Leo is 1 standard deviation above the mean.

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

In most scenarios the further away a standard deviation is from the mean, the better. So in this case we would have thought that Leo would have a better ranking than Mary since his standard deviation is much greater than Mary. However, because this Z score is about the quickest time a triathlete can finish a competion, than the further away the standard deviation is below to the mean the better. Thus, in this scenario Mary will rank better because she is further below the average running time than Leo's in her grouop.

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

```
# Load standard libraries
leoAbovePercent = 1 - pnorm(ZscoreLeo)
leoAbovePercent
```

```
## [1] 0.1380342
```

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

```
# Load standard libraries
MaryAbovePercent = 1 - pnorm(ZscoreMary)
MaryAbovePercent
```

```
## [1] 0.3774186
```

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

For part b, the Z-score value will not be different because the Z-score does not depend on a specific distribution. As long as, a mean,and standard deviation are known then a Z-score can easily be calculate.

For part e, the percentiles will not change. This is because a percentiles mainly identifies what percent of a specific condition is above or below a percentiles.

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'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?

a) Sampling with replacement With replacement implies that an observation can be selected more than once. But in the case of selecting a person, he or she can only be selected once teven with replacemen. In other words, after being selected a person is removed from the selection pool. That said, if we are selectig two distinct females, we cannot add the same person back in the pool again. We won't be truly selecting two disctinct females. We do now that selecting one female should be 5/10, and ideally if we were to add the same female again then the probability would have been 5/10*5/10 or .25. But the fact that we want two different females then the probability will only be 1/2

```
# Load standard libraries
pWithReplacement= 5/10
pWithReplacement
```

```
## [1] 0.5
```

```
# sample function to be used throughout since there are multiple questions that might utilize this.
comb = function(n, x) {
  return(factorial(n) / (factorial(x) * factorial(n-x)))
}
```

Response

b) sampling without replacement

I followed the same principles as this tutorial in youtube at: https://www.youtube.com/watch?v=PHajAlsahW0 - up to the 4:46 time length. The instructor uses a combination model to find the probability.

Choosing two females out ten is the same as choosing 2 females or 3 males since the group is divided in group of 5. Thus, I calculated the combination of choosing two from ten from the entire set, and choosing the combination of two from five or the combination of three from five.

```r
# Load standard libraries

denominator = comb(10, 2) #choosing 2 people out of 10
denominator
```

```

```
[1] 45
```

```r
numerator = comb(5,3) + comb(5,2) # since there are 5 people per groups, choosing 2 females is really,
numerator
```

```
## [1] 20
```

```r
probabilityWithoutReplacement  = numerator / denominator
probabilityWithoutReplacement
```

```
## [1] 0.4444444
```

(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?

same as the previous example, selecting two different females means that we cannot add the same female back in the pool even though this is with replacement. Thus, I am using the same principiles I have used above and have a ratio of 5000/10000 or 5/10

```r
# Load standard libraries
pWithReplacement= 5000/10000
pWithReplacement
```

```
## [1] 0.5
```

b) sampling without replacement

Same as section b in the previous example, I would have calculated the Combination, but since factorial in R is failing for large numbers. And 10000 is considered a large number. I ended up simplifing my response based on the previous question posed above. The previous example came out to be the ratio of 9/49. Thus, if following the same principle as prvevious similar question, it will be the ratio of 4999/9999 for this example.

```r
pWithoutReplacement= 4999/9999 # ratio for choosing 2 females
pWithoutReplacement
```

```
## [1] 0.49995
```
```

(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, it is a reasonable assumption. As the size of the population gets bigger compare to the sampling size the difference between sampling with replacement and sampling without replacement tends to be close to zero.

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.

Response

With a p-value = 0.05 means we have 5% of being wrong if we reject the null hypothesis. Thus, we have 95% of confidence level. Given this confidence level, I can calculate the z score by using the qnorm function. Given the size and the standard deviation and the z score, I can calculate the margin of error using the standard error of the mean formula. knowing that it's a right tailed with mu > 34, i will take the upper bound and add the mean to the calculated margin of error. A value greater than 34 means that this is located beyond the mean.

```r
# calculate the standard error
zScore = qnorm(.95)
margineOfError = zScore * (10/sqrt(65))
SampleMean = margineOfError + 34

SampleMean
```

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