ANLY 500 Laboratory #1 (Part 4) – Descriptive Statistics

Evans Chapter 6

“Performance Lawn Equipment Case Study” from Evans, **Business Analytics**

Contents

[Introduction 1](#_Toc489866776)

[Chapter 6 2](#_Toc489866777)

[Part 1 2](#_Toc489866778)

[Step 1 2](#_Toc489866779)

[Step 2 3](#_Toc489866780)

[Step 3 4](#_Toc489866781)

[Step 4 5](#_Toc489866782)

[Step 5 6](#_Toc489866783)

[Step 6 8](#_Toc489866784)

# Introduction

This laboratory follows the exercises in the book, specifically the Performance Lawn Equipment Case Study Chapters 6, except this laboratory requires that you use R to complete the exercises. That is, you should answer all questions in the textbook exercises and when necessary to complete computations use R. Each laboratory in ANLY 500 will build on the laboratories you have completed before. So, you will want to set-up a folder or file to keep your work in so that you can refer back to previous laboratories if necessary.

In your report, you are required to answer the questions carefully using your analysis.

# Chapter 6

## Part 1

For Chapter 6’s Performance Lawn Equipment (PLE) you are tasked with conducting analyses to answer the following questions:

1. What proportion of customers rate the company with “top box” survey responses (which is defined as scale levels 4 and 5) on quality, ease of use, price, and service in the 2014 Customer Survey worksheet? How do these proportions differ by geographic region?
2. What estimates, with reasonable assurance, can PLE give customers for response times to customer service calls?
3. Engineering has collected data on alternative process costs for building transmissions in the worksheet Transmission Costs. Can you determine whether one of the proposed processes is better than the current process?
4. What would be a confidence interval for an additional sample of mower test performance as in the worksheet Mower Test?
5. For the data in the worksheet Blade Weight, what is the sampling distribution of the mean, the overall mean, and the standard error of the mean? Is a normal distribution an appropriate assumption for the sampling distribution of the mean?
6. How many blade weights must be measured to find a 95% confidence interval for the mean blade weight with a sampling error of at most 0.2? What if the sampling error is specified as 0.1?

### Step 1

First, as usual, you’ll need to import the data into R/RStudio. Don’t forget to pay attention to whether or not there are header rows in the file; and, whether or not there may be abbreviations that conflict with R/RStudio’s predefined symbols such as “NA”. If there are such conflicts be sure to either uncheck the “Strings as Factors” box in the Import window or use the parameter “stringsAsFactors=FALSE” in your R command to read.csv().

Count data, such as associated with Binomial and Poisson distributions, are used in frequencies and proportions. Be sure you understand these distributions and what constitute frequencies and proportions. In R/RStudio we can simply create tables that give us the counts, for example:

> countByRegion <- table(CustomerSurvey2014$Region, CustomerSurvey2014$Quality)

> countByRegion

1 2 3 4 5

China 0 1 2 5 2

Eur 0 1 6 12 11

NorthA 1 0 3 30 66

Pac 0 0 1 4 5

SA 1 0 4 24 21

Which gives us, for example, that the North American region has a total of 96 counts of Level 4 or 5 (“Top box” survey responses) for Quality. *You may find a command for this part.*

We can also sum by region (row) to see that there were 100 responses from the North American region. Then it is a simple calculation to find the proportion of “Top box” responses by region. We can also compute the proportion of all customers with “Top box” ratings, which is a simple calculation. You can use these simple sums to compute all the requested information. Of course there are many more sophisticated commands in R that will return or compute and save in a data object the values you want. If you are not used to working with loops, particularly in conjunction with data tables, matrices and/or arrays this simple example will help you sort out it all out by computing the various values required.

### Step 2

To determine how to respond to customers with regard to service response time we need to move more deeply into statistics. The first thing you will want to do before you begin any analysis is to thoroughly understand the question, what it is you are going to do. To completely answer this question we’ll need to estimate, with reasonable assurance, the response time for customer service calls. Therefore, the first thing is to establish what we mean by reasonable assurance. For this particular question, we can just use Evans’ confidence level, i.e. 95% for this. As we’ve seen from the textbook, that means that our level of significance, or α, is 0.05. For other analyses, you will want to make sure that you have really determined what is meant by “reasonable assurance”. As we have seen in other situations, what is reasonable often depends on details inherent to the specific situation.

Because we want to give customers a time range rather than an exact time for response, we will need to calculate the average time as well as the margin of error for the computed average time. Either you can calculate the Response Time for each quarter or, you can determine the answers for the overall data set.

First, we will have to import our data, ResponseTime.csv. Looking at the data using the str() command we see that we have 50 observations of eight variables, one variable for each quarter over a two year period. We can use the entire 50 observations per quarter to calculate the mean and standard deviation per quarter as follows:

> Q1.13mean <- mean(ResponseTime$Q1.2013)

> Q1.13std <- sd(ResponseTime$Q1.2013)

> Q1.13mean

[1] 3.9152

> Q1.13std

[1] 1.482202

where mean() and sd() are the functions required to find the mean and standard deviation of the specified variable.

In R/RStudio, we can use the Basic Statistics and Data Analysis (BSDA) package to find the confidence interval given the standard deviation and knowing that 95% is the default confidence interval. Don’t forget to use library() to attach BSDA after you have installed the package. *Compute the z-test for first and second quarter.*

We didn’t need to compute the mean separately first. The z.test returns the mean as part of the output with the 95% confidence interval. Also, it is very important to note that the output returns the P value indicating that the analysis is statistically significant in both quarters. *Use the result of z-test to answer the question.*

### Step 3

Now you are asked to evaluate the proposed processes for building transmissions and compare those to the current process. The question you are asked to answer is whether or not you can determine if one of the proposed processes is better than the current process.

The data you have to do this with is the TransmissionCosts.csv data. You’ll need to import it and look at it. The data look straightforward with 30 observations of the three variables about the processes; current, Process.A, and Process.B. Again we use a 95% confidence interval for each process then all can be compared. So, we can use the same process as we did for the last question as follows:

> current.sd <- sd(TransmissionCosts$Current)

> z.test(TransmissionCosts$Current, sigma.x = current.sd)

One-sample z-Test

data: TransmissionCosts$Current

z = 34.939, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

273.3542 305.8458

sample estimates:

mean of x

289.6

*You’ll need to complete the computations for all three processes and include your comparison in your lab report.*

### Step 4

This question asks what the confidence interval would be for an additional sample of mower test performance. This involves a bit more than the last two questions. Now we are asked for the confidence interval for an additional sample of mower test performance. Because we do not have data for the entire population, i.e. all tests on all mowers, we cannot calculate the standard deviation for the entire population of mowers. We need to use the t-distribution.

We started looking at the Mower Test data in Chapter 5. There we found that the mean fraction of failures were 54/3000 or 0.018. If we wanted to we could use the mean to find the standard deviation using the formula:

This gives us the value 0.0024273..., which we can round to 0.0024. But, since we cannot compute the mean of the entire population this doesn’t help us a lot here.

The t.test() function parameters we will need include the mower test data and the mean, which we already have. But before we can execute the t.test() function we will need to get the mower test data in the proper format. There are many ways you can do this. What you want to have in the end is essentially a single vector with all the data in it. You could play around with unlist() and other functions. :

1. Convert the Pass/Fail levels to numeric values

> A <- data.frame(lapply(MowerTest, function(x) as.numeric(x)))

> str(A)

'data.frame': 100 obs. of 31 variables:

$ Observation: num 1 2 3 4 5 6 7 8 9 10 ...

$ Sample.1 : num 2 2 2 2 2 2 2 2 2 2 ...

$ Sample.2 : num 1 1 2 2 2 2 2 2 2 2 ...

$ Sample.3 : num 2 2 2 2 2 2 2 2 2 2 ...

$ Sample.4 : num 1 1 1 1 1 1 1 1 1 1 ...

$ Sample.5 : num 2 2 2 2 2 2 2 2 2 2 ...

1. If you check the str() of the database, you will see that those columns which only have one state (either fail or pass), are not correctly converted. For example, Sample.4 looks like it only contains “Fail” tests. You may fix this by assigning the correct values manually. The columns are 5, 10, 27 and 30. So, you may write:

> A[,5] <- 2

1. Then correct the data so that Pass = 1 and Fail = 0 (rather than Pass = 2 and Fail = 1) and call the result B.
2. To check your data you may check the total number of failures which is:

> length(which(B == 0))

[1] 55

Find the mean of B to use it in t.test computation. If you use directly mean() command, you will see an error because part of the database are not numbers. *It is your responsibility to find a way to get rid of those extra parts.*

> mean(B)

[1] 0.982

Now use t.test command and you should see results as below:

One Sample t-test

data: B

t = 0, df = 2999, p-value = 1

alternative hypothesis: true mean is not equal to 0.982

95 percent confidence interval:

0.9772398 0.9867602

sample estimates:

mean of x

0.982

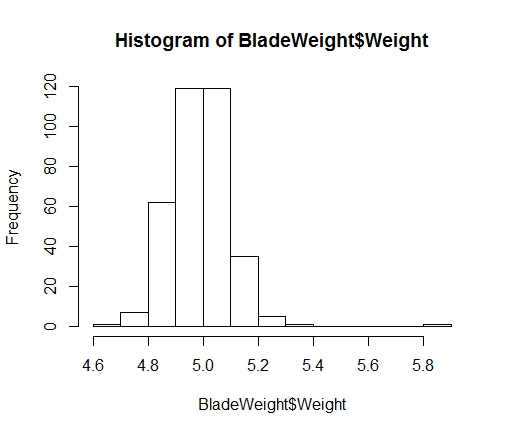
Having completed the analysis this way there is one last thing to do to get the final answer. The output from the t.test() function is the answer for “Pass”. Keeping in mind that we want the interval representing the failures so *compute it and report the values here.*

## Step 5

If you have exited R/RStudio and need to you should import the data set BladeWeight.csv again. There are several questions that we need to answer to complete this. First, we need to determine what the sampling distribution of the mean is. This is pretty easy to do using the hist() function. Simply enter:

> hist(BladeWeight$Weight)

and the plot reveals:



which appears to be a normal distribution. The mean is easy to compute as follows:

> mean(BladeWeight$Weight)

[1] 4.9908

*Compute the standard error of the mean for our blade weight data using given formula:*

> sem

[1] 0.005841666

If we want to determine if the normal distribution is an appropriate assumption one thing we can do is look at a normal probability plot also called a Q-Q plot. Remember that a normal probability plot is a routine way of testing normality in statistics. To make this easy I’m going to apply a linear model to the blade weight data and use R/RStudio’s built in plotting functions to create the normal probability plot. So:

> A <- lm(BladeWeight$Weight ~ ., BladeWeight)

where I established a data object “A” for the output from the lm() function, i.e. the linear model function in R/RStudio. Assuming the data are normal and linear or close to linear that should give us a good fit on a normal probability plot. So I:

> plot(A)

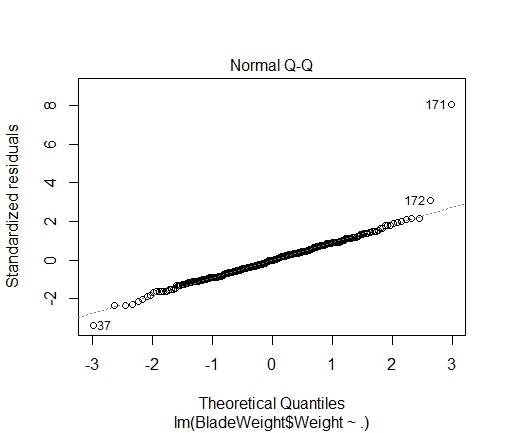
Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

and since I’m most interest in the normal probability plot here it is:



*How do you think now? Is it good? Explain your findings.*

## Step 6

You are asked to find the answers to two questions in this last part of Chapter 6, i.e. “How many blade weights must be measured to find a 95% confidence interval for the mean blade weight with a sampling error of at most 0.2? What if the sampling error is specified as 0.1?” The equation to determine n for a given sampling error is in the textbook. We had already found that the average blade weight was 4.99 and the standard deviation was 0.11. To apply this equation we need to determine the value of z for the specified level of significance, . But, in actuality the calculation for this value is well understood and the value is already known. It is 1.96. If this doesn’t sound familiar be sure to review because this is one of a set of standard values that will very likely appear on exams.

*Compute n, the number of samples required, using the formula provided in lecture.*

Remember we can’t have a partial sample and we have to round up the result. You can use this same approach to find the reduced sample error = 0.1. You should take a minute to think about this problem in general. For example, if you are reducing the sampling error would you expect a larger number of samples to be required or a smaller number?