STATS 3042

Homework 2

Yevhen Myroniuk

A01005177 4r

2018

**Homework 2 – Numerical Descriptive Statistics  
Due: Friday, October 1, at beginning of lecture, as a hard copy**

In this assignment, you will using the commands from your labs along with the theory from the lecture notes on numerical descriptive statistics to make sense of the data in the file milkcosts.txt, which gives fuel, repair and capital costs associated with transporting milk from farms to dairy plants for gasoline trucks. You will need to import into R as a **text file**.

To import the dataset, first save it to your local drive. Then, in RStudio, go **to File->Import Dataset->From Text (base)**. RStudio will prompt you and allow you to preview the dataset. If it looks OK (it should), go ahead and click **Import*.*** You will see this data in the top left window of RStudio, just like we viewed the built-in datasets. (If this doesn’t work, you may have to try again, possibly closing RStudio and reopening it.)

**To submit: answers to all of the following questions. If you wrote commands and/or generated plots to arrive at your answer, include those as well.**

1. You will notice that the columns in this dataset do not have names. You can give them names with a command along the lines of

> names(milkcosts)=c("col1", "col2", "col3", "col4")  
Rename the columns with descriptive names.

> names(milkcosts) = c("fuel\_cost", "repair\_cost", "capital\_cost", "engine\_type")

1. Which of the following sets of costs – fuel, repair, and capital – is the most skewed? Which is the most symmetric?

> mean(milkcosts$fuel\_cost)-median(milkcosts$fuel\_cost)

[1] 1.114915

> mean(milkcosts$repair\_cost)-median(milkcosts$repair\_cost)

[1] 0.2554237

> mean(milkcosts$capital\_cost)-median(milkcosts$capital\_cost)

[1] 1.704068

Most skewed set of costs is capital cost. Most symmetric is repair cost.

1. Find the percentage of fuel costs that are within 1, 2, and 3 standard deviations of the mean fuel cost. (You should answer this question using only commands in R; do not count anything by hand.) Does the fuel cost data satisfy Chebyshev’s Theorem? Does it satisfy the Empirical Rule?

> fuel\_sd = sd(milkcosts$fuel\_cost)

> fuel\_mean = mean(milkcosts$fuel\_cost)

> data\_count = count(milkcosts)

> s1 = fuel\_mean + fuel\_sd\*1

> s2 = fuel\_mean + fuel\_sd\*2

> s3 = fuel\_mean + fuel\_sd\*3

> s1b = fuel\_mean - fuel\_sd\*1

> s2b = fuel\_mean - fuel\_sd\*2

> s3b = fuel\_mean - fuel\_sd\*3

> options(digits = 2)

> (count(filter(milkcosts, milkcosts$fuel\_cost < s1 & milkcosts$fuel\_cost > s1b)))/data\_count

n

1 0.81

> (count(filter(milkcosts, milkcosts$fuel\_cost < s2 & milkcosts$fuel\_cost > s2b)))/data\_count

n

1 0.97

> (count(filter(milkcosts, milkcosts$fuel\_cost < s3 & milkcosts$fuel\_cost > s3b)))/data\_count

n

1 0.97

According to Empirical rule ~68% of all data stays within 1sd from the mean, ~95% from 2sd and ~99.7% from 3sd. From observed results we can see that we have overflow of data for the given margins. 81% within 1sd, 97% within 2sd and 97% within 3sd.

As for Chebyshev’s Theorem, our results are consistent sinse for 2sd we get 97%>75%, and 3sd we have 97% > 88.89%.

1. Find the percentage of repair costs that are within 1, 2, and 3 standard deviations of the mean repair cost. (You should answer this question using only commands in R; do not count anything by hand.) Does the repair cost data satisfy Chebyshev’s Theorem? Does it satisfy the Empirical Rule?

> repair\_sd = sd(milkcosts$repair\_cost)

> repair\_mean = mean(milkcosts$repair\_cost)

> t1 = repair\_mean + repair\_sd\*1

> t2 = repair\_mean + repair\_sd\*2

> t3 = repair\_mean + repair\_sd\*3

> t1b = repair\_mean - repair\_sd\*1

> t2b = repair\_mean - repair\_sd\*2

> t3b = repair\_mean - repair\_sd\*3

> (count(filter(milkcosts, milkcosts$repair\_cost < t1 & milkcosts$repair\_cost > t1b)))/data\_count

n

1 0.68

> (count(filter(milkcosts, milkcosts$repair\_cost < t2 & milkcosts$repair\_cost > t2b)))/data\_count

n

1 0.98

> (count(filter(milkcosts, milkcosts$repair\_cost < t3 & milkcosts$repair\_cost > t3b)))/data\_count

n

1 1

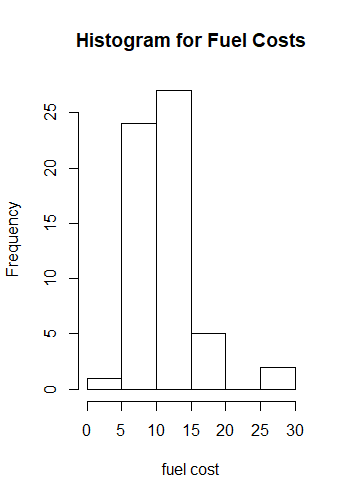
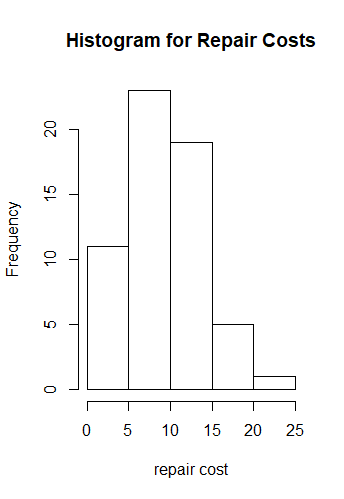
According to Empirical rule ~68% of all data stays within 1sd from the mean, ~95% from 2sd and ~99.7% from 3sd. From observed results we can see that we have exactly 68% of data within 1sd, little overflow for 2sd – 98%, and 100% of data withing 3sd. Our dataset satisfies Empirical rule.

As for Chebyshev’s Theorem, our results are consistent sinse for 2sd we get 98%>75%, and 3sd we have 100% > 88.89%.

1. Based on your answers to Questions 3 and 4, which set of data appears to be closest to having a normal (bell-shaped) distribution – fuel costs, or repair costs? Explain. Create histograms of both sets of data. Do your histograms support your answer?

Based on Q3 and Q4, repair cost is closest to have bell-shaped distribution, because judging by Empirical rule results repair costs are better aligned than fuel costs.

> hist(milkcosts$fuel\_cost, main="Histogram for Fuel Costs", xlab="fuel cost")  
> hist(milkcosts$repair\_cost, main="Histogram for Repair Costs", xlab="repair cost")

The histograms support the observation – we can see that repair cost is normally distributed into bell-shape compared to fuel cost.

1. Which of the three types of costs has the most variation? Explain.

> var(milkcosts$fuel\_cost)

[1] 17

> var(milkcosts$repair\_cost)

[1] 22

> var(milkcosts$capital\_cost)

[1] 44

Capital cost has the most variation. If we observe the histogram we can see that there are more bar distributions which means that there are more different data entries.

1. Call the **favstats** function three times to produce six sets of statistics: fuel costs by fuel type, repair costs by fuel type, and capital costs by fuel type. Of the three numerical categories, which one seems to depend the most on the type of fuel used? Explain, making reference to the numbers you found.

> favstats(milkcosts$fuel\_cost~milkcosts$engine\_type)

milkcosts$engine\_type min Q1 median Q3 max mean sd n missing

1 diesel 6.5 8.8 9.8 11 16 10 2.1 23 0

2 gasoline 4.2 9.4 11.2 14 29 12 4.8 36 0

> favstats(milkcosts$repair\_cost~milkcosts$engine\_type)

milkcosts$engine\_type min Q1 median Q3 max mean sd n missing

1 diesel 2.9 6.0 11.8 13 22 10.8 5.1 23 0

2 gasoline 1.4 5.1 7.7 11 17 8.1 4.2 36 0

> favstats(milkcosts$capital\_cost~milkcosts$engine\_type)

milkcosts$engine\_type min Q1 median Q3 max mean sd n missing

1 diesel 6.0 14.1 17.4 21 35 18.2 6.8 23 0

2 gasoline 3.3 6.8 9.6 12 18 9.6 3.7 36 0

Out of three data sets it seems like capital cost is the most dependent on the gas type, because the Q’s, min/max, medians and means for diesel type are 2 times higher than gasoline.

1. Using a single command, create side-by-side boxplots of capital costs grouped by fuel type. Based on your boxplots, approximately what percentage of the time do capital costs for diesel trucks exceed the **maximum** capital costs for gasoline trucks? Now find the exact percentage using one or more R commands.

|  |
| --- |
| > boxplot(milkcosts$capital\_cost~milkcosts$engine\_type) |
| From observation we can assume that cost exceed about 50% of the times since the max value of gasoline are at the level of the median of the diesel. |
| |  | | --- | |  | |

> gas = (filter(milkcosts, engine\_type == 'gasoline'))

> max\_gas = max(gas$capital\_cost)

> dsl = (filter(milkcosts, engine\_type == 'diesel'))

> count(filter(dsl, dsl$capital\_cost > max\_gas))/count(dsl)

n

1 0.48

The exact percentage is 48% and it is very close to our original observation.

1. Suppose the company reimburses truck drivers for their costs. Instead of having truck drivers submit receipts for each trip, the company gives every driver a flat amount that is expected to exceed their total costs 55% of the time. How much should the company give drivers of diesel trucks? Of gasoline trucks? Show all your computations.

> total\_cost = milkcosts$fuel\_cost + milkcosts$repair\_cost + milkcosts$capital\_cost

> quantile(data=total\_cost, total\_cost,probs=0.55)

55%

32.47

> total\_cost\_g = gas$fuel\_cost + gas$repair\_cost + gas$capital\_cost

> total\_cost\_d = dsl$fuel\_cost + dsl$repair\_cost + dsl$capital\_cost

> quantile(data=total\_cost\_g, total\_cost\_g,probs=0.55)

55%

28.74

>

> quantile(data=total\_cost\_d, total\_cost\_d,probs=0.55)

55%

38.84

On average, company must give $32.47 to any driver. Separated by the types, diesel drivers should receive $38.84, and gasoline drives should get $28.74.

1. The company boss explains to its drivers that since it overpays them for their operating costs 55% of the time and underpays them 45% of the time, the drivers come out ahead. If you were a driver working for this company, would you accept this explanation?

> quantile(data=total\_cost\_d, total\_cost\_d,probs=seq(0.1,0.9,0.05))

10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60%

29.73 30.17 30.88 31.64 32.95 34.05 35.35 35.81 36.12 38.84 39.90

65% 70% 75% 80% 85% 90%

41.15 43.54 44.80 45.85 47.40 48.75

>

> quantile(data=total\_cost\_g, total\_cost\_g,probs=seq(0.1,0.9,0.05))

10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60%

18.40 21.12 22.66 23.58 25.18 26.40 26.94 28.05 28.37 28.74 29.83

65% 70% 75% 80% 85% 90%

30.37 32.12 34.67 37.04 40.09 40.58

Based on detailed quantiles we can say that the overpay system does benefit diesel drivers, but it definitely does not benefit gasoline drivers. The difference between the 45% and 55% is not significant in the gasoline pay, making some drivers benefit of it much more than the others.