**Lab 3 – Numerical Descriptive Statistics**

**To submit before your next lab: answers to all numbered questions. When the question asks you to generate output in R, such as a graph, submit the output in the Word document as part of your answer. Also submit all commands you used to generate your output.**

We are going to analyze data about student absences from school in New South Wales, Australia. This data can be found in the built-in dataset **quine**, which is part of the **MASS** library. Load and view this dataset, and also take a couple of minutes to read its help file. Note: the functions we will be using are part of the **mosaic** package, so make sure you have that package loaded.

# Commands using the “mosaic” Package

The “mosaic” package was developed by a number of college and university professors to make the syntax of R more approachable. Since R is open-source software, anyone who wants to can write their own functions, which has resulted in a lot of commands having very different syntax. The commands available with the “mosaic” package are more consistent.

**General function syntax:**

goal ( y ~ x , data = mydata, …)

The components in boxes are the things you will need to enter to make the command work.

* The **goal** is what you want R to do.
* *y* and *x* are the **variables** that R will need to achieve the goal that you want to achieve.
* **mydata** is the dataframe in which these variables are stored.
* The **…** refers to additional options that you can include in the command. We will discuss these later (often includes formatting options, axis titles, etc.)

**One-variable function syntax:**

When you are only working with one variable, it appears as the *x* in the general syntax, and the *y* is omitted:

goal ( ~ x , data = mydata, …)

**One Quantitative Variable Commands**

Like last week, we will type our commands in the console.

We can find the mean number of days of absence among the students:

> mean(~Days, data=quine)

[1] 16.4589

That is, the mean number of days absent was 16.4589.

Similarly (replacing “mean” in the code above with your desired function name or goal):

|  |  |
| --- | --- |
| goal (function name) | What it does |
| Mean | calculates the mean |
| median | determines the median |
| Sd | determines the sample standard deviation |
| Var | determines the sample variance (sample standard deviation squared) |
| Min | determines the minimum |
| Max | determines the maximum |
| Sum | computes the sum of all the values for that variable |
| IQR | determines the inter-quartile range |
| favstats | a special function designed by the “mosaic” programmers as their “favourite” statistics. Computes: minimum, first quartile, median, third quartile, maximum, mean, s.d., sample size (*n*), and how many values were missing. |
| histogram | Creates a histogram giving the distribution of the variable of interest. |
| bwplot | Creates a box-and-whisker plot (or boxplot) of the variable of interest. |
| boxplot | Another way of creating boxplots |

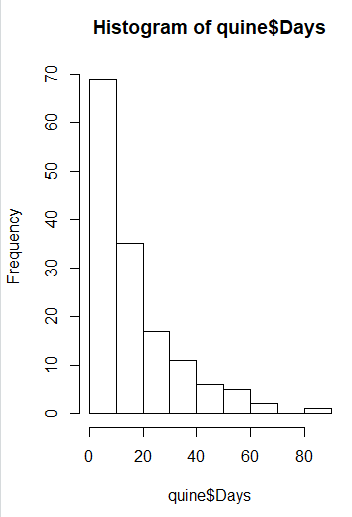
The output for calculating summary statistics will appear in the **console window**, but the graphs will appear in the **miscellaneous window**, in the *plots* tab (RStudio will automatically switch to the plots tab when a plot is generated).

1. Using commands above, find the Pearsonian coefficient of skewness for the number of days of absence. Include the command(s) you used to find this information. Is this data symmetric, skewed left, or skewed right?

> (3\*(mean(mean(~Days, data=quine)) - median(~Days, data=quine)))/sd(~Days, data=quine)

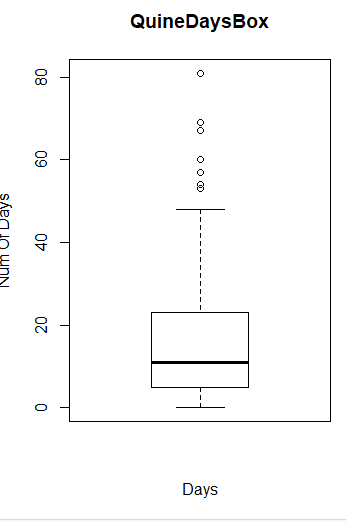
[1] 1.007598

1. Create a histogram for the number of days of absence. (No need to get fancy with the labels or bins, but it may be good practice to do so anyway.) Is your histogram consistent with your answer to Question 1? Explain in a sentence.



It is consistent because the mean 16.45 is greater than the median 11 which skews to the right.

1. Create a boxplot for the number of days of absence. Give your boxplot descriptive labels. Based on your boxplot, how many outliers does your data have? What is a typical number of days of absence, where “typical” excludes the largest and smallest 25% of absences?



7 outliers

Just estimating the range I would get about 23-50 days.

1. Using one of the commands listed above, compute the typical number of days of absence again. Does it agree with your answer to Question 3?

> fav\_stats(quine$Days)

min Q1 median Q3 max mean sd n missing

0 5 11 22.75 81 16.4589 16.25322 146 0

This means that the typical range is between 5-22.75, aka the interquartile range.

This fits the image of the boxplot’s q1-q3 so yes it agrees.

# One Quantitative Variable and One Categorical Variable Commands

We may be interested in comparing absences by category. For instance, we might want to see if boys and girls have different patterns of absence. We can do this by computing statistics for the **Days** field, grouped by **Sex**.

> favstats(Days~Sex, data=quine)

Sex min Q1 median Q3 max mean sd n missing

1 F 0 5.00 10 20.25 81 15.22500 15.93100 80 0

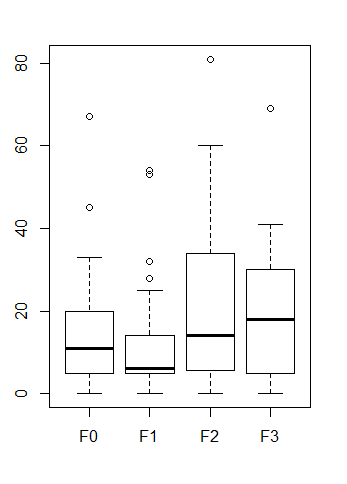
2 M 0 5.25 14 27.00 69 17.95455 16.63401 66 0

Here we notice a few things. For instance:

* there were both boys and girls who had perfect attendance (zero days of absence)
* In the first three quartiles, boys had more absences than girls. For instance, three quarters of girls had fewer than 20.25 days of absence, whereas three quarters of boys had fewer than 27 days of absence.
* However, the student with the most absences was female.

This works the same for all the other descriptive statistics functions given in the table on the previous page.

1. Create side-by-side boxplots for the number of days of absence, grouped by **age**. Make at least three observations.



* In general absences go up with ages
* In general the median absences rise with age
* Group f1 appears to have the least amount of absences
* 50 percent of absences of f1 fit under the median of f2

We can also group data by multiple factors; for instance, we can find the mean number of absences, grouped by ethnicity and sex.

> mean(Days~Eth+Sex, data=quine)

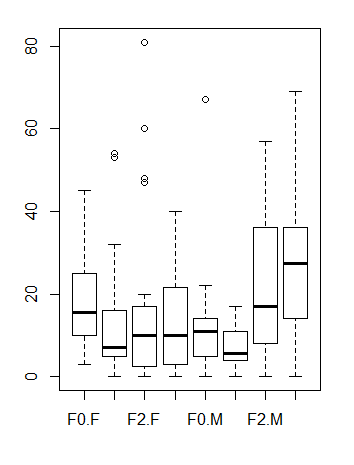
A.F N.F A.M N.M

20.92105 10.07143 21.61290 14.71429

1. Use appropriate commands to determine which group of students, grouped by age and sex, is most consistent with regards to absences, and which is least consistent.

> boxplot(Days~Age+Sex, data=quine)

The least consistent is the final group, which has the largest deviation, the most consistent if group 6, the smallest deviation.



# Measures of Position

The **favstats** command returned the values corresponding to the quartiles: 25%, 50% (median), and 75%. The **quantile** command allows us to generalize to other percentages.

The **quantile** command with default settings has similar syntax to the other commands we have encountered in this lab, and returns the quartiles. For instance, if we are interested in the quartiles for days of absence:

> quantile(data=quine, ~Days)

0% 25% 50% 75% 100%

0.00 5.00 11.00 22.75 81.00

As before, we see that 75% of students were absent for fewer than 22.75 days of the year.

(Note: you may get an error after this output, because R is missing a package. Install the **Rcpp** package and run the command again and the error should disappear.)

We can obtain different percentile values with the **probs** argument. For instance, suppose we are interested in the 90th percentiles of absences:

> quantile(data=quine, ~Days, probs=0.9)

90%

40

90% of students had fewer than 40 days of absence.

We can find multiple percentiles by listing them as sequences:

> quantile(data=quine, ~Days, probs=c(0.1,0.9))

10% 90%

2 40

> quantile(data=quine, ~Days, probs=seq(0.1,0.9, 0.1))

10% 20% 30% 40% 50% 60% 70% 80% 90%

2 5 5 7 11 14 20 27 40

1. Find the 20th, 40th, 60th, and 80th percentiles for days of absence, grouped by:

- learner status  
- sex  
- ethnicity

(ie, you will be running three separate commands). If you were to estimate the number of days a student was going to be absent from class, which of learner status, sex, and ethnicity would allow you to make the most accurate prediction? Explain.

> quantile(data=quine, Days~Lrn, probs=seq(0.2, 0.8, 0.2))

Lrn 20% 40% 60% 80%

1 AL 5 8.8 16.0 27

2 SL 5 6.0 13.2 28

> quantile(data=quine, Days~Sex, probs=seq(0.2, 0.8, 0.2))

Sex 20% 40% 60% 80%

1 F 5 6.6 13 23.2

2 M 5 10.0 16 30.0

> quantile(data=quine, Days~Eth, probs=seq(0.2, 0.8, 0.2))

Eth 20% 40% 60% 80%

1 A 6 13 20.0 36.8

2 N 3 5 10.6 19.6

Ethnicity is the biggest predictor of absence as it has the largest distances between the percentiles.

We can also go in the other direction: that is, we can find the percentile rankings for each data value. We use the **percent\_rank** command to do this. This command is part of the **dplyr** library, which you will need to install.

Once you have installed the **dplyr** library, you can apply it to an entire column of data. Its syntax is a bit different from that of the commands in the **mosaic** library (sorry, I don’t make the rules), and you need to feed it the data as a single argument. Here we obtain the percentile rankings for the number of days of absence. Only the first line of output is given below.

> percent\_rank(quine$Days)

[1] 0.08965517 0.47586207 0.55862069 0.17931034 0.17931034 0.53793103

This tells us that the first value in the **Days** column represents the 8.965517th percentile, ie, it is larger than 8.965517% of the data.

Percentile rankings are usually given as whole numbers, so we can format our output accordingly:

> options(digits=1)

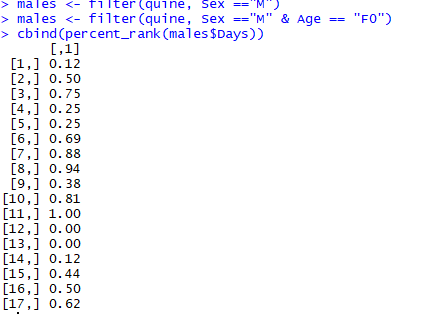
> percent\_rank(quine$Days)

[1] 0.09 0.48 0.56 0.18 0.18 0.54 0.70 0.72 0.31 0.31 0.61

This tells us that the first entry in the table is from a student who was absent more often than 9% of students.

(Note: R will display this number of digits from now on unless you change the **options** setting   
again.)

1. Give a list of percentile rankings (two decimal places) for the number of days absent for the male F0 students only. Give all the commands you used, as well as your output **as a column**. Note: you will have to use some of the commands from your last lab.



# Z-scores, Chebyshev’s Theorem, and the Empirical Rule

We saw in class that it is often useful to know a data value’s **z-score**, or how many standard deviations a data value is from the mean. Like percentile rankings, z-scores give a measure of how large or small a data value is relative to its neighbours. Recall that a value that is more than two standard deviations from the mean is considered unusual.

In R, the **scale** function provides z-scores of all data in a list. Applying this function to the **Days** column of our table (only the first four results are displayed):

> scale(quine$Days)

[,1]

[1,] -0.890

[2,] -0.336

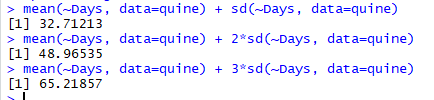
[3,] -0.151

[4,] -0.705

This tells us that the first student was absent for a number of days that was 0.890 standard deviations less than the mean.

1. Using commands from a previous lab, find the number of days of absence that are   
   - at least 1 standard deviation from the mean  
   - at least 2 standard deviations from the mean

- at least 3 standard deviations from the mean



> filter(quine, Days < 32.7 )

Eth Sex Age Lrn Days

1 A M F0 SL 2

2 A M F0 SL 11

3 A M F0 SL 14

4 A M F0 AL 5

5 A M F0 AL 5

6 A M F0 AL 13

7 A M F0 AL 20

…………………………

119 N F F3 AL 3

120 N F F3 AL 5

121 N F F3 AL 15

122 N F F3 AL 18

123 N F F3 AL 22

123/ 146 = 84%

> filter(quine, Days < 48.9)

Eth Sex Age Lrn Days

1 A M F0 SL 2

2 A M F0 SL 11

3 A M F0 SL 14

4 A M F0 AL 5

5 A M F0 AL 5

…………………….

135 N F F3 AL 15

136 N F F3 AL 18

137 N F F3 AL 22

138 N F F3 AL 37

138/ 146 = 95%

> filter(quine, Days < 99.7)

Eth Sex Age Lrn Days

1 A M F0 SL 2

2 A M F0 SL 11

3 A M F0 SL 14

4 A M F0 AL 5

5 A M F0 AL 5

…….

143 N F F3 AL 15

144 N F F3 AL 18

145 N F F3 AL 22

146 N F F3 AL 37

146/ 146 = 100%

1. Do your results from Question 9 satisfy Chebyshev’s Theorem? Do they satisfy the Empirical Rule? If not, explain why this is with reference to answers to earlier questions.

Empirical Rule is proven by question nine as

1. 123/ 145 = 84%
2. 138/ 146 = 95%
3. 146/ 146 = 100%

Only #1 is off, the other two are very close to the predicted 95% and 99.7%

Chebyshev says 75% of data lies WITHIN two standard deviations of the mean, and this

Is borne out by question nine as over this amount (84%) does.