---
title: 'CWRU DSCI351-351m-451: Lab Exercise LE1b'
author: "Prof.:Roger French, TAs: Raymond Wieser, Sameera Nalin Venkat"
author: "Roger H. French, Sameera Nalin Venkat, Raymond Weiser"
date: "07 September, 2021"
output:
pdf_document:
latex_engine: xelatex
toc: TRUE
number_sections: TRUE
toc_depth: 6
highlight: tango
html_notebook:
html_document:
df_print: paged
toc: yes
toc_depth: 6
urlcolor: blue
always_allow_html: true

LE1, 7 points, 7 questions.

In every assignment file,

- CHANGE "NAME" to your Name!!!!
  - Otherwise you have git "Merge Conflicts"

Lab Exercise (LE) 1

- Due Tuesday September 8th
  - Before Class
- Answers to these problems may be in the back of the book (OIStats-v4),
  - so you can check your work.

- The grading is done on how you show your thinking,
  - explain yourself and
  - show your R code and
  - the output you got from your code.
- Code style is important
  - Follow Rstudio code diagnostics notices
  - And the Google R Style Guide
  - Which is related to the Tidyverse Style Guide
  - Also available in your class repo, cheat sheets

To be done as an Rmd file,

- where you turn in
  - the Rmd file and
  - the compiled pdf showing your work.
  - and the R script of IntroR.R

You will want to produce a report type format

- (html and pdf type document) to turn in.
- And not an ioslides or beamer (slide type) compiled output.

  - These are presentation formats, and can be fussy

Also are you backing up your git repo

- in a second and third location,
- to avoid corruption problems?

So by now I believe everyone has

Logged into your ODS VDI, (or your VUVlab VDI).

```
- If not send the help@case.edu helpdesk an email,
  - directed to CSE-IT, saying you are in DSCI351/351M/451
  - and should have access to the ODS VDI,
  - that you use Citrix REceiver to connect to.
```

Your H: drive is big enough

- so that you can Git clone your personal fork of the Prof repo,
- from Bitbucket down to `H:\Git\` folder.

Ask Questions in CWRU-DSCI Slack Channel for DSCI351-351M-451

- This is the easier way to ask and answer questions
- You can use @Raymond Wieser and @Roger French
  - To direct a question to us
  - But anyone can answer the questions

351, 351M and 451 students

- Will all do the last part of the homework,
  - where you are doing some R coding,
  - inside the R code blocks of the Rmd file
  - (between the "‘*r* and the "‘ that closes the R code block in the Rmd file.

And 451 students

- will start writing about what they are considering for their Semester Project.
  - Read about the 451 Semester Project in 1-assignments>SemProj-451>1808-451-SemProj-Overview.pdf
- Your SemProj will have 3 in-class report outs on progress, and a final full report.
- It should ideally be related to your thesis research,
  - and be a data analysis project that will help advance your research.

- We will be defining and refining what you will do your Semester Project on,
  - in the next few weeks.

Here are answers to a few questions we usually get about HW1 A. "I am having trouble converting the Rmd file containing lab exercise 1 into a PDF. I was able to save it as a .txt file–is it okay if I submit that instead of a PDF?"

Once you have made a *.Rmd file, you compile it to make the pdf, by hitting the Knit button at the top of the Rstudio text editor, or you can click on the Knit button to choose Knit to PDF from the choices.

You can also use the keyboard shortcut Cntrl+Shift+K. (You can find lots of keyboard shortcuts help with Alt+Shift+K).

And if you open the LE1b Rmd file named "2008-351-351M-451-LE1b-NAME.Rmd" and change NAME to your own name. Then you can immediately compile that Rmd file to make the pdf. This way you'll know that its not some error in what you have added to the file's text. Compiling to pdf, uses the LaTeX publishing distribution on your VDI. So if you are trying this on your own personal computer, it won't work, since you probably don't have a LaTeX distribution, such as MikTeX (for windows), MacTeX (for Macs), or TexLive for Linux, installed, so can't produce a LaTeX pdf output.

B. "I was also wondering where we are supposed to submit our homework assignments. Are we supposed to upload them onto BitBucket?"

You will upload your *.Rmd file (so we can see your coding style and commenting), and your compiled Pdf file to our Canvas Assignment page in canvas.case.edu for the DSCI351-351M-451 class.

Summary of LE1 (quick tip: use Ctrl+Shift+O for viewing all the sub-questions):-

LE1-1: 1/2 pt. LE1-2: 1 pt. LE1-3: 1 pt. LE1-4: 1 pt. LE1-5: 1 pt. LE1-6: 1/2 pt. LE1-7: 2 pt.

Remember to include text-based answers of the questions next to 'ANSWER ->'. Text-based answers carry points.

LE1-1. (½ pt.) R Calculator:

In the 1-Assignments/LE1 folder in your repo,

- You will find an R script with some basic R variable problems.
  - 2108-351-351m-451-LE1a-introR.R
- Complete these using proper R commands in your R script file
  - and submit the solution.

- Don't forget attribution, versioning and licensing.

LE1-2. (1 pt.) Structure of a Data Analysis

Read about the Structure of a Data Analysis, by Jeff Leek

- Located in .3-readings/0-Leek-DataAnalysisStructure-slides/
  - Leek-ADataAnalysisStructureAndOrganizing.pdf
  - 1503LeekDataAnalyticStyle-outline.txt
- And take a look at Leek's book in ./readings/Texts/Leek-DataAnalyticStyle.pdf
- This is the approach we take in Applied Data Science
  - Focus on the Question, the Dataset, and the Analysis and Reporting it

On organizing a data analysis

- Jeff Leek is a biostatistician
- At Johns Hopkins School of Public Health

Steps in a data analysis

- Define the question
- Define the ideal data set

- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Basic R operations

- Show an example of addition, subtraction, multiplication, division, and an exponential below

```
3+2
```

```
## [1] 5
```

```
3-2
```

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

```
3*2
```

```
## [1] 6
```

```
3^2
```

```
## [1] 9
```

Exploratory Data Analysis (EDA) using data frames

- Data frames are an important data format in R
- Example data can be loaded from base R
- Run the code below to load the iris dataset into your environment
- This data set will be used for the later problems

```
data("iris")
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

- Give the class of each of the columns in the iris data set
- Explain what is a factor and how it differs from a character

```
?factor
lapply(iris, class) # lapply(dataset, function)
```

```
## $Sepal.Length
## [1] "numeric"
##
## $Sepal.Width
## [1] "numeric"
```

```
##
## $Petal.Length
## [1] "numeric"
##
## $Petal.Width
## [1] "numeric"
##
## $Species
## [1] "factor"
```

ANSWER -> factors are used to describe/represent a categorical data set. for eg: whether a course is paid: it will either have yes or no values. such a data can be described by factor, it takes the unique values and stores as levels. Now, you can also use character class to store such values, but you do not get the benefit of visualising it for what it is, which is a category rather than something unique to each row. Character can be used for columns where each row has its own individual identity, whereas columns which signify a group identity of the row can be stored as a factor, for eg. a house name can be used as a character, but its zip-code can be stored as a factor.

- Use the table() function to determine how many species there are
- and how many observation each one has (Species column in the data frame)

```
?table

# data_species <- iris[, c('Species')]
data_species <- iris$Species# isolates Species column
table(data_species)
```

```
## data_species
##     setosa versicolor  virginica
##         50         50         50
```

ANSWER -> There are 3 species - setosa ,versicolor , virginica : each having 50 observations

- Use the subset() function create a new data frame of only versicolor flower data

```
?subset

dt_versicolor <- subset(iris, Species == "versicolor")
dt_versicolor
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 51           7.0         3.2          4.7         1.4 versicolor
## 52           6.4         3.2          4.5         1.5 versicolor
## 53           6.9         3.1          4.9         1.5 versicolor
## 54           5.5         2.3          4.0         1.3 versicolor
## 55           6.5         2.8          4.6         1.5 versicolor
## 56           5.7         2.8          4.5         1.3 versicolor
## 57           6.3         3.3          4.7         1.6 versicolor
## 58           4.9         2.4          3.3         1.0 versicolor
## 59           6.6         2.9          4.6         1.3 versicolor
## 60           5.2         2.7          3.9         1.4 versicolor
## 61           5.0         2.0          3.5         1.0 versicolor
## 62           5.9         3.0          4.2         1.5 versicolor
## 63           6.0         2.2          4.0         1.0 versicolor
## 64           6.1         2.9          4.7         1.4 versicolor
## 65           5.6         2.9          3.6         1.3 versicolor
## 66           6.7         3.1          4.4         1.4 versicolor
## 67           5.6         3.0          4.5         1.5 versicolor
```

```
## 68          5.8         2.7          4.1         1.0 versicolor
## 69          6.2         2.2          4.5         1.5 versicolor
## 70          5.6         2.5          3.9         1.1 versicolor
## 71          5.9         3.2          4.8         1.8 versicolor
## 72          6.1         2.8          4.0         1.3 versicolor
## 73          6.3         2.5          4.9         1.5 versicolor
## 74          6.1         2.8          4.7         1.2 versicolor
## 75          6.4         2.9          4.3         1.3 versicolor
## 76          6.6         3.0          4.4         1.4 versicolor
## 77          6.8         2.8          4.8         1.4 versicolor
## 78          6.7         3.0          5.0         1.7 versicolor
## 79          6.0         2.9          4.5         1.5 versicolor
## 80          5.7         2.6          3.5         1.0 versicolor
## 81          5.5         2.4          3.8         1.1 versicolor
## 82          5.5         2.4          3.7         1.0 versicolor
## 83          5.8         2.7          3.9         1.2 versicolor
## 84          6.0         2.7          5.1         1.6 versicolor
## 85          5.4         3.0          4.5         1.5 versicolor
## 86          6.0         3.4          4.5         1.6 versicolor
## 87          6.7         3.1          4.7         1.5 versicolor
## 88          6.3         2.3          4.4         1.3 versicolor
## 89          5.6         3.0          4.1         1.3 versicolor
## 90          5.5         2.5          4.0         1.3 versicolor
## 91          5.5         2.6          4.4         1.2 versicolor
## 92          6.1         3.0          4.6         1.4 versicolor
## 93          5.8         2.6          4.0         1.2 versicolor
## 94          5.0         2.3          3.3         1.0 versicolor
## 95          5.6         2.7          4.2         1.3 versicolor
## 96          5.7         3.0          4.2         1.2 versicolor
## 97          5.7         2.9          4.2         1.3 versicolor
## 98          6.2         2.9          4.3         1.3 versicolor
## 99          5.1         2.5          3.0         1.1 versicolor
## 100         5.7         2.8          4.1         1.3 versicolor
```

- Give the mean and median of each of the numeric columns for the versicolor data frame
- Why might the mean and median of the entire iris dataset be misleading?

```
# ?median
# ?mean
med_sep_l <- median(iris[, c('Sepal.Length')])
print( paste("median sepal length -> " ,med_sep_l))
```

```
## [1] "median sepal length ->  5.8"
```

```
mean_sep_l <- mean(iris[, c('Sepal.Length')])
print( paste("mean sepal length -> " ,mean_sep_l))
```

```
## [1] "mean sepal length ->  5.84333333333333"
```

```
med_sep_w <- median(iris[, c('Sepal.Width')])
print( paste("median sepal width -> " ,med_sep_w))
```

```
## [1] "median sepal width ->  3"
```

```
mean_sep_w <- mean(iris[, c('Sepal.Width')])
print( paste("mean sepal width -> " ,mean_sep_w))
```

```
## [1] "mean sepal width ->  3.05733333333333"
```

```
med_pet_l <- median(iris[, c('Petal.Length')])
print( paste("median petal length -> " ,med_pet_l))
```

```
## [1] "median petal length ->  4.35"
```

```
mean_pet_l <- mean(iris[, c('Petal.Length')])
print( paste("mean petal length -> " ,mean_pet_l))
```

```
## [1] "mean petal length ->  3.758"
```

```
med_pet_w <- median(iris[, c('Petal.Width')])
print( paste("median petal width -> " ,med_pet_w))
```

```
## [1] "median petal width ->  1.3"
```

```
mean_pet_w <- mean(iris[, c('Petal.Width')])
print( paste("mean petal length -> " ,mean_pet_w))
```

```
## [1] "mean petal length ->  1.19933333333333"
```

ANSWER -> There are 4 numrical columns: [1] “median sepal length -> 5.8” [1] “mean sepal length -> 5.84333333333333” [1] “median sepal width -> 3” [1] “mean sepal width -> 3.05733333333333” [1] “median petal length -> 4.35” [1] “mean petal length -> 3.758” [1] “median petal width -> 1.3” [1] “mean petal length -> 1.19933333333333”

Mean is useful only when the data is mostly symmetrical, i.e. most of the observations are in the middle of the data. Median is useful when data is skewed on either of the extremeties. Depending on the distribution of this numerical data mean or median can be helpful. It is also possible that data has multiple modes, ie, there is a high frequency f a certain kind of sepal length/ petal length, in which mode would be the best way to describe the data

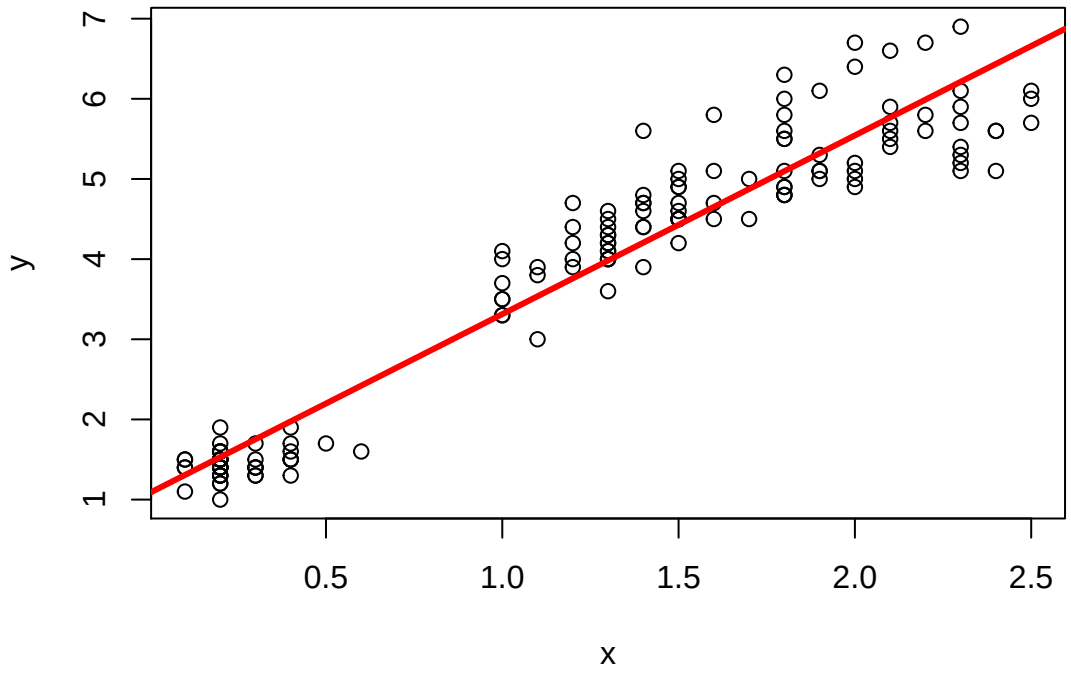
Modeling and plotting

- Use the lm() and plot() functions to build a simple linear model
  - of versicolor petal length as a function of petal width
- What are the dependent and independent variables in this case?
- Add the model to the plot with abline()

```
?stats::lm
# ?plot
x <-iris[, c('Petal.Width')]
y <-iris[, c('Petal.Length')]
model1 <- lm(y ~ x)

plot(x,y)

abline(model1, lwd =3, col ="red")
```



ANSWER -> since petal length is a function of petal width - dependent variable - petal length and - independent variables - petal width

- Print the summary of this model

```
# ?Summary
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <2e-16 ***
## x            2.22994    0.05140   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

LE1-3. (1 pt.) Data Basics: OpenIntroStats Exercise 1.9 1n Chapter 1, pg. 20.

Answer this in this Rmd file and

- explain what you are doing,
- i.e. show your R code and work.

OIS Exercise 1.9 Fisher's irises, OISv4 page 20. Sir Ronald Aylmer Fisher was

- An English statistician, evolutionary biologist, geneticist
  - and unfortunately a eugenicist
- Who worked on a data set that contained
  - sepal length and width, and petal length and width
  - from three species of iris flowers
    - (setosa, versicolor and virginica).
- There were 50 flowers from each species in the data set.

(3a) How many cases were included in the data?

Show you R code!

```
# length of dataset
dim(iris)[1]
```

```
## [1] 150
```

ANSWER -> 150

(3b) How many numerical variables are included in the data?

- Indicate what they are, and
- if they are continuous or discrete.

```
lapply(iris, class)
```

```
## $Sepal.Length
## [1] "numeric"
##
## $Sepal.Width
## [1] "numeric"
##
## $Petal.Length
## [1] "numeric"
##
## $Petal.Width
## [1] "numeric"
##
## $Species
## [1] "factor"
```

ANSWER -> 4 numeric variables and all are continuous and not discrete, meaning they are not integral. $Sepal.Length [1] "numeric"

$Sepal.Width [1] "numeric"

$Petal.Length [1] "numeric"

$Petal.Width [1] "numeric"

(3c) How many categorical variables are included in the data,

- and what are they?
- List the corresponding levels (categories).

```
iris$Species
```

```
##   [1] setosa     setosa     setosa     setosa     setosa     setosa
##   [7] setosa     setosa     setosa     setosa     setosa     setosa
##  [13] setosa     setosa     setosa     setosa     setosa     setosa
##  [19] setosa     setosa     setosa     setosa     setosa     setosa
##  [25] setosa     setosa     setosa     setosa     setosa     setosa
##  [31] setosa     setosa     setosa     setosa     setosa     setosa
##  [37] setosa     setosa     setosa     setosa     setosa     setosa
##  [43] setosa     setosa     setosa     setosa     setosa     setosa
##  [49] setosa     setosa     versicolor versicolor versicolor versicolor
##  [55] versicolor versicolor versicolor versicolor versicolor versicolor
##  [61] versicolor versicolor versicolor versicolor versicolor versicolor
##  [67] versicolor versicolor versicolor versicolor versicolor versicolor
##  [73] versicolor versicolor versicolor versicolor versicolor versicolor
##  [79] versicolor versicolor versicolor versicolor versicolor versicolor
##  [85] versicolor versicolor versicolor versicolor versicolor versicolor
##  [91] versicolor versicolor versicolor versicolor versicolor versicolor
##  [97] versicolor versicolor versicolor versicolor virginica  virginica
## [103] virginica  virginica  virginica  virginica  virginica  virginica
## [109] virginica  virginica  virginica  virginica  virginica  virginica
## [115] virginica  virginica  virginica  virginica  virginica  virginica
## [121] virginica  virginica  virginica  virginica  virginica  virginica
## [127] virginica  virginica  virginica  virginica  virginica  virginica
## [133] virginica  virginica  virginica  virginica  virginica  virginica
## [139] virginica  virginica  virginica  virginica  virginica  virginica
## [145] virginica  virginica  virginica  virginica  virginica  virginica
## Levels: setosa versicolor virginica
```

ANSWER -> Species is the only categorical data and it has 3 levels : Levels: setosa versicolor virginica

LE1-4. (1 pt.) Examining Numerical Data:. Distributions and Appropriate Statistics

OISv4 Exercise 2.16 p. 59.

For each of the following, answer three things

i) State whether you expect the distribution to be

- symmetric,
- right skewed,
- or left skewed.

ii) Also specify whether the mean or median

- would best represent a typical observation in the data,

iii) and whether the variability of observations

- would be best represented
- using the standard deviation or IQR.

Explain your reasoning.

(LE1-4a) Housing prices in a country

- where 25% of the houses cost below $350,000,
- 50% of the houses cost below $450,000,
- 75% of the houses cost below $1,000,000
- and there are a meaningful number of houses that cost more than $6,000,000.

Answer -> if we plot a histogram of this plot, where the x axis denotes housing prices, and y-axis denotes percentage of houses , we notice that data in uniform at 25% till $1,000,000 and then it reduces indicating decline, but then increases again after $6,000,0000 indicating outliers. Therefore, data set is **RIGHT-SKEWED**, and can be best represented using a **MEDIAN** and **IQR**

(LE1-4b) Housing prices in a country where

- 25% of the houses cost below $300,000,
- 50% of the houses cost below $600,000,
- 75% of the houses cost below $900,000
- and very few houses that cost more than $1,200,000.

Answer -> if we plot a histogram of this plot, where the x axis denotes hosusing prices, and y-axis denotes percentage of houses, we notice that the dataset is fairly uniform throughout the range 0 - 1,200,000, with 25% and very few houses beyond 1,200,000. This shows that data is **symmetric**, and best represented by a **mean** and variability - **standard deviation**.

(LE1-4c) Number of alcoholic drinks consumed by college students in a given week.

Assume that most of these students don't drink

- since they are under 21 years old,
- and only a few drink excessively.

Answer -> if we plot a histogram of this plot, where the x axis denotes no of drinks consumed by college students in a given week, and y-axis denotes percentage of students drinking, we observe that a large percentage of students have close to 0 no of drinks and then 1 or 2% has 10 drinks per week, this means the data is **right-skwed**, and **median and IQR**, are best ways to represent this

(LE1-4d) Annual salaries of the employees at a Fortune 500 company

- where only a few high level executives earn much higher salaries
- than all the other employees.

Answer -> if we plot a histogram of this plot, where the x axis denotes salaries , and y-axis denotes No of employees receiving the salaries, lower salaries will be received by many and higher salaries will be received by a few,the data is **right skewed** and best represented by **median and IQR.**

LE1-5. (1 pt.) Examining Numerical Data: OpenIntroStats Exercise 2.9

in OISv4 Chapter 2, pg 57.

Exercise 2.9 Means and SDs. For each part, compare distributions (1) and (2)

- based on their means and standard deviations.

You do not need to calculate these statistics;

- simply state how the means and the standard deviations compare.

Make sure to explain your reasoning.

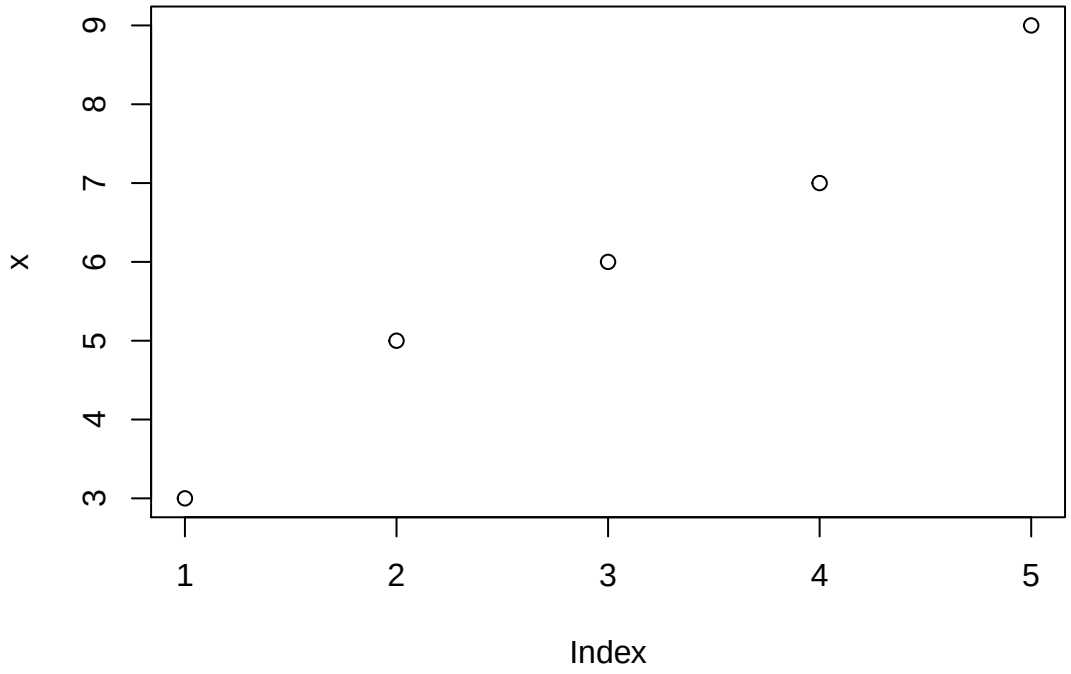
- Hint: It may be useful to sketch dot plots of the distributions.

LE1-5a)

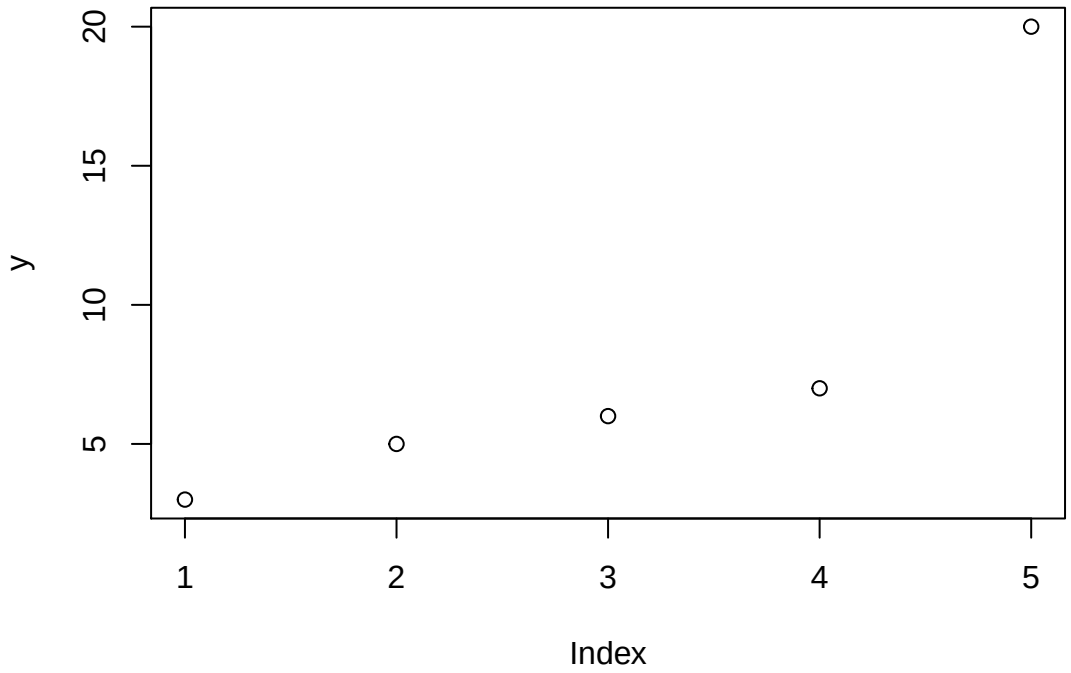
(1) 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20

```
?plot
?vector
x <- c(3, 5, 6, 7, 9)
y <- c(3, 5, 6, 7, 20)

plot(x)
```




```
plot(y)
```



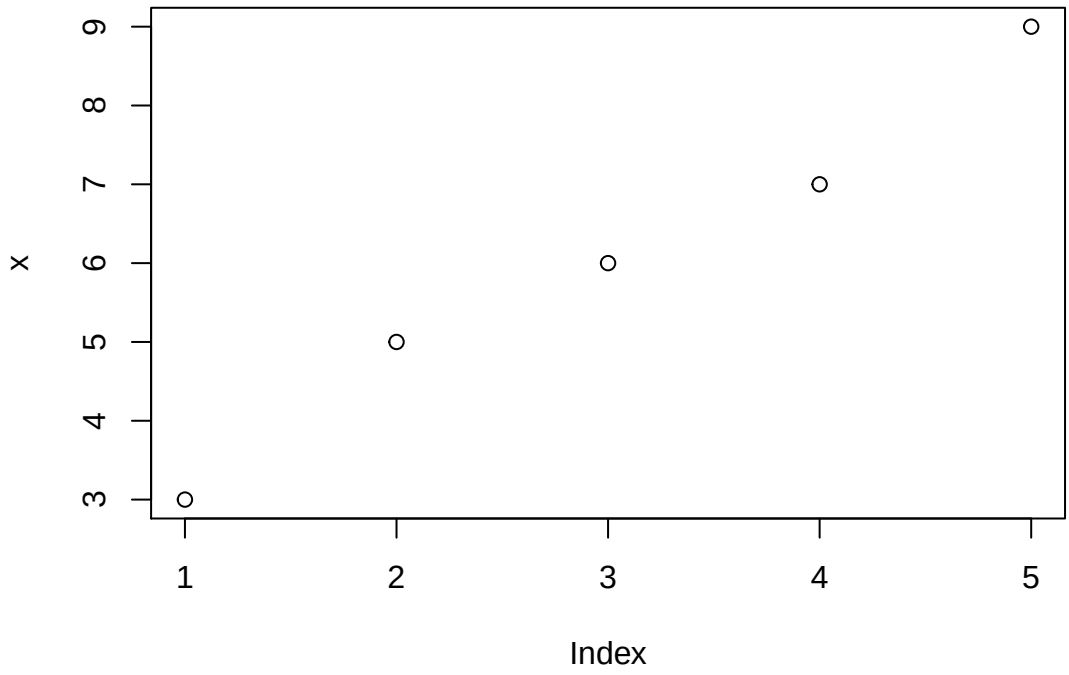
ANSWER -> (1) is more uniform than (2), (1) has a lower mean than (2), (1) has lower variance/std deviation than (2)

LE1-5b)

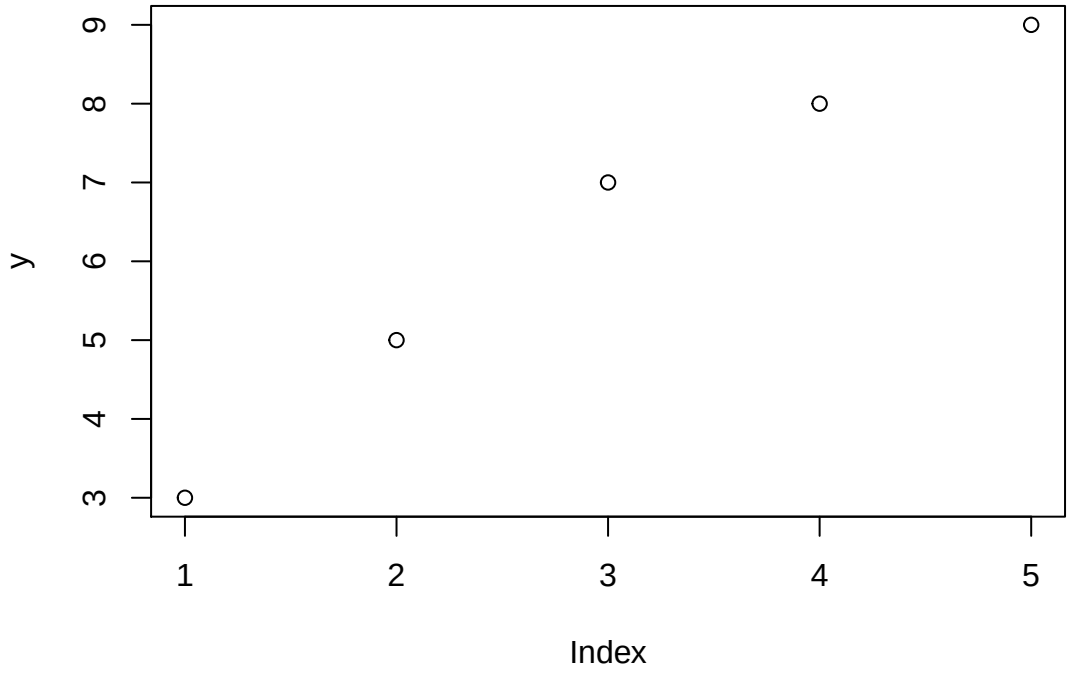
```
(1) 3, 5, 6, 7, 9
(2) 3, 5, 7, 8, 9
```

```
x <- c(3, 5, 6, 7, 9)
y <- c(3, 5, 7, 8, 9)

plot(x)
```




```
plot(y)
```



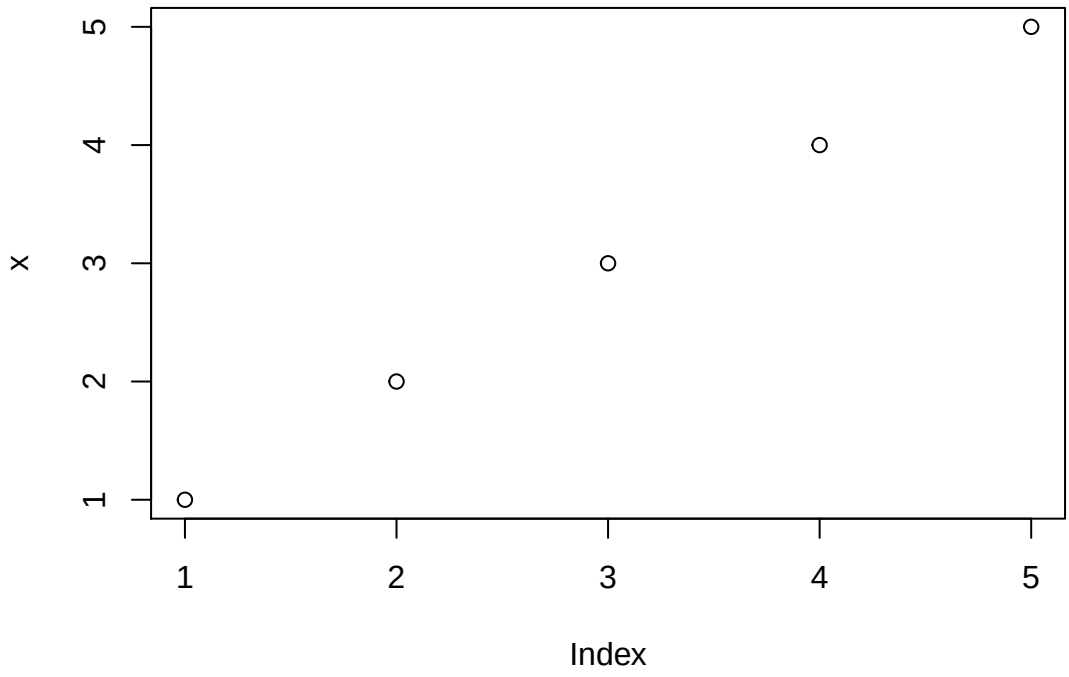
ANSWER ->(1) and (2) are uniformly spread, (2) has a higher mean than (1), both have about the same variance and standard deviation

LE1-5c)

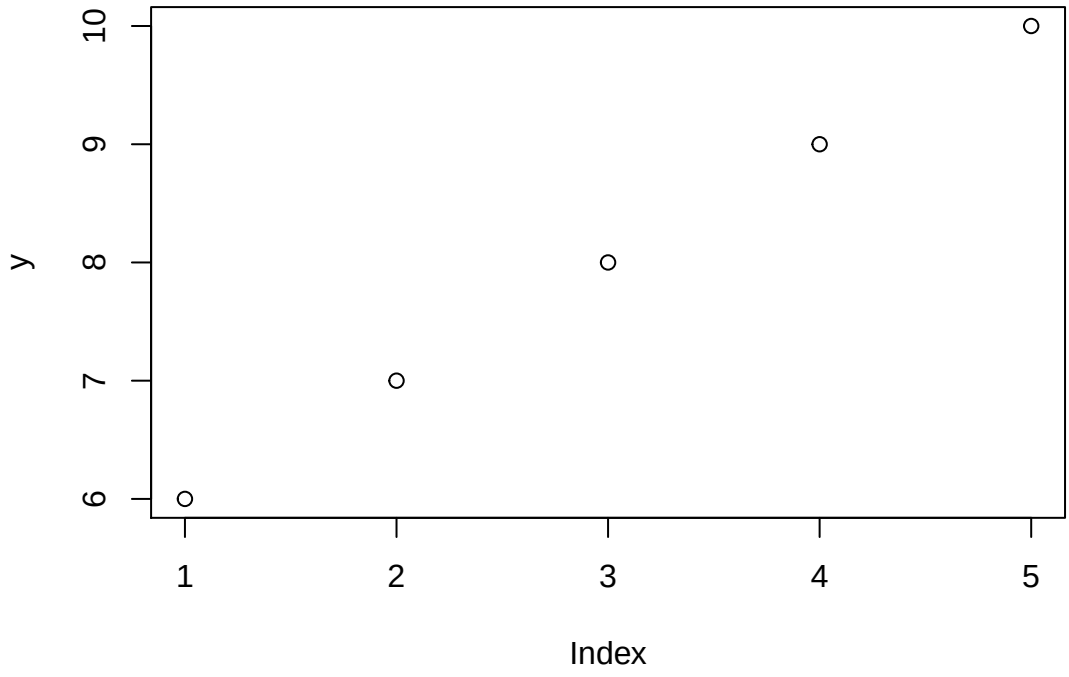
```
(1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10
```

```r
x <- c(1, 2, 3, 4, 5)
y <- c(6, 7, 8, 9, 10)

plot(x)
```




```r
plot(y)
```



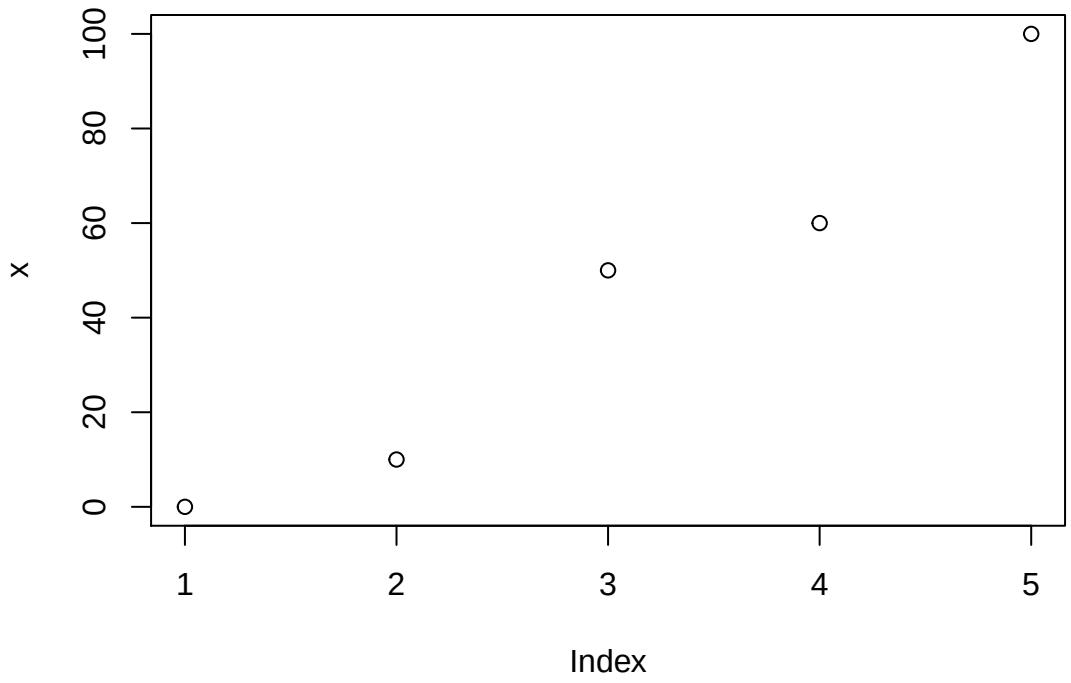
ANSWER -> they have similar spread, so same standard deviation, but (2) has a higher mean.

LE1-5d)

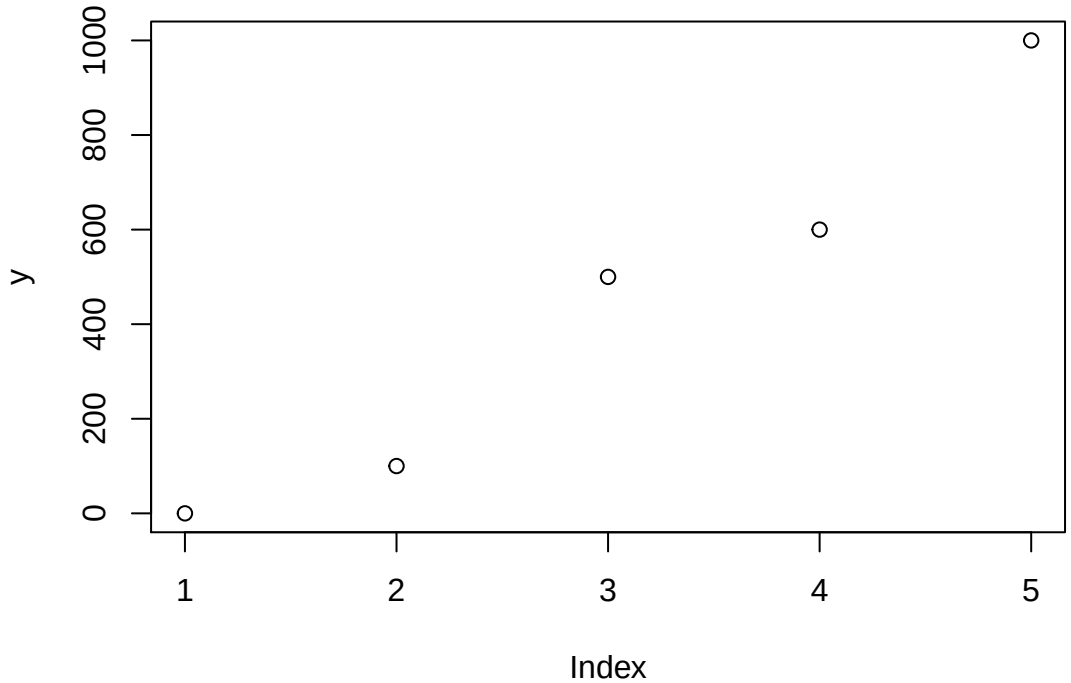
(1) 0, 10, 50, 60, 100
(2) 0, 100, 500, 600, 1000

```
x <- c(0, 10, 50, 60, 100)
y <- c(0, 100, 500, 600, 1000)

plot(x)
```




```
plot(y)
```



ANSWER -> same spread : same standard deviation, mean of (2) is higher than (1)

LE1-6. (½ pt.) For Loops

Using a for loop

- complete the problem below in the given code space
- Create a data frame of
  - the average temperature (Temp) and
  - wind speeds (Wind) for each month
- The data frame must have 3 columns -
  - average temperature,
  - average wind speed, and
  - month number (5, 6, etc.),
- colnames are up to you

You may only use one for loop - You may not hard code (i.e. type in manually) - the number of each month -Hint: you may find the unique() function useful

```
data("airquality")
# head(airquality)
# str(airquality)

#check air quality columns for nonusable -  NA data
 is.na(airquality$Wind)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
 is.na(airquality$Temp)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
 is.na(airquality$Month)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
# no cleaning required data has no Na values

# find the months for which data is available
dt_month = unique(airquality$Month)

# make an empty list to fill in the desired values
avg_temp_per_month <- vector(mode = "list", length = 0)
wind_speed_per_month <- vector(mode = "list", length = 0)

#iterate through each month whoes data is available
for(i in dt_month){
    aqPm <- subset(airquality, Month == i) # break aq into monthly dtset
    sz <- dim(aqPm)[1]# find the number of rows

    wind_sum <- sum(aqPm$Wind) # find the sum of wind
    avg_wind <- wind_sum/sz # average of sum
    wind_speed_per_month <-append(wind_speed_per_month, avg_wind)  # append to wind/month

    temp_sum <-sum(aqPm$Temp)
    avg_temp <- temp_sum/sz
    avg_temp_per_month <- append(avg_temp_per_month,avg_temp)

}
wind_speed_per_month <- unlist(wind_speed_per_month)
avg_temp_per_month <- unlist(avg_temp_per_month)


airquality_daughter.data <- data.frame(wind_speed_per_month, avg_temp_per_month, dt_month)

# View(airquality_daughter.data)
head(airquality_daughter.data)
```

```
##   wind_speed_per_month avg_temp_per_month dt_month
## 1            11.622581           65.54839        5
## 2            10.266667           79.10000        6
## 3             8.941935           83.90323        7
## 4             8.793548           83.96774        8
## 5            10.180000           76.90000        9
```

```
# ?unique
```

LE1-7. (2 pts.) Heart Transplants

Heart Transplants, Chapter 2, Exercise 2.26, p. 76. The Stanford University Heart Transplant Study was conducted to determine

- whether an experimental heart transplant program increased lifespan.
- Each patient entering the program was designated
  - an official heart transplant candidate,
  - meaning that they were gravely ill
  - and would most likely benefit from a new heart.
- Some patients got a transplant and some did not.

The variable transplant indicates

- which group the patients were in;
  - patients in the treatment group got a transplant and
  - those in the control group did not.

Of the 34 patients in the control group,

- 30 died.

Of the 69 people in the treatment group,

- 45 died.

Another variable called survived was used

- to indicate whether or not
- the patient was alive at the end of the study.[1]

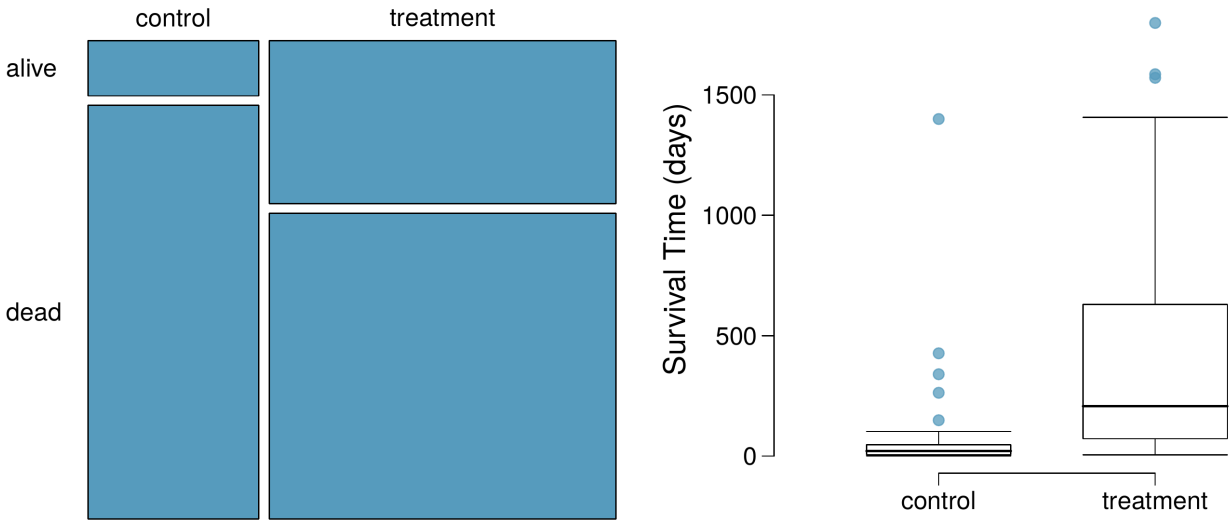


Figure 1: figures

LE1-7a) Based on the mosaic plot,

- is survival independent of whether or not the patient got a transplant?

- Explain your reasoning.

ANSWER -> Looking at the mosaic plot : 24/69 = 0.347 survived in the treatment group and 4/34 = 0.117 survived in the control group. Survival is more than twice likely in the treatment group, therefore there is a dependence of survival on taking transplant.

LE1-7b) What do the box plots above

- suggest about the efficacy (effectiveness) of the heart transplant treatment.

ANSWER -> The box plot gives a variable survival time in days for both controland treatment group. The meadian survival days for contol groupis 10 and treatment grup is 250, this means half of the people in the control group died in 10 days whereas half of the people in the treatment group died after 250 days, which certainly shows that transplant pushed the survival days for most people in the group. The box edges give us an idea about the rest of the 50 %, in each group, most of these in treatmet group lived for 550 days, but most of the 50% left in the control group lived only for like 20days. The box plot also tells us about outliers in each group, the once that did not follow the trend, there were many outliers in control group but their survival days were equivalent to the upper edge of the box plot for treatment, which means the outliers in control group did not outlive the treatment group, and the outliers in treatment group outlived the outliers in control group.

Holistically, this plot suggests that the heart transplant patients lived more than the once who did not have transplant.

LE1-7c) What proportion

- of patients in the treatment group and what proportion of patients in the control group
- died?

ANSWER -> 30/34 = 0.882 died in the control group and 45/69 = 0.652 died in the treatment group

LE1-7d) One approach for investigating whether or not the treatment is effective

- is to use a randomization technique.

LE1-7d-i. What are the claims being tested? ANSWER -> Transplant and survival are related or independent

LE1-7d-ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software.

- Fill in the blanks with a number or phrase, whichever is appropriate.

We write alive on **28**_____ cards representing patients who were alive at the end of the study, and dead on ***75***_ cards representing patients who were not.

Then, we shuffle these cards and split them into two groups:

- one group of size _**69** representing treatment,
- and another group of size **34**_ representing control.

We calculate the difference between the proportion of dead cards in the treatment and control groups (treatment - control) and record this value.

We repeat this 100 times to build a distribution centered at_**0** .

Lastly, we calculate the fraction of simulations where the simulated differences in proportions are *larger*____ .

If this fraction is low,

- we conclude that it is unlikely to have observed such an outcome by chance
  - and that the null hypothesis should be rejected
  - in favor of the alternative hypothesis.

LE1-7d-iii. What do the simulation results shown below suggest

- about the effectiveness of the transplant program?

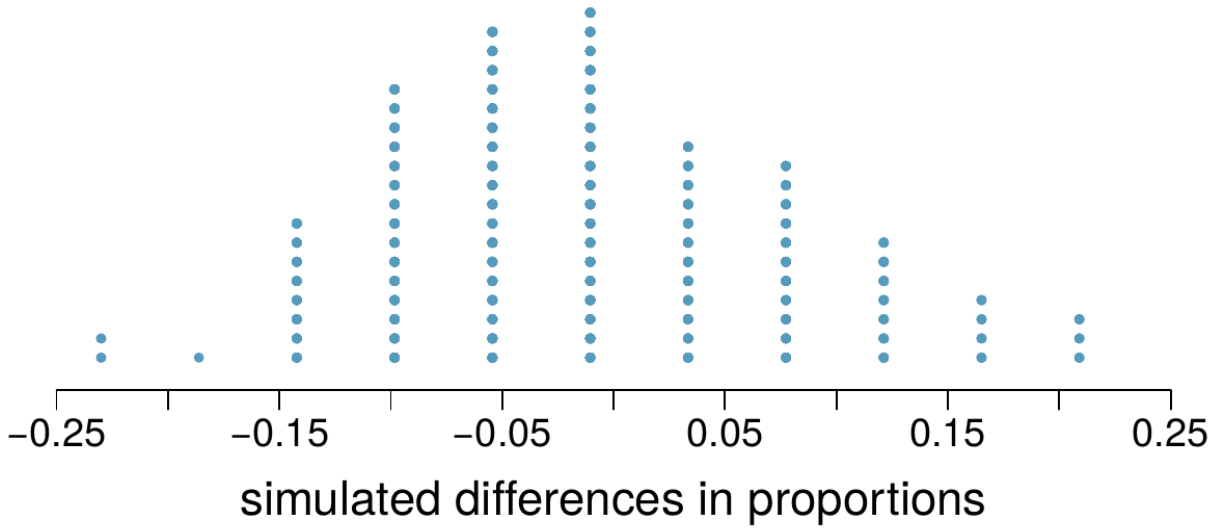


Figure 2: simulation results

ANSWER -> The observed difference is (45/69) - (30/34) = 0.184. This is what is observed. Next we see, what is the occurence of this difference in the simulation, since this difference did not even occur in the simulation, it is a rare event, and has not occurred merely by chance, and therefore, there is a relationship between status of transplant and survival . #### Links

1. B. Turnbull et al. “Survivorship of Heart Transplant Data”. In: Journal of the American Statistical Association 69 (1974), pp. 74–80. https://www.jstor.org/stable/pdf/2285502.pdf
2. http://www.r-project.org
3. http://rmarkdown.rstudio.com/
4. Open Intro Statistics version 4