ISE 3293/5013-140

Assignment #1 (Guide)

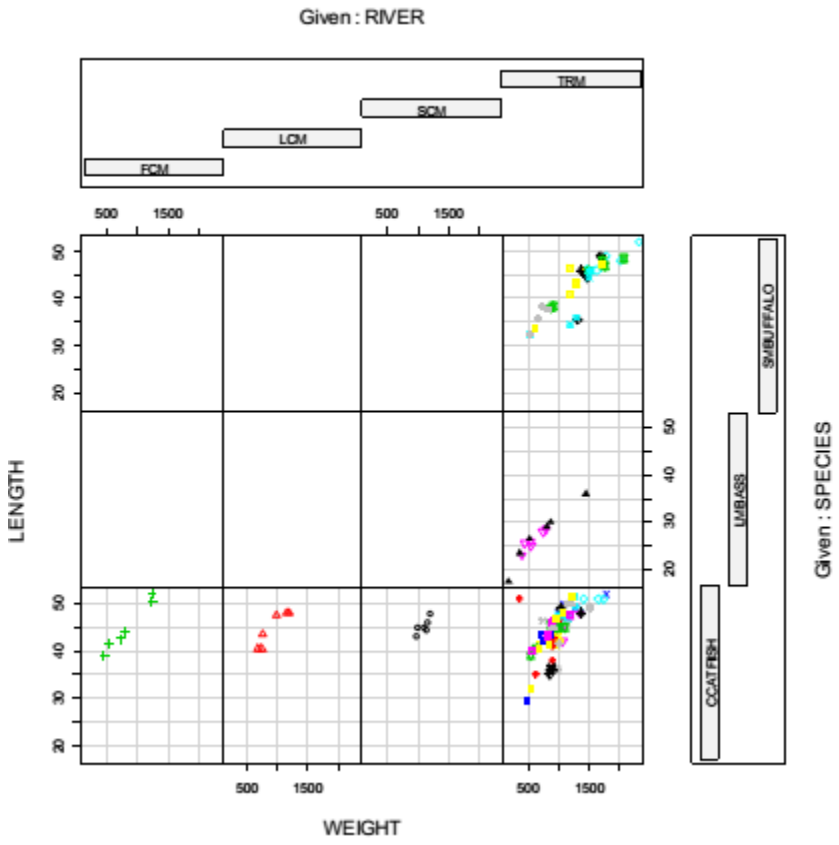
Due May 20, 2016

1) Reading **Assessment-ISE-SUMMER-2016** PDF file which was uploaded on D2L and make a summary about different parts of assessment in this course such as Assignments, Laboratories, etc.

2)

(a) Make the coplot as the biologist required **Hint:** Use coplot(), Lab 1, the code provided, and plotting options pch and col to differentiate the MILE variable. You should be able to produce something like what is shown below.

(use coplot similar to what you had in lab1)



You will have a plot similar to the above plot.

(b) Interpret the lower left three conditional plots.

(c) What does line A do?

(d) What does line B do?

(e) Why are the top six plots empty?

(f) What is the mean value of DDT found in the sample of CCATFISH caught in the FCM river? (**Final answer is 45)**

Hint:

```
ddt=read.csv("..\\CSV\\DDT.csv")
head(ddt)
subset(ddt,RIVER=="FCM" & SPECIES=="CCATFISH",) #or
ddt[ddt$RIVER=="FCM" & ddt$SPECIES=="CCATFISH",]
```

3) MS 1.14 Page 8

(For those who do not have the textbook)

1.14 *National Bridge Inventory.* All highway bridges in the United States are inspected periodically for structural deficiency by the Federal Highway Administration (FHWA). Data from the FHWA inspections are compiled into the National Bridge Inventory (NBI). Several of the nearly 100 variables maintained by the NBI are listed below. Classify each variable as quantitative or qualitative.

a. Length of maximum span (feet)
b. Number of vehicle lanes
c. Toll bridge (yes or no)
d. Average daily traffic
e. Condition of deck (good, fair, or poor)
f. Bypass or detour length (miles)
g. Route type (interstate, U.S., state, county, or city)

Definition 1.9

Quantitative data are those that are recorded on a naturally occurring numerical scale, i.e., they represent the quantity or amount of something.

Definition 1.10

Qualitative data are those that cannot be measured on a natural numerical scale, i.e., they can only be classified into categories.

Example 1.2

Characteristics of Water Pipes

The *Journal of Performance of Constructed Facilities* reported on the performance dimensions of water distribution networks in the Philadelphia area. For one part of the study, the following variables were measured for each sampled water pipe section. Identify the data produced by each as quantitative or qualitative.

a. Pipe diameter (measured in inches)
b. Pipe material (steel or PVC)
c. Pipe location (Center City or suburbs)
d. Pipe length (measured in feet)

Solution

Both pipe diameter (in inches) and pipe length (in feet) are measured on a meaningful numerical scale; hence, these two variables produce quantitative data. Both type of pipe material and pipe location can only be classified—material is either steel or PVC; location is either Center City or the suburbs. Consequently, pipe material and pipe location are both qualitative variables.

4) MS Page 12&13

a) Read the bottom of page 12 and just name the four random sampling designs.

b) Read page 13 and summarize it for each four random sampling design.

5) MS 1.15 Page 15

1.15 *Groundwater contamination in wells. Environmental Science & Technology* (Jan. 2005) published a study of methyl *tert*-butyl ether (MTBE) contamination in 223 New Hampshire wells. The data for the wells is saved in the **MTBE** file. Suppose you want to sample 5 of these wells and conduct a thorough analysis of the water contained in each. Use a random number generator to select a random sample of 5 wells from the 223. List the wells in your sample.

1)mtbe=read.table(file.choose( ),sep=",",header=TRUE) (Unzip Dataxls and find the MTBE file to address)

2)head(mtbe) # First six lines

3)dim(mtbe) # number of rows and columns

4)ind=sample(1:223,5,replace=FALSE) # random indices

5)mtbe[ind,]

(Answers will vary. For example my sample includes these wells: 98, 117, 61, 19 and 123)

(i) Remove all the rows in mtbe that contain one or more NA's mtbeo=na.omit(mtbe)

mtbeo=na.omit(mtbe)

(ii) Now calculate the standard deviation (sd() in R) of the depth of wells which have "Bedrock" as the Aquifier (this is using the entire mtbeo data frame)

depth=mtbeo[mtbeo$Aquifier=="Bedrock",]$Depth

sd(depth)

(Final Answer for standard deviation is : **56.45357**)

6) MS 1.16 Page 15

The first part of this problem is similar to the first part of the previous problem (1-5). The only difference is the sample size. In this problem we want to generate a sample of 30 rather than 5 and the total number of rows are 2929 rather than 223.

Your answers will vary however, it must include 30 aftershocks.

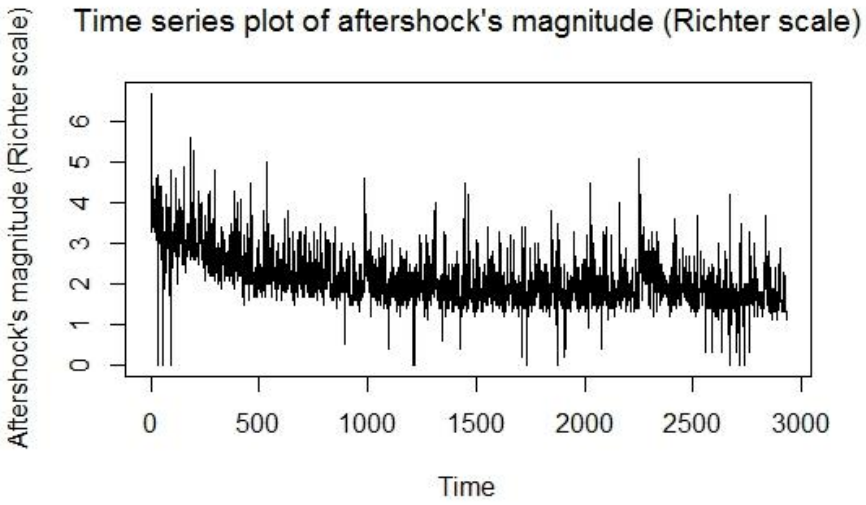
(i) Make the following plot plot(ts(eq$MAG)) (You can add some details about the title of plot, axis, etc.)

(ii) Using the entire eq data frame find the median (median()) of the MAGNITUDE variable.

magnitude = eq$MAG

median(magnitude) ⟹ the final answer is 2

1.16 *Earthquake aftershock magnitudes.* Seismologists use the term *aftershock* to describe the smaller earthquakes that follow a main earthquake. Following a major earthquake in the Los Angeles area, the U.S. Geological Survey recorded information on 2,929 aftershocks. Data on the magnitudes (measured on the Richter scale) for the 2,929 aftershocks are saved in the **EARTHQUAKE** file. Use a random number generator to select a random sample of 30 aftershocks from the **EARTHQUAKE** file. Identify the aftershocks in your sample.



Your plot must be similar to above plot for part (i).

7) MS Statistics in action. Read the Page 18 and answer the following:

STATISTICS IN ACTION REVISITED

DDT Contamination of Fish in the Tennessee River — Identifying the Data Collection Method, Population, Sample, and Types of Data

We now return to the U.S. Army Corps of Engineers study of the level of DDT contamination of fish in the Tennessee River (Alabama). Recall that the engineers collected fish specimens at different locations along the Tennessee River (TR) and three tributary creeks: Flint Creek (FC), Limestone Creek (LC), and Spring Creek (SC). Consequently, each fish specimen represents the *experimental unit* for this study. Five *variables* were measured for each captured fish: location of capture, species, weight (in grams), length (in centimeters), and DDT concentration (ppm). These data are saved in the **DDT** file. Upon examining the data you will find that capture location is represented by the columns "River" and "Mile". The possible values of "River" are TR, FC, LC, and SC (as described above), while "Mile" gives the distance (in miles) from the mouth of the river or creek. Three species of fish were captured: channel catfish, largemouth bass, and smallmouth buffalofish. Both capture location and species are categorical in nature, hence they are *qualitative* variables. In contrast, weight, length, and DDT concentration are measured on numerical scales; thus, these three variables are *quantitative*.

The data collection method is actually a *designed experiment*, one involving a stratified sample. Why? The Corps of Engineers made sure to collect samples of fish at each of the river and tributary creek locations. These locations represent the different strata for the study. The MINITAB printout shown in Figure SIA1.1 shows the number of fish specimens collected at each river location. You can see that 6 fish were captured at each of the three tributary creeks, and either 6, 8, 10, or 12 fish were captured at various locations (miles upstream) along the Tennessee River, for a total of 144 fish specimens. Of course the data for the 144 captured fish represent a *sample* selected from the much larger *population* of all fish in the Tennessee River and its tributaries.

The U.S. Army Corps of Engineers used the data in the **DDT** file to compare the DDT levels of fish at different locations and among different species, and to determine if any of the quantitative variables (e.g., length and weight) are related to DDT content. In subsequent chapters, we demonstrate several of these analyses.

(a) What is the data collection method? Read second paragraph carefully.

(b) What is the population?

(c) Give the names of all the **qualitative** variables.

Definition 1.4

A statistical **population** is a data set (usually large, sometimes conceptual) that is our target of interest.

Definition 1.5

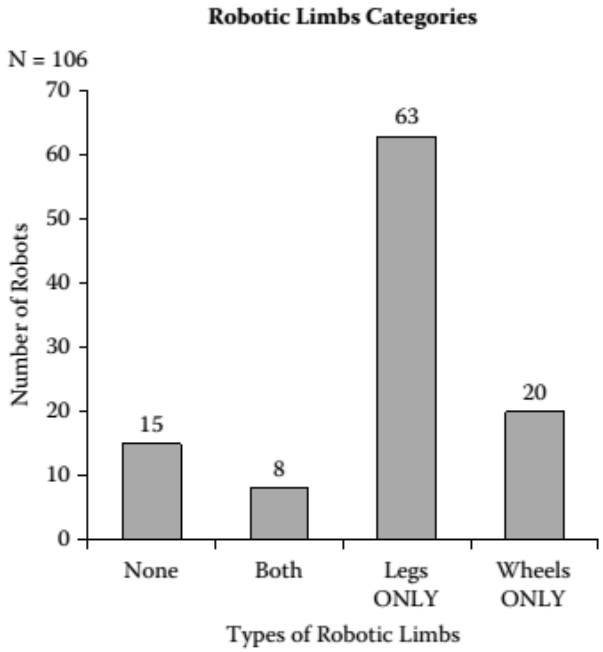
A **sample** is a subset of data selected from the target population.

8) MS 2.1 Page 26 Use pareto() **Hint:**

```
freq=c(15,8,63,20)
RL=c("None","Both","LegsO","WheelsO")
l=rep(RL,freq)
```

2.1 Do social robots walk or roll? According to the United Nations, social robots now outnumber industrial robots worldwide. A social (or service) robot is designed to entertain, educate, and care for human users. In a paper published by the *International Conference on Social Robotics* (Vol. 6414, 2010),

design engineers investigated the trend in the design of social robots. **Using a random sample of 106 social robots obtained through a web search, the engineers found that 63 were built with legs only, 20 with wheels only, 8 with both legs and wheels, and 15 with neither legs nor wheels**. This information is portrayed in the accompanying figure.



a. What type of graph is used to describe the data? Read the summary from the textbook then write the correct answer.

Summary of Graphical Descriptive Methods for Qualitative Data

Bar Graph: The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency, or class percentage.

Pie Chart: The categories (classes) of the qualitative variable are represented by slices of a pie (circle). The size of each slice is proportional to the class relative frequency.

Pareto Diagram: A bar graph with the categories (classes) of the qualitative variable (i.e., the bars) arranged by height in descending order from left to right.

b. Identify the variable measured for each of the 106 robot designs.

c. Use graph to identify the social robot design that is currently used the most.

d. Compute class relative frequencies for the different categories shown in the graph. Read example 2.1 in the textbook (and the solution) to review how to calculate relative frequency for a class or category.

Example 2.1

Graphing Qualitative Data Characteristics of Ice Meltponds

PONDICE

The National Snow and Ice Data Center (NSIDC) collects data on the albedo, depth, and physical characteristics of ice meltponds in the Canadian Arctic. Environmental engineers at the University of Colorado are using these data to study how climate impacts the sea ice. Data for 504 ice meltponds located in the Barrow Strait in the Canadian Arctic are saved in the PONDICE file. One variable of interest is the type of ice observed for each pond. Ice type is classified as first-year ice, multiyear ice, or landfast ice. Construct a summary table and a horizontal bar graph to describe the ice types of the 504 meltponds. Interpret the results.

Solution

The data in the **PONDICE** file were analyzed using SAS. Figure 2.4 shows a SAS summary table for the three ice types. Of the 504 meltponds, 88 had first-year ice, 220 had multiyear ice, and 196 had landfast ice. The corresponding proportions (or relative frequencies) are 88/504 = .175, 220/504 = .437, and 196/504 = .389. These proportions are shown in the "Percent" column in the table and in the accompanying SAS horizontal bar graph in Figure 2.4. The University of Colorado researchers used this information to estimate that about 17% of meltponds in the Canadian Arctic have first-year ice.

The FREQ Procedure

ICETYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
First-year	88	17.46	88	17.46
Landfast	196	38.89	284	56.35
Multi-year	220	43.65	504	100.00

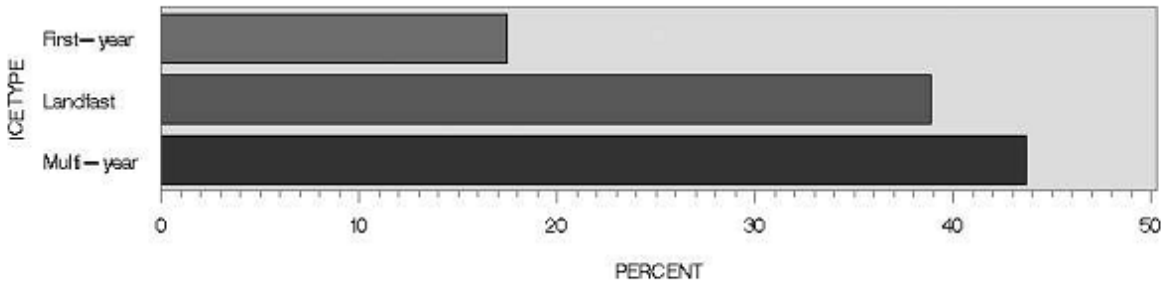
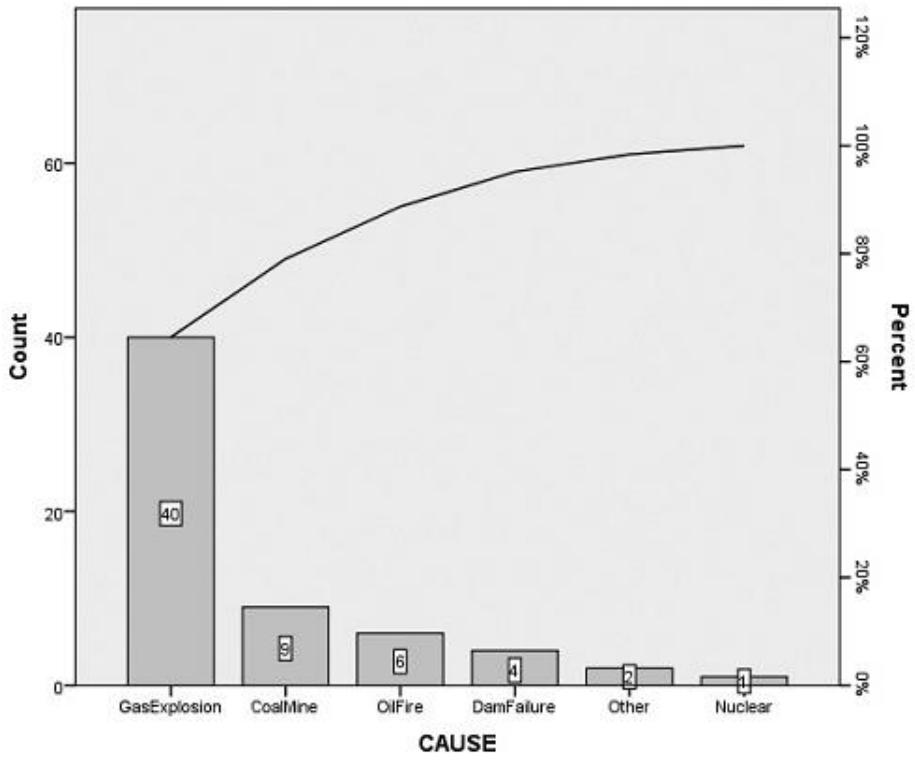


FIGURE 2.4
SAS analysis of ice types for meltponds

e. Use the results, part d, to construct a Pareto diagram for the data.

Read the textbook page 24 to review Pareto diagram.

Vertical bar graphs like Figure 2.1 can be enhanced by arranging the bars on the graph in the form of a **Pareto diagram**. A Pareto diagram (named for the Italian economist Vilfredo Pareto) is a frequency bar graph with the bars displayed in order of height, starting with the tallest bar on the left. Pareto diagrams are popular graphical tools in process and quality control, where the heights of the bars often represent frequencies of problems (e.g., defects, accidents, breakdowns, and failures) in the production process. Because the bars are arranged in descending order of height, it is easy to identify the areas with the most severe problems.




```
# to construct a Pareto diagram in R for problem 2.1 in textbook :
freq=c(15,8,63,20)
RL=c("None","Both","LegsO","WheelsO")
l=rep(RL,freq)
pareto(l)
```

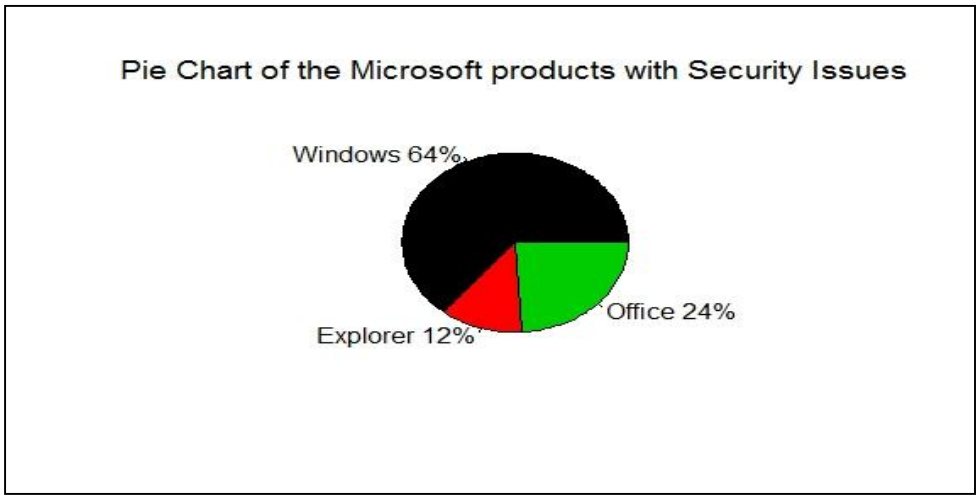
9) MS 2.4 - Page 27 - Please use the pareto() function I made.

2.4 Microsoft program security issues. The dominance of Microsoft in the computer software market has led to numerous malicious attacks (e.g., worms, viruses) on its programs. To help its users combat these problems, Microsoft periodically issues a Security Bulletin that reports the software affected by the

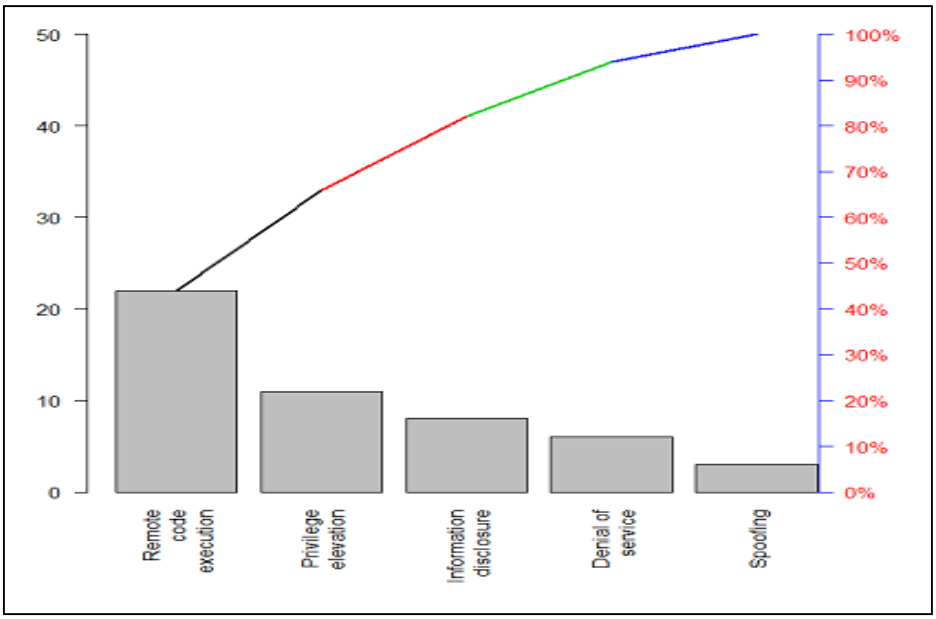
vulnerability. In *Computers & Security* (July 2013), researchers focused on reported security issues with **three Microsoft products: Office, Windows, and Explorer**. In a sample of **50** security bulletins issued in 2012, **32 reported a security issue with Windows, 6 with Explorer, and 12 with Office**. The researchers also categorized the security bulletins according to the expected repercussion of the vulnerability. Categories were **Denial of service**, **Information disclosure**, **Remote code execution**, **Spoofing,** and **Privilege elevation**. Suppose that of the 50 bulletins sampled, the following numbers of bulletins were classified into each respective category: **6, 8, 22, 3, 11**.

a. Construct **a pie chart** to describe the **Microsoft products** with security issues. Which product had the lowest proportion of security issues in 2012?

Your final plot should be something like this. Which product had the lowest proportion of security issues in 2012?



b. Construct a **Pareto diagram** to describe the expected repercussions from security issues. Based on the graph, what repercussion would you advise Microsoft to focus on? similar to previous problem.



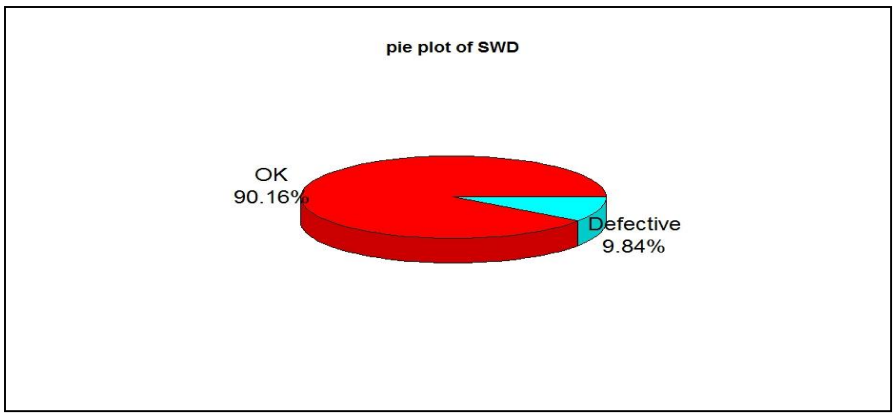
Your diagram will be similar to the above diagram.

10) MS 2.10 - Page 28 – Use pie3D() from plotrix package (may need to install it) **Hint:**
swd=read.csv("..//CSV//SWDEFECTS.csv", header=TRUE)
head(swd)
library(plotrix)
tab=table(swd$defect)
rtab=tab/sum(tab)
round(rtab,2)
pie3D(rtab,labels=list("OK","Defective"),main="pie plot of SWD")

2.10 Software defects. The PROMISE Software Engineering Repository is a collection of data sets available to serve researchers in building predictive software models. One such data set, saved in the **SWDEFECTS** file, contains information on 498 modules of software code. Each module was analyzed for defects and classified as "true" if it contained defective code and "false" if not. Access the data file and produce a pie chart for the defect variable. **Use the pie chart to make a statement about the likelihood of defective software code.**

```
swd=read.table(file.choose(),sep=",",header=TRUE)

head(swd)
library(plotrix)
tab=table(swd$defect)
rtab=tab/sum(tab)
round(rtab,2)
pie3D(rtab,labels=list("OK","Defective"),main="pie plot of SWD") (you can add more details to plot)
```



Your final plot will be similar to this. Based on chart, make a statement about the likelihood of defective software code.

11) MS 2.72 - Page 70 (This problem is long it is better to solve the remaining problems first and then solve this one)

2.72 Process voltage readings. A Harris Corporation/University of Florida study was undertaken to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the old and new locations and voltage readings on the process were obtained. A "good process" was considered to be one with voltage readings of at least 9.2 volts (with larger readings being better than smaller readings). The table contains voltage readings for 30 production runs at each location.

VOLTAGE

Old Location			New Location		
9.98	10.12	9.84	9.19	10.01	8.82
10.26	10.05	10.15	9.63	8.82	8.65
10.05	9.80	10.02	10.10	9.43	8.51
10.29	10.15	9.80	9.70	10.03	9.14
10.03	10.00	9.73	10.09	9.85	9.75
8.05	9.87	10.01	9.60	9.27	8.78
10.55	9.55	9.98	10.05	8.83	9.35
10.26	9.95	8.72	10.12	9.39	9.54
9.97	9.70	8.80	9.49	9.48	9.36
9.87	8.72	9.84	9.37	9.64	8.68

Source: Harris Corporation, Melbourne, FL.

When answering this question you will need to do most of the construction by hand. Unlike other questions please follow parts a) -m) in conjunction with MS as I have given below. For constructing the histogram and table below use the left end point as 8.0 and right end point as 10.6, with 9 classes. After constructing table 1 make the graph in **R** using barplot(...,space=0), use the classes as names to the vector containing the frequencies.

(a) Fill out the table when constructing the Histogram in pt a). Then plot the histogram by first creating a vector, 'v' say, of relative frequencies, then use names(v) and assign class names to each component, finally using barplot(v,space=0) make your plot.

(b) Use the stem() function in **R** for part b).
(c) Use **R** to make the histogram. Do NOT use hist()

Class	Class Interval	Data Tabulation	Frequency	Relative Frequency
1	8.0000-8.2889			
2				
3				
4				
5				
6				
7				
8				
9				
Total				

Table 1: Histogram table

Read section 2.2 (Graphical Methods for Describing Quantitative Data)

a. Construct a relative frequency histogram for the voltage readings of the old process.

b. Construct a stem-and-leaf display for the voltage readings of the old process. Which of the two graphs in parts **a** and **b** is more informative about where most of the voltage readings lie?

Use stem() but before that define a vector.

c. Construct a relative frequency histogram for the voltage readings of the new process.

d. Compare the two graphs in parts **a** and **c**. (You may want to draw the two histograms on the same graph.) Does it appear that the manufacturing process can be established locally (i.e., is the new process as good as or better than the old)?

e. Find and interpret the mean, median, and mode for each of the voltage readings data sets. Which is the preferred measure of central tendency? Explain. (Calculate the mean, median and mode similar to previous problems.)

f. Calculate the *z*-score for a voltage reading of 10.50 at the old location.

g. Calculate the *z*-score for a voltage reading of 10.50 at the new location.

Key Formulas

$\frac{\text{Category frequency}}{n}$	Category relative frequency 23
$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	Sample mean 39
$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$	Sample variance 46
$s = \sqrt{s^2}$	Sample standard deviation 46
$z = \frac{y - \bar{y}}{s}$	Sample z-score 52
$z = \frac{y - \mu}{\sigma}$	Population z-score 52
$\text{IQR} = Q_U - Q_L$	Interquartile range 56
$Q_L - 1.5(\text{IQR})$	Lower inner fence 56
$Q_U + 1.5(\text{IQR})$	Upper inner fence 56
$Q_L - 3(\text{IQR})$	Lower outer fence 57
$Q_U + 3(\text{IQR})$	Upper outer fence 57

h. Based on the results of parts **f** and **g**, at which location is a voltage reading of 10.50 more likely to occur? Explain.
i. Construct a box plot for the data at the old location. Do you detect any outliers?
j. Use the method of *z*-scores to detect outliers at the old location.
k. Construct a box plot for the data at the new location. Do you detect any outliers?
l. Use the method of *z*-scores to detect outliers at the new location.
m. Compare the distributions of voltage readings at the two locations by placing the box plots, parts **i** and **k**, side by side vertically.

12) MS 2.73 - Page 70

2.73 Surface roughness of pipe. Refer to the *Anti-corrosion Methods and Materials* (Vol. 50, 2003) study of the surface roughness of coated oil field pipes, Exercise 2.20 (p. 37). The data (in micrometers) are repeated in the table. Give an interval that will likely contain about 95% of all coated pipe roughness measurements.

ROUGHPIPE

1.72	2.50	2.16	2.13	1.06	2.24	2.31	2.03	1.09	1.40
2.57	2.64	1.26	2.05	1.19	2.13	1.27	1.51	2.41	1.95

Chapter Summary Notes

- Graphical methods for qualitative data: **pie chart, bar graph**, and **Pareto diagram**
- Graphical methods for quantitative data: **dot plot, stem-and-leaf display**, and **histogram**
- Numerical measures of central tendency: **mean, median**, and **mode**

- Numerical measures of variation: **range, variance**, and **standard deviation**
- Sample numerical descriptive measures are called **statistics**.
- Population numerical descriptive measures are called **parameters**.
- Rules for determining the percentage of measurements in the interval (mean) $\pm$ 2 (std. dev.): **Chebyshev's Rule** (at least 75%) and **Empirical Rule** (approximately 95%)
- Measures of relative standing: **percentile score** and ***z*-score**
- Methods for detecting outliers: **box plots** and ***z*-scores**

The Empirical Rule

If a data set has an approximately mound-shaped, symmetric distribution, then the following rules of thumb may be used to describe the data set (see Figure 2.12a):

1. Approximately 68% of the measurements will lie within 1 standard deviation of their mean (i.e., within the interval $\bar{y} \pm s$ for samples and $\mu \pm \sigma$ for populations).
2. Approximately 95% of the measurements will lie within 2 standard deviations of their mean (i.e., within the interval $\bar{y} \pm 2s$ for samples and $\mu \pm 2\sigma$ for populations).
3. Almost all the measurements will lie within 3 standard deviations of their mean (i.e., within the interval $\bar{y} \pm 3s$ for samples and $\mu \pm 3\sigma$ for populations).

According to Empirical Rule we know that approximately **95%** of all observations will be **within 2 standard deviations of the mean**.

```
roughpipe=read.table(file.choose(),sep=",",header=TRUE)
roughpipe = with(roughpipe, ROUGH)      # make a vector with ROUGH values
mean_roughpipe = mean(roughpipe)        # calculate the mean of ROUGH
std_roughpipe = 2*sd(roughpipe)         # calculate 2*std of ROUGH
Interval95Beg = round(mean_roughpipe -  std_roughpipe,4)     # calculate the lower band
Interval95End = round(mean_roughpipe +  std_roughpipe,4)     # calculate the upper band
```

The final answer for interval is: **(0.8332, 2.9288)**

13) MS 2.80 - Page 72

2.80 Mongolian desert ants. The *Journal of Biogeography* (Dec. 2003) published an article on the first comprehensive study of ants in Mongolia (Central Asia). Botanists placed seed baits at 11 study sites and observed the ant species attracted to each site. Some of the data recorded at each study site are provided in the table at the top of p. 73.

GOBIANTS

Site	Region	Annual Rainfall (mm)	Max. Daily Temp. (°C)	Total Plant Cover (%)	Number of Ant Species	Species Diversity Index
1	Dry Steppe	196	5.7	40	3	.89
2	Dry Steppe	196	5.7	52	3	.83
3	Dry Steppe	179	7.0	40	52	1.31
4	Dry Steppe	197	8.0	43	7	1.48
5	Dry Steppe	149	8.5	27	5	.97
6	Gobi Desert	112	10.7	30	49	.46
7	Gobi Desert	125	11.4	16	5	1.23
8	Gobi Desert	99	10.9	30	4	.
9	Gobi Desert	125	11.4	56	4	.76
10	Gobi Desert	84	11.4	22	5	1.26
11	Gobi Desert	115	11.4	14	4	.69

a. Find the mean, median, and mode for the number of ant species discovered at the 11 sites. Interpret each of these values.

2.4 Measures of Central Tendency

The three most common measures of central tendency are the **arithmetic mean**, the **median**, and the **mode**. Of the three, the arithmetic mean (or **mean**, as it is commonly called) is used most frequently in practice.

Definition 2.6

The **arithmetic mean** of a set of n measurements, $y_1, y_2, \ldots, y_n$, is the average of the measurements:

$$\frac{\sum_{i=1}^{n} y_i}{n}$$

Typically, the symbol $\bar{y}$ is used to represent the **sample mean** (i.e., the mean of a sample of n measurements), whereas the Greek letter μ represents the **population mean**.

To illustrate, we will calculate the mean for the set of $n = 5$ sample measurements: 4, 6, 1, 2, 3. Substitution into the formula for $\bar{y}$ yields

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{4 + 6 + 1 + 2 + 3}{5} = 3.2$$

Definition 2.7

The **median** of a set of n measurements, $y_1, y_2, \ldots, y_n$, is the middle number when the measurements are arranged in ascending (or descending) order, i.e., the value of y located so that half the area under the relative frequency histogram lies to its left and half the area lies to its right. We will use the symbol m to represent the *sample median* and the symbol τ to represent the *population median*.

If the number of measurements in a data set is odd, the median is the measurement that falls in the middle when the measurements are arranged in increasing order. For example, the median of the $n = 5$ sample measurements of Example 2.3 is $m = 3$. If the number of measurements is even, the median is defined to be the mean of the two middle measurements when the measurements are arranged in increasing order. For example, the median of the $n = 6$ measurements, 1, 4, 5, 8, 10, 11, is

$$m = \frac{5 + 8}{2} = 6.5$$

Calculating the Median of Small Sample Data Sets

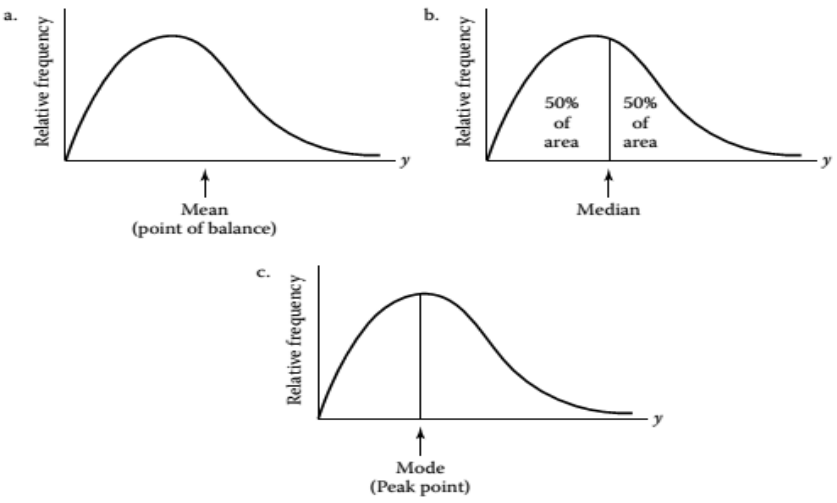
Let $y_{(i)}$ denote the ith value of y when the sample of n measurements is arranged in ascending order. Then the sample median is calculated as follows:

$$m = \begin{cases} y_{[(n+1)/2]} & \text{if } n \text{ is odd} \\ \dfrac{y_{(n/2)} + y_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Definition 2.8

The **mode** of a set of n measurements, $y_1, y_2, \ldots, y_n$, is the value of y that occurs with the greatest frequency.

FIGURE 2.9
Interpretations of the mean, median, and mode for a relative frequency distribution



If the outline of a relative frequency histogram were cut from a piece of plywood, it would be perfectly balanced over the point that locates its mean, as illustrated in Figure 2.9a. As noted in Definition 2.6, half the area under the relative frequency distribution will lie to the left of the median, and half will lie to the right, as shown in Figure 2.9b. The mode will locate the point at which the greatest frequency occurs, i.e., the peak of the relative frequency distribution, as shown in Figure 2.9c.

gobiants=read.table(file.choose(),sep=",",header=TRUE)

Define a vector for AntSpecies values. Similar to previous problem.

Use mean( ) and median( ) to calculate mean and median of ant species.

The final answer for mean is **12.8182** and the final answer for median is **5.**

According to the definition of mode, mode is the value with the greatest frequency.

You can make a table of then sort it and then use names( ).

The final answer for mode is **4.**

b. Which measure of central tendency would you recommend to describe the center of the number of ant species distribution? Explain. **To answer part b read the textbook pages 41 and 42 and write your conclusions**.

c. Find the mean, median, and mode for the total plant cover percentage **at the 5 Dry Steppe sites only**.

It is similar to the part a however, we must calculate the mean, median and mode for the plant cover percentage (PlantCov column in file) for 5 sites with Dry Steppe region.

Make table but only for Region=="Dry Steppe”

Define a vector wiht Total Plant Cover (only Dry Steppe)

Then use codes for part a to calculate mean,median and mode.

The final answers are:

Mean= 40.4, Median= 40 and Mode= 40

d. Find the mean, median, and mode for the total plant cover percentage at the 6 Gobi Desert sites only.

The steps are the same as part c.

The final answers are: **Mean=28, Median = 26 and Mode = 30**

e. Based on the results, parts **c** and **d**, does the center of the total plant cover percentage distribution appear to be different at the two regions? Draw conclusion from part c and d.

14) MS 2.84 - Page 74

2.84 Speed of light from galaxies. Astronomers theorize that cold dark matter (CDM) caused the formation of galaxies. The theoretical CDM model requires an estimate of the velocity of light emitted from the galaxy. *The Astronomical Journal* (July, 1995) published a study of galaxy velocities. One galaxy, named A1775, is thought to be a *double cluster*; that is, two clusters of galaxies in close proximity. Fifty-one velocity observations (in kilometers per second, km/s) from cluster A1775 are listed in the table.

GALAXY2

22922	20210	21911	19225	18792	21993	23059
20785	22781	23303	22192	19462	19057	23017
20186	23292	19408	24909	19866	22891	23121
19673	23261	22796	22355	19807	23432	22625
22744	22426	19111	18933	22417	19595	23408
22809	19619	22738	18499	19130	23220	22647
22718	22779	19026	22513	19740	22682	19179
19404	22193					

a. Use a graphical method to describe the velocity distribution of galaxy cluster A1775.

(For example you can make histogram)

b. Examine the graph, part **a**. Is there evidence to support the double cluster theory? Explain.
c. Calculate numerical descriptive measures (e.g., mean and standard deviation) for galaxy velocities in cluster A1775. Depending on your answer to part **b**, you may need to calculate two sets of numerical descriptive measures, one for each of the clusters (say, A1775A and A1775B) within the double cluster.

Final answers are: **Mean= 19462.24** and **sd= 532.2868**

Use mean() and sd() to calculate mean and standard deviation.
d. Suppose you observe a galaxy velocity of 20,000 km/s. Is this galaxy likely to belong to cluster A1775A or A1775B? Explain

For A1775B the answers are: **Mean= 22838.47** and **sd= 560.9767**

Make conclusions about results.