Lesson 1 Part A
Author: Atul Roy

Remember to read the definitions of
Numerical or Quantitative Data
Categorical or Qualitative data
Descriptive Statistics
Inferential Statistics
Distribution of the data

Topic 1.
Organization and Display of Numerical Data

Example 1.

The following data shows the prices of the townhomes on Truffle Lane in Montgomery County, Maryland. The data is obtained from the real property search data base of the State of Maryland.

The townhomes were built and sold in the year 2000-2001 and the prices are the initial purchase prices of the homes. Since all the houses were over $200,000 the list shows only the amount that was paid in excess of $200,000 rounded to the nearest thousandth.

price paid (over $200,000)

67	25	38	36	31	24	45	72	30	5
13	1	2	63	53	21	5	77	34	16
15	39	45	19	23	23	12	31	25	12
23	33	58							

That is the number 67 means that the house was sold for $267,000 (rounded.)
One way to organize and display such data is called Stem and Leaf Diagram

Since an ordered Stem and Leaf Diagram is more meaningful, let us first put the data in order

1	2	5	5	12	12	13	15	16	19
21	23	23	23	24	25	25	30	31	31
33	34	36	38	39	45	45	53	58	63
67	72	77							

Treating all the numbers as two digit numbers, that is 1 as 01 etc., we may use the first digit as stem and the second digit as leaf and the display will be

0*1255
1*223569
2*1333455
3*01134689
4*55
5*38
6*37
7*27

Note that a stemplot will give an idea of the shape of a distribution. It will help you visually locate the outliers (we shall see a well formulated approach to outliers in the lesson 2.)

It helps us locate the vital numbers, for example

0*1255
1*223569
2*1333455
3*01134689
4*55
5*38
6*37
7*27

you may see that 25 is in the center of the distribution in the sense that it is right in the middle, telling us that half of the homes were sold at 225,000 or below.

Such a value is called the median of the data, we shall take median up more in detail in the lesson 2.

Even though the time of purchase and the number of options had influence on the price, all the townhouses $245,000 or over are the end units.

An advantage of a stemplot is that we can see each and every individual entry but we can not use such an advantage when the data set is large.

Frequency Distribution

Example 2.

We are going to see how to organize the data in a table called a frequency distribution. Such an organization is very useful when the data set is large.

The following data shows the points per game (total points scored/ # of games played) by the players of the teams in the playoffs in the year 2003 that were available on www.nba.com
for 175 players.

Points per Game									
27.1	17.3	15.8	9.9	9.6	6.6	4.7	1.0	0.3	25.3
19.5	18.3	16.1	8.0	6.1	5.7	5.7	3.1	2.9	2.5
1.2	20.1	18.9	14.1	10.8	7.8	6.7	6.5	5.9	2.5
1.8	0.6	0.5	31.7	9.4	4.9	4.0	4.0	3.0	2.8
0.9	0.8	27.0	23.5	14.5	8.3	7.0	5.2	4.8	4.7
2.8	2.3	1.0	31.7	13.6	11.5	10.4	7.8	2.6	1.9
0.8	0.0	24.7	14.7	12.8	9.4	9.3	7.8	6.9	5.2
2.6	2.2	1.8	1.3	19.6	14.8	11.6	11.2	9.2	7.2
5.0	4.8	4.5	4.4	3.0	2.0	32.1	27.0	12.8	8.0
6.0	5.6	3.8	3.3	3.2	2.8	2.1	0.4	22.5	18.0
9.4	9.3	8.9	7.8	6.1	5.5	5.0	4.3	2.7	0.5
22.0	18.5	14.2	12.7	6.0	5.3	4.0	2.5	1.0	0.7
22.8	19.0	9.2	8.7	8.5	6.5	5.7	4.2	4.0	3.3
3.0	2.3	19.0	17.4	15.3	13.9	10.0	10.0	7.5	5.8
3.7	3.0	1.8	1.0	17.8	17.2	14.8	9.7	5.0	3.8
1.0	0.5	23.7	23.1	14.3	12.7	11.4	11.3	9.1	5.3
4.6	3.0	1.5	0.9	24.8	20.4	13.2	11.5	10.2	7.0
5.8	4.3	3.8	3.6	3.5					

Note that the smallest observation is 0 and the largest is 32.1. We may choose to split the data in classes 4 units wide.

Remember that while making the classes, we have to make sure that

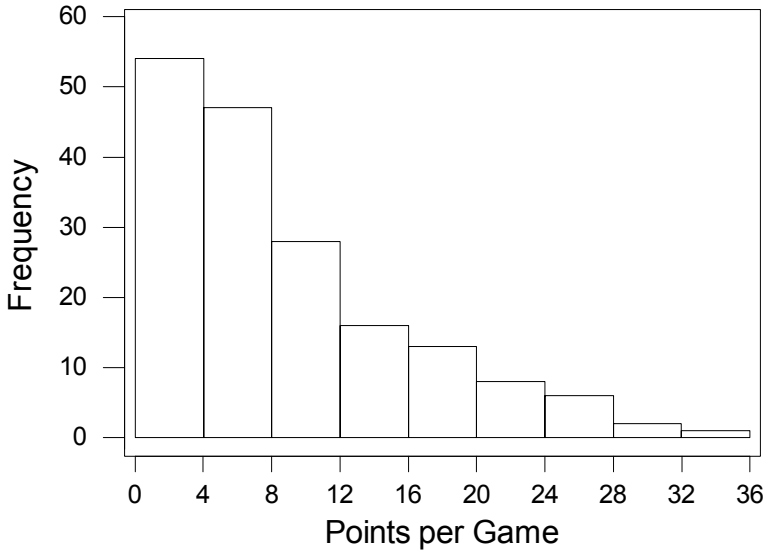
1. The classes are disjoint
2. They cover the entire data

Note that Relative F r e q = $\frac{\text{Frequency}}{\text{Total}}$

Class	Frequency	Relative Frequency
$0 \le x < 4$	54	$\frac{54}{175} = 0.30857$
$4 \le x < 8$	47	$\frac{47}{175} = 0.26857$
$8 \le x < 12$	28	$\frac{28}{175} = 0.16$
$12 \le x < 16$	16	$\frac{16}{175} = 0.09143$
$16 \le x < 20$	13	$\frac{13}{175} = 0.07429$
$20 \le x < 24$	8	$\frac{8}{175} = 0.04571$
$24 \le x < 28$	6	$\frac{6}{175} = 0.03429$
$28 \le x < 32$	2	$\frac{2}{175} = 0.01143$
$32 \le x < 36$	1	$\frac{1}{175} = 0.0057$

A graphical display of a frequency table is a histogram in which we display the classes on the horizontal axis and rectangles with altitudes equal to the frequency of the class.

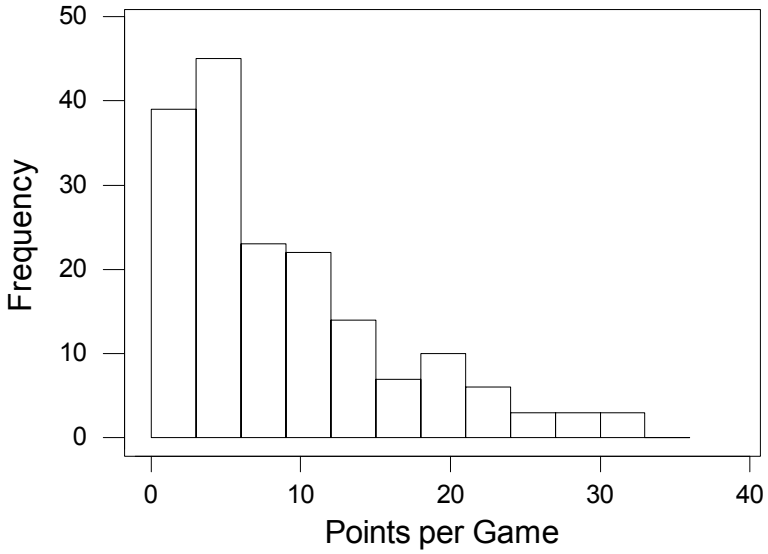
The histogram to display the above frequency table is shown below.



Note that the bars to display a histogram have to be adjacent (unless there is a class with 0 frequency) and the classes must be equally wide.

In the above display, you may note that most of the data clusters in the first two classes and the distribution is skewed to the right.

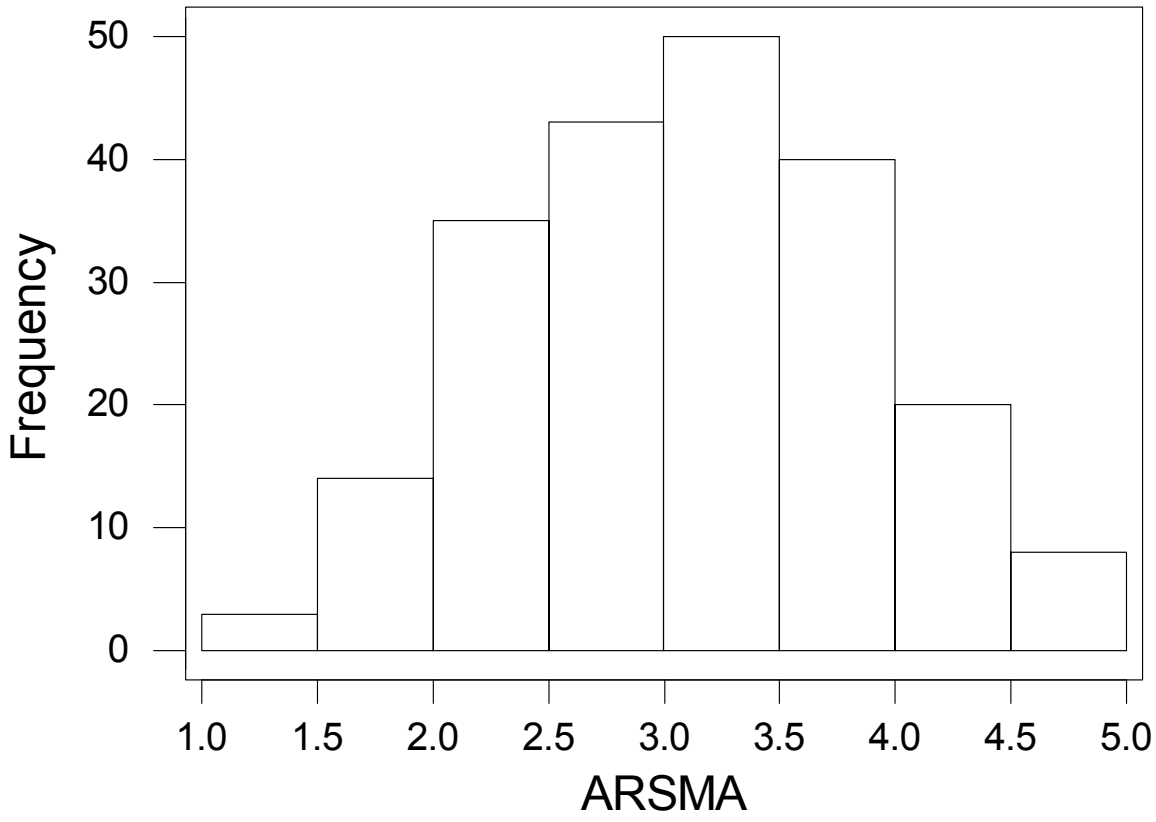
Changing the class length will show a different view of the distribution, for example if we took a class width of 3 in this data, the histogram will look like



When the data set is very large, a histogram may be made with the use of relative frequencies.

Example 4:

This example relates to reading information back from a given graph.
Suppose that the following histogram shows a distribution of the ARSMA (Acculturation Rating Scale for Mexican Americans) for a group of 213 people.



The right end point is excluded in counting the frequency.
That is, the first class shows the number of people in the range $1 \leq x < 1.5$. The altitude of the rectangle representing the class interval $1 \leq x < 1.5$ is approximately 3, therefore the numer of people (frequency) belonging to this particular class is 3. The tallest rectangle is that of the class $3 \leq x < 3.5$, note that the altitude of this class is 50, therefore 50 people scored in the interval $3 \leq x < 3.5$.

Let us answer the following questions.
a) approximately how many people scored below 2 points?
b) approximately how many people in this group scored 2 or more points.

For part a)
Note that the frequency of the class $1 \leq x < 1.5$ is approximately 3

and that the frequency of the class $1.5 \leq x < 2$ is approximately 14

Therefore the combined frequency OR the cumulative frequency of these two classes is $3 + 14 = 17$

the answer is $\boxed{17}$

for part b)

Instead of counting and combining the frequencies of all the classes for 2 or above, we may subtract from 213 (the total number of people in this group) the

total for the classes below 2 and the answer will be
213 – 17 = 196

Therefore 200 people in this group scored 2 points above.

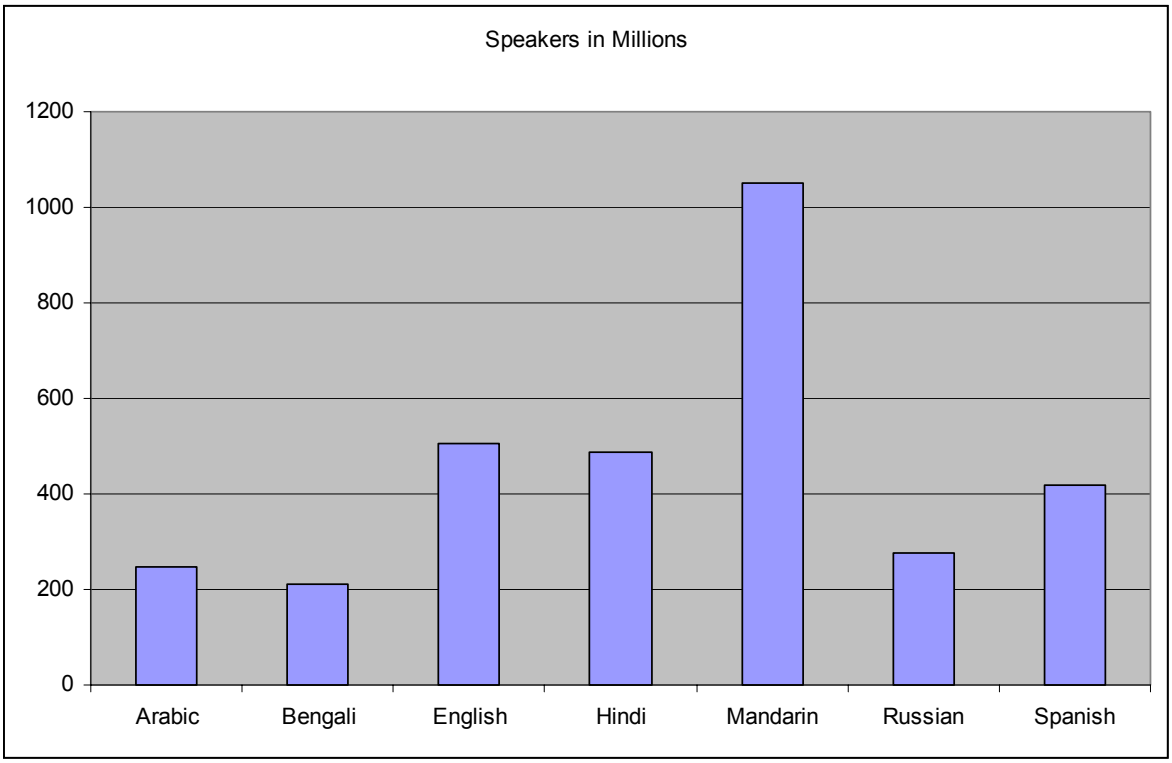
Display of Categorical Data

Example 5:

The following table is from Introductory Statistics sixth edition by Neil Weiss published by Addison Wesley.

Language	Speakers in Millions
Arabic	246
Bengali	211
English	506
Hindi	487
Mandarin	1052
Russian	277
Spanish	417

A bar graph to display such data will have the categories listed on the horizontal line and rectangles with altitude equal to the frequencies as shown below.
To stress the categorical nature of the data, here, leave gaps between the bars.

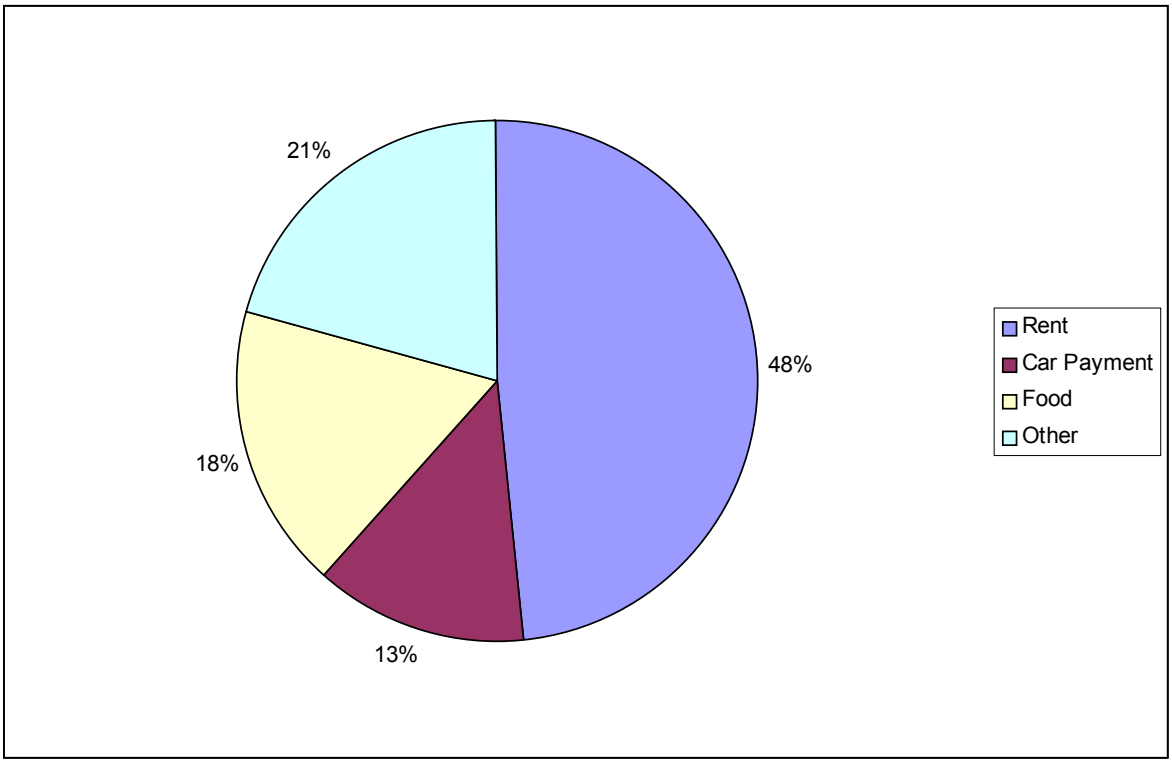


Example 4:

A family has take home family income of $3100 and the following are their expense categories.

Category	Expense
Rent	1500
Car Payment	400
Food	550
Other	650
Total	3100

Even though we can use a bar graph for a display of this data, still a pie chart will communicate the idea much better.



Example 5.

The following table will show the importance of displaying the data according to different categories.

	Alaska Airline		America West	
	On Time	Delayed	On Time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Franscisco	503	102	320	129
Seattle	1841	305	201	61

(The data is from the text The Basic Practice of Statistics by David Moore, second edition.)

Let us look at the on time rates of the two airlines overall as well as at each of the above airports.

	Alaska Airline			America West		
	Total	On Time	% on time	Total	On Time	% o
Los Angeles	559	497	88.90876565	811	694	85.57
Phoenix	233	221	94.84978541	5255	4840	92.10
San Diego	232	212	91.37931034	448	383	85.49
San Franscisco	605	503	83.14049587	449	320	71.26
Seattle	2146	1841	85.78751165	262	201	76.71
Total	3775	3274	86.72847682	7225	6438	89.10

Note that if we look at the overall percentage of on time flights it appears that America West flies more on time than Alaska Airline but if we look at each airport, Alaska seems to have a better on time rate. Such a reversal in the trend when combining the data from seveal different categories is called Simpson's Paradox. One reason for such a reversal of the trend this example is that America West flies mostly into the airports with less weather related problems.

Think of more examples in which such a paradox can happen.

Finish theAssignment 1 on time.