BNAD 276
Lecture 1
Data and Descriptive Statistics I

Phuong Ho

May 14, 2017

Outline

Motivation

- Statistics
  - *statistics is defined as the art and science of collecting, analyzing, presenting, and interpreting data.*
- It can be considered as a study of “skills” to extract meaningful and useful information from data.
- Data usually does not give us useful or meaningful information by it self.
- We need to “tailor” and analyze data to get useful and meaningful information.
- We will study skills that allow us to extract information from data in this course.

Data

- A statistician always works with data.
  - Thus, we need to understand what data is.

Data

the facts and figures collected, analyzed and summarized for presentation and interpretation.

A Data Set

all the data collected in a particular study

Example

- The following table is a data set for restaurants in Tucson.

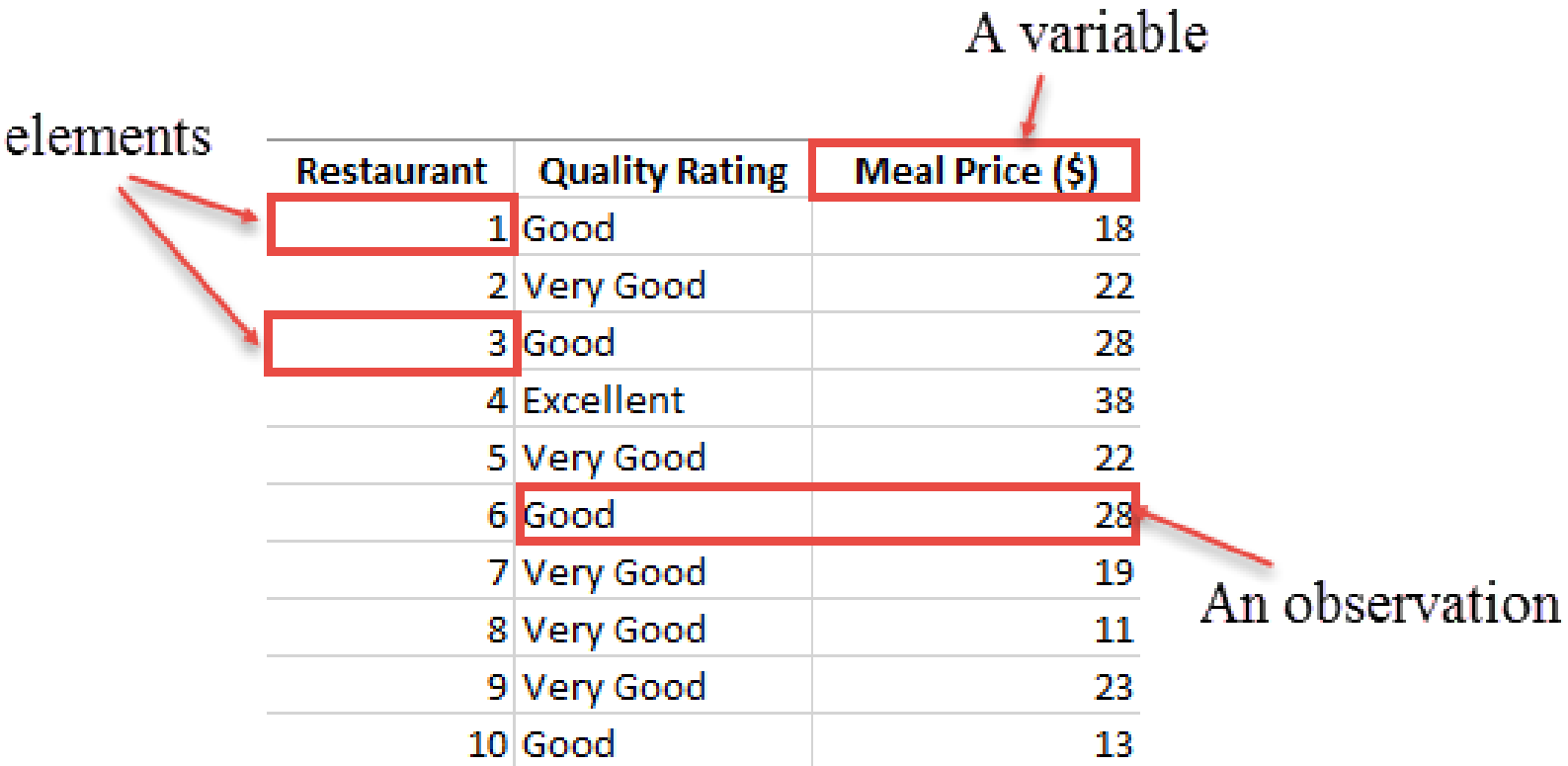
Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13

Some Terminologies

In a data set, we will find..

1. Elements: the entities on which information is collected.
2. A variable: a characteristic of interest for the elements.
3. An observation: the set of values that all variables take for a particular element. (The set of measurements obtained for a particular element.)

Example cont'd



Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	22
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13

Classifying Data

Data can be classified in two ways:

1. Categorical Data vs. Quantitative Data,
2. Cross-sectional Data vs. Time Series Data vs. Panel Data.

Categorical Data vs. Quantitative Data

Categorical Variable

A variable that can be easily grouped by specific categories

Categorical Data

A set of values that a categorical variable takes.
(or, Data that can be grouped by specific categories.)

Quantitative Variable

A variable that uses numerical values to indicate how much or how many

Quantitative Data

A set of values that a quantitative variable takes

Example

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13

- Quality Rating is a categorical variable.
- Meal Price is a quantitative variable.

Cross-sectional, Time Series, and Panel Data

Cross-sectional Data

Data collected from multiple elements at the same point in time

Time Series Data

Data collected from a single element over several time periods.

Panel Data

Data collected from multiple elements over several time periods.
Also called longitudinal data in biostatistics.

Example of Cross-section Data

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13

Example of Time Series Data

- The following data set is sample data from a particular restaurant in Tucson.

Week	Number of Commercials	Sales
1	2	50
2	5	57
3	1	49
4	3	52
5	4	58
6	1	50
7	2	52
8	3	54
9	4	56
10	5	60

Example of Panel Data

- The following data set is a sample data from students in a class.

Student Name	Period	AVG HW Score	Test Score
A	1	7.09	36
A	2	6.77	66
B	1	4.42	50
B	2	7.55	58
C	1	9.88	40
C	2	9.35	66
D	1	9.19	52
D	2	7.50	56
E	1	9.13	38
E	2	9.65	56

Population vs. Sample

Population

The set of all elements of interest in a particular study

Sample

A subset of the population

Why study sample?

- It is extremely hard to have a population data.
- Most of the time, all data we can get is a sample, not a population.

However, statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process. That process is referred **statistical inference**.

Example

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13

- Our interests are meal prices and quality ratings of restaurants in Tucson.
- Is the above data a sample or population?
- Is it possible to get the population to do our study?

What is descriptive statistics?

Let's consider the data set for restaurants in Tucson.

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13

- Can we get useful or meaningful information from the above data set?
- Data does not give us useful or meaningful information by itself.
- One job a statistician does is to "tailor" and summarize data to clearly show information contained in a data set.

What is descriptive statistics?

Descriptive Statistics refers to the data summaries that may be tabular, graphical, or numerical so that readers can easily understand the data.

Tools to provide descriptive statistics:

- Tabular and graphical displays:
  - for a Categorical variable: frequency distribution, bar charts, pie charts
  - for a Quantitative variable: frequency distribution & cumulative distribution, dot plot, histogram, stem-and-leaf display
  - for Two Variables: crosstabulation, scatter diagrams
- Numerical measures: measures of location, measures of variability, measures of association between two variables, etc.

Frequency Distribution

- One common way to summarize a data set is to use the frequency distribution.

Frequency Distribution

A frequency distribution is a **tabular summary** of data showing the number (frequency) of items in each of several **nonoverlapping** classes.

- A frequency distribution is presented in a table.
- The best way to understand frequency distribution is to construct the frequency distribution of a data set.

Example of a frequency distribution

- To warm up, we start with constructing frequency distribution of a categorical variable, ie. quality rating variable.
- Let's consider the restaurant data set so far.

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13
11	Very Good	33
12	Very Good	44
13	Excellent	42
14	Excellent	34
15	Good	25

Steps to construct frequency distribution

1. Determine **Nonoverlapping** Classes
   - In our example, nonoverlapping classes are already given.
   - i.e. we have Good, Very Good, and Excellent class.
2. Count the number of elements that fall into each class.
   - In our example, the table we will fill out will take the following form.

Class	Frequency
Good	
Very Good	
Excellent	
Total	

- Let's work out by looking at our data set.

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13
11	Very Good	33
12	Very Good	44
13	Excellent	42
14	Excellent	34
15	Good	25

- If we fill out the numbers we counted in the previous table, that table is the frequency distribution of quality rating.
- The following is the completed frequency distribution.

Class	Frequency
Good	5
Very Good	7
Excellent	3
Total	15

- Now, it tells us something.
- We can see immediately
  1. how many Good restaurants in Tucson
  2. The majority of restaurants are Very Good.

Relative and Percent Frequency

- We can offer clearer picture of it by doing a little more.
- We can calculate Relative and Percent Frequency.

Relative Frequency

$$\text{Relative Frequency of a class} = \frac{\text{Frequency of the class}}{n(= \text{Total Frequency})}$$

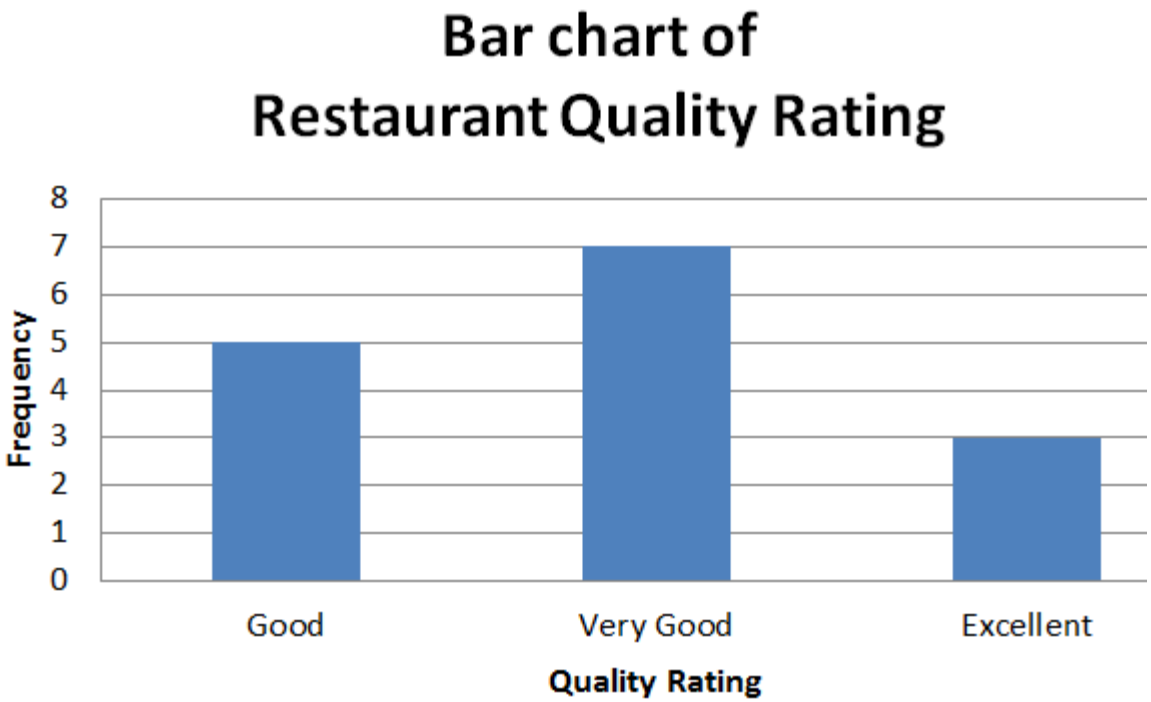
Percent Frequency

$$\text{Percent Frequency of a class} = \frac{\text{Frequency of the class}}{n(= \text{Total Frequency})} \times 100$$

Example

Class	Frequency	Relative Frequency	Percent Frequency
Good	5	0.33	?
Very Good	7	0.47	47%
Excellent	3	?	20%
Total	15		100%

Bar Chart & Pie Chart



Frequency Distribution of a Quantitative Variable

- We can do the similar exercise to construct frequency distribution of a quantitative data.
- However, we have more steps to do since a quantitative data does not give us classes for free.
- Thus, at first, we need to determine classes that we will use.
- There is no consensus about how to determine classes of a quantitative data.
- In this lecture, we will follow a general popular guide line to determine classes.

A Guide Line to Determine Nonoverlapping Classes

1. Determine the number of classes. As a general guide line, using between 5 and 20 classes is recommended.

2. Find the minimum value and the maximum value in a data.

3. Calculate approximate class width with the following formula.

$$\text{Approx. class width} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Number of Classes}}$$

▸ data

4. Round up or down the approx. class width so that we can work with an integer number.

5. Using the rounded approx. class width, construct nonoverlapping classes starting from the smallest value until the last class includes the largest value.

e.g. The resulted classes in our example are as follows.

1st	2nd	3rd	4th	5th
11-18	18.1-25	25.1-32	32.1-39	39.1-46

▸ data

1st	2nd	3rd	4th	5th
11-18	18.1-25	25.1-32	32.1-39	39.1-46

- Once we have the above classes, what we need to do is just same as what we did in the case of a categorical variable.

Class	Frequency	Relative Frequency	Percent Frequency
11-18	3	0.2	?
18.1-25	?	?	16%
25.1-32	2	?	13%
32.1-39	4	0.27	27%
39.1-46	?	0.13	13%
Total	15	1.00	100%

▸ data

Dot Plot

- The horizontal axis shows the range for the data
- Each data value is represented by a dot placed above the axis.
- Dot plots show the details of the data and are useful for comparing the distributions of variables.
- Problem of a dot plot?

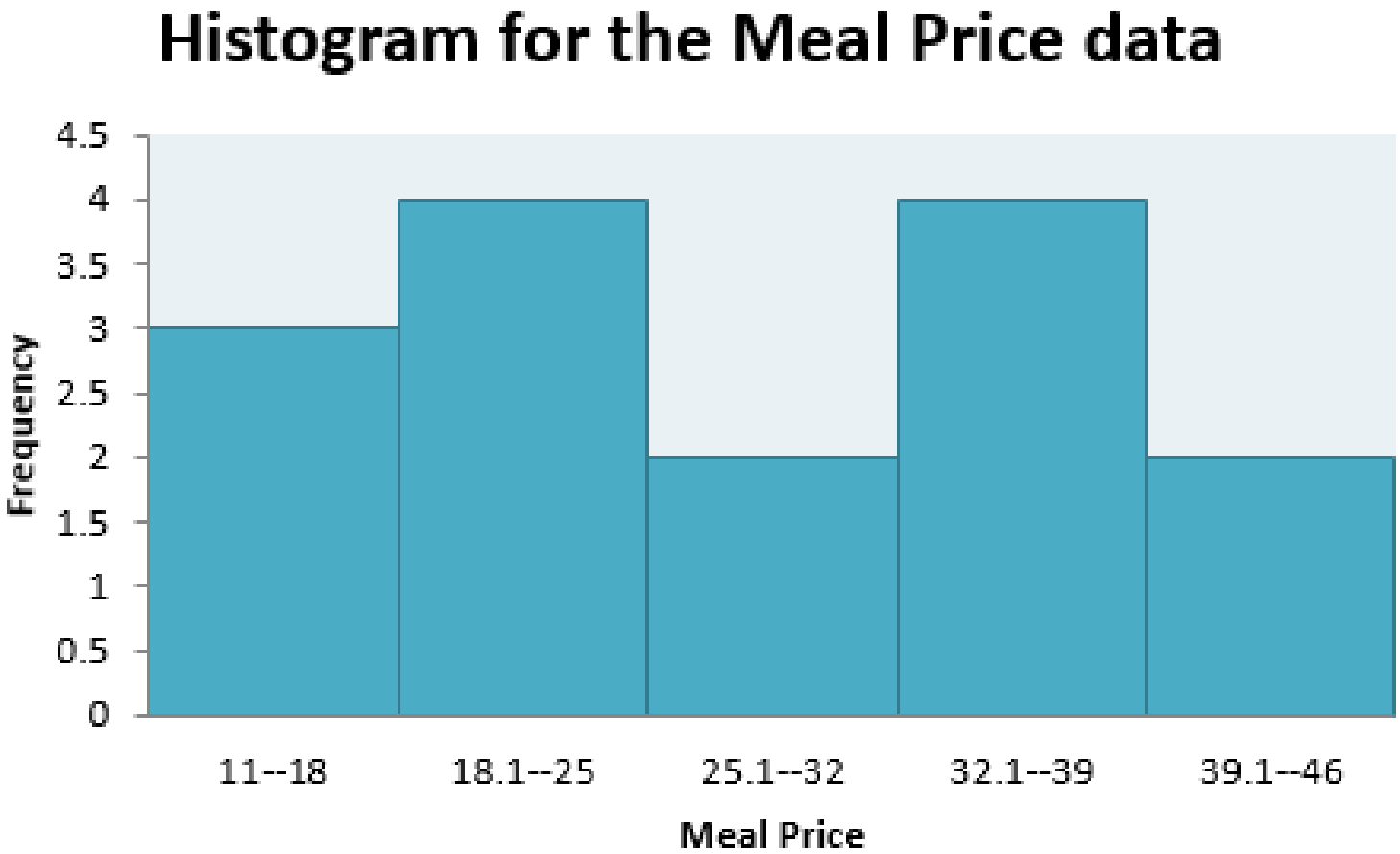
Histogram

- Once we have frequency, relative and percent frequency distribution, we can represent those in a chart.

Histogram

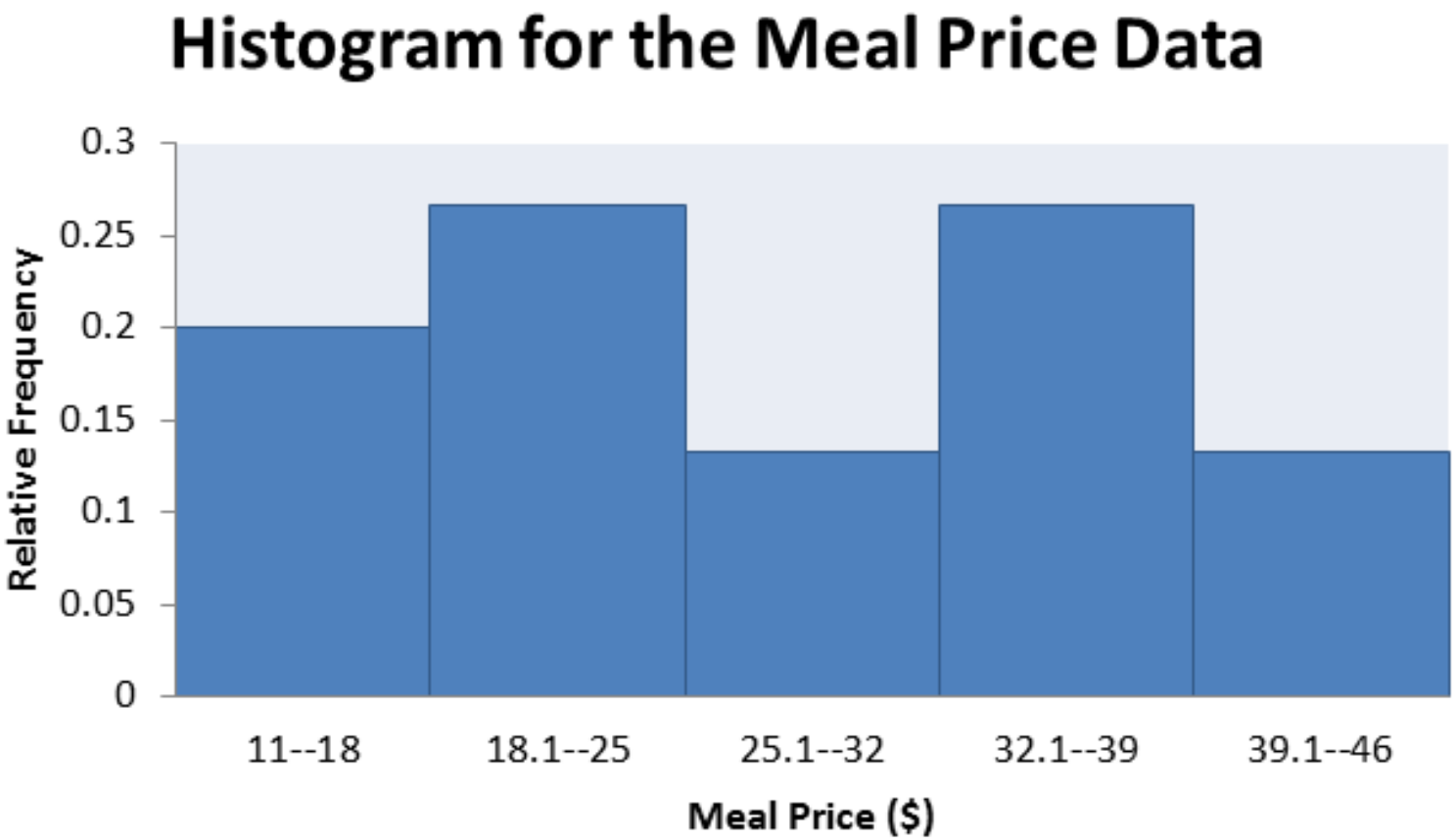
A common graphical presentation of **quantitative data**.
Usually, we put class on **horizontal axis** and frequency, or relative frequency, or percent frequency on **vertical axis**

Example 1. Histogram using frequency



Unlike a bar graph, a histogram contains no natural separation between the rectangles of adjacent classes.

Example. Histogram using relative frequency



Useful information about the shape, or form of a distribution.

Cumulative Distribution

- Sometimes, the cumulative distribution of data is used to provide summary of that data.
- The cumulative distribution shows frequencies that are “piled” up to each class.
- When we construct the cumulative distribution, we count the number of values that is no greater than the upper limit of each class.
- Example will help us to understand the cumulative distribution.

Example

- Recall that we have 5 classes in our previous example of meal prices.
- For each class, we will count the number of values that are no greater than the upper limit of each class.
- We can directly use the frequency distribution to construct the cumulative distribution.

Meal Price	C. Frequency	Frequency (Class)
less than or equal to 18		3 (11-18)
less than or equal to 25		4 (18.1-25)
less than or equal to 32		2 (25.1-32)
less than or equal to 39		4 (32.1-39)
less than or equal to 46	15	2 (39.1-46)
Total	-	15

Stem-and-Leaf Diagrams

Optional material, refer to JK book pg. 41

Two Variables: Crosstabulations and Scatter Diagrams

- So far, we only summarize data for one variable.
- When we summarize data for only one variable, we ignore the relation of that one variable with the other variables.
- However, sometimes, a relation between variables can be our interests.
- We will study Crosstabulations and Scatter Diagrams that help us to understand relation between two variables.

Crosstabulations

Crosstabulations

A Crosstabulation is a tabular summary of data for two variables

- We will use the same concept of the frequency distribution.
- Let's consider the example we used before.

Example: Data from 15 Restaurants in Tucson

- Recall that we have two variables, quality rating and meal price.
- Also recall that we have 3 classes for quality rating and 5 classes for meal price.
- We will put 3 classes for quality rating on the left and 5 classes for meal price on the top of table.

	Meal Prices				
Quality Rating	11-18	18.1-25	25.1-32	32.1-39	39.1-46
Good					
Very Good					
Excellent					

- Now we have 15 classes by combining 3 classes for QR and 5 classes for MP.

	Meal Prices				
Quality Rating	11-18	18.1-25	25.1-32	32.1-39	39.1-46
Good					
Very Good					
Excellent					

- The idea is the same but only difference is that we have more classes than before.
- Let's fill out some of cells to construct our Crosstabulation.

	Meal Prices					
Quality Rating	11-18	18.1-25	25.1-32	32.1-39	39.1-46	Total
Good				0	0	
Very Good	1	3	0		1	
Excellent	0	0	0	2		
Total						15

◂ Data

	Meal Prices					
Quality Rating	11-18	18.1-25	25.1-32	32.1-39	39.1-46	Total
Good	2	1	2	0	0	
Very Good	1	3	0	2	1	
Excellent	0	0	0	2	1	
Total						15

1. With the cross tabulation, we can see a relation between two variables.
    - Good restaurants have low prices and Excellent restaurants have high price.

	Meal Prices					
Quality Rating	11-18	18.1-25	25.1-32	32.1-39	39.1-46	Total
Good	2	1	2	0	0	5
Very Good	1	3	0	2	1	7
Excellent	0	0	0	2	1	3
Total	3	4	2	4	2	15

2. We just recover frequency distributions of both QR and PM variables at the margins of above table.
   - At the right margin, we have the frequency distribution of QR.
   - At the bottom margin, we have the frequency distribution of PM.

Scatter Plot

The other way to see relation between two variables is to construct a scatter plot and trendline.

Scatter Plot

A graphical presentation of the relationship between two variables

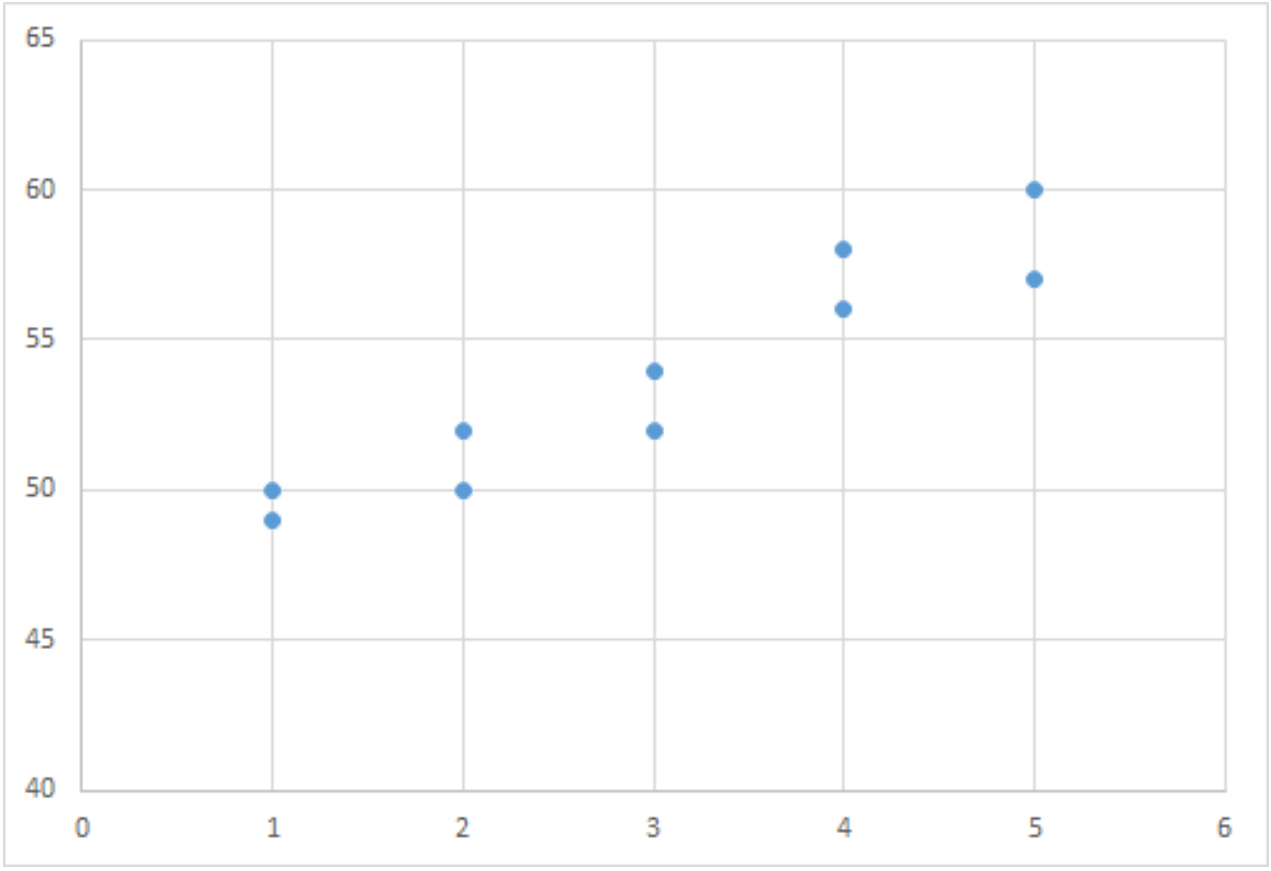
- We will use the following data set for this section.

Week	Number of Commercials	Sales
1	2	50
2	5	57
3	1	49
4	3	52
5	4	58
6	1	50
7	2	52
8	3	54
9	4	56
10	5	60

Scatter Plot

- The way to draw a scatter plot is simple.

1. Put one variable at horizontal axis and the other variable at vertical axis of graph.
2. Then, each pair of observation will have a coordinate on the constructed graph.
3. Mark coordinates of all observations on the graph.



- From the above figure, we can clear see that the number of commercials is positively related to sales.

Crosstabulations vs. Scatter Plots

- Scatter plots usually gives us clearer picture of relation between two variables.
- However, we can't get anything except the picture.
- Crosstabulation gives us less clear information about relation between two variables.
- However, it gives us the frequency distributions of both variables.

These two are complements not substitutes.

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13
11	Very Good	33
12	Very Good	44
13	Excellent	42
14	Excellent	34
15	Good	25

◂ Go Back:cw1 ◂ Go Back:cw2 ◂ Go Back: MPF ◂ Go Back:CRT

Exercise 1

Suppose that we are interested in studying some characteristics of people whose age is 40. Especially, we are interested in their years of schooling and annual income. We decided to conduct a survey for 10 people. We randomly chooses 10 people whose age is 40. We asked them to offer their years of schooling and annual income. We classify the years of schooling into three classes: Bachelor degree (B), Master degree (M), and Ph.d. degree (P). Finally, we could get the following data set.

People	Years of Schooling	Annual Income ($1000s)
1	B	42
2	M	50
3	B	47
4	B	49
5	M	52
6	P	55
7	B	51
8	B	40
9	B	46
10	P	53

1. What are the variables?
2. Which variable is a categorical variable and which variable is a quantitative variable?
3. Construct the percent distribution for years of schooling.
4. Construct the percent distribution for annual income (the number of classes is chosen to be 4.)
   a. Calculate the width of the classes
   b. Construct frequency distribution for annual income.
5. Construct Crosstabulation and Scatter plot.

6. Check if margins in crosstabulations (right and bottom parts in the table) are coincided with the frequency distribution you've obtained from number 3 and 4.

7. Is there any relation between years of schooling and annual income? If it does, state clearly what is the relation.