Ch1 - Descriptive Statistics

(*Graphical and Numerical Methods for Summarizing Data)*

ECO 104 - Statistics For Business and Economics - I

Shaikh Tanvir Hossain

East West University, Dhaka

June 3, 2025

Outline

1. **What is Statistics**

2. **Data and Variables**

3. **Population and Sample**

4. **Descriptive Statistics - Tabular and Graphical Methods**
   - Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts
   - Single Variable Methods (Quantitative): Histogram
   - Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

5. **Descriptive Statistics - Numerical Measures**
   - 1. Measures of Location - Mean, Median and Percentile
   - 2. Measures of Variability - Range, Interquartile Range, and Variance
   - 3. Five-Number Summaries and Boxplots

6. **Two Variable Measures For Association : Covariance, Correlation and ScatterPlot**
   - 1. Numerical Measures: Covariance and Correlation
   - 2. Graphical Measure: Scatterplot

7. **Chapter Recap For the Descriptive Statistics Part**

What is Statistics?

- In one line perhaps we can say ....

  Statistics is the language which we use to collect, analyze and interpret a data.

- What is data?

  Data is a set of information presented in a systematic way.

- In this world, the use of data is almost everywhere, and often it happens so fast that we don't even realize it.

Figure 1: A title of an article of the Forbes magazine, November 4, 2020

- Let's see some real life examples.

What is Statistics?

- **Location data:** Think about you are driving a car with your friend. Then if you are using GPS, Google collects data about where are you, where you going and and how much time it takes to reach the destination. Then Google uses this data to give you a *prediction* about the traffic. Just open Google map, and you will probably see it.

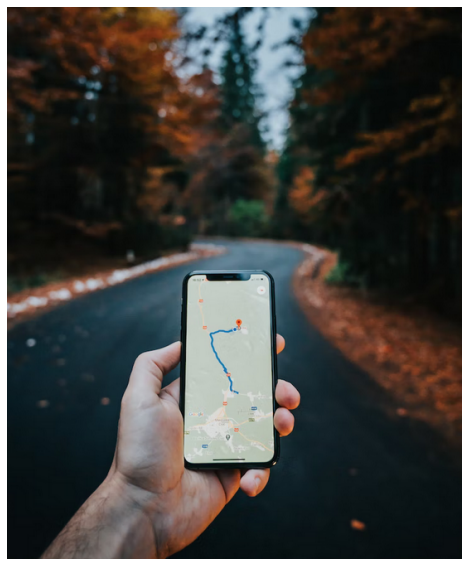


Figure 2: GPS trackers

What is Statistics?

- **Browsing data:** Google and Facebook collects many information when you are on your web browser searching for different things. These are all data. Google continuously collects this and then shows you contents that you might find interesting and useful. Sometimes they also sell this data to other companies (e.g., as a part of marketing). Interestingly Google also shares different search patterns across the world, look at google trend website *.

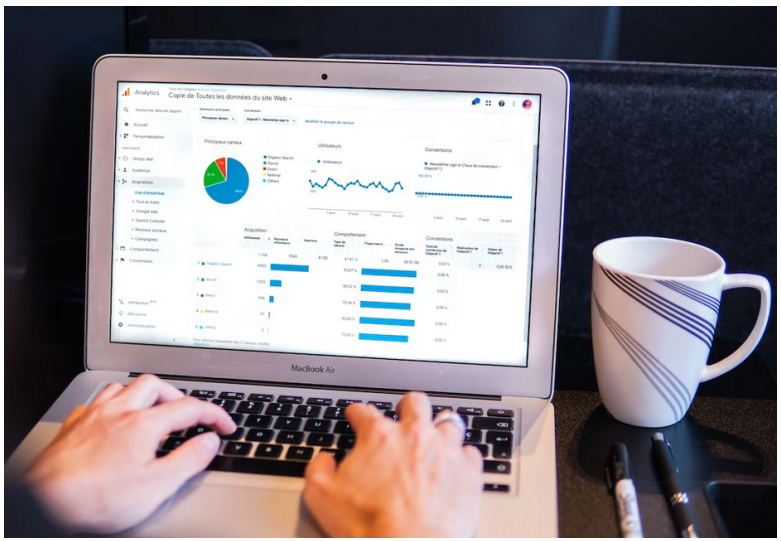
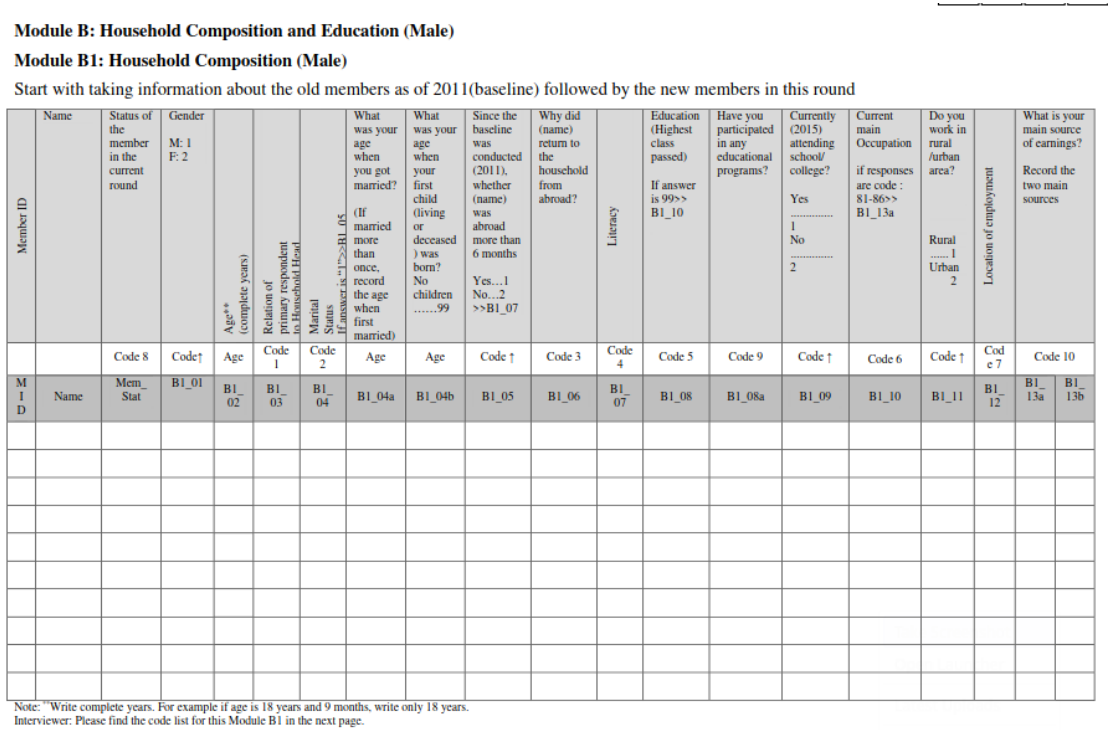


Figure 3: Browsing data

*https://trends.google.com/trends/?geo=BD

What is Statistics?

- **Household Surveys:** Sometimes Government or other research organizations collects important information from families or households using surveys. This is known as *household surveys*. This has information about family income, expenditure, education levels and many more. With this data we can get many useful information, for example average income or expenditures of different families in Bangladesh.



Module B: Household Composition and Education (Male)

Module B1: Household Composition (Male)

Start with taking information about the old members as of 2011(baseline) followed by the new members in this round

Member ID	Name	Status of the member in the current round	Gender M: 1 F: 2	Age** (complete years)	Relation of primary respondent to Household Head	Marital Status If answer is "1">>B1_05	What was your age when you got married? (If married more than once, record the age when first married)	What was your age when your first child (living or deceased) was born? No children 99	Since the baseline was conducted (2011), whether (name) was abroad more than 6 months Yes....1 No....2 >>B1_07	Why did (name) return to the household from abroad?	Literacy	Education (Highest class passed) If answer is 99>> B1_10	Have you participated in any educational programs?	Currently (2015) attending school/ college? Yes ..... 1 No 2	Current main Occupation if responses are code : 81-86>> B1_13a	Do you work in rural /urban area? Rural ..... 1 Urban 2	Location of employment	What is your main source of earnings? Record the two main sources	
		Code 8	Code↑	Age	Code 1	Code 2	Age	Age	Code ↑	Code 3	Code 4	Code 5	Code 9	Code ↑	Code 6	Code ↑	Code 7	Code 10	
MID	Name	Mem_ Stat	B1_01	B1_ 02	B1_ 03	B1_ 04	B1_04a	B1_04b	B1_05	B1_06	B1_ 07	B1_08	B1_08a	B1_09	B1_10	B1_11	B1_ 12	B1_ 13a	B1_ 13b

Note: **Write complete years. For example if age is 18 years and 9 months, write only 18 years.
Interviewer: Please find the code list for this Module B1 in the next page.

Figure 4: A page from the questionnaire of the Bangladesh Integrated Household Survey (BIHS) 2015, IFPRI

What is Statistics?

- **Weather Forecasts:** We all know about whether forecasts. Now this is also an application of Statistics. Whether it is going to be raining tomorrow or not, we can try to predict this using historical weather data in the specific location.
- **Financial data:** If you are an expert with financial data, perhaps you might get very rich! Actually it's not that simple. but it's true that financial data is possibly the earning source of many people.
- **Data analyst in companies:** Many companies are now looking for good data analysts. You might have heard about "Machine Learning" or "Data Science"[†]. Usually any company has lots of data about its different activities, and if we can analyze these data properly, this might be very beneficial for the company, because companies can use them for different tasks, for example maximizing the sales, minimizing the costs, planning, and perhaps many more.

[†] If not, just google them.

What is Statistics?

Figure 5: About the Statistics Career, taken from https://thisisstatistics.org/

What is Statistics?

- So now you got some idea about different type of data sets, and the work of Statistics is to analyze and extract useful information from data sets.
- Statistics uses the concept *"randomness"*, more concretely when we have a data in Statistics we say we have a *"random sample"*. Question is what does *randomness* mean (you should always ask questions)?
- To understand and explain this properly scholars from past developed a new language known as *Probability Theory*, with Probability we can explain uncertainty.
- At the first part of this part of this course we will learn some *Descriptive Statistics* and *Probability Theory*, but don't worry this course is going to be fun and challenging, so get prepared

Some Advice

- Our goal is to understand the concepts, not just doing crazy math problems.
- Sometimes there are simple math concepts, so please do not be scared just because this is math. When it comes to studying Mathematics often people fall into some mental traps. So please try to avoid following traps -
  - *"Everyone else has been doing math for so long and there is no way I'll ever be as good as them." (NO! and please stop thinking this!)*
  - *"A small minority of people are math geniuses and everyone else has no chance at being good at math" (Everyone has more or less same brain, so use it, you can be genius too!)*
  - *Being good at math means being able to instantly solve any math problem thrown at you. (Not necessarily, Math needs both understanding the concepts and practice!)*
  - *"Being good at Math means one can solve crazy calculations pretty fast (Not necessarily, crazy calculations is not the art of the math.)"*

Some Advice

- *Here are some advice -*
  - *Have a Growth Mindset!*
  - *Question everything!*
  - *Attend lectures consistently.*
  - *Write and think, and also think and write!*
  - *If you find any mistakes in my explanation, that is good news :), means you are thinking critically, please let me know!*
  - *Study strategically and with motivation, not mechanically!*

"If people do not believe that Mathematics is simple, it is only because they do not realize how complicated life is."- John von Neumann (1903 - 1957)

According to Franz L, this is remark made from the podium by von Neumann as keynote speaker at the first national meeting of the Association for Computing Machinery in 1947.

Again some motivation

Figure 6: fear can be a barrier for success, so stop being scared and work hard, https://www.forbes.com/sites/carolinecastrillon/2021/08/22/top-10-reasons-you-have-a-fear-of-success/?sh=54dde6da1c15

1. What is Statistics

2. **Data and Variables**

3. Population and Sample

4. Descriptive Statistics - Tabular and Graphical Methods
   - Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts
   - Single Variable Methods (Quantitative): Histogram
   - Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

5. Descriptive Statistics - Numerical Measures
   - 1. Measures of Location - Mean, Median and Percentile
   - 2. Measures of Variability - Range, Interquartile Range, and Variance
   - 3. Five-Number Summaries and Boxplots

6. Two Variable Measures For Association : Covariance, Correlation and ScatterPlot
   - 1. Numerical Measures: Covariance and Correlation
   - 2. Graphical Measure: Scatterplot

7. Chapter Recap For the Descriptive Statistics Part

Data and Variables

Data and Variables

- Statistics starts from data, so let's consider the following hypothetical data set,

	Gender	Monthly Income (tk)	ECO-101 Grade	# Retakes
1.	Male	3615	B-	3
2.	Female	49755	A	2
3.	Male	44758	A	1
4.	Female	3879	B	0
5.	Male	22579	A+	2

- Here the columns are called *variables* and the rows are called *observations* or *units*. We have 5 observations and 4 variables. Total number of observations is called *sample size*, here the sample size is equal to 5.

Data and Variables

- We can classify the variables in different ways,

§. Whether we can group them or categorize them!

1. *Categorical Variables*, also called *Qualitative Variables*
   - **Key Characteristic:** Values can be grouped
   - **Examples:** Gender: Male / Female, Grade: Pass / Fail, Marital Status: Married / Unmarried
2. *Numeric Variables*, also called *Quantitative Variables*
   - **Key Characteristic:** Values are numeric and represent how much or how many
   - **Examples:** Monthly Income, Height, Weight, Exam Scores.

§. Whether the values are discrete or continuous

1. *Discrete Variables*
   - **Key Characteristic:** Only certain values (often integers) are possible, countable
   - **Examples:** Number of children - 0, 1, 2, 3, Number of times a student retakes a course 0,1,2,....
2. *Continuous Variables*
   - **Key Characteristic:** Values can lie in a real interval or in set in $\mathbb{R}$
   - **Examples:** Income, Height, Weight, Time to finish a task.

§. How the values are compared or what is the scale of measurement

1. *Nominal Scale*

   - *Categorical Variables* (e.g., gender, blood type) with *NO intrinsic order*
   - Example: Gender - Male/Female, Fruits - Apple/Orange/Banana.
   - Important: You cannot say one category is higher or greater than another; only comment they are different.

2. *Ordinal Scale*

   - *Categorical Variables* that can be *ranked or ordered*, but differences between ranks may not be uniform.
   - Examples: Letter grades A/B/C, or Likert scale: Strongly Disagree $<$ Disagree $<$ Neutral $<$ Agree $<$ Strongly Agree)
   - Important: The difference between Strongly Disagree and Disagree may not be the same as the difference between Neutral and Agree.

3. *Interval Scale*

   - *Numeric scale* where *differences* are meaningful, but zero does not indicate absence of the quantity.
   - Typical For: Numeric variables (Quantitative) measured on a scale without a true zero.
   - Example: Temperature in Celsius or Fahrenheit (0°C is not no heat).
   - Key Property: We can add or subtract values meaningfully, but a ratio (e.g., twice as much) is not valid.

4. *Ratio Scale*

   - This is a numeric scale that *includes a true zero*, which means that both differences and ratios are meaningful (e.g., income, height, weight).
   - Examples: Income, height, weight, time to finish a task.

Collection of data and different type of data

- Now let's talk about *how do we collect data*, roughly we can collect data using any of the following three methods,
  - by existing sources (e.g., administrative data sets)
  - by conducting an observational study (e.g., by taking surveys)
  - or by conducting an experiment.(e.g., giving training programs)
- Depending upon *space and time* we can also categorize the data in following three ways,
  - Cross-sectional data (data at same point in time, but for different units, e.g., household surveys at any fixed time)
  - Time Series data (will have a time component, e.g., looking at GDP of Bangladesh for 10 consecutive years)
  - It is also possible to have a combination of both, which we call Panel data.

1. What is Statistics

2. Data and Variables

3. **Population and Sample**

4. Descriptive Statistics - Tabular and Graphical Methods
   - Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts
   - Single Variable Methods (Quantitative): Histogram
   - Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

5. Descriptive Statistics - Numerical Measures
   - 1. Measures of Location - Mean, Median and Percentile
   - 2. Measures of Variability - Range, Interquartile Range, and Variance
   - 3. Five-Number Summaries and Boxplots

6. Two Variable Measures For Association : Covariance, Correlation and ScatterPlot
   - 1. Numerical Measures: Covariance and Correlation
   - 2. Graphical Measure: Scatterplot

7. Chapter Recap For the Descriptive Statistics Part

Population and Sample

Population Data Vs. Sample Data

- Before we go further, we need to talk about what do we mean by the word "population"and "sample" in Statistics? Consider the data set we saw before, someone may ask

  "Why did we collect the data of 5 students studying currently at EWU?"

- One answer could be

  "Maybe we are interested to get some idea about all of the students studying currently at EWU."

- In this case, we say the *population* is the set of all current students at EWU. And the set of *5 students* is a sample of that population.

Definition 1.1: (Population and Sample)

The collection / set of *all observations* in a particular study is called the population. A sample is a subset of the population.

Population Data Vs. Sample Data

- Usually collecting population data is very time consuming and often impossible, for example ???
- So we collect *a sample* from the population.
- Note that a sample is supposed to be a good representative of the whole population.
- If the sample is not a good representative, then we say we have a *biased sample*.
- Biased sample is bad, why?... because any conclusion from a biased sample might lead to incorrect conclusion regarding the population (can you think about an example?)

Population Data Vs. Sample Data

- One way to get a good sample is - *Simple Random Sampling!* (details on board)
- What if the population is infinite ??? Is it even possible in reality ???
- One of the major tasks of statistical analysis is, using a sample to make some conclusions regarding the population. This process is called *Statistical Inference*, or we call this - *Inferential Statistics*.
- In this course, we will not talk about Inferential Statistics. But in ECO 204, you will learn some techniques related to Inferential Statistics.
- However let's do an example on board! (Population Mean Vs. Sample Mean).
- Here we are making inference about Population Mean using Sample Data!

1. What is Statistics

2. Data and Variables

3. Population and Sample

4. **Descriptive Statistics - Tabular and Graphical Methods**
   - Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts
   - Single Variable Methods (Quantitative): Histogram
   - Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

5. Descriptive Statistics - Numerical Measures
   - 1. Measures of Location - Mean, Median and Percentile
   - 2. Measures of Variability - Range, Interquartile Range, and Variance
   - 3. Five-Number Summaries and Boxplots

6. Two Variable Measures For Association : Covariance, Correlation and ScatterPlot
   - 1. Numerical Measures: Covariance and Correlation
   - 2. Graphical Measure: Scatterplot

7. Chapter Recap For the Descriptive Statistics Part

Descriptive Statistics - Tabular and Graphical Methods

Descriptive Statistics

- Now we will start with descriptive statistics, which is the first major topic of this course. Recall in *Inferential Statistics*, we try to use the sample data to *comment / infer* about the population.
- *Descriptive Statistics* is simpler, in descriptive statistics we don't make any conclusion about population, the only thing we have is data, so with different procedures we will comment on data, that's it. Essentially the idea of a descriptive statistical analysis is to *summarize the data* so that the key information *in the data* are clear, whether we do it *visually or numerically*
- There are two ways we can do descriptive statistical analysis (or descriptive statistics!)
  - Tabular and Graphical Methods
  - Numerical Methods

Descriptive Statistics

- We will cover following Tabular and Graphical methods! Tabular and Graphical Methods are often discussed together since most of the times, first we have some tables and then we convert the table into some graphs.
  - **For Single Variable:**
    1. *Frequency Distribution Table (Tabular)* and *Bar Chart and Pie Chart (Graphical)* (All are applicable only for *categorical* variables.)
    2. *Grouped Frequency Distribution Table (Tabular)* and *Histogram (Graphical)* (Only applicable for *numeric* variables.)
  - A side note: The Frequency Distribution (or Relative Frequency or Percent Frequency Distribution) is a *tabular method*. There is no graphical element here; rather, it provides the table used to create bar charts and pie charts, which are graphical methods.

- **For Two Variables:**

  1. *Cross-tabulation or Contingency Table or Joint Frequency Distribution (Tabular)* and *Side-by-Side Bar Chart or Stacked Bar Chart (Graphical)*
     (Applicable for *two categorical* variables.)

  2. *Bivariate Frequency Distribution Table (Tabular)* and *Joint Histogram (3D Histogram) (Graphical)*
     (Applicable for *two numeric* variables.)

  3. *Scatter Plot (Graphical)*
     (Applicable for *two numeric* variables to understand association.)

- The cross-tabulation can also be applied for a combination of *categorical and numeric variables*, we will see some examples.

Descriptive Statistics - Tabular and Graphical Methods

Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts

Frequency Distribution and Bar Charts

- We will start with *frequency distribution*, a tabular method for describing categorical data.

Definition 1.2: (Frequency and Frequency Distribution)

- **Frequency:** The number of observations in a given category.
- **Frequency Distribution:** A *frequency distribution* is a *tabular summary* that displays the frequency (i.e., the number of observations) for each category.

We will also look at two related concepts

- *Relative Frequency Distribution*: This shows the *fraction or proportion* of observations for each category.
- *Percent Frequency Distribution*: Shows the *percentage* of observations for each category.

Lets consider an example. Suppose we have sales data from a supershop that shows which soft drink was sold in the last 50 transactions. Here is how the data looks:

Frequency Distribution and Bar Charts

	A	B	C	D
1	sales number	**Brand Purchased**	sales number	**Brand Purchased**
2	1	Coca-Cola	26	Coca-Cola
3	2	Diet Coke	27	Coca-Cola
4	3	Pepsi	28	Coca-Cola
5	4	Diet Coke	29	Pepsi
6	5	Coca-Cola	30	Coca-Cola
7	6	Coca-Cola	31	Sprite
8	7	Dr. Pepper	32	Dr. Pepper
9	8	Diet Coke	33	Pepsi
10	9	Pepsi	34	Diet Coke
11	10	Pepsi	35	Pepsi
12	11	Coca-Cola	36	Coca-Cola
13	12	Dr. Pepper	37	Coca-Cola
14	13	Sprite	38	Coca-Cola
15	14	Coca-Cola	39	Pepsi
16	15	Diet Coke	40	Dr. Pepper
17	16	Coca-Cola	41	Coca-Cola
18	17	Coca-Cola	42	Diet Coke
19	18	Sprite	43	Pepsi
20	19	Coca-Cola	44	Pepsi
21	20	Diet Coke	45	Pepsi
22	21	Coca-Cola	46	Pepsi
23	22	Diet Coke	47	Coca-Cola
24	23	Coca-Cola	48	Dr. Pepper
25	24	Sprite	49	Pepsi
26	25	Pepsi	50	Sprite
27				

Frequency Distribution and Bar Charts

- The frequency distribution, percent frequency distribution and relative frequency distribution in this case is,

Brand	Frequency	Relative Frequency	Percent Frequency
Coca-Cola	19	0.38	38
Diet Coke	8	0.16	16
Dr. Pepper	5	0.1	10
Pepsi	13	0.26	26
Sprite	5	0.1	10
Grand Total	50	1	100

Frequency Distribution and Bar Charts

An important side note:

Notice that the term *distribution* is used here because the relative and percent frequency distributions actually represent the *probability distribution* of each category. In other words, they show how the overall probability of sales is allocated among the different categories.

Roughly probability gives us a measure of *likelihood / chance / percentage of happening certain events*, for example if someone asks,

what is the probability that the next sale will be Coca-Cola?

Based on the data, we might conclude that,

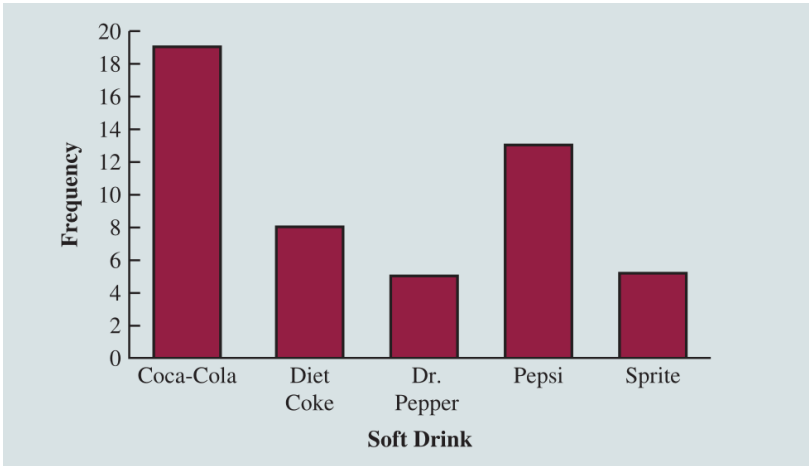
Well, we have actually 38% chance that the next sales will be Coca-Cola

Well explore probability in more depth in the next chapter, but for now, the key takeaway is that percent frequency distributions provide a practical approximation of the probability distribution of sales of all categories. This understanding will be very important to understand probability distribution of a random variable later... so keep this in mind.

Frequency Distribution and Bar Charts

- Now the bar chart is,

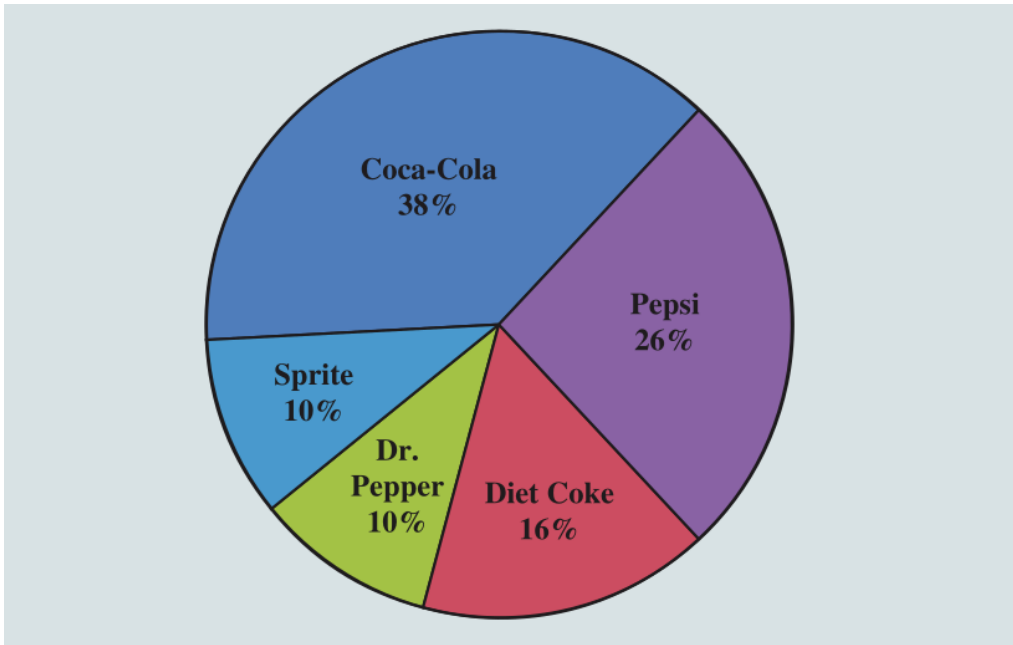
Figure 7: Bar Chart of Soft Drink Purchases



- It is possible to sort the table in ascending order to create a more visually appealing bar chart in Excel. What does the bar chart convey, and how would you interpret it
- Ans: The bar chart shows that the highest sales occur for Coca-Cola and Pepsi, indicating that consumer preferences in that region favor these brands. In contrast, the lowest sales are recorded for Dr. Pepper and Sprite, suggesting these brands are less popular among consumers.

Pie Chart

- From the frequency distribution table, we can also construct a pie chart as well. Here is how Pie Chart looks like (we will do this in Excel, note the message is same!)



Descriptive Statistics - Tabular and Graphical Methods

Single Variable Methods (Quantitative): Histogram

Histogram

- If we are looking for a summary measure of a numeric variable, then it is not a good idea to use bar chart / pie chart directly, Why? because there are just too many different values in the data set tp treat them as categories. So we need a *new* graphical method to understand the numeric variables, this is where *histogram* comes.Consider the following data of an audit firm which shows the number of days required to audit different companies. Let's create a histogram from this data.

Figure 8: Data From Anderson et al. (2020)'s book

	Audit Time
2	12
3	15
4	20
5	22
6	14
7	14
8	15
9	27
10	21
11	18
12	19
13	18
14	22
15	33
16	16
17	18
18	17
19	23
20	28
21	13

Histogram

- To create a histogram, we first craete *bins* or *classes*. In this example, we create 5 bins: $10-14$, $15-19$, $20-24$, $25-29$, and $30-34$. We then count how many data points fall into each bin, which gives us a frequency distribution similar to what we saw beforebut now for a numeric variable. This is what we call *grouped frequency distribution*,

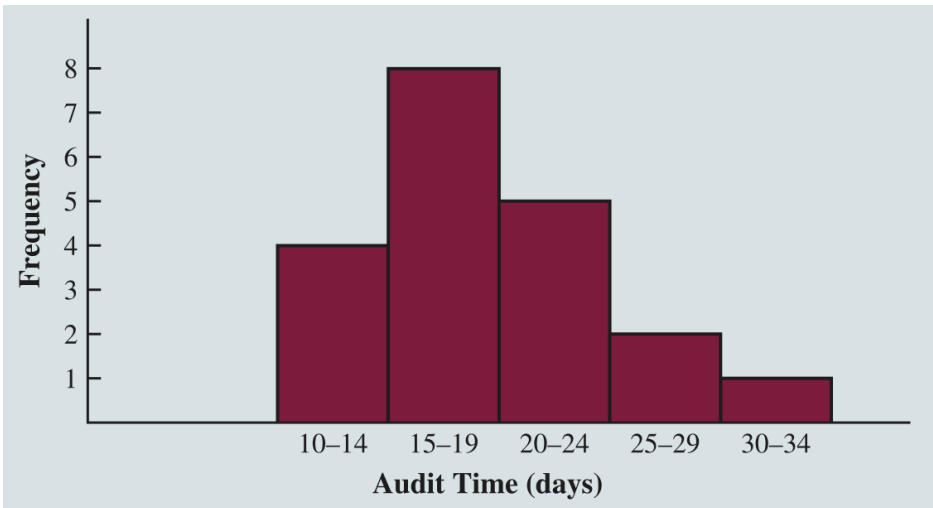
Audit Time (Days)	Frequency
10 -14	4
15 - 19	8
20 - 24	5
25 - 29	2
30 - 34	1
Total	20

We can also calculate the *relative frequency* and *percent frequency* for each bin, as well as the *cumulative percent frequency*, which shows the running total of percentages as we include more data.

Audit Time (Days)	Frequency	Relative Freq	Percent Freq	Cum Percent Freq
10 -14	4	.20	20	20
15 - 19	8	.40	40	60
20 - 24	5	.25	25	85
25 - 29	2	.10	10	95
30 - 34	1	.05	5	100
Total	20			

Histogram

- Next, we convert the first table into a graph like a bar chart, which is known as a *histogram*. Here is the resulting histogram:



Histogram

- You might notice that a histogram looks similar to a bar chart. While this is true, there is a fundamental difference:

  bar charts display categorical variables, whereas histograms display quantitative variables

- The visual representation is similar, but the nature of the data is quite different, also unlike the bar chart, the bars here have no space between them, this is because we have a continuous numeric data, not a discrete / categorical one.
- Can you interpret the histogram? It appears that the majority of the data falls within the $15 - 19$ range. This suggests that most audits are completed within 15 to 19 days.

Histogram

- In the last example, we first picked the number of bins $k = 5$, then we calculated the bin-width h using the formula (here we let k be the number of bins then, h be the bin-width)

$$h \approx \frac{\text{maximum} - \text{minimum}}{k} = \frac{35 - 10}{5} = 5$$

- After that, we get the bins, $10 - 14$, $15 - 19$, $20 - 24$, $25 - 29$, and $30 - 34$.
- Note 33 and 12 is the max and min but we picked 35 and 10 just to get a round figure when we divide the difference by 5.

Histogram

- Two million dollar questions are, 1) Why did we choose 5 as the number of bins, and 2) What happens if we increase the number of bins?
- Let's answer the second question first, if increase the number of bins then each bin cover a smaller range, so yes we might see more details about the datas distribution.
- However, if we use too many bins, the histogram might become too "noisy" with many bins having very few observations, making it harder to see overall patterns in the data.
- Conversely, too few bins can oversimplify the data, hiding important patterns also. The key is to find a balance that provides meaningful insight without excessive detail or oversimplification.

Histogram

- Now we answer the first one, there different techniques to select number of bins or bin-width, but there is no single technique that works best always.
- Here are two common techniques,
  - 1. Square Root Rule, which says

$$k = \sqrt{n}$$

    where n is the sample size. This rule is straightforward and often used for a quick estimate.

  - 2. Freedman-Diaconis Rule, which says
    Instead of directly giving the number of bins, it provides a bin width:

$$h = 2 \times \frac{\text{IQR}}{n^{1/3}}$$

    where IQR means inter-quartile range (we will see how to calculate this in the next section), then the number of bins can be calculated as:

$$k = \frac{\text{Range}}{h}$$

    This rule is useful when you want to balance the bin width against the variability in the data.

Descriptive Statistics - Tabular and Graphical Methods

Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

Cross Tabulation

- So far we only talked about graphical methods using a single variable, now we will consider methods using two variables together.
- The idea of cross tabulation is very similar to frequency distribution, but here we have two variables.
- We can construct cross tabulation in three ways,
  - Both variables can be categorical, in this case this is often known as *contingency table*.
  - One variable can be categorical and the other can quantitative.
  - Both variables can be quantitative.
- Let's see some examples (we will see how to create this tables in Microsoft Excel in the lab class.)

Cross Tabulation

Both Categorical

- Consider following data (Right-H means Right Handed, and Left-H means Left-Handed)

Observation	*Handed*	*Male/Female*	Observation	*Handed*	*Male/Female*
1.	Right-H	Male	16.	Left-H	Female
2.	Left-H	Male	17.	Left-H	Male
3.	Left-H	Male	18.	Left-H	Male
4.	Right-H	Female	19.	Right-H	Male
5.	Left-H	Male	20.	Left-H	Male
6.	Right-H	Female	21.	Right-H	Female
7.	Left-H	Male	22.	Left-H	Male
8.	Left-H	Female	23.	Left-H	Female
9.	Right-H	Male	24.	Right-H	Male
10.	Left-H	Male	25.	Left-H	Male
11.	Right-H	Male	26.	Left-H	Female
12.	Left-H	Male	27.	Left-H	Female
13.	Left-H	Female	28.	Right-H	Male
14.	Left-H	Female	29.	Left-H	Male
15.	Left-H	Female	30.	Left-H	Female

- With this table we can construct following crosstabulation or contingency table,

	Female	*Male*	Total Result
Left-H	9	12	21
Right-H	3	6	9
Total Result	12	18	30

Cross Tabulation

Both Categorical

- Here each variable in the data set has 2 categories, so in total we have 4 categories,
  - Right Handed Male (count - 6)
  - Right Handed Female (count - 3)
  - Left Handed Male (count - 12)
  - Left Handed Female (count - 9)
- We can also calculate the percentages? How ?
- We can also calculate Row total or Row Percentages, this means
  - Total Number of Left Handed people, or Total Percentage of Left Handed people.
  - Total Number of Right Handed people, or Total Percentage of Right Handed people.
- We can also calculate Column total or Column Percentages, this means
  - Total Number of Male, or Total Percentage of Male.
  - Total Number of Female, or Total Percentage of Female.
- We will also see how to calculate these numbers using MS-Excel in our lab class.

Cross Tabulation

One Categorical and One Quantitative

- It is also possible to construct a crosstabulation when one variable is categorical and the other is quantitative.
- For example, consider following data which is collected from 300 restaurants (only partially shown, the data is available in the file `Restaurant.xlsx` file in chapter 2)

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
⋮	⋮	⋮
⋮	⋮	⋮
300.	Very Good	31

- Here we have two kinds of variables, `Quality Rating` is a categorical variable and `Meal Price` is a quantitative variable.

Cross Tabulation

One Categorical and One Quantitative

- It is also possible to construct cross tabulation using this data, here is a crosstabulation that we can make using this data.

	Meal Price				
Quality Rating	$\$10-19$	$\$20-29$	$\$30-39$	$\$40-49$	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

- Notice the columns are like histogram categories, where a qunatitative variable is categorized in different categories, and then we calculated the frequency of each combination of categories.
- In the lab class we will try to do this using MS-Excel, but now you know how it looks like.
- What is the interpretation of 42?
- Similarly we can construct crosstabulation for two quantitative variables as well.

Side by side bar chart or stcaked bar chart

- Once we have the crosstabulation we can also plot the side by side bar chart oor stacked bar chart

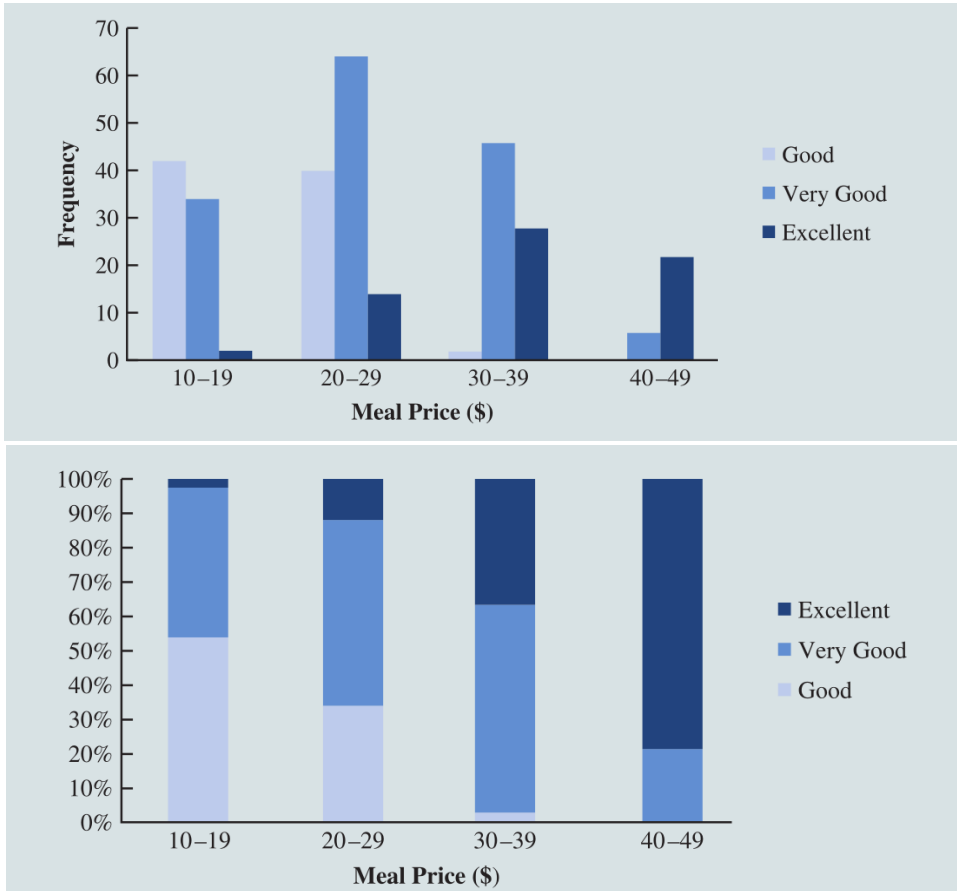


Figure 9: The top one is the side by side bar chart and the bottom one is called stacked bar chart

Side by side bar chart or stcaked bar chart

- Note that these are also bar charts, but the important thing is these type of bar charts are for two variables.

1. What is Statistics

2. Data and Variables

3. Population and Sample

4. Descriptive Statistics - Tabular and Graphical Methods
   - Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts
   - Single Variable Methods (Quantitative): Histogram
   - Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

5. **Descriptive Statistics - Numerical Measures**
   - 1. Measures of Location - Mean, Median and Percentile
   - 2. Measures of Variability - Range, Interquartile Range, and Variance
   - 3. Five-Number Summaries and Boxplots

6. Two Variable Measures For Association : Covariance, Correlation and ScatterPlot
   - 1. Numerical Measures: Covariance and Correlation
   - 2. Graphical Measure: Scatterplot

7. Chapter Recap For the Descriptive Statistics Part

Descriptive Statistics - Numerical Measures

Descriptive Statistics - Numerical Measures

1. Measures of Location - Mean, Median and Percentile

Sample Mean (Arithmetic Mean)

- In the last section we discussed some *tabular* and *graphical methods* to summarize a data, now we will see some *numerical measures* that provide more ways for summarizing a data. It is possible to have numerical summary measures for a single variable or for more variables, first let's focus on one variable.
- We will start with a measure of *central location* of any data. The most common measure of the central location is the *sample average* or *sample mean*. Think about a variable *height* and assume we have some kind of height data for 10 students, then the following is a sample.

student	heights
1	5.5
2	4.8
3	5.2
4	4.5
5	6
6	5.9
7	6.2
8	5.3
9	6
10	5

Sample Mean (Arithmetic Mean)

- Usually we will write $x_1, x_2, \ldots, x_n$ rather than numbers $5.5, 4.8, 5.2, \ldots, 5$. This is just to generalize the writings. Now the sample mean can be calculated as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(5.5 + 4.8 + \ldots + 5) = 5.44$$

- You can calculate this by hand, or using calculator, or using `MS-Excel`!
- So the *sample mean* (another name is *arithmetic mean*) is simply the average value, it gives an idea of the center location of the data.

Sample Mean (Arithmetic Mean)

Issues with sample mean

- There is one issue with the Mean, that is, it changes drastically with *extreme values*, for example, if we replace the 10th observation with 20 (which is unrealistic, but just to give you an example), then the sample mean would be 6.94. So what happens is, if we change one value, the sample mean changes a lot, we say sample mean is not a *robust* measure!
- Later we will see another measure of central location which is *sample median*, which is more robust than sample mean!

Weighted Mean

- In the formulas for the sample mean each x_i is given equal importance or weight. For instance, the formula for the sample mean can be written as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{n}\left(x_1\right) + \frac{1}{n}\left(x_2\right) + \cdots + \frac{1}{n}\left(x_n\right)$$

- This shows that each observation in the sample is given a weight of $1/n$.
- Although this is very common, but sometimes we might give different weights to different observations depending upon its relative importance. A mean computed in this manner is referred to as a *weighted mean*. The formula is following

$$\bar{x}_n = \sum_{i=1}^{n} w_i x_i \tag{1}$$

- Note that, the weights have to be summed to 1. Here is an example,

Weighted Mean

- Suppose we have following data of different costs of purchase, but notice each time the cost changes with number of pounds

Number of Purchase	Cost per pound	Number of Pounds
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

- Now simple arithmetic mean would be average of all costs. If you calculate this you will get,

$$\text{Arithmetic Mean } = \frac{3.00 + 3.40 + 2.80 + 2.90 + 3.25}{5} = \frac{15.35}{5} = 3.07.$$

- Now we will see the weighted average of costs will be different. We do here weighted average because the costs are not always same for every pound since amount of purchase is different. For the weighted average, first we calculate all weights for the purchase,

Weighted Mean

- Total number of pounds, is $1200 + 500 + 2750 + 1000 + 800 = 6250$, we use this to calculate weights,

$$w_1 = \frac{1200}{6250} = 0.192, \qquad w_2 = \frac{500}{6250} = 0.08$$
$$w_3 = \frac{2750}{6250} = 0.44, \qquad w_4 = \frac{1000}{6250} = 0.16$$
$$w_5 = \frac{800}{6250} = 0.128$$

- Note in this case, the weights are summed to 1, i.e., $\sum_{i=1}^{5} w_i = 1$. Now the weighted average or weighted mean will be

$$\bar{x}_n = \sum_{i=1}^{n} w_i x_i = (0.192 \times 3) + (0.08 \times 3.4) + (0.44 \times 2.8) + \\ + (0.16 \times 2.9) + (0.128 \times 3.25) = 2.96$$

Weighted Mean

As a side note: if you read Anderson et al. (2020), the formula that you will find is,

$$\bar{x}_n = \sum_{i=1}^{n} \frac{w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

This is actually same thing, except, in this formula the weights are total quantity (not the proportion).

For example, if you apply this formula, then the weights will be $w_1 = 1200, w_2 = 500, w_3 = 2759, w_4 = 1000$ and $w_5 = 800$ (In this case the weights doesn't have to be add up to 1). Here $\sum_{i=1}^{n} w_i = 6250$. How do we know the two formulas are same, notice

$$\bar{x}_n = \sum_{i=1}^{n} \frac{w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3}$$

$$= \sum_{i=1}^{n} \tilde{w}_i x_i \text{ where } \tilde{w}_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{4}$$

I personally prefer the first way of writing because it shows you proportions and is easy to understand, but it's your choice which formulas you want to use. In the book, when you solve problems if the weights are not in fractions, you can also apply the first formula in 1 and convert the weights into fractions.

Quantiles and Percentiles

- A quantile or percentile provides information about how the data are spread from the smallest value to the largest value. For a data set containing n observations, the pth quantile (or we $100 \times p$th percentile) divides the data into two parts:
  - Approximately $100 \times p\%$ of the observations are less than the p th quantile,
  - and approximately $100 \times (1 - p)\%$ of the observations are greater than the pth quantile.
- Colleges and universities frequently report admission test scores in terms of quantiles. For instance, suppose an applicant obtains a score of 630 on the math portion of an admissions test. How this applicant performed in relation to others taking the same test may not be readily apparent.
- However, if the score of 630 corresponds to the .82 nd quantile, we know that approximately that 82% of the applicants scored lower than this individual and approximately 18% of the applicants scored higher than this individual.

Quantiles and Percentiles

- Let's see how to calculate the p^{th} quantile (where $0 < p < 1$) for a data set with n observations,

$$x_1, x_2, \ldots, x_n$$

Now follow these steps:

1. **Arrange the Data:** Sort the data in ascending order (from smallest to largest). The smallest value is in position 1, the next smallest in position 2, and so on, so get the data

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

this is the sorted data set,

2. **Find the Location:** Compute the location of the p^{th} quantile, denoted by i_p, using the formula:

$$i_p = p\,(n+1).$$

3. **Interpret the Location:** If i_p is not an integer, it indicates that the p^{th} quantile lies between two data points. For example, if $p = 0.8$ and $i_{0.8} = 10.4$, then the 0.8 quantile (equivalently, the 80th percentile) is 40% of the way between the value in position 10 and the value in position 11.

4. **Calculate the Quantile:** Suppose the value in position k is $x_{(k)}$ and in position $k+1$ is $x_{(k+1)}$, and let γ be the fractional part of i_p. Then the p^{th} quantile is computed as:

$$q(p) = x_{(k)} + \gamma \left(x_{(k+1)} - x_{(k)} \right).$$

Quantiles and Percentiles

- **Example:** Find the 0.3 quantile (equivalently, the 30^{th} percentile) of the following heights:

$$5.5,\ 4.8,\ 5.2,\ 4.5,\ 6.0,\ 5.9,\ 6.2,\ 5.3,\ 6.0,\ 5.0.$$

1. **Step 1: Arrange the Data.** Sorted in ascending order, the heights are:

$$4.5,\ 4.8,\ 5.0,\ 5.2,\ 5.3,\ 5.5,\ 5.9,\ 6.0,\ 6.0,\ 6.2.$$

   Here, $n = 10$.
2. **Step 2: Find the Location.** For $p = 0.3$, compute:

$$i_{0.3} = 0.3\,(10 + 1) = 0.3 \times 11 = 3.3.$$

3. **Step 3: Interpret the Location.** Since $i_{0.3} = 3.3$, the 0.3 quantile lies 30% of the way between the 3rd and 4th values in the ordered list.
4. **Step 4: Calculate the Quantile.** The 3rd value is 5.0 and the 4th value is 5.2. Thus, the 0.3 quantile is:

$$Q(0.3) = 5.0 + 0.3 \times (5.2 - 5.0) = 5.0 + 0.3 \times 0.2 = 5.0 + 0.06 = 5.06.$$

Quartiles

- It is often desirable to divide a data set into four parts, with each part containing approximately one-fourth, or 25%, of the observations. These division points are referred to as the *quartiles* and are defined as follows:
  - Q_1 = first quartile, or 25 th percentile
  - Q_2 = second quartile, or 50 th percentile (also the median)
  - Q_3 = third quartile, or 75 th percentile
- Because *quartiles* are just specific percentiles, the procedure for computing percentiles can be used to compute the quartiles.
- MS-Excel actually gives direct function to calculate both percentiles and quartiles.

Sample Mode

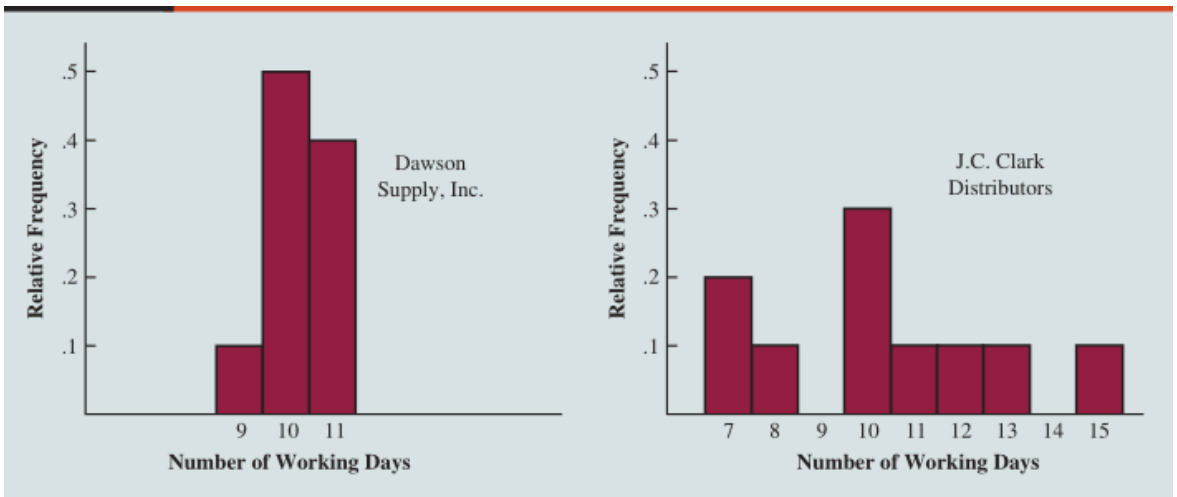
- Another measure of location is the Mode. The mode is the value that occurs with greatest frequency.
- Situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists.
- If the data contain exactly two modes, we say that the data are bimodal. If data contain more than two modes, we say that the data are multimodal.
- In multimodal cases the mode is almost never reported because listing three or more modes would not be particularly helpful in describing a location for the data.
- Anderson et al. (2020) has more details about Mean, Median and Mode, so please read chapter 3.1.
- What is the sample mode of the 10 heights? - Ans - 6 right?

Descriptive Statistics - Numerical Measures

2. Measures of Variability - Range, Interquartile Range, and Variance

Understanding Variability or Dispersion

- In addition to measures of location, it is often desirable to consider measures of variability, or dispersion.
- This is actually very important to understand the variability of the data? Why do we care of variability? Because more variability means more uncertainty, and we always prefer less uncertainty!
- For example, suppose that you are a purchasing agent for a large manufacturing firm and that you regularly place orders with two different suppliers. After several months of operation, you find that the *mean* number of days required to fill orders is 10 days for both of the suppliers. The histograms summarizing the number of working days required to fill orders from the suppliers are shown below



- Question - do the two suppliers demonstrate the same degree of reliability in terms of making deliveries on schedule? Note the dispersion, or variability, in delivery times indicated by the histograms. Which supplier would you prefer? The left one right? Why?

Range, Interquartile Range and Variance

- Anderson et al. (2020) discussed three measures of dispersion
  - 1. Range (Largest value - Smallest Value)
  - 2. Interquartile Range ($Q_3 - Q_1$)
  - 3. Sample Variance (or Sample Standard Deviation)
- Range is simply the difference between largest and the smallest value. So this is easy to compute. We can use direct Excel functions for it.
- Interquartile Range ($Q_3 - Q_1$) is the difference between two quartiles.
- And sample variance can be calculated using following formula.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- s^2 is the notation for sample variance.
- There is another object, called *standard deviation*, which is the square root of the sample variance, so we will write s for standard deviation.
- Let's do a simple example using MS-Excel.

Descriptive Statistics - Numerical Measures

3. Five-Number Summaries and Boxplots

Five number Summaries and Boxplot

- We already know the following numerical methods for quantitative variable,
  - 1. Smallest value
  - 2. First quartile (Q_1)
  - 3. Median (Q_2)
  - 4. Third quartile (Q_3)
  - 5. Largest value
- A box plot is just a visual representation of these numbers.
- Consider following dataset,

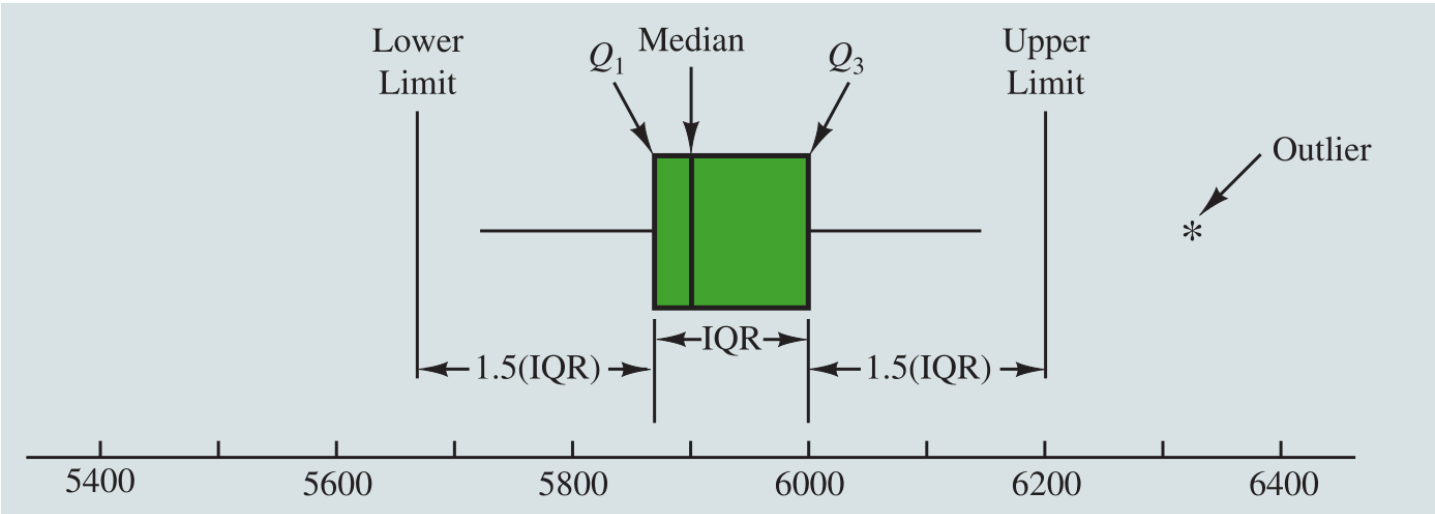
5710	5755	5850	5880	5880	5890
5920	5940	5950	6050	6130	6325

- The smallest value is 5710 and the largest value is 6325.
- We already learned how to compute the quartiles $(Q_1 = 5857.5; Q_2 = 5905$; and $Q_3 = 6025)$ in Section 3.1.
- Thus, the five-number summary for the monthly starting salary data is

$$\text{smallest } = 5710, \quad Q_1 = 5857.5, \quad Q_2 = 5905, \quad Q_3 = 6025, \quad \text{largest } = 6325$$

- The five-number summary indicates that the starting salaries in the sample are between 5710 and 6325 and that the median or middle value is 5905 ; and, the first and third quartiles show that *approximately* 50% *of the starting salaries are between 5857.5 and 6025*.
- Now using this information we can draw a box plot,

Five number Summaries and Boxplot



- It is very easy to draw boxplot using Excel, we will do this in the lab class.

Geometric Mean

- There is another concept known as *Geometric Mean*, which is often used to calculate the *average growth rate*. Here is an example, suppose you invested 100$ in a stock and following (see next page) is the yearly return of last 10 years.
- You want to know what is the average annual growth rate of this stock based on this data? This is taken from Anderson et al. (2020)
- Why we want to calculate average growth rate? Because if you know this then we can predict what will be the value of the stock after next 5 or 10 or 15... years.
- For example if the average annual growth rate is 5% or 0.05, then we can roughly say that the stock's value will be $100(1+0.05)^{10}$ after 10 years (why???).
- This is because,

$$
\begin{aligned}
\text{after one year} &= 100(1+0.05) \\
\text{after two years} &= 100(1+0.05)\times(1+0.05) = 100(1+0.05)^2 \\
&\vdots \\
\text{after ten years} &= 100\underbrace{(1+0.05)\times(1+0.05)\times\ldots\times(1+0.05)}_{10} = 100(1+0.05)^{10}
\end{aligned}
$$

Geometric Mean

Year	Return (%)
1	-22.1
2	28.7
3	10.9
4	4.9
5	15.8
6	5.5
7	-37.0
8	26.5
9	15.1
10	2.1

- To calculate the average annual growth rate, we need to first calculate the yearly growth rates.

Geometric Mean

Year	Return (%)	Yearly Growth Rate (or Factor)
1	-22.1 or -0.221	$1+(-0.221)=.779$
2	28.7 or 0.287	$1+0.287=1.287$
3	10.9 or 0.109	$1+.109=1.109$
4	4.9 or 0.049	$1+0.049=1.049$
5	15.8 or 0.158	$1+0.158=1.158$
6	5.5 or 0.055	$1+0.055=1.055$
7	-37.0 or -0.370	$1+(-0.370)=.630$
8	26.5 or 0.265	$1+0.265=1.265$
9	15.1 or 0.151	$1+0.151=1.151$
10	2.1 or -0.021	$1+(-0.021)=1.021$

- In Anderson et al. (2020) yearly growth rate is same as Growth Factor.

Geometric Mean

- Once we have the yearly growth rate, then we can take the geometric mean of yearly growth rate, and this is calculated as,

$$\begin{aligned}
&\sqrt[10]{[(.779)(1.287)(1.109)(1.049)(1.158)(1.055)(.630)(1.265)(1.151)(1.021)]} \\
&= [(.779)(1.287)(1.109)(1.049)(1.158)(1.055)(.630)(1.265)(1.151)(1.021)]^{1/10} \\
&= (1.3344)^{1/10} \\
&= 1.029 = 1 + 0.029
\end{aligned}$$

- So 2.9% is the average annual growth rate.
- If you compare and contrast with Arithmetic mean, then the idea of the Geometric mean is rather than taking sum, we are taking product, and rather than dividing we are taking the power $1/n$ (notice $1/n$ is same as $\sqrt[n]{}$)
- So the geometric mean is a measure of location that is calculated by finding the nth root of the product of n values. The general formula for the geometric mean, denoted $\bar{x}_g$, follows.

$$\bar{x}_g = \sqrt[n]{(x_1)\,(x_2)\cdots(x_n)} = [(x_1)\,(x_2)\cdots(x_n)]^{1/n}$$

- If you calculate the arithmetic mean of returns from the table, you should get 5.04%, and we calculated the geometric mean and we have 2.9%.
- Now we can ask, what might be average return after 10 years???
- This problem is taken from Anderson et al. (2020) so you should do it on your own.

Sample Median

- The median is another measure of central location.
- The median is the value in the *middle* when the data are arranged in *ascending order (smallest value to largest value)*.
- With an odd number of observations or samples, the median is the middle value.
- With an even number of observations there is no single middle value. In this case, we follow convention and define the median as the average of the values for the middle two observations. So here is how you can find median,
- Arrange the data in ascending order (smallest value to largest value).
  - (a) For an odd number of observations, the median is the middle value.
  - (b) For an even number of observations, the median is the average of the two middle values.
- Calculating Median by hand is very easy, we will also see an excel in the lab class.
- Note that median is where the 50% data is on the left and 50% is on the right!

1. What is Statistics

2. Data and Variables

3. Population and Sample

4. Descriptive Statistics - Tabular and Graphical Methods
   - Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts
   - Single Variable Methods (Quantitative): Histogram
   - Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

5. Descriptive Statistics - Numerical Measures
   - 1. Measures of Location - Mean, Median and Percentile
   - 2. Measures of Variability - Range, Interquartile Range, and Variance
   - 3. Five-Number Summaries and Boxplots

6. Two Variable Measures For Association : Covariance, Correlation and ScatterPlot
   - 1. Numerical Measures: Covariance and Correlation
   - 2. Graphical Measure: Scatterplot

7. Chapter Recap For the Descriptive Statistics Part

Two Variable Measures For Association : Covariance, Correlation and ScatterPlot

1. Numerical Measures: Covariance and Correlation

Measure of Association between two variables

Numerical Measures: Covariance and Correlation

- Now we will see some numerical and graphical methods which show association between two variables. The idea here is slightly different than before, here we are interested in *association*...
- What do we mean by "association"? Roughly this means *how two variables behaves together*.
- It's important to mention that *association does not imply causality*,... however *causality implies association*.
- This means that if the two variables are associated, we cannot say one variable causes the other to change, maybe there is a third variable which causes both of them to move together.
- Two important sample *numerical measures are important*, one is *covariance* and other is *correlation*
- Here is the formula for covariance

$$s_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- If you do the calculation in excel, then here is how you can do

Measure of Association between two variables

Numerical Measures: Covariance and Correlation

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	-1	-1	1
	5	57	2	6	12
	1	41	-2	-10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	-2	-13	26
	5	63	2	12	24
	3	48	0	-3	0
	4	59	1	8	8
	2	46	-1	-5	5
Totals	30	510	0	0	99

and then the covariance is

$$s_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{10-1} = 11$$

Now how do we interpret the covariance? The idea is *positive covariance means positive association, this happens when both deviations are either positive or negative*, and similarly, *negative covariance means negative association, and this happens when one deviation is positive and other is negative*

- We will see some applications in Excel,

Measure of Association between two variables

Numerical Measures: Covariance and Correlation

- The formula for correlation is just a modification of the covariance formula, but here we just need standard deviation of each of the variables, so here are the covariance of x and y and standard deviation of x and y respectively

$$s_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

- The correlation formula is the following,

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

- Interestingly the correlation between x and y, which we denoted with r_{xy} will always lie between -1 and 1, so

$$-1 \leq r_{x,y} \leq 1$$

- Interpretation is similar to covariance, however here we can also understand the strength of the association.

Measure of Association between two variables

- If the value of $r_{x,y}$ is *close to* 1 *indicates strong positive correlation*, which means strong positive association between two variables,
- Similarly if the value of $r_{x,y}$ is *close to* -1 *indicates strong negative correlation*, which means strong positive association between two variables,
- And finally if the value is close to 0 indicates, zero correlation and there is no association.

Two Variable Measures For Association : Covariance, Correlation and ScatterPlot

2. Graphical Measure: Scatterplot

Measure of Association between two variables

Graphical Measures: Scatterplot

- scatterplot is a plot where we plot (x, y) pairs in the $(x - y)$ co-ordinate, for example here is a data, which has two variables, the number of commercials which is x and sales (in 100\$) which is y

Week	Number of Commercials $\boldsymbol{x}$	Sales Volume (\$100s) $\boldsymbol{y}$
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

we can plot the (x, y) pairs as follows,

Measure of Association between two variables

Graphical Measures: Scatterplot

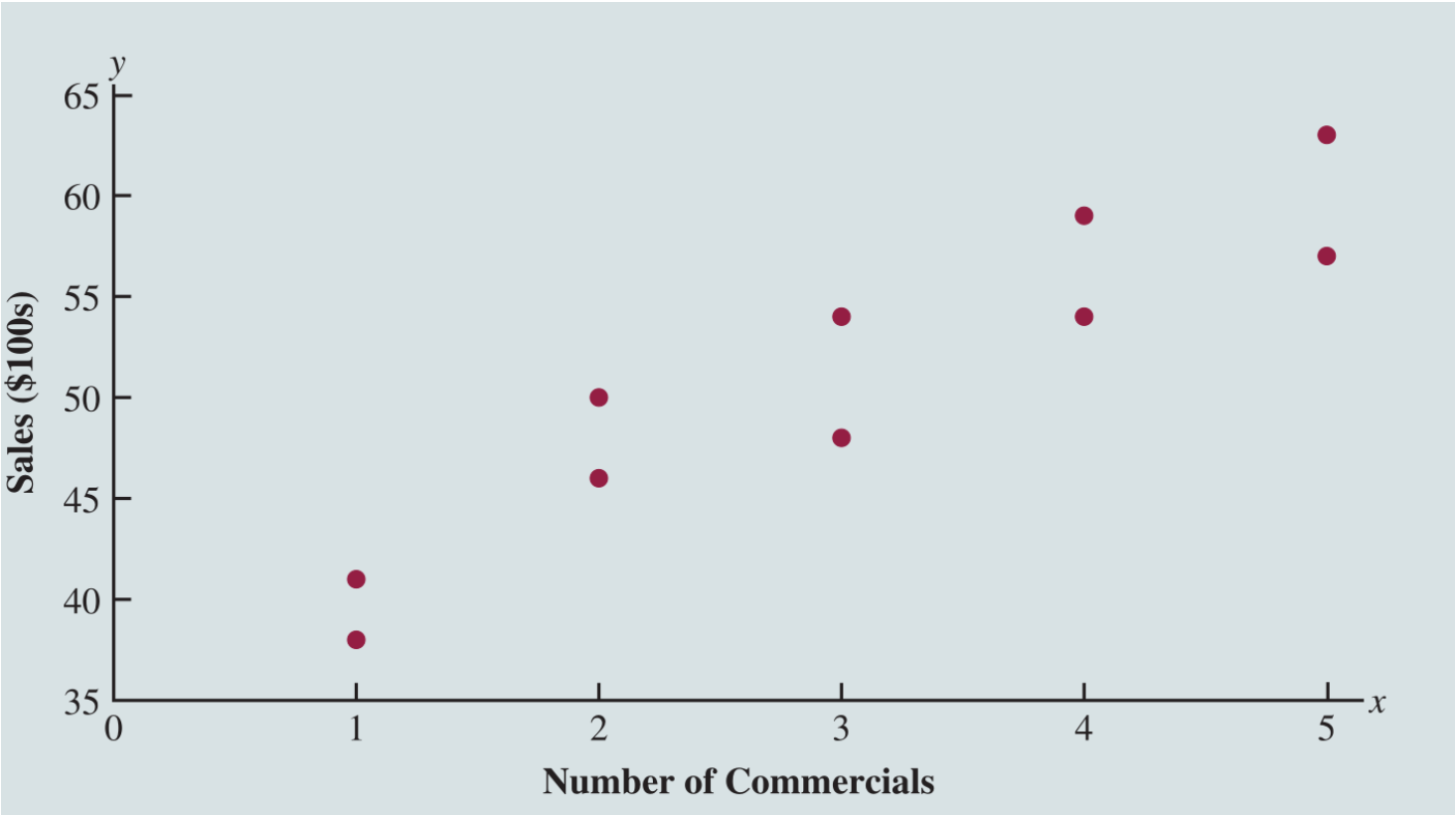


Figure 10: Scatter plot between x and y

- We will usually use MS-Excel to produce the scatter plot, and it's quite easy using Excel.

Measure of Association between two variables

Graphical Measures: Scatterplot

- It's important to note that, scatter-plot also shows positive and negative association, following picture is useful to understand this,

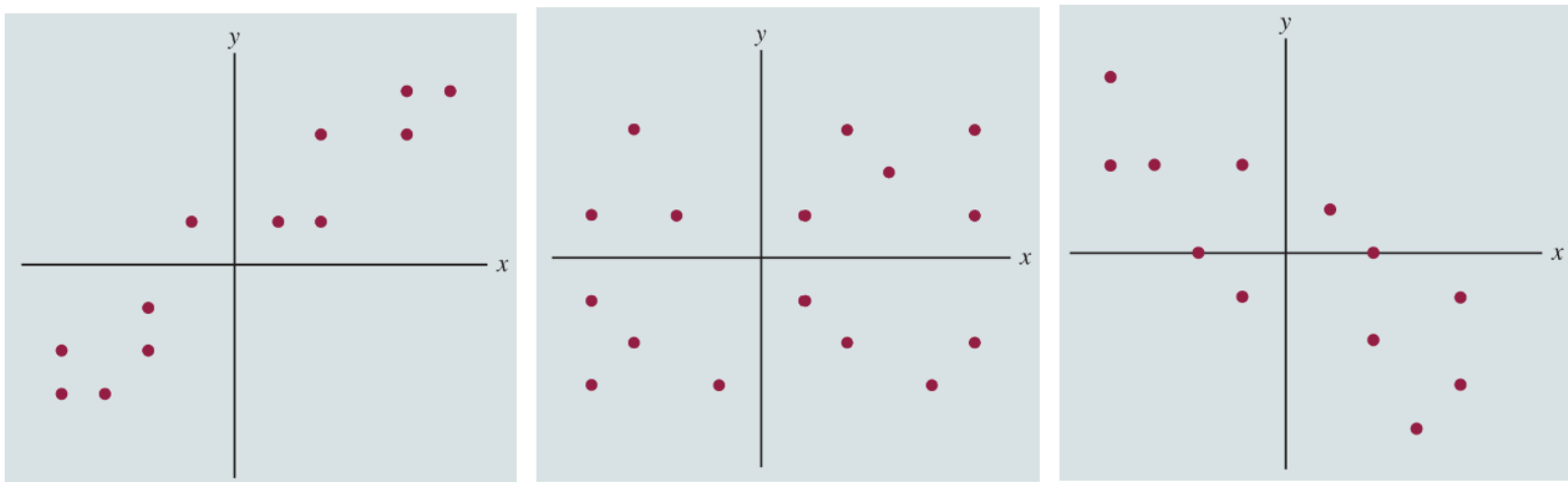


Figure 11: On the *Left*: *Positive Association*, On the *Middle*: Almost *No Association*, and On the *Right*: *Negative Association*

- The easy way to remember this is to notice - *when x is increases whether y increases or decreases...*
- Clearly on the picture, on the left if we calculate $r_{x,y}$ it should be close to 1, on the middle it should be close 0 and on the right it should be close to -1

1. **What is Statistics**

2. **Data and Variables**

3. **Population and Sample**

4. **Descriptive Statistics - Tabular and Graphical Methods**
   - Single Variable Methods (Qualitative): Frequency Distribution, Bar Charts and Pie Charts
   - Single Variable Methods (Quantitative): Histogram
   - Two Variable Methods (Qualitative): Cross Tabulation, Two Variate Bar Charts

5. **Descriptive Statistics - Numerical Measures**
   - 1. Measures of Location - Mean, Median and Percentile
   - 2. Measures of Variability - Range, Interquartile Range, and Variance
   - 3. Five-Number Summaries and Boxplots

6. **Two Variable Measures For Association : Covariance, Correlation and ScatterPlot**
   - 1. Numerical Measures: Covariance and Correlation
   - 2. Graphical Measure: Scatterplot

7. **Chapter Recap For the Descriptive Statistics Part**

Chapter Recap For the Descriptive Statistics Part

Recap For the Descriptive Statistics Part

- Since we have learned a lot of things it's a good idea to make a list of what we have learned in The Descriptive Statistics part,
- Here is what we have learned in the Tabular and Graphical Side
  - Frequency Distribution Table, Bar Chart and Pie Chart
  - Grouped Frequency Distribution Table and Histogram
  - Cross-tabulation or Contingency Table or Joint Frequency Distribution, and Side-by-Side Bar Chart or Stacked Bar Chart
  - Bivariate Frequency Distribution Table and Joint Histogram
  - Scatter Plot For Association
  - Boxplot
- Here is what we have learned in the Numerical Side
  - Mean, Weighted Mean, Percentile (Quantile), Median and Mode
  - Largest Value, Smallest Value, Range, Interquartile Range, Variance
  - Covariance and Correlation
  - You should know the excel functions of the numerical measures...

References

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J., and Ohlmann, J. W. (2020). *Statistics for Business & Economics*. Cengage, Boston, MA, 14th edition.