

# GE2262 Business Statistics

Lecturer: Dr. Iris Yeung  
Room : LAU-7239  
Tel No.: 34428566  
E-mail: [msiris@cityu.edu.hk](mailto:msiris@cityu.edu.hk)

# GE2262 Business Statistics

## Aim

- With today's widespread use of statistics in the media, academic and business firms, this course aims to provide students with a good understanding of basic statistical concepts so as to facilitate their decision making.

## Intended Learning Outcomes (ILO)

- Explain concepts in numerical descriptive measures, sampling distributions, confidence interval estimation, hypothesis testing, and simple linear regression model.
- Select appropriate statistical methods to analyze real-life business data, interpret the results and give recommendations for business decisions.
- Apply standard statistical software, such as Microsoft Excel, to analyze data arising from real-life business problems.
- Provide recommendations / innovations based on statistical data.

# GE2262 Business Statistics

## Course Outline

Introduction to Statistics. Basic Probability. Discrete and Continuous Probability Distributions. Sampling Distribution. Confidence Interval Estimation for the Population Mean. Hypothesis Testing for the Population Mean. Confidence Interval Estimation and Hypothesis Testing for the Population Proportion. Simple Linear Regression

## Assessment

Coursework : 50%

Two assignments : 25%

(Deadline: 5pm, March 5 [Week 7] and April 2 [Week 11])

Test : 25% (5:05 pm- 6:45 pm, March 18 [Week 9])

Written Examination : 50% (two hours)

# GE2262 Business Statistics

## Textbooks

1. Levine, D.M., Kathryn, A.S. and David, F.S. Business Statistics: A First Course, 2020, Pearson Education Limited.
2. Liu, K. I., To K. M., Speaking of Statistics, 2014, Pearson Education Ltd
3. Newbold, P., Carlson, W.L. and Thorne, B. Statistics for Business and Economic. Prentice Hall

# GE2262 Business Statistics

<i>Week</i>	<i>Content</i>	<i>Reading</i>
1-2	Topic 1: Introduction to Statistics	Le Ch 1-3
2-3	Topic 2: Basic Probability	Le Ch 4
3-4	Topic 3: Discrete and Continuous Probability Distributions	Le Ch 5-6
5	Topic 4: Sampling Distribution	Le Ch 7
6	Topic 5: Confidence Interval Estimation for the Population Mean	Le Ch 8
7-8	Topic 6: Hypothesis Testing for the Population Mean	Le Ch 9
10	Topic 7: Confidence Interval Estimation and Hypothesis Testing for the Population Proportion	Le Ch 9
11-12	Topic 8: Simple Linear Regression	Le Ch 12
9	Test	

# GE2262 Business Statistics

## Topic 1 Introduction to Statistics

Lecturer: Dr. Iris Yeung

Room : LAU-7239

Tel No.: 34428566

E-mail: [msiris@cityu.edu.hk](mailto:msiris@cityu.edu.hk)

# Outline

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- Descriptive Statistics
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - Measures of Variation
  - Distribution Shape
  - Use of Excel in Organizing Categorical Data
  - Use of Excel in Calculating Summary Statistics for Numeric Data

# Part 1a

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- Descriptive Statistics
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - Measures of Variation
  - Distribution Shape
  - Use of Excel in Organizing Categorical Data
  - Use of Excel in Calculating Summary Statistics for Numeric Data



# What is Statistics?

- **Statistics** is a field of study that deals with collection, analysis, interpretation and presentation of masses of data.
- Two types of Statistics
  - **Descriptive Statistics** (Topic 1)
    - to describe, summarize and present data via tables, graphs, and summary measures
  - **Inferential Statistics** (Topics 5 – 7)
    - to infer, conclude, and make decisions about a large group (population) from a small group (sample).

# Example for Descriptive and Inferential Statistics

## ■ Raw Data

- Age of 30 students in a class are 20, 19, 18, 19, ... .

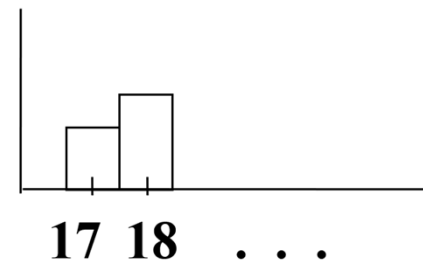
## ■ Descriptive Statistics

mean age = 18.5

minimum age = 17

maximum age = 23

range = 6



## ■ Inferential Statistics

- Based on the mean age of 30 students in a class, we want to estimate the mean age of all students in a class (around 200)

# Population and Sample

- A **population** contains **all** the items or individuals about which we want to study
  - Example: all batteries produced by a manufacturer in a day (around 1000); all students in a class (around 200)
- A **sample** contains only a **portion** of the population of items or individuals
  - Example: 50 batteries, 30 students
- A **variable** is a characteristic of an item or individual
  - Example: a battery's lifetime, weight,...; a student's gender, age, ...
- A **population parameter** summarizes the value of a specific variable for a population
  - Example: mean lifetime of all batteries; mean age of all students
- A **sample statistic** summarizes the value of a specific variable for sample data
  - Example: mean lifetime of 50 batteries; mean age of 30 students

# Census and Sampling

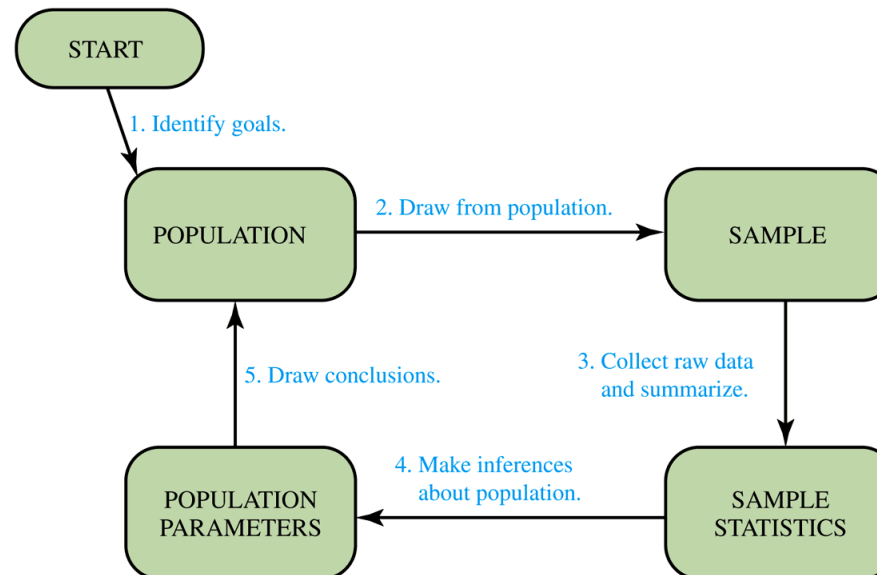
- When all the items or individuals in a population are studied, we are taking a **census**.
- If only a portion of the items or individuals in the population are studied, we are doing **sampling**.
- For census, we need only descriptive statistics to describe and present census data.
- But for sampling, we need both descriptive statistics to describe and present sample data and inferential statistics to infer about the population based on sample results. The inference is not exact and will involve uncertainty.
- The language of uncertainty (probability, probability distribution and sampling distribution) will be studied in Topics 2-4.

# Why Sampling?

- Resource limitation (labour, money, time etc.)
  - Example: students' age
- Destructive Testing
  - Example: lifetime of batteries, light bulb, ..., blood test for covid19 antibodies, ...

# Process of a Statistical Study

- Step 1: State the goal(s) of your study precisely -- determine the **population** we want to study and what we'd like to learn about it (population parameters)
- Step 2: Choose a **sample** from the population (use an appropriate sampling technique)
- probability sampling** – select items based on known probability, eg: simple random sampling
  - non-probability sampling** – select items without knowing their probability of selection, eg: convenience sampling
- Step 3. **Collect** raw data from the sample and **summarize** these data by finding sample statistics of interest
- Step 4. Use the sample statistics to **make inference** about the population
- Step 5. Draw **conclusions**, determine what we learned and whether we achieved your goal(s)



# Types of Variables and Data

- A **variable** is a characteristic of an item or individual
  - Example: a person's gender, age, education level, income, ...
- **Data** are the set of values associated with one or more variables
  - Example: Age of 30 students in a class are 20, 19, 18, 19, ... .
- Variables/Data can be categorical (qualitative) or numeric (quantitative)
  - **Categorical/Qualitative**: data are non-numeric and categorical
    - Example: a person's gender, education level, occupation, district of living, colour preference ...
  - **Numeric/Quantitative**: data are measured on a numerical scale or counted
    - Example: age (in number of years), height, weight, salary, number of children, ...

# Data Coding

Respondent	Age	Gender	Opinion on next year's house price
1	1	1	1
2	1	0	3
3	2	1	2
4	5	0	3
...			

Age	Code
18-24	1
25-34	2
35-44	3
45-54	4
55 & over	5

Gender	Code
Male	1
Female	0

House price	Code
Decrease	1
No change	2
Increase	3



# Part 1b

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- Descriptive Statistics
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - Measures of Variation
  - Distribution Shape
  - Use of Excel in Organizing Categorical Data
  - Use of Excel in Calculating Summary Statistics for Numeric Data

# Application of Statistics

## Accounting

### Scope of this ISA

1. This International Standard on Auditing (ISA) applies when the auditor has decided to use audit sampling in performing audit procedures. It deals with the auditor's use of statistical and non-statistical sampling when designing and selecting the audit sample, performing tests of controls and tests of details, and evaluating the results from the sample.
2. This ISA complements ISA 500,<sup>1</sup> which deals with the auditor's responsibility to design and perform audit procedures to obtain sufficient appropriate audit evidence to be able to draw reasonable conclusions on which to base the auditor's opinion. ISA 500 provides guidance on the means available to the auditor for selecting items for testing, of which audit sampling is one means.

### Objective

4. The objective of the auditor, when using audit sampling, is to provide a reasonable basis for the auditor to draw conclusions about the population from which the sample is selected.



### ■ Audit Sampling

- The application of audit procedures to **less than 100%** of items within a population of audit relevance such that all sampling units have a chance of selection in order to provide the auditor with a reasonable basis on which to **draw conclusions about the entire population**
- The auditor shall determine a **sample size** sufficient to **reduce sampling risk to an acceptably low level**

Source: <http://www.ifac.org/sites/default/files/publications/files/A028%202012%20IAASB%20Handbook%20ISA%20530.pdf>

# Application of Statistics

## Information Technology

### ■ IT Auditing

- ❑ Almost all computer-assisted audit tools (CAATs) have a command for Benford's Law
- ❑ Benford's Law holds true for a data set that grows **exponentially**, but also appears to hold true for many cases in which an exponential growth pattern is not obvious
- ❑ Beneficial tool for **fraud detection**, e.g. credit card transactions, purchase orders, loan data, customer refunds, ...
- ❑ Example
  - If a bank's policy is to refer loans at or above US \$50,000 to a loan committee, looking just below that approval threshold gives a loan officer the potential to discover loan frauds.
  - Note that 4 is aberrantly high in occurrence, and 5 is too low, indicating the possible manipulation of the natural occurrence of loans beginning with 5 (US \$50,000 loans) possibly being switched to just under the cutoff or indicating that the suspect could be issuing a lot of \$49,999.99 loans fictitiously to embezzle funds.

Figure 1—Benford's Law Distribution Leading Digit

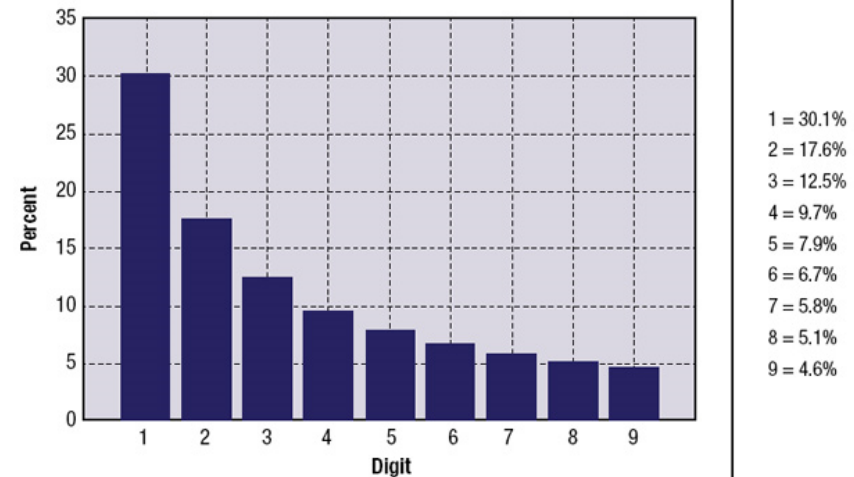
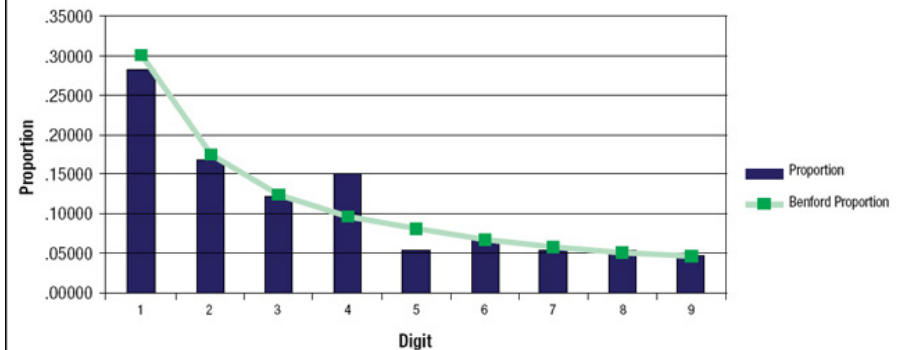


Figure 2—Benford's Law Test/Comparison



Source: <http://www.isaca.org/Journal/Past-Issues/2011/Volume-3/Pages/Understanding-and-Applying-Benford's-Law.aspx>

# Application of Statistics

## Marketing

### ■ Marketing Research

- Construction of questionnaires and scales
- **Understand the needs** of individuals in marketplace, to create **marketing strategies and plans**

#### Interesting facts on mobile website users expectations

\*The following stats are compiled from a study from Google (conducted by Sterling Research and SmithGeiger, independent market research firms). The report surveyed 1,088 US adult smartphone Internet users in July 2012.

**Friendly** = More likely to buy

**Unfriendly** = More likely to leave

**67%**

"A mobile-friendly site makes me more likely to buy a product or use a service."



**61%**

"If I don't see what I'm looking for right away on a mobile site, I'll quickly move on to another site."



#### Turning visitors into customers\*

If your site offers a great mobile experience your chance of getting new customers that visit your site increases dramatically.

- 74% of people say they are more likely to return to a website if it is mobile friendly.

#### Hurting your business and helping your competition\*

If you have a poor mobile experience and your competitors have built a great experience for mobile users chances are they will benefit and you will be hurt.

- 61% of users said if they don't find what they are after right away on a mobile site, they will

#### Non mobile sites can damage your reputation\*

If a site isn't designed for mobile users it can leave users feeling frustrated.

- 48% of users say they feel frustrated and annoyed when they get to a site that is not mobile friendly.
- 52% of users said that a bad mobile experience made them less likely to engage with a company.
- 48% said that if a site didn't work well on their smartphones, it made them feel like the company didn't care about their business.

While you may agree or disagree with some of the responses the thing to take away from this is that it's imperative to your business to ensure you provide a great mobile experience.

Source: <http://www.onlinemarketing.fuelgroup.com.au/mobile-websites>

# Application of Statistics

## Economics

- Analyse and predict **economic performance**

( July 1997 = 100 )

### Centa-City Leading Index CCL

Announced every Friday, latest on 2014/08/15; reflecting secondary private residential property price from 2014/08/04 to 2014/08/10 (based on scheduled formal sale & purchase date; on average, formal S&P are signed within 14 days after preliminary S&P)

	This Week	Previous Week	Previous Month
[Centa-City Leading Index]	125.66	↑ 1.34 %	↑ 1.86 %
[Centa-City (large units) Leading Index]	131.41	↑ 2.03 %	↑ 3.53 %
[Centa-City (small/medium units) Leading Index]	124.01	↑ 1.21 %	↑ 1.55 %
[Mass Centa-City Leading Index]	125.64	↑ 1.14 %	↑ 1.43 %

### [Centa-City Leading Sub-index]

	This Week	Previous Week	Previous Month
HK	135.79	↑ 1.23 %	↑ 1.7 %
KLN	125	↑ 1.86 %	↑ 1.7 %
NT (East)	126.35	↑ 0.76 %	↑ 0.42 %
NT (West)	107.26	↑ 0.13 %	↑ 1.31 %



Source: [http://hk.centadata.com/cci/cci\\_e.htm](http://hk.centadata.com/cci/cci_e.htm)

# Application of Statistics

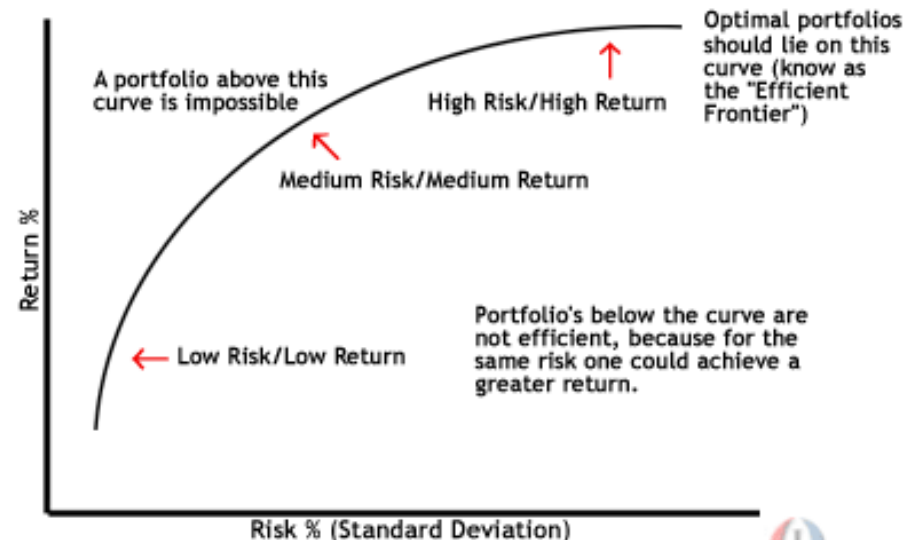
## Finance

### ■ Risk and Portfolio Management

- Use statistical models to analyze the market
- Efficient frontier of portfolio (a basket of stocks)

**Modern portfolio theory (MPT)** is a theory of finance that attempts to maximize portfolio expected return for a given amount of portfolio risk, or equivalently minimize risk for a given level of expected return, by carefully choosing the proportions of various assets. Although MPT is widely used in practice in the financial industry and several of its creators won a Nobel memorial prize for the theory,<sup>[1]</sup> in recent years the basic assumptions of MPT have been widely challenged by fields such as behavioral economics.

More technically, MPT models an asset's return as a normally distributed function (or more generally as an elliptically distributed random variable), defines risk as the standard deviation of return, and models a portfolio as a weighted combination of assets, so that the return of a portfolio is the weighted combination of the assets' returns. By combining different assets whose returns are not perfectly positively correlated, MPT seeks to reduce the total variance of the portfolio return. MPT also assumes that investors are rational and markets are efficient.



Source:

[http://en.wikipedia.org/wiki/Modern\\_portfolio\\_theory#The\\_efficient\\_frontier\\_with\\_no\\_risk-free\\_asset](http://en.wikipedia.org/wiki/Modern_portfolio_theory#The_efficient_frontier_with_no_risk-free_asset)



# Application of Statistics

## Big Data Analysis

- Using big data analysis, Walmart found from checkout-counter data an unexpected correlation – diapers and beer are usually bought together on Fridays
- Why? -- young fathers pick up diapers and beer after work on Fridays for the weekend
- Capitalizing on the discovery, the store
  - ❑ placed these two disparate items together
  - ❑ placed high-price diapers besides beer (as males don't concern the price)
- Sales zoomed!!!



# Application of Statistics

News  
report

## HP to Lay Off 9,000 in Enterprise Services Revamp

by Douglas McIntyre  Jun 1st 2010 9:35AM

Updated Jun 1st 2010 9:43AM

Hewlett-Packard (HPQ) announced Tuesday that it would put \$1 billion into its enterprise services division -- and lay off 9,000 workers in the process. In the last quarter for which it has reported results, which ended on Jan. 31, HP had revenue of \$21 billion and net income of \$2.3 billion. The enterprise unit provides consulting, outsourcing and technology services.



Paul Sakuma, AP

HP's 10-Q shows that the revenue from the division was \$8.7 billion, down slightly from the same period a year ago. Operating income was \$1.3 billion. So, the business is critical to HP's success, but it isn't doing terribly well.

The enterprise operation is part of HP's plan to diversify beyond hardware and become more competitive with rival IBM (IBM). HP bought information-technology consulting firm EDS in May 2008 for \$13.1 billion. In March 2009, HP said it would eliminate 24,600 jobs during the integration of EDS -- but it's not entirely clear whether the new layoffs are a subset of those or in addition to them.

In the case of the HP layoff, there were actually 324,600 staff members employed by HP at the time. Therefore, the percentage change in total employees as a result of the layoff was just:

$$\frac{9,000}{324,600} \times 100\% \approx 2.77\%$$

which suddenly diminishes the severity of the issue



## Part 2a

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- Descriptive Statistics
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - Measures of Variation
  - Distribution Shape
  - Use of Excel in Organizing Categorical Data
  - Use of Excel in Calculating Summary Statistics for Numeric Data

# Organizing and Visualizing Data

- Categorical Data
  - Summary table
  - Bar chart
  - Pie chart
- Numeric Data
  - Frequency distribution
  - Histogram

# Categorical Data – Summary Table

- Suppose you asked 60 customers to pick which of the three colours, say green, red, or blue they like best for a product and obtained the following raw data:
  - green, red, green, green, red, red, blue, blue, green, red, green, blue, red, blue, green, green, blue, green, green, blue, green, blue, green, red, blue, green, green, green, green, red, red, red, blue, green, green, green, green, blue, red, red, green, green, red, blue, green, red, green, green, blue, red, green, red, green, blue, blue, blue, green, green, green, green

- Summary table:

Customers' Favorite Colour		
Colour	Number of Customers	Percent of Customers
blue	15	25% (= 15/60)
green	30	50%
red	15	25%
Total	60	100%

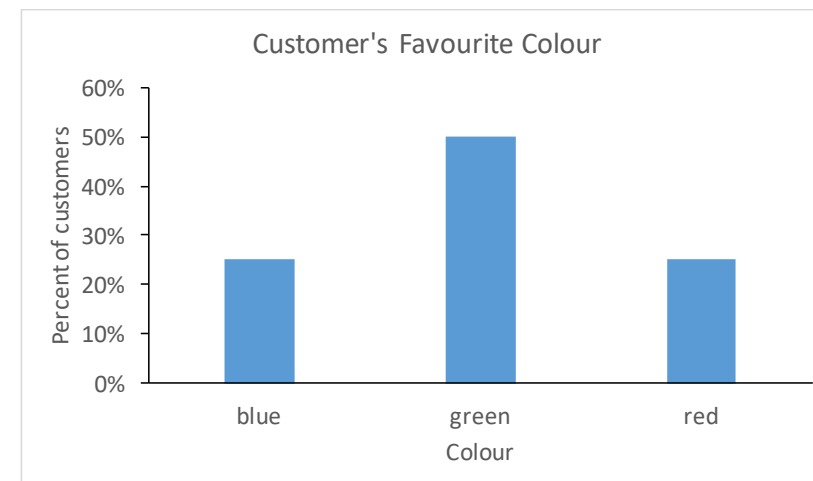
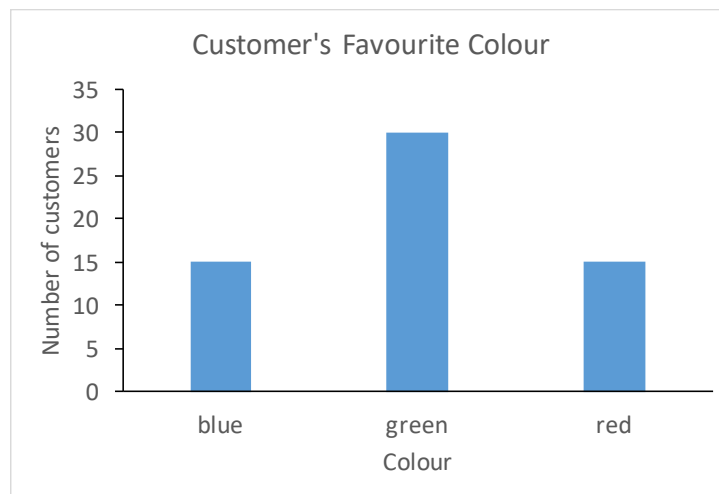
- It is accustomed to list the values of the variable in alphabetical order of the category, or in descending (or ascending) order of the number of customers (called **frequency**) or percent of customers (called **relative frequency**)
- The table tells us
  - 15 (25%) customers picked blue, 30 (50%) customers picked green, and 15 (25%) customers picked red
  - More customers picked green colour rather than blue or red

# Categorical Data – Bar Chart

## ■ Features of a Bar Chart

- ❑ It is accustomed to arrange the bars in the alphabetical order of the categories of the variable, or in descending (or in ascending) order of the frequency or relative frequency
- ❑ It is up to you to decide the gap between two bars and the width of each bar, so long as the gaps are the same and the bars have the same width
- ❑ The height of each bar represents the frequency or relative frequency (%) in that category

Colour	Number of Customers	Percent of Customers
blue	15	25% (= 15/60)
green	30	50%
red	15	25%
Total	60	100%

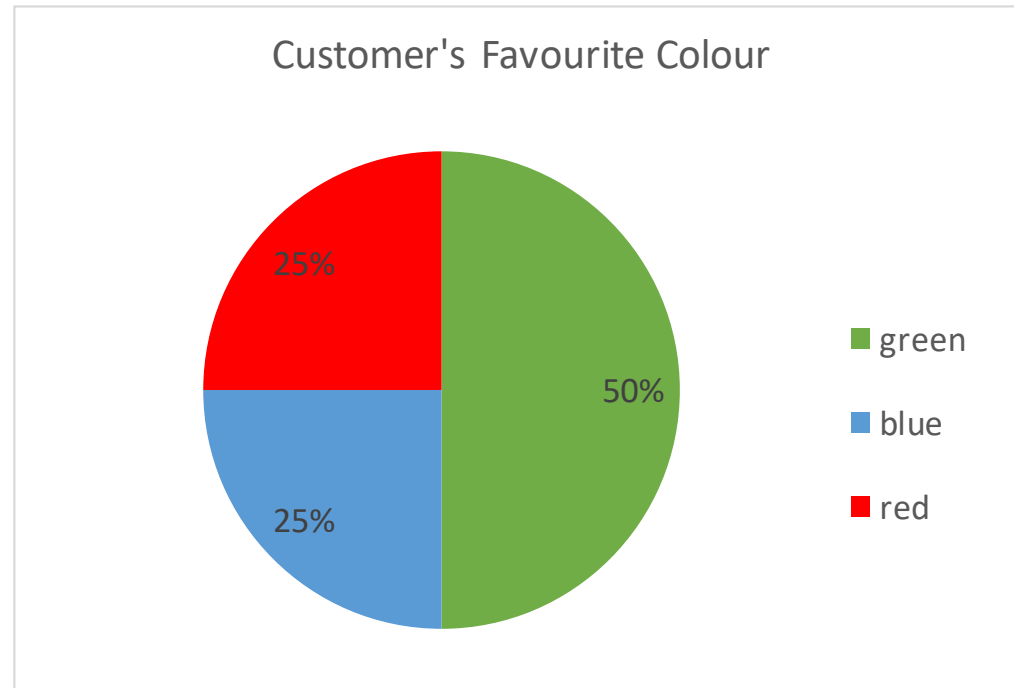


# Categorical Data – Pie Chart

## ■ Features of a Pie Chart

- ❑ It is accustomed to arrange the slices in the alphabetical order of the categories of the variable, or preferably in **descending order** of the frequency or relative frequency
- ❑ Slices with very small percentage may need to be combined with others
- ❑ Numbers (frequency or relative frequency) should be shown as it is difficult to compare slices of similar size

Colour	Number of Customers	Percent of Customers
blue	15	25% (= 15/60)
green	30	50%
red	15	25%
Total	60	100%



# Numeric Data – Frequency Distribution

- Suppose you asked 100 people about the amount they spent in their last visit to supermarket and obtained the following raw data:
  - ▣ 44.8, 230.5, 303.6, 70.8, 534.4, 166.2, 466, 85.1, 63, 47.8 36.5, 35.7, 12.7, 11.9, 297.5, 74.1, 77.1, 251.2, 127.1, 118.6, 211.2, 221.9, 49.1, 349.1, 556.6, 768, 231.7, 247.2, 87.4, 304.3, 311.3, **825.8**, 15.9, 526, 5.2, 156.7, 65.2, 143.3, 138.5, 478.4, 124.2, 205.1, 90.8, 3.1, 334.8, 7.4, 113.8, 79.2, 128.8, 26.6, 15.2, 554.4, **2.9**, 70.2, 540.7, 36.4, 588.9, 151.5, 14.2, 235.7, 13.7, 187.4, 817.8, 140.3, 114.9, 219.5, 31.4, 99.4, 47.3, 111.8, 230.2, 478.2, 4.6, 783.5, 483.5, 99.3, 92.8, 464.2, 172.9, 380.1, 234.5, 120.2, 100.3, 109.8, 276.1, 157.7, 192.9, 13.1, 62.2, 44.2, 35.9, 239.9, 193.8, 591.9, 249.1, 17.9, 89.3, 369.1, 38.2, 154.3
- The frequency distribution /relative frequency distribution is a summary table in which the data are arranged into numerically ordered classes

## Frequency distribution

Amount Spent (\$)	Frequency
0 - < 100	40
100 - < 200	22
200 - < 300	15
300 - < 400	7
400 - < 500	5
500 - < 600	7
600 - < 700	0
700 - < 800	2
800 - < 900	2
900 - < 1000	0
Total	100

## Relative frequency distribution

Amount Spent (\$)	Frequency	Relative Frequency
0 - < 100	40	0.40
100 - < 200	22	0.22
200 - < 300	15	0.15
300 - < 400	7	0.07
400 - < 500	5	0.05
500 - < 600	7	0.07
600 - < 700	0	0.00
700 - < 800	2	0.02
800 - < 900	2	0.02
900 - < 1000	0	0.00
Total	100	1.00

# Steps to Construct a Frequency Distribution

1. If the number of observations is small, it is good to sort data in ascending order first
2. Find the **range**:  $825.8 - 2.9 = 822.9$
3. Select the **number of classes**: 10
4. Compute the **class interval** (width):  $822.9 / 10 = 82.29$ 
  - ❑ Round up to a convenient number, say 100
5. Determine **class boundaries** (limits):
  - ❑ Class 1: 0 but less than 100
  - ❑ Class 2: 100 but less than 200
  - ❑ ...
6. Assign the observation to each class and count the number of observations

# Features of Frequency Distribution

- Exact value of each observation is lost
- The number of classes depends on the number of observations and the data range. In general, 5 to 15 classes will be sufficient
- The width of a class depends on the number of classes adopted and the data range
  - To determine the width of a class, you divide the range (highest value – lowest value) of the data by the number of classes desired
- The width of each class is equal
  - Width can be unequal. However, it should be done so only under very special circumstances, such as when the data are sparsely distributed, or have a very long tail at one or both ends
- The lower value of the first class interval is often the smallest value in the data, or a smaller value which is selected for the reason of convenience, such as 0
- Class boundaries include the left and right endpoint, but may not especially for the first and last class interval
  - endpoint policy need to be consistent

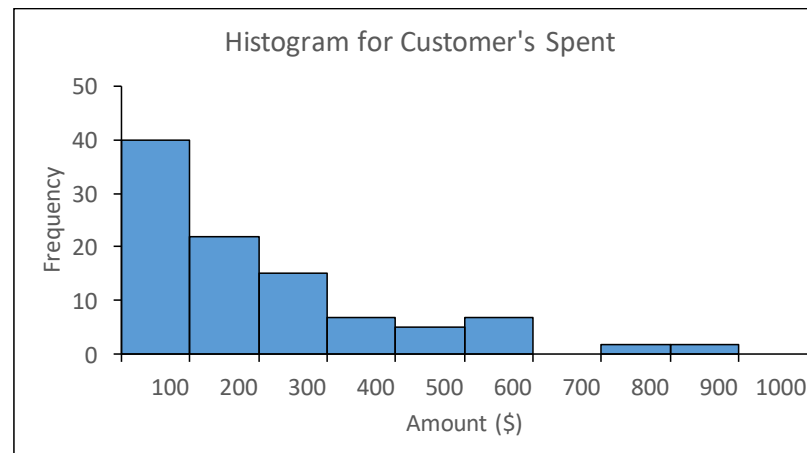


# Numeric Data – Histogram

- A histogram is a bar chart for grouped numeric data in which the frequency of each group of numerical data is represented as a bar
- Features of a histogram
  - The height of the bars represents the frequency (or relative frequency)
  - There is **no gap** between bars
    - If an interval has 0 frequency, the height of the bar in the histogram is 0
  - The width of the bars must be identical if the width of the intervals are identical
  - The bar must be drawn in the same sequence as of the intervals in the frequency distribution

## Frequency distribution

Amount Spent (\$)	Frequency
0 - < 100	40
100 - < 200	22
200 - < 300	15
300 - < 400	7
400 - < 500	5
500 - < 600	7
600 - < 700	0
700 - < 800	2
800 - < 900	2
900 - < 1000	0
Total	100

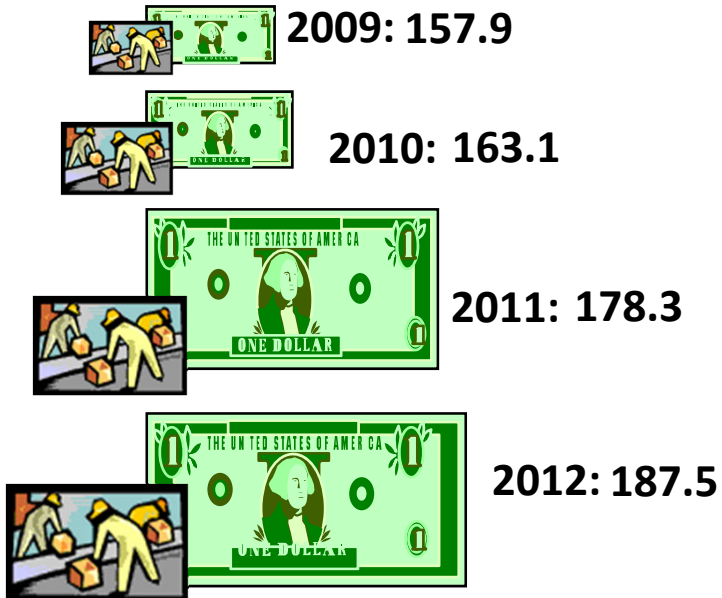


# Bad Graphs: Chart Junk



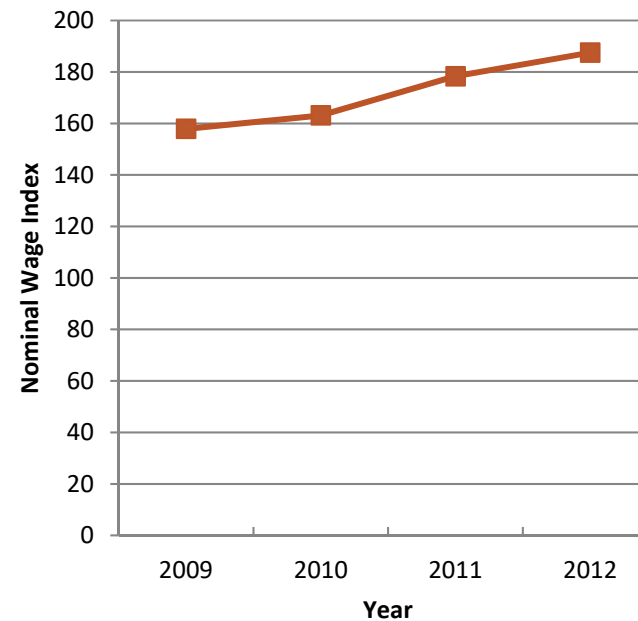
## Bad Presentation

Nominal Wage Index in Hong Kong  
(Sept 1992=100)



## Good Presentation

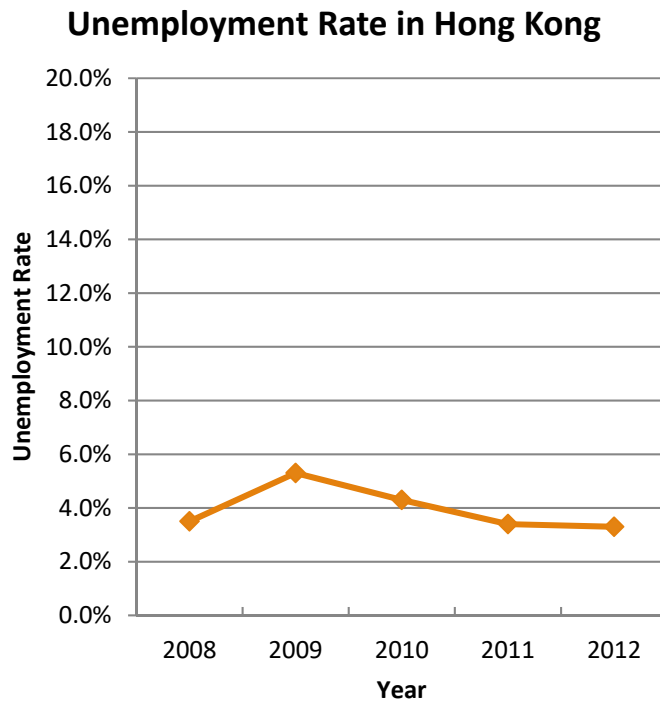
Nominal Wage Index in Hong Kong  
(Sept 1992=100)



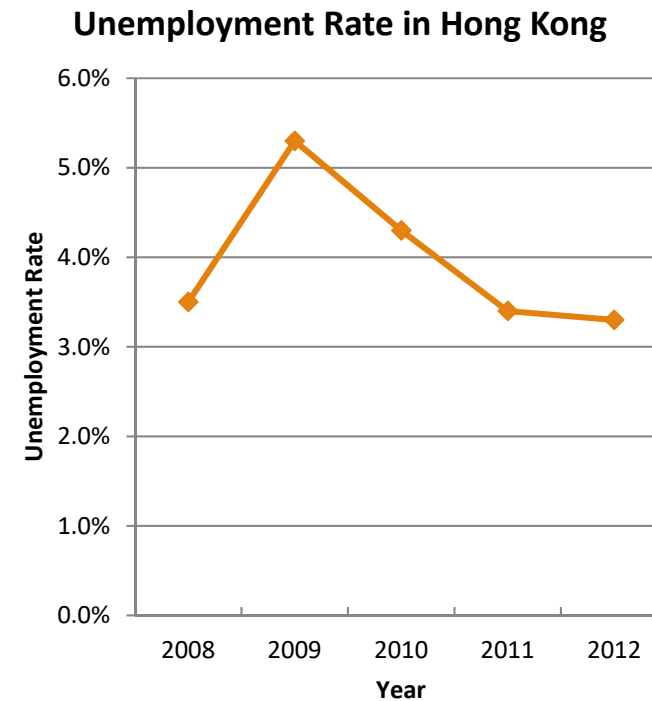
# Bad Graphs: Compressing the Vertical Axis



**Bad Presentation**



**Good Presentation**

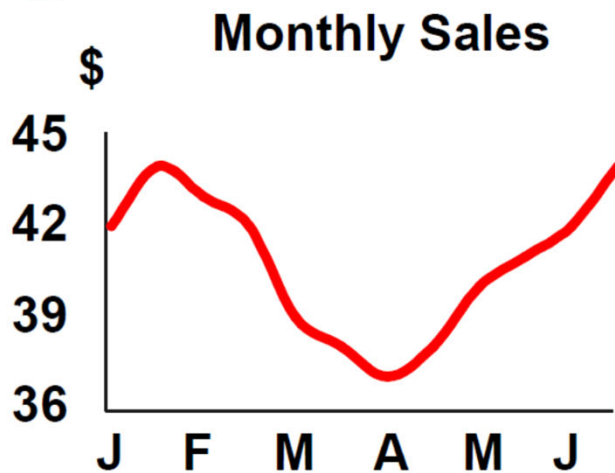


## Bad Graphs:

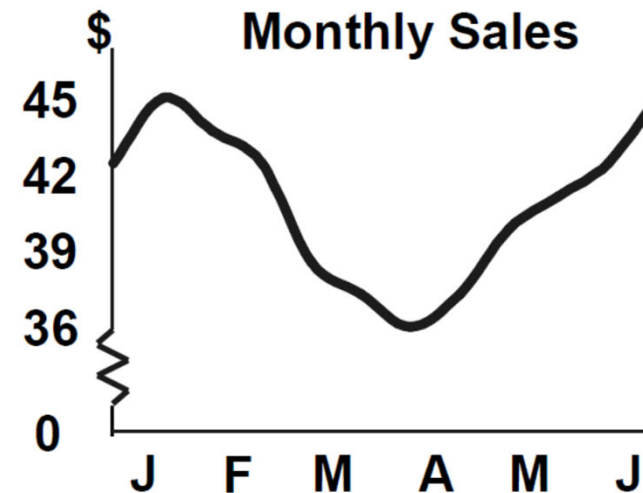
### No Zero Point on the Vertical Axis



**Bad Presentation**



**✓ Good Presentations**



# Principles of Excellent Graphs

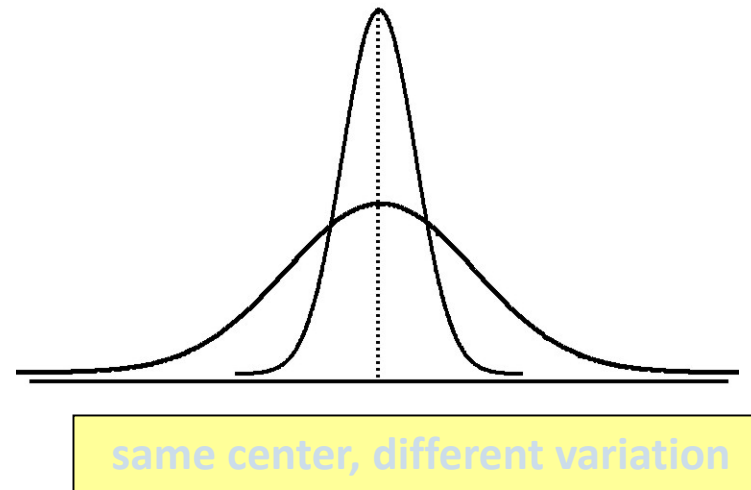
- The graph should **not contain unnecessary adornments** (sometimes referred to as chart junk)
- The graph should **not distort** the data
- The scale on the vertical axis should **begin at zero**
- The graph should contain a **title**
- All **axes** should be properly **labeled**
- The **simplest** possible graph should be used for a given set of data

## Part 2b

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- **Descriptive Statistics**
  - Organizing and Visualizing Data
  - **Measures of Central Tendency**
  - Measures of Variation
  - Distribution Shape
  - Use of Excel in Organizing Categorical Data
  - Use of Excel in Calculating Summary Statistics for Numeric Data

# Numeric Summary Measures

- Measures of **central tendency** gives information on the central value of the data (i.e. what value do the data center around? What is the typical value?)
  - ❑ Mean
  - ❑ Median
  - ❑ Mode
- Measures of **variation** gives information on the dispersion of the data
  - ❑ Range
  - ❑ Interquartile range
  - ❑ Variance
  - ❑ Standard deviation



# Mean

- Population mean (parameter)

pronounced mu

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{X_1 + X_2 + \dots + X_N}{N}$$

Population Size

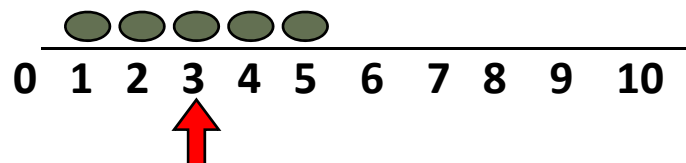
- Sample mean (statistic)

pronounced x-bar

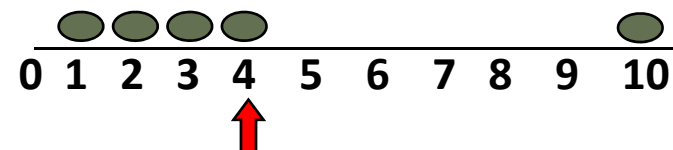
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Sample Size

- Affected by extreme values (outliers)



$$\frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

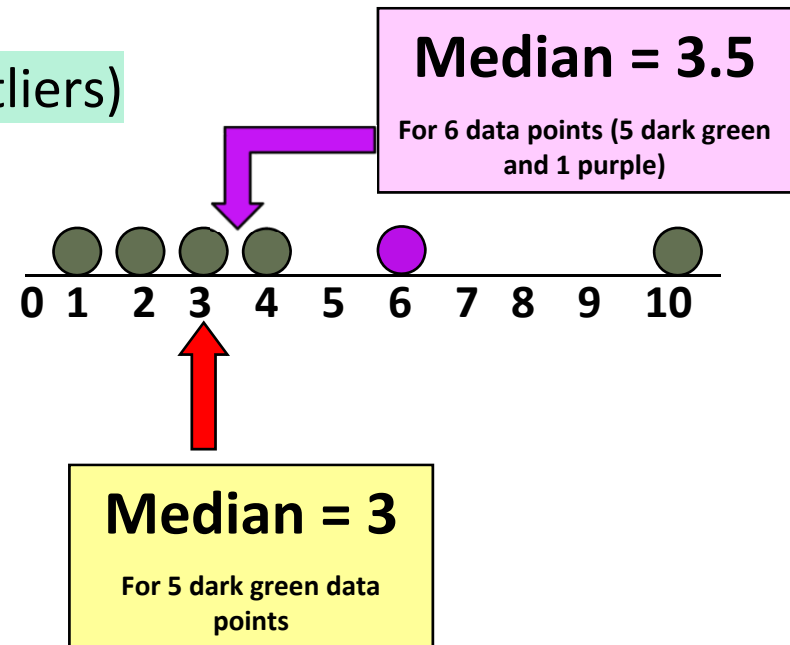
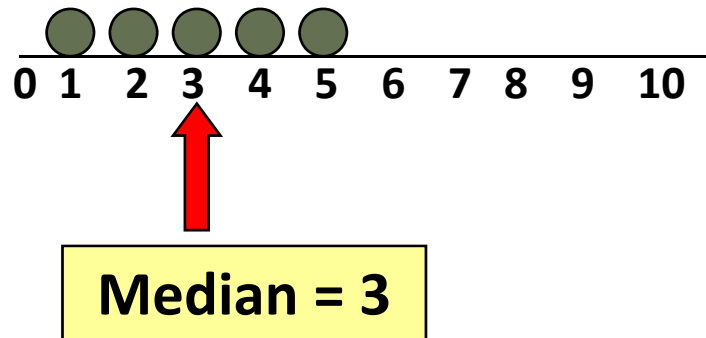
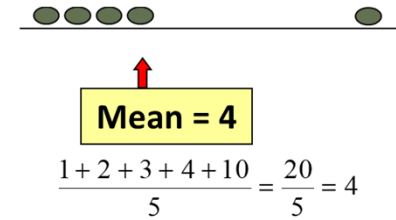
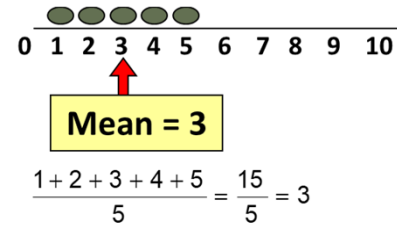


$$\frac{1 + 2 + 3 + 4 + 10}{5} = \frac{20}{5} = 4$$



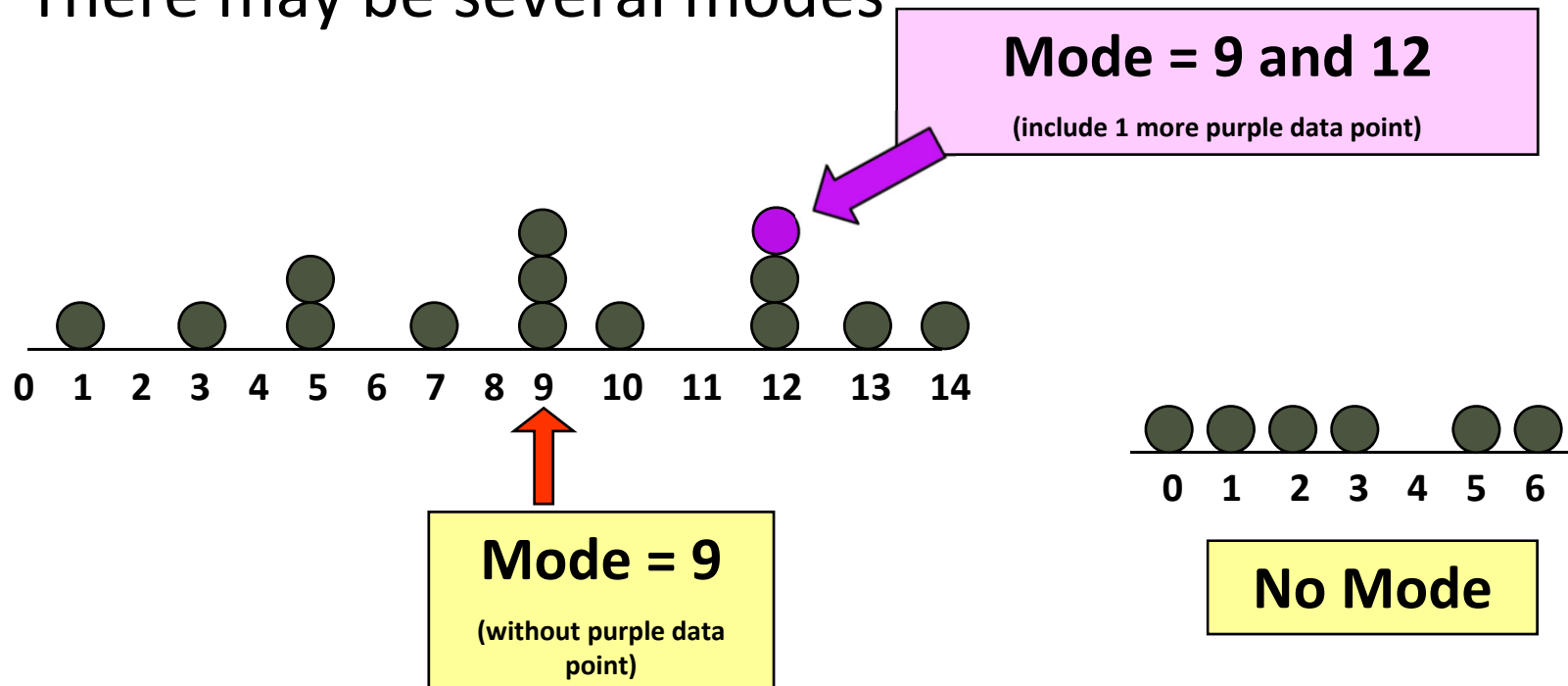
# Median

- Sort data in ascending order
- In an **ordered array**, the median is the “middle” number (50% above, 50% below). The **median position** in the ordered array is :  $\frac{(n+1)}{2}$ 
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the 2 middle numbers
- Not affected by extreme values (outliers)



# Mode

- Value that occurs most often
- Not affected by extreme values (outliers)
- Used for both numeric and categorical data
- There may be no mode
- There may be several modes



# Effects of Outliers

Imagine that the five graduating seniors on a college basketball team receive the following first-year contract offers to play in the National Basketball Association (zero indicates that the player did not receive a contract offer):

0    0    0    0    \$3,500,000

The mean contract offer is:

$$\text{mean} = \frac{0 + 0 + 0 + 0 + \$3,500,000}{5} = \$700,000$$

Is it fair to say that the average senior on this basketball team received a \$700,000 contract offer?

Note:

When the Hong Kong Housing Authority revises the rent of public housing, they look at median income of the tenants rather than average income

# Comparison of Mean, Median & Mode

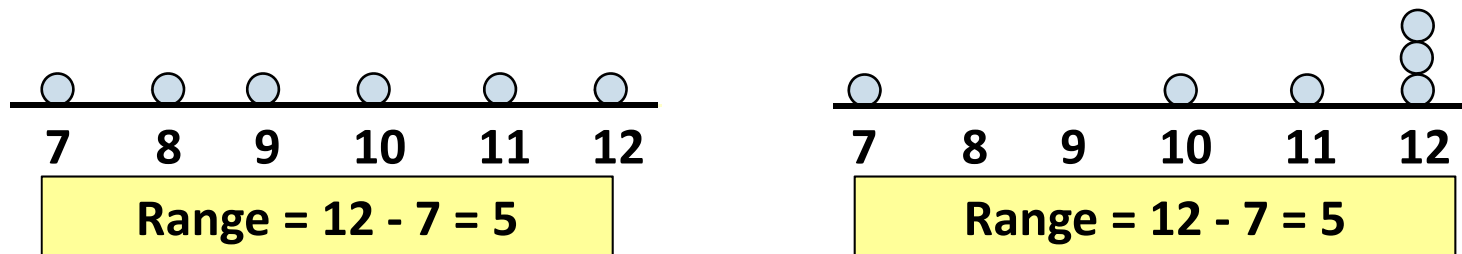
Measure	Definition	How common?	Existence	Takes every value into account?	Affected by outliers?	Advantages
Mean	$\frac{\text{sum of all values}}{\text{total number of values}}$	most familiar "average"	always exists	yes	yes	commonly understood; works well with many statistical methods
Median	middle value	common	always exists	no (aside from counting the total number of values)	no	when there are outliers, may be more representative of an "average" than the mean
Mode	most frequent value	sometimes used	may be no mode, one mode, or more than one mode	no	no	most appropriate for qualitative data (see Section 2.1)

## Part 2c

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- **Descriptive Statistics**
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - **Measures of Variation**
  - Distribution Shape
  - Use of Excel in Organizing Categorical Data
  - Use of Excel in Calculating Summary Statistics for Numeric Data

# Range

- Simplest measure of variation
- Difference between the largest and the smallest values  
$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$
- Ignores the way in which data are distributed



- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

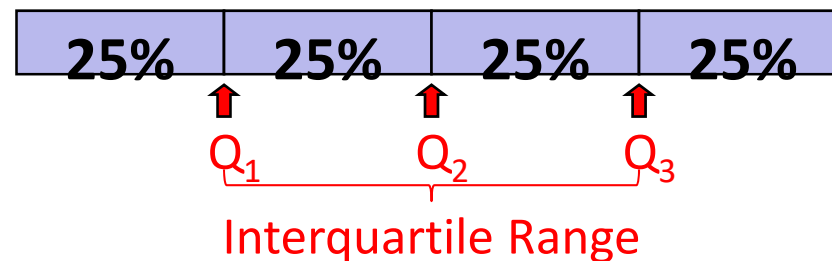
$$\text{Range} = 5 - 1 = 4$$

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

$$\text{Range} = 120 - 1 = 119$$

# Quartiles and Interquartile Range

- Quartiles split the **ranked** data into 4 segments with 25% of the data values in each segment



- The **first quartile**,  $Q_1$ , is the value for which 25% of the observations are smaller and 75% are larger than it
- $Q_2$  is the same as the **median** (50% of the observations are smaller and 50% are larger than it)
- Only 25% of the observations are greater than the **third quartile**,  $Q_3$
- **Interquartile range** is  $Q_3 - Q_1$  and measures the spread in the middle 50% of the data
  - Interquartile range is also called the **midsread** because it covers the middle 50% of the data
- The quartiles and interquartile range are not influenced by outliers or extreme values

# Quartiles

$Q_1$  position:  $\frac{n+1}{4}$  ranked value

$Q_2$  position:  $\frac{2(n+1)}{4}$  ranked value

$Q_3$  position:  $\frac{3(n+1)}{4}$  ranked value

where  $n$  is the number of observed values

When calculating the ranked position, use the following rules:

- If the result is a **whole number**, it is the ranked position to use
- If the result is a **fractional half** (e.g. 2.5, 8.5, ...), average the two corresponding data values
- If the result is **not a whole number or a fractional half**, round the result to the nearest integer to find the ranked position



# Example

Rank: 1 2 3 4 5 6 7 8 9

Sample data in an ordered array (n=9): 11 12 13 16 16 17 18 21 22



$Q_1$  is in the  $(9+1)/4 = 2.5$  position of the ranked data,  
so  $Q_1 = (12+13)/2 = 12.5$

$Q_2$  is in the  $(9+1)/2 = 5^{\text{th}}$  position of the ranked data,  
so  $Q_2 = \text{median} = 16$

$Q_3$  is in the  $3(9+1)/4 = 7.5$  position of the ranked data,  
so  $Q_3 = (18+21)/2 = 19.5$

**Interquartile range =  $19.5 - 12.5 = 7$**

# Variance and Standard Deviation

- Variance is an average of squared deviation of values from the mean

- Population Variance  
(parameter)

pronounced  
sigma squared

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

- Sample Variance  
(statistic)

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

- Standard deviation is the square-root of variance and has the same units as the original data

- Population Standard Deviation  
(parameter)

pronounced  
sigma

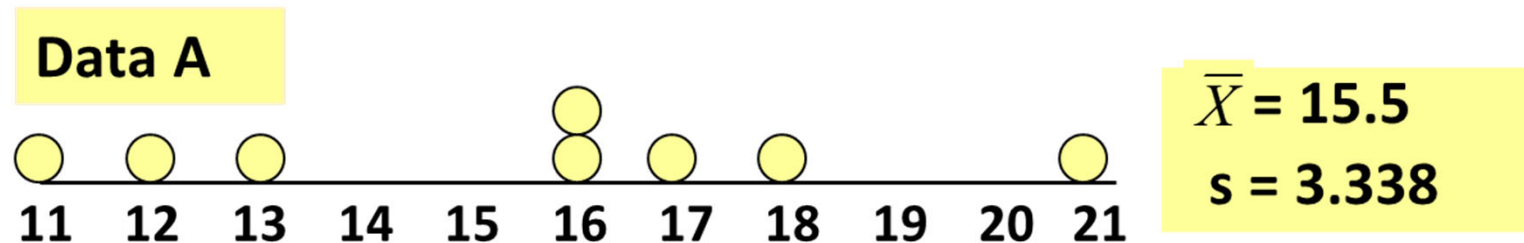
$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

- Sample Standard Deviation  
(statistic)

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

## Example for Calculating Sample Standard Deviation

The data set below is a random sample taken from the population. Calculate the mean, variance and standard deviation.



$$n = 8$$

$$\bar{x} = \frac{(11+12+13+16+16+17+18+21)}{8} = 15.5$$

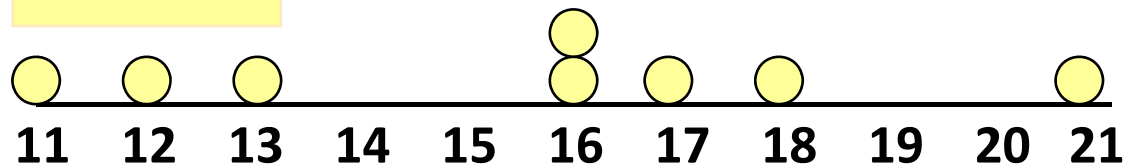
$$s^2 = \frac{(11-15.5)^2 + (12-15.5)^2 + \cdots + (21-15.5)^2}{8-1} = \frac{78}{7} = 11.143$$

$$s = \sqrt{11.143} = 3.338$$

# Comparing Standard Deviations

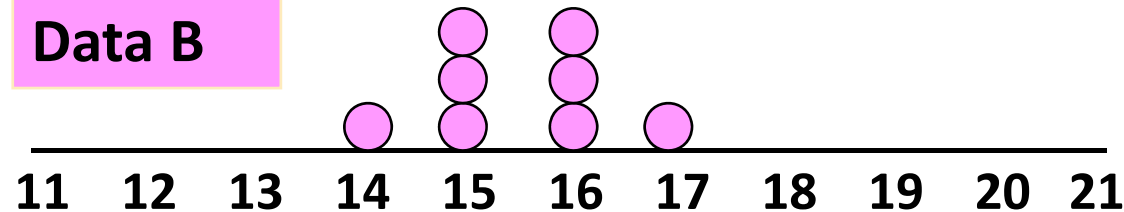
Note: All the data set are random samples from the population

**Data A**



$$\bar{X} = 15.5$$
$$s = 3.338$$

**Data B**



$$\bar{X} = 15.5$$
$$s = 0.926$$

**Data C**

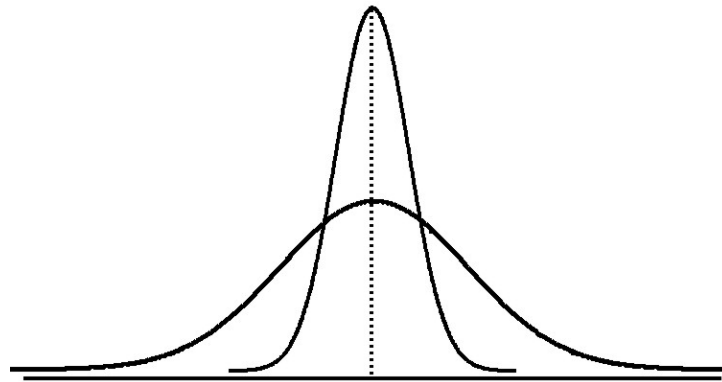


$$\bar{X} = 15.5$$
$$s = 4.570$$

# Summary about Measures of Variation

- The more the data are **spread out**, the **greater** the range, variance, and standard deviation
- The more the data are **concentrated**, the **smaller** the range, variance, and standard deviation
- If the values are all the same (no variation), all these measures will be **zero**
- None of these measures are **ever negative**

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$



# Important Population Parameters and Sample Statistics

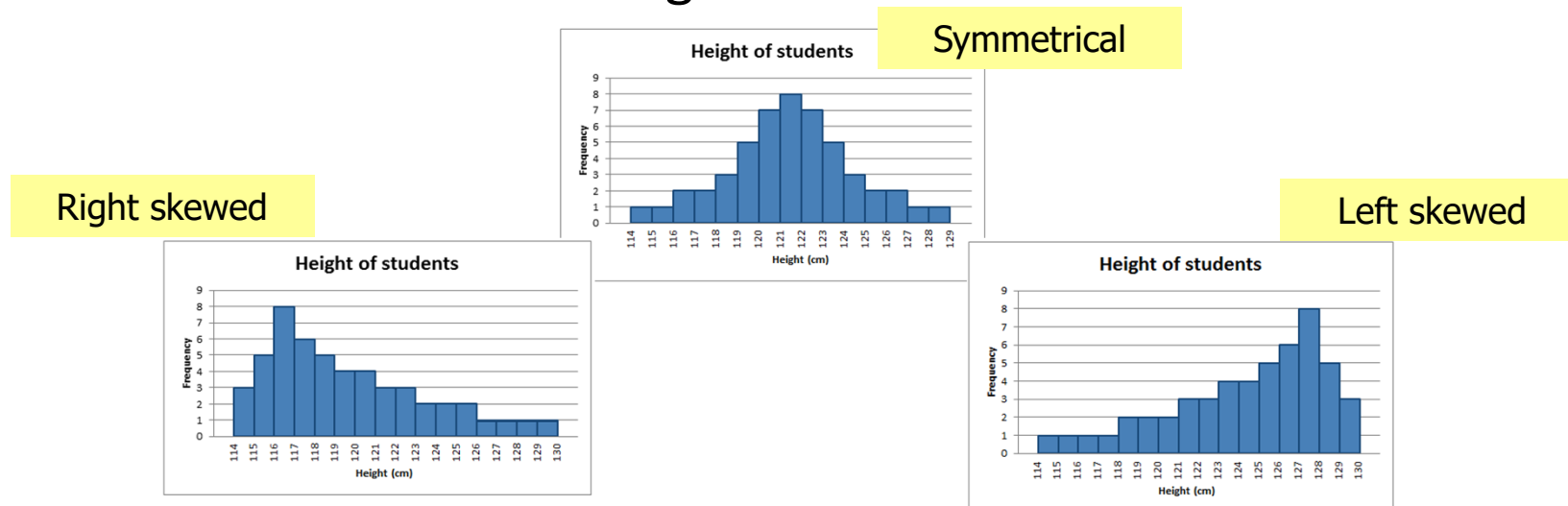
Measure	Population parameter	Sample statistic
Mean	$\mu$	$\bar{X}$
Variance	$\sigma^2$	$s^2$
Standard deviation	$\sigma$	$s$

## Part 2d

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- **Descriptive Statistics**
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - Measures of Variation
  - **Distribution Shape**
  - Use of Excel in Organizing Categorical Data
  - Use of Excel in Calculating Summary Statistics for Numeric Data

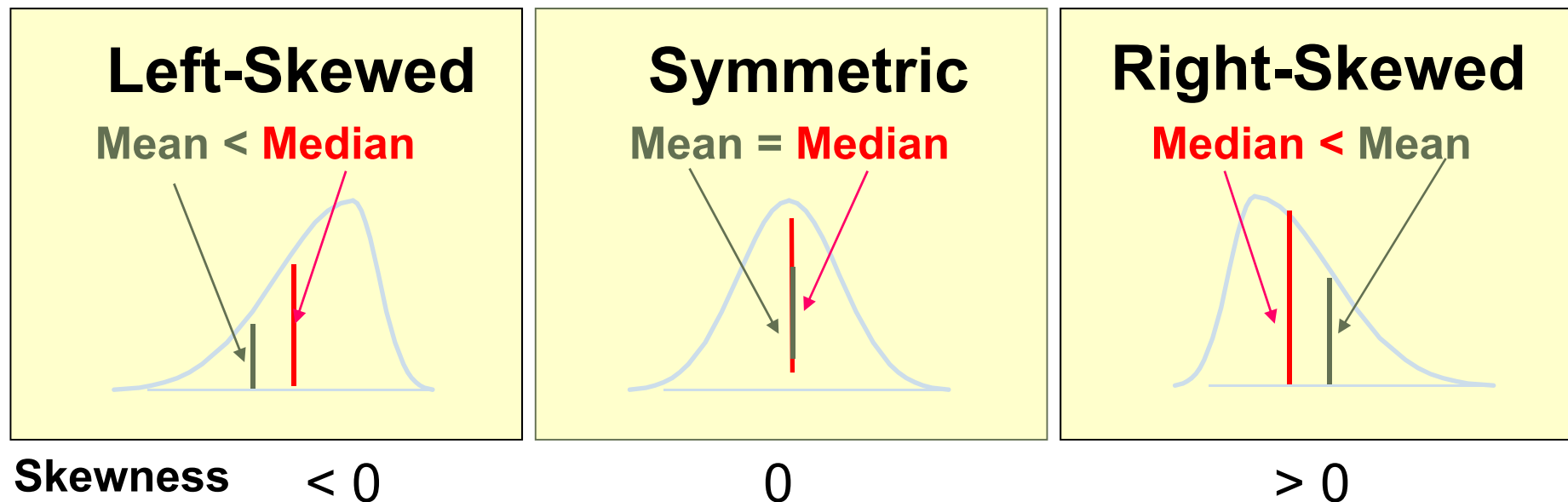
# Distribution Shape -- Skewness

- Data sets may have similar central tendency measures, similar standard deviations, but different distribution shape
- Skewness measures the extent to which data values are not symmetrical
- Skewness equals to 0 if the distribution of the variable is symmetrical
- Skewness is negative if the distribution is left skewed, positive if the distribution is right skewed





# Distribution Shape -- Skewness



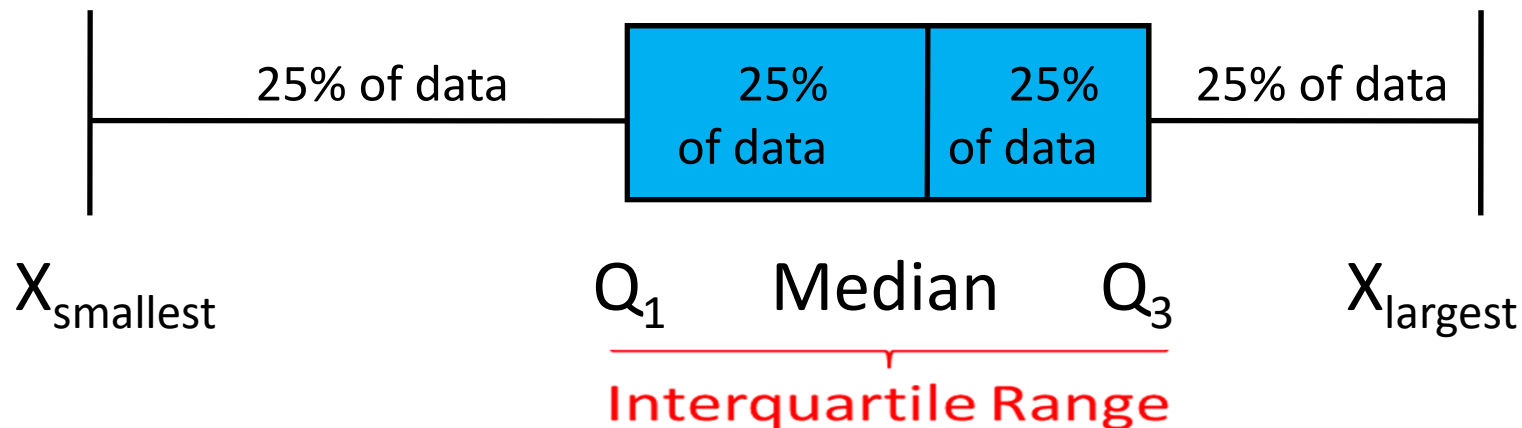
- If data are skewed, the median may be a more appropriate measure of central tendency

# The Five Number Summary and Boxplot

- The five numbers that help describe the center, spread and shape of data are

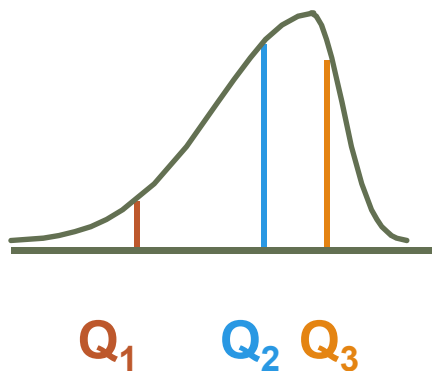
$X_{\text{smallest}}$  --  $Q_1$  -- Median --  $Q_3$  --  $X_{\text{largest}}$

- Boxplot

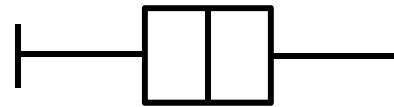
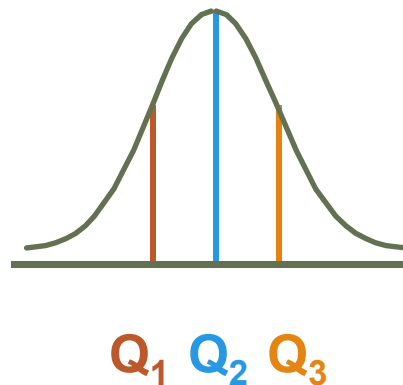


# Relationship between Skewness and Boxplot

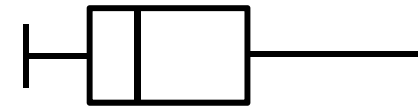
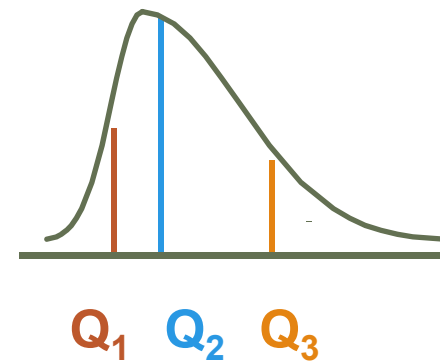
Left-Skewed



Symmetric



Right-Skewed



# Boxplot Example

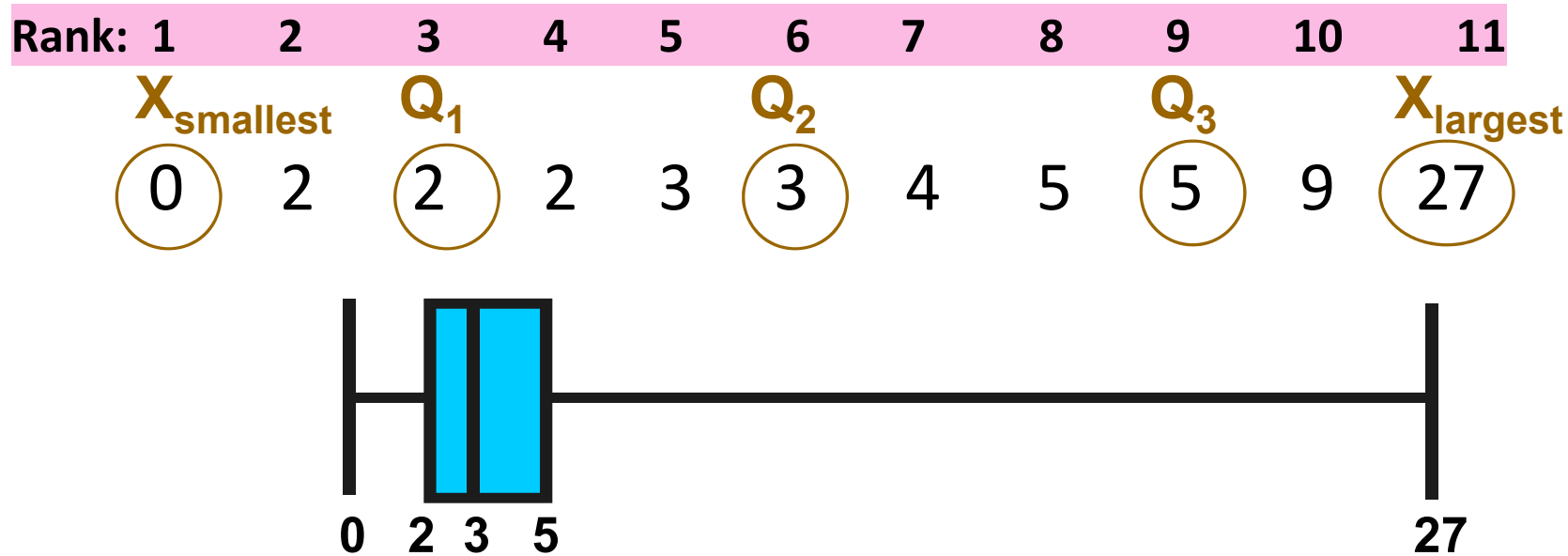
$n=11$

$$Q_1 = (11+1)/4 = 3^{\text{th}} = 2$$

$$Q_2 = (11+1)/2 = 6^{\text{th}} = 3$$

$$Q_3 = 3(11+1)/4 = 9^{\text{th}} = 5$$

$$\text{Interquartile range} = Q_3 - Q_1 = 5 - 2 = 3$$



The data are right skewed

## Part 2e

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- **Descriptive Statistics**
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - Measures of Variation
  - Distribution Shape
  - **Use of Excel in Organizing Categorical Data**
  - Use of Excel in Calculating Summary Statistics for Numeric Data

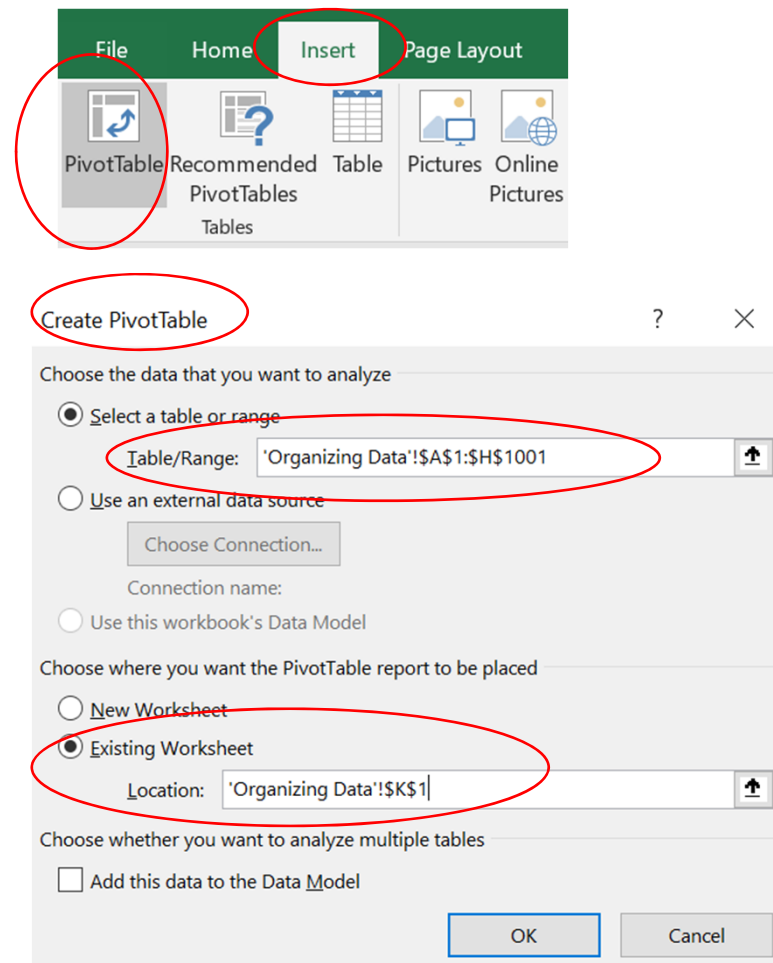
# Use of Excel in Organizing Categorical Data

## ■ PivotTable

- ❑ PivotTable can be used to create summary table for categorical data

- ❑ Steps to create a pivot table

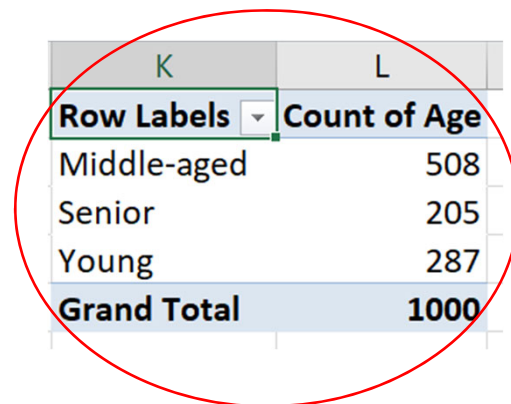
1. In the ribbon, click “Insert” and then “Pivot Table”. In “Create PivotTable” dialog box, click the box in “Table/Range” and select the data that you want to analyze (the cell address is \$A\$1:\$H\$1001). Choose “Existing worksheet” and specify a location (say, K1) for the PivotTable report to be placed



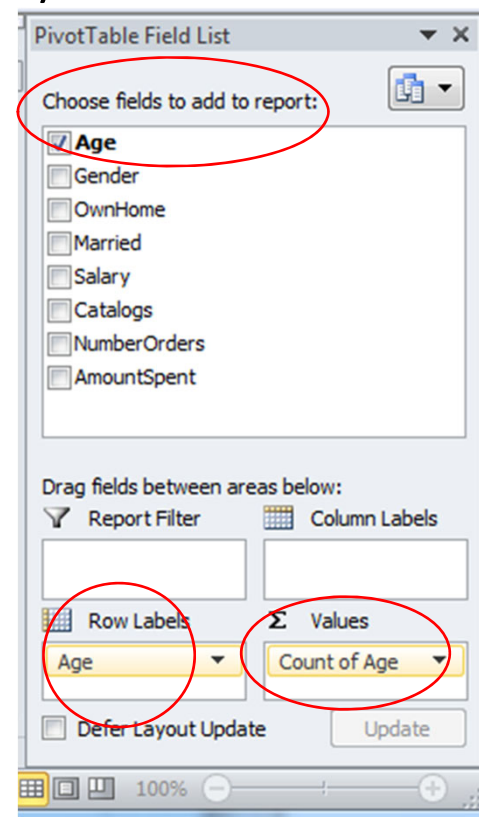
# Use of Excel in Organizing Categorical Data

2. Click the box next to Age (Age will appear in the “Row” area). Drag Age to “Values” area. A summary table appears in K1 box. The default summary value for categorical data is “Count”. This can be changed by clicking “value field settings” in the dropdown list of “Values” area

- ❑ Grand Total is included in the table by default



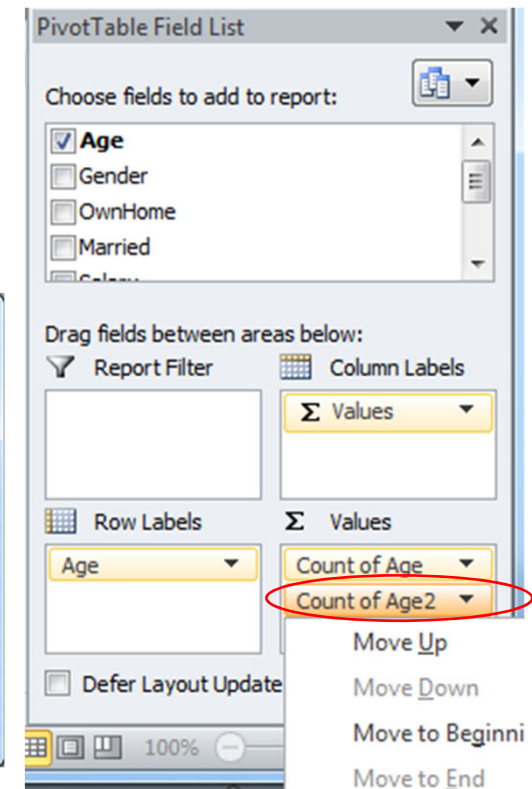
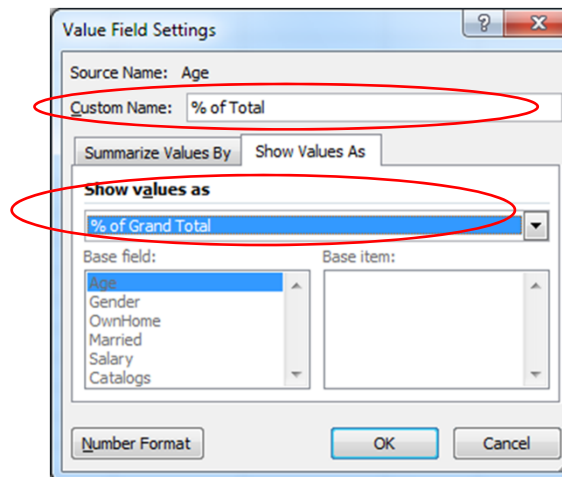
K	L
Row Labels	Count of Age
Middle-aged	508
Senior	205
Young	287
<b>Grand Total</b>	<b>1000</b>



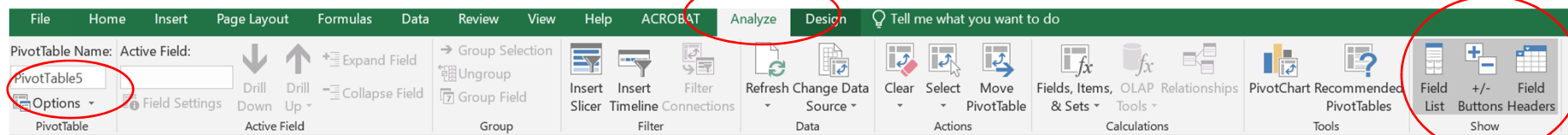
# Use of Excel in Organizing Categorical Data

3. If the relative frequency is also wanted, drag Age into the “Values” area again. Click the dropdown list next to “Count of Age 2”, select “Value Field Settings”, then select “Show Value As % of Grand Total”. Enter “% of Total” into the Custom Name box

	K	L	M
Row Labels	Count of Age	% of Total	
Middle-aged	508	50.80%	
Senior	205	20.50%	
Young	287	28.70%	
Grand Total	1000	100.00%	



- If you do not see the “Field List”, click “FieldList” in “Show” group ( which is a “PivotTable” tool in the “Analyze” menu bar)





## Part 2f

- Introduction to Statistics
  - Introduction
  - Application of Statistics
- **Descriptive Statistics**
  - Organizing and Visualizing Data
  - Measures of Central Tendency
  - Measures of Variation
  - Distribution Shape
  - Use of Excel in Organizing Categorical Data
  - **Use of Excel in Calculating Summary Statistics for Numeric Data**

# Calculating Summary Statistics for Numeric Data in Excel

- The preparation time for the examination of 12 randomly selected students (in days):

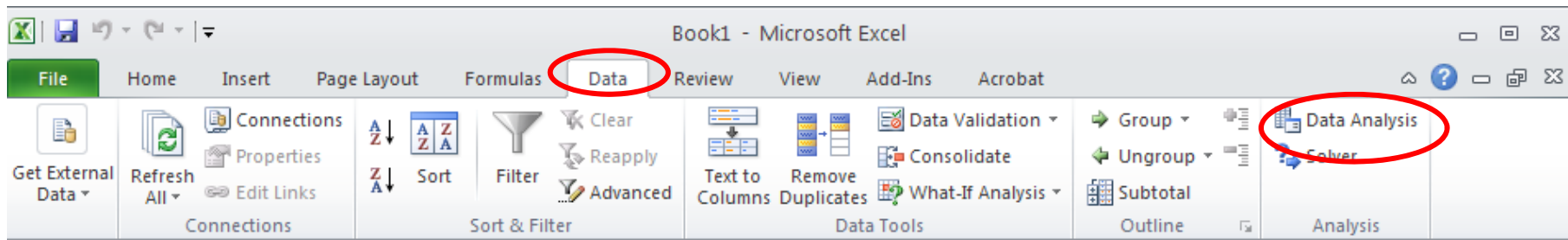
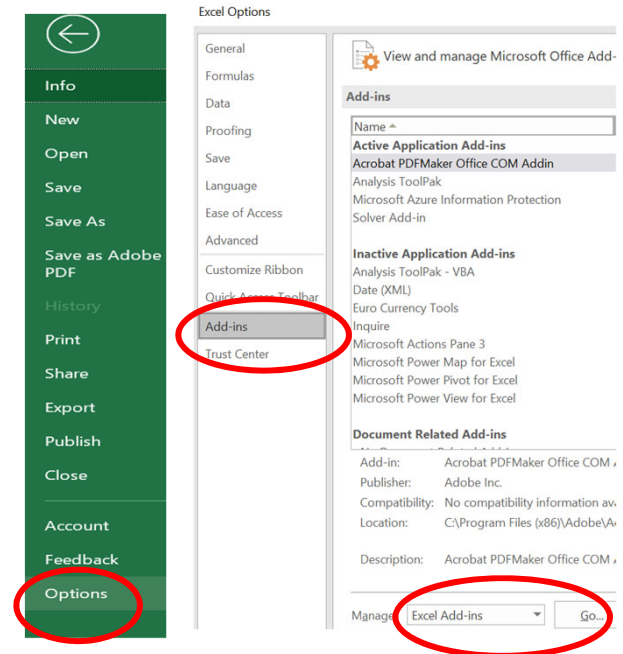
**5 21 18 9 4 17 11 28 19 2 18 22**

D5		fx =STDEV(A1:A12)		
	A	B	C	D
1	5			
2	21		Mean	14.5
3	18		Median	17.5
4	9		Mode	18
5	4		Sample Standard Deviation	8.151966
6	17		Sample Variance	66.45455
7	11		Minimum	2
8	28		Maximum	28
9	19		Range	26
10	2		Sum	174
11	18		Count	12
12	22			

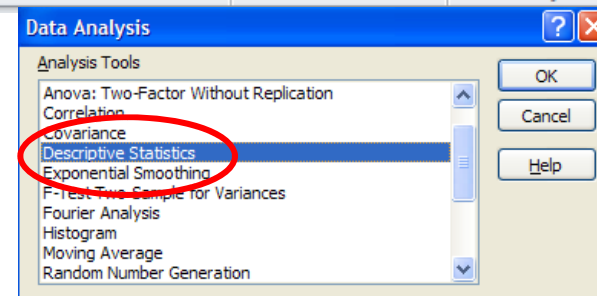
Mean	=average(A1:A12)
Median	=median(A1:A12)
Mode	=mode(A1:A12)
Sample Standard Deviation	=stdev.s(A1:A12)
Sample Variance	=var.s(A1:A12)
Population Standard Deviation	=stdev.p(A1:A12)
Population Variance	=var.p(A1:A12)
Maximum	=max(A1:A12)
Minimum	=min(A1:A12)
Range	=max(A1:A12)-min(A1:A12)
Sum	=sum(A1:A12)
Count	=count(A1:A12)

# Calculating Summary Statistics for Numeric Data in Excel

- Use Excel “Data Analysis” Add-Ins tool to find descriptive measures
  - ❑ File → Options → Add-Ins → Click “Go” at the bottom → Check “Analysis ToolPak” and click “OK”
  - ❑ You can find “Data Analysis” in the “Data” menu bar



- ❑ Choose “Descriptive Statistics” in “Data Analysis” dialog box



# Calculating Summary Statistics for Numeric Data in Excel

- Use Excel “Data Analysis” Add-Ins tool to find Summary statistics

The screenshot shows the 'Descriptive Statistics' dialog box in Excel. The 'Input Range' is set to '\$A\$1:\$A\$12', which is circled in blue with a red arrow pointing to the text 'Data Cells'. The 'Grouped By' option is 'Columns'. The 'Output options' section has 'Output Range' set to '\$F\$1' (circled in blue with a red arrow pointing to 'Output Cell'), 'New Worksheet Ply' selected, and 'Summary statistics' checked (circled in blue with a red arrow pointing to 'Generate Descriptive Statistics'). Other options like 'Confidence Level for Mean' (95%), 'Kth Largest', and 'Kth Smallest' are visible but not selected.

Column1	
Mean	14.5
Standard Error	2.35327
Median	17.5
Mode	18
Standard Deviation	8.151966
Sample Variance	66.45455
Kurtosis	-1.01139
Skewness	-0.16674
Range	26
Minimum	2
Maximum	28
Sum	174
Count	12