

# **SC2000/CZ2100/CE2100**

## **Probability & Statistics**

### **Week 1**

# Course Logistics

**Part 1:** Asst Prof. Themis Gouleakis

Email: [themis.gouleakis@ntu.edu.sg](mailto:themis.gouleakis@ntu.edu.sg)

**Part 2:** Assoc Prof. Kong Wai Kin Adams

Email: [AdamsKong@ntu.edu.sg](mailto:AdamsKong@ntu.edu.sg)

Room: LT2A-01-01 (LEVEL 1, NEAR MAE)

~ 24 Sessions of lectures

~ 11 Sessions of tutorials

~ 50 min per session

## Accessing the Materials on NTU Learn


▼ 24S2-SC2000-CE2100-  
CZ2100-PROB & STAT FOR  
COMPUTING

Announcements

Information

Content

### Content



**Part 1: Teaching Materials (Lecture and Tutorial)**  
This folder contains the materials for Part 1.  
Instructor: Asst Prof Themis Gouleakis

## Quiz Logistics

Two quizzes for part 1 – 25% each

Quizzes are closed-book and in person, during lecture hours.

You can do them using your own laptop. Labs will also be made available.

Quiz one: **11 February,**

Quiz two: **25 February.**

# What we will talk about in Part 1 (Week 1-7):

Ch1 Introduction to Statistics

Ch2 Presenting Data

Ch3 Summarizing Distributions

Ch4 Bivariate Data

Ch5 Probability Theory

Ch6 Probability Distribution

- Random Variables
- Discrete Distribution
- Continuous Distribution

- TEL supporting Materials for Ch1-4 (short video clips and notes) are available in NTULearn.

Good to watch the video clips before attending the lectures for Part 1 (Ch1-4).

- Lectures (Week 1 to 7)
  - Discussion & worked examples on topics covered in the TEL materials
  - Additional topics not covered in TEL materials (particularly those in Ch 5 and 6)

- Recess week (Self-study non-exam)
  - Use of R programming for analysis and presentation of statistical data (Practice materials will be available in NTULearn)
- Tutorial – One session per week, to begin in Week 3

All Course Materials are available in [NTULearn](#).

Part Two starts from week 8.

# Ch 1. Introduction to Statistics

- Descriptive & Inferential Statistics
- Types of Variables
- Percentiles
- Types of Measurement Scale
- Distributions
- Linear Transformations



Statistics involve gathering, organizing, analyzing, interpreting and presenting **data**.

Statistics are being used everywhere:

- Number of students enrolled in this course
- Index measuring stock market
- Singapore household income
- Live data of new Covid-19 cases
- Opinion poll, benchmarks poll, tracking polls, etc.

Statistics are obviously important.

- Predicting the spreading of diseases
- Weather forecasting based on statistics
- Provide informed choice on investment decision
- AI or machine learning based on past statistical data

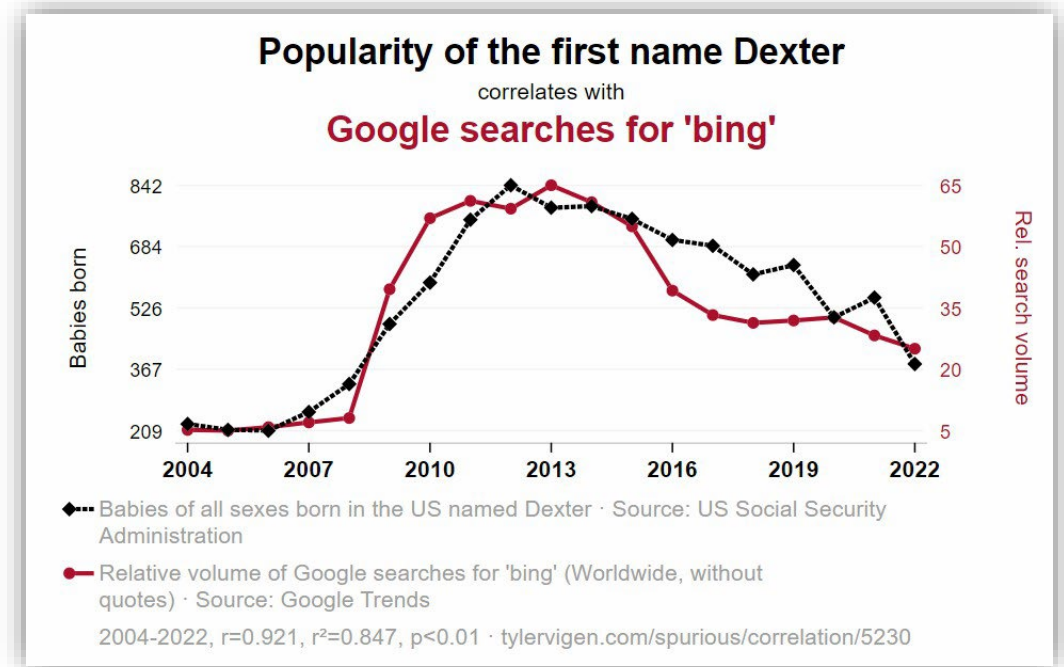
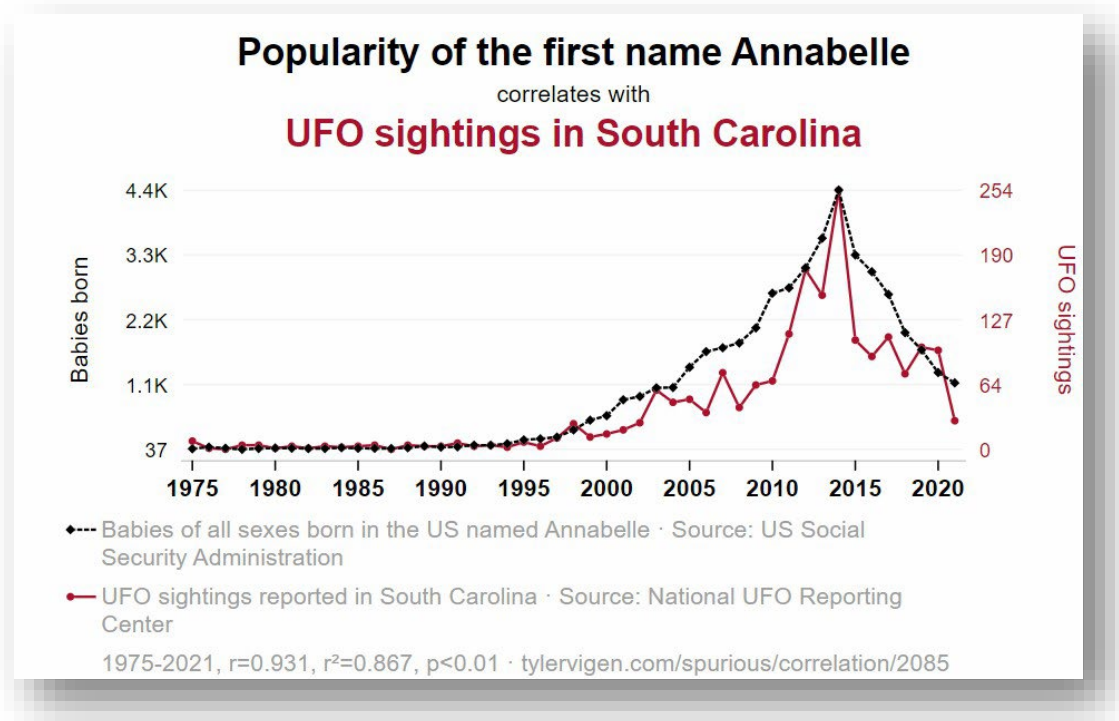
But statistics can be **misleading**.

Example: A toothpaste manufacturer claims that more than 80% of Dentists recommend a particular brand of toothpaste. This was based on surveys of dentists which allow selection of one or more brands.

Why?

Because it may be understood that 80% of dentists recommend this brand over the others. It should be noted that other brands were also recommended and may be as much as that particular brand.

# More examples: **spurious correlations**



# Ice Cream Sales and Shark Attacks

- **Observation:** There is a positive correlation between ice cream sales and shark attacks.
- **Explanation:** These two variables are not causally linked. Instead, a **third variable**, hot weather, drives both. During warmer months:
  - More people buy ice cream.
  - More people swim in the ocean, increasing the likelihood of shark encounters.This is a classic example of a spurious correlation, where the observed relationship is due to a shared underlying factor rather than a direct connection between the two variables.

## ■ Descriptive & Inferential Statistics

**Descriptive** statistics – summarize and describe important features of the **data collected**. Does not generalize beyond the data collected.

**Inferential** statistics - collection of sample to draw inferences about the **population**, i.e. formal guesses of statistical parameters about the population by looking at the samples.

A teacher wishes to know whether the males in his class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared.

Is this an example of descriptive or inferential statistics?

A cognitive psychologist is interested in comparing two ways of presenting stimuli on subsequent memory. Twelve subjects are presented with each method and a memory test is given.

What would be the roles of descriptive and inferential statistics in the analysis of these data?

**Descriptive** statistics – we describe and analyze the data from the **sample**.

**Inferential** statistics – we use the data from the sample to generalize to a larger **population** of people.



## ■ Types of Variables

In statistic, we can broadly group variables into two categories: **Qualitative** and **Quantitative**.

Examples:

Qualitative: categorical variables (e.g., gender, marital status, province).

Quantitative: numerical values (e.g., age, height). Can be discrete or continuous.

## ■ Percentile

In certain experiments, it is more meaningful to compare the outcomes obtained.

Eg: if you know that your quiz marks is 80 out of 100, you may not know how well you have done compared to others in your class.

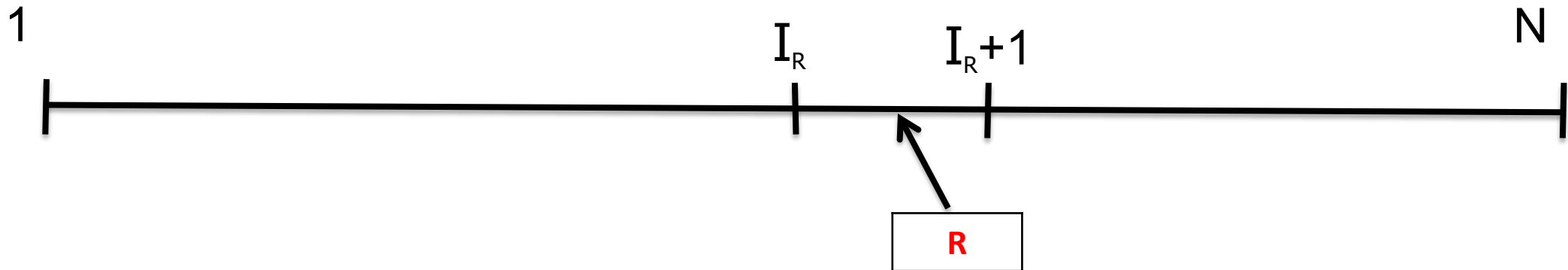
A **percentile** is a comparison score between a particular score and the scores of the rest of a group.

## ■ Percentile - calculation

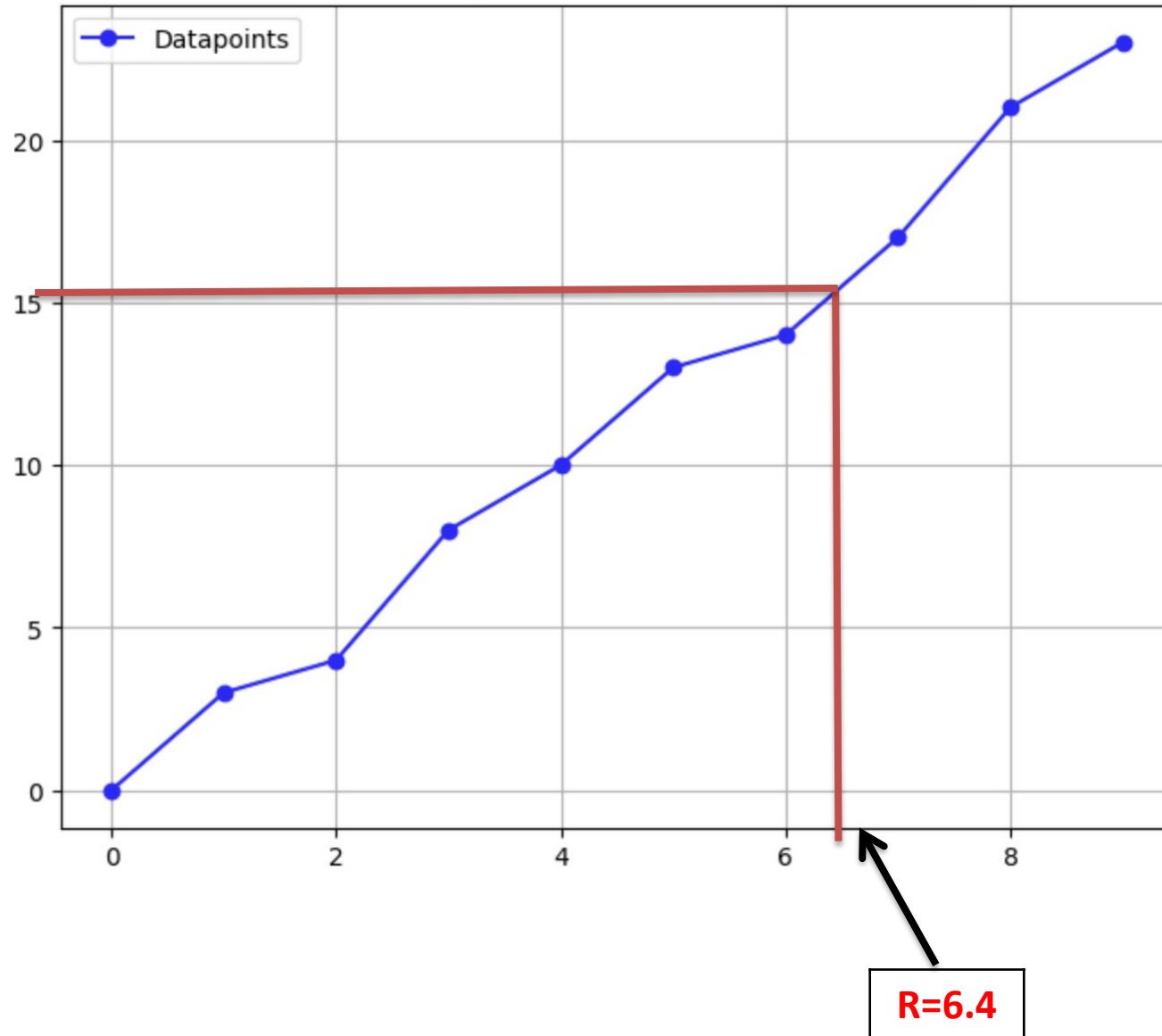
Calculation of  $P^{\text{th}}$  Percentile for a set of  $N$  data:

← Data arranged in the order of magnitude

1. Compute the rank  $R = (P/100) \times (N - 1) + 1$
2. Let  $I_R$  = Integer part of  $R$  and  $F_R$  = Fractional part of  $R$
3.  $P^{\text{th}}$  Percentile = Data at rank  $I_R$  +  
( Data at rank  $(I_R+1)$  – Data at rank  $I_R$  )  $\times F_R$



## ■ Percentile - calculation



Eg: Given data: [3, 5, 7, 8, 9, 11, 13, 15], compute the 25th and 75th percentile.

For 25<sup>th</sup> percentile-

**Step 0:** Are the numbers in ascending order? Yes.

**Step 1:**  $R = 25/100 \times (8-1)+1 = 2.75$

**Step 2:**  $I_R = 2$   $F_R = 0.75$

**Step 3:** 25<sup>th</sup> percentile =  $5 + (7-5) \times 0.75 = 6.5$

Practice at home: You should get 11.5 for 75<sup>th</sup> percentile.

Calculation of P<sup>th</sup> Percentile for a set of N data:

1. Compute the rank  $R = P/100 \times (N - 1) + 1$
2. Let  $I_R$  = Integer part of R and  $F_R$  = Fractional part of R
3. P<sup>th</sup> Percentile = Data at rank  $I_R$  +  
( Data at rank  $(I_R+1)$  – Data at rank  $I_R$  )  $\times F_R$

Compute the 25th percentile for the data set of 20 values:

<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>5</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>7</b>
<b>7</b>	<b>8</b>	<b>8</b>	<b>8</b>	<b>9</b>	<b>9</b>	<b>9</b>	<b>10</b>	<b>10</b>	<b>10</b>

For 25<sup>th</sup> percentile-

**Step 0:** Are the numbers in ascending order? Yes.

**Step 1:**  $R = 25/100 \times (20-1) + 1 = 5.75$

**Step 2:**  $I_R = 5$   $F_R = 0.75$

**Step 3:** 25<sup>th</sup> percentile =  $5 + (5-5) \times 0.75 = 5$

Calculation of P<sup>th</sup> Percentile for a set of N data:

1. Compute the rank  $R = P/100 \times (N - 1) + 1$
2. Let  $I_R$  = Integer part of R and  $F_R$  = Fractional part of R
3. P<sup>th</sup> Percentile = Data at rank  $I_R$  +  
( Data at rank  $(I_R+1)$  – Data at rank  $I_R$  )  $\times F_R$

Compute the 85th percentile for the data set of 20 values:

<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>5</b>	<b>10</b>	<b>10</b>	<b>10</b>	<b>6</b>
<b>6</b>	<b>9</b>	<b>9</b>	<b>9</b>	<b>7</b>	<b>7</b>	<b>7</b>	<b>8</b>	<b>8</b>	<b>8</b>

**Step 0:** Are the numbers in ascending order?

No! You need to re-arrange them first.

<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>5</b>	<b>6</b>	<b>6</b>	<b>7</b>	<b>7</b>
<b>7</b>	<b>8</b>	<b>8</b>	<b>8</b>	<b>9</b>	<b>9</b>	<b>9</b>	<b>10</b>	<b>10</b>	<b>10</b>

Practice at home: You should get 9.15 for the 85<sup>th</sup> percentile.

## Practical example:

Graduate Management Admissions Test - a standardized test for application to graduate-level business programs.

Report online

**UNOFFICIAL GMAT® SCORE REPORT**

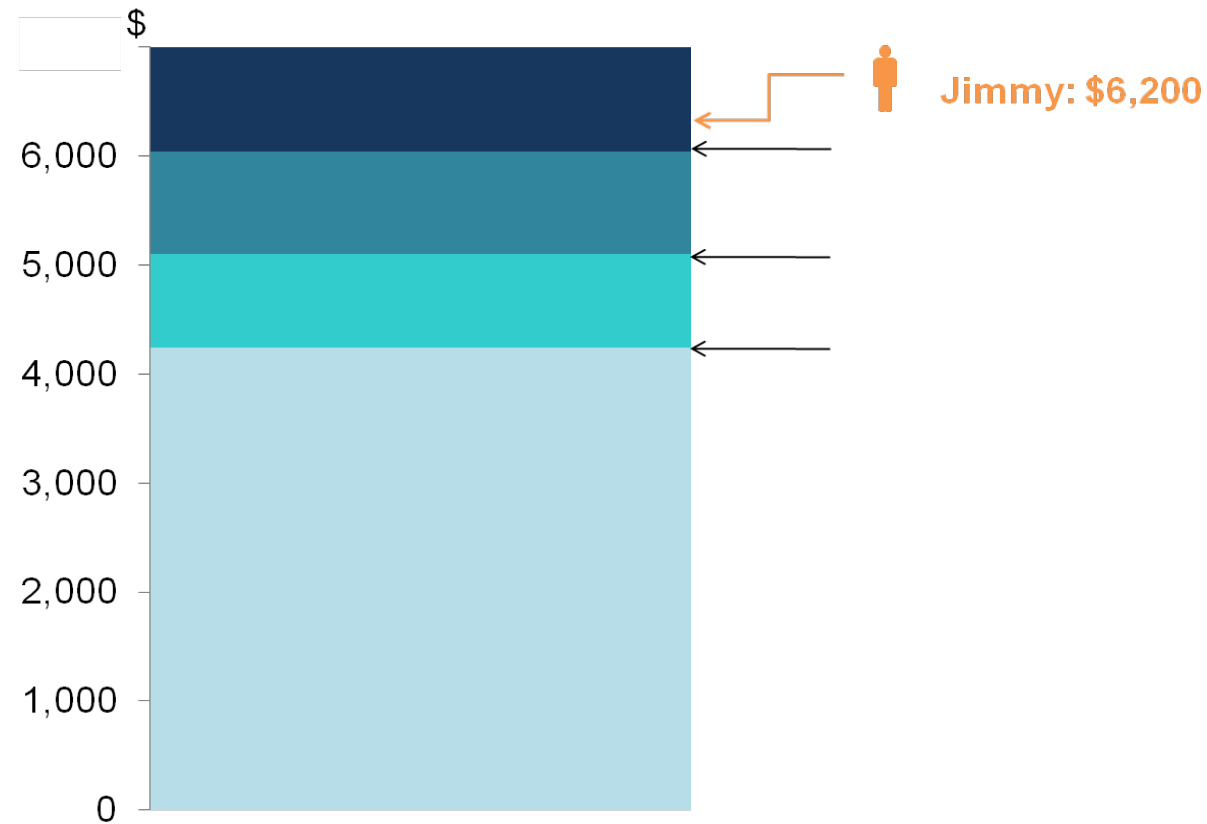
	Scaled Score	
Integrated Reasoning	6	
Quantitative	47	
Verbal	47	
Total*	750	

total score is derived from your Verbal and Quantitative scores.

Official Score Report, including your Analytical Writing Assessment score, writing score and percentile ranking.



Example: Percentile for benchmarking.  
Suppose Jimmy earns a monthly wage of \$6,200. He would like to compare his wages with those working in the same occupation and industry.

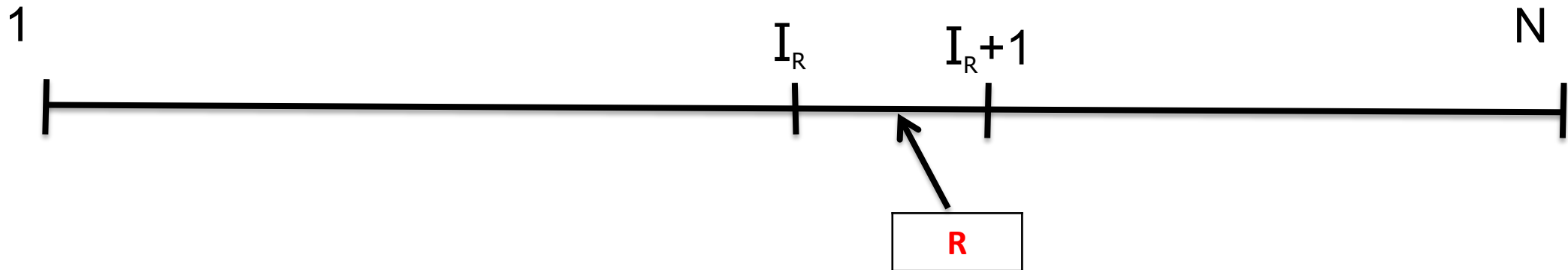


## ■ Percentile – calculation (revisited)

Calculation of  $P^{\text{th}}$  Percentile for a set of  $N$  data:

← Data arranged in the order of magnitude

1. Compute the rank  $R = (P/100) \times (N - 1) + 1$
2. Let  $I_R$  = Integer part of  $R$  and  $F_R$  = Fractional part of  $R$
3.  $P^{\text{th}}$  Percentile = Data at rank  $I_R$  +  
( Data at rank  $(I_R+1)$  – Data at rank  $I_R$  )  $\times F_R$



## A special case: Median (50th percentile)

N is odd

1	2	3	4	5	6	7
3	8	15	23	42	56	78

Median=23

N is even

1	2	3	4	5	6
2	5	11	19	34	51

Median=(11+19)/2=15

# Alternative definitions of Pth-Percentile

- NO universally accepted definition.
- **Example 1:** "smallest number in the dataset that is greater than P% of the elements"  
**Problem:** 50th percentile does not equal the median.
- **Example 2:** "smallest number in the dataset that is greater **or equal** than P% of the elements"  
**Problem:** 50th percentile does not equal the median if the number of elements is **even**.
- **Example 3:** Follow the procedure in slide 19 with rank  $R=(P/100)*(N+1)$

With this definition **median=50th percentile**:

$P^{\text{th}}$  Percentile = Data at rank  $(N+1)/2$  if N is **odd** (= median)

$P^{\text{th}}$  Percentile =  $\frac{1}{2} * [(\text{Data at rank } N/2) + (\text{Data at rank } N/2+1)]$  if N is **even** (= median)

**Problem:** Pth percentiles, where  $P < 100/(N+1)$  or  $P > N/(N+1) * 100$  are **undefined**.

Definition we use in **this course** (see slide 19):  $R=(P/100)*(N-1)+1$

- ✓ - median=50th percentile
- ✓ - Pth percentile defined **for any P** (0th = smallest datapoint, 100th = largest datapoint)
- ✓ - **Default** definition used by R programming language

**SC2000/CZ2100/CE2100**

**Probability & Statistics**

**Ch 1. Introduction to Statistics**

- Descriptive & Inferential Statistics
- Types of Variables
- Percentiles
- Types of Measurement Scale
- Distributions
- Linear Transformations

## ■ Types of Measurement Scales

Four basic levels of measurement scales:

**Nominal** – names or labels with no specific order (e.g description of taste: sweet, sour, bitter etc...)

**Ordinal** – variables in a specific order (e.g military ranks: soldier, lieutenant, general)

**Interval** – numerical scales in which intervals have the same interpretation throughout, but no true zero (e.g temperature in Celcius/Fahrenheit)

**Ratio** – include all the characteristics of interval scale, plus it has zero position indicating the absence of the quantity being measured (e.g number of eggs in a basket)

Specify the level of measurement used for the following variables:

1. Calendar years
2. Colors
3. Amount of money in your pocket
4. Police ranks

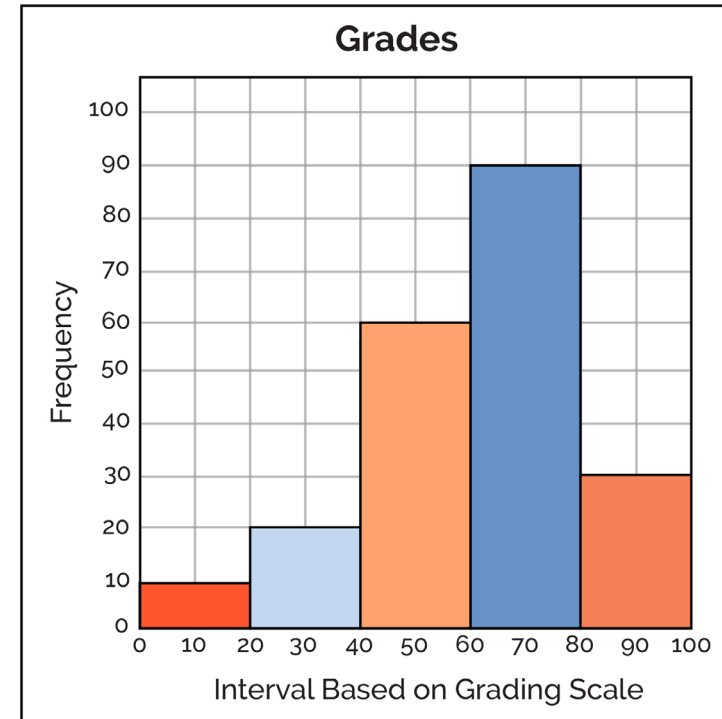
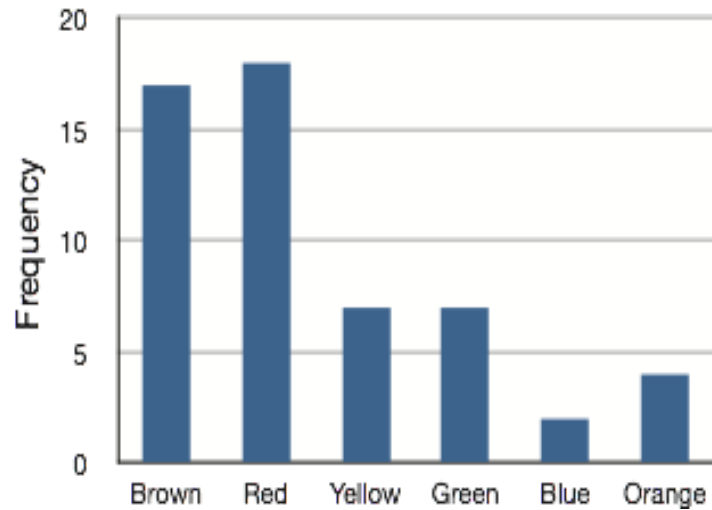


<https://app.wooclap.com/events/AGLILC/questions/6789fca786383bb338ed8fd8>

## ■ Distributions

Frequency distribution

- Discrete variables
- Continuous variables

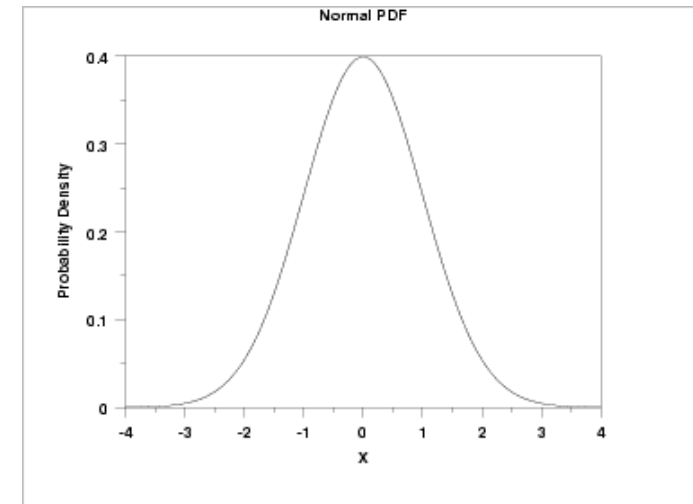
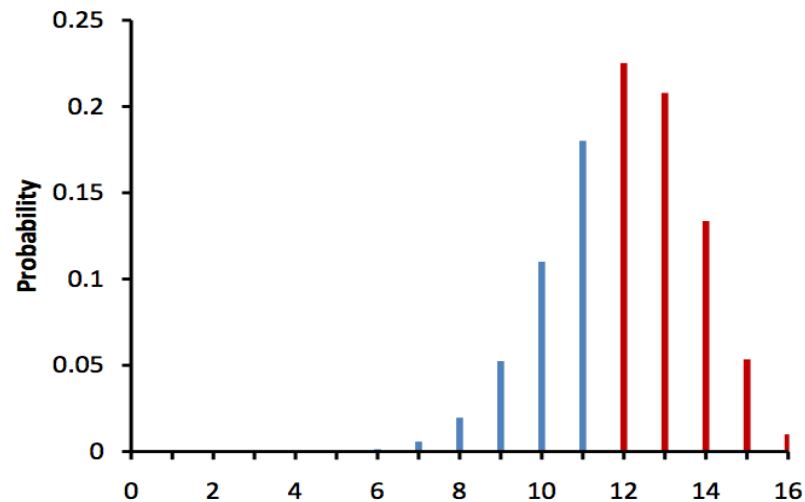




## ■ Distributions

### Probability distribution

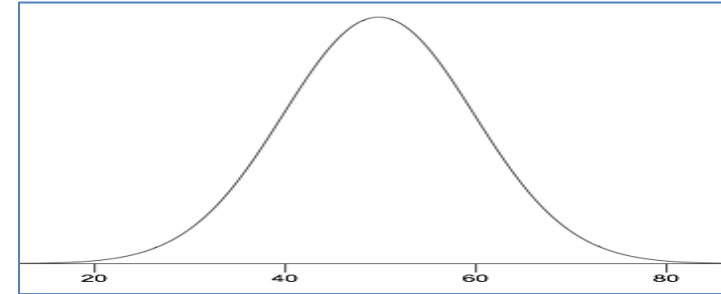
- Probability mass function (pmf) for discrete variables
- Probability density function (pdf) for continuous variables



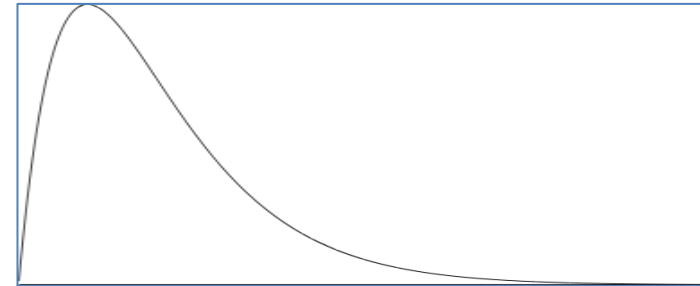
## ■ Distributions

### Shapes of distributions

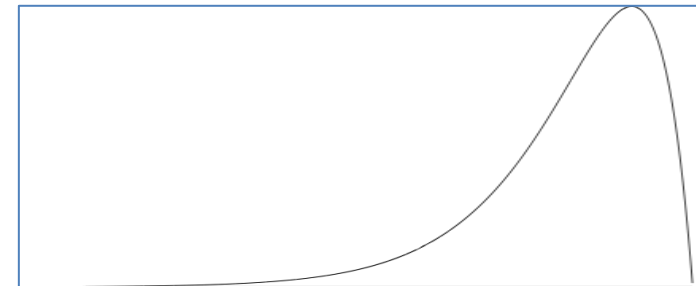
- symmetric
- skewed to the right
- skewed to the left



Symmetric

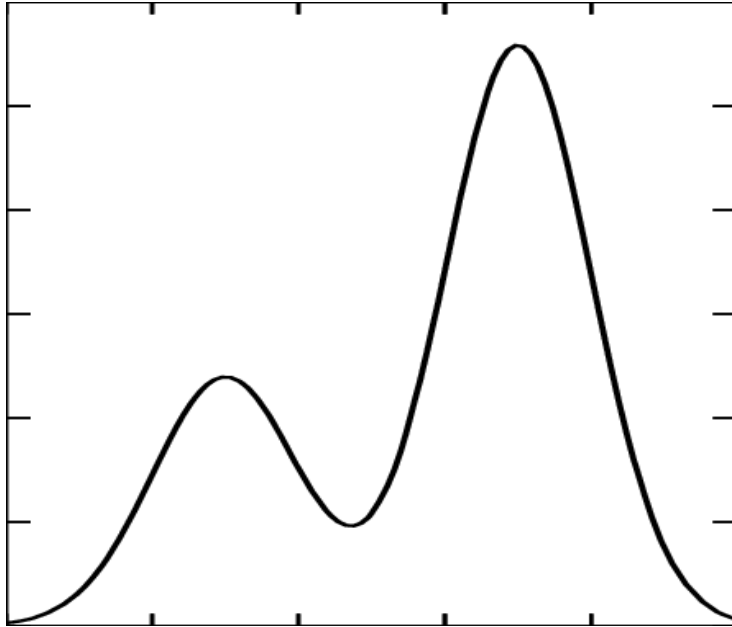


Positive skew (to the right)



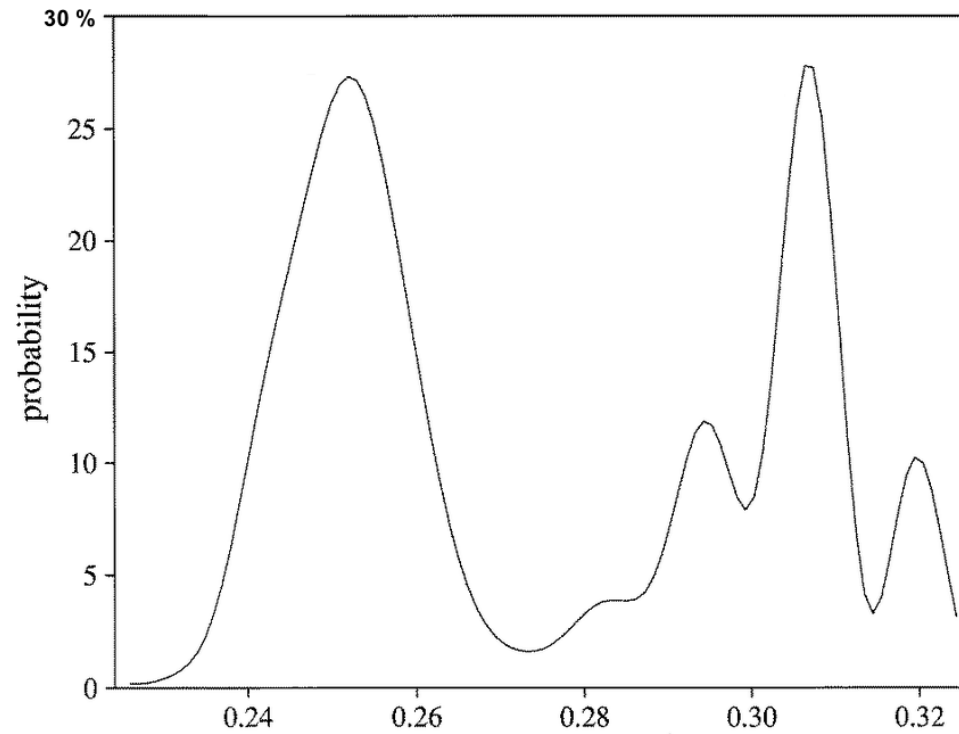
Negative skew (to the left)

- Distributions – other shapes:

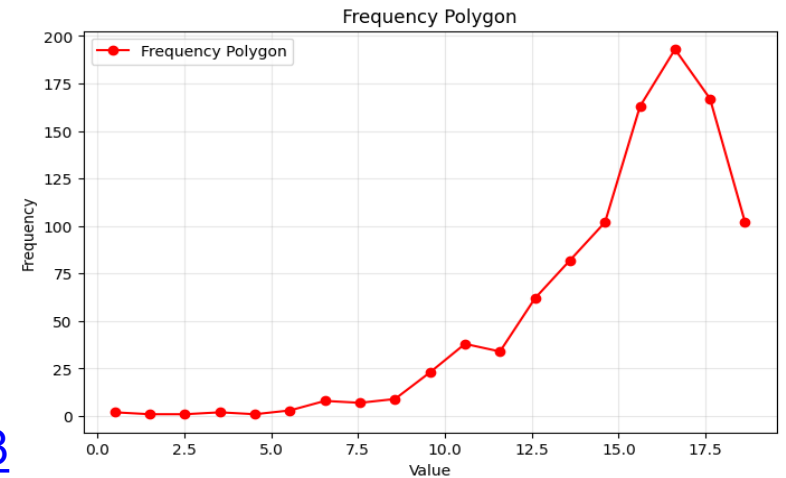
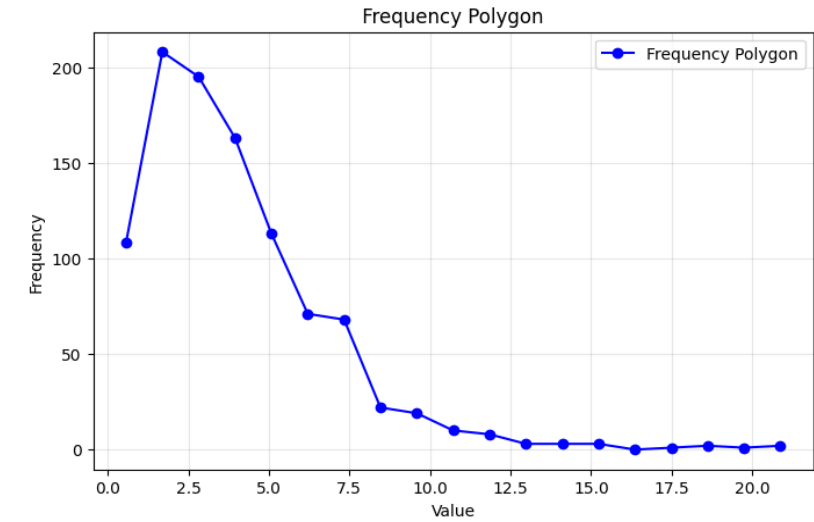
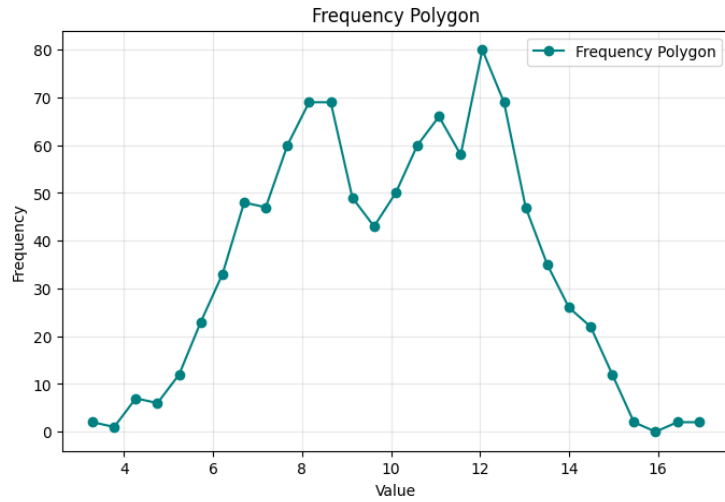


Bimodal distribution

Multi-modal distribution



# Which of the Frequency polygons has a large **positive** skew?



<https://app.wooclap.com/events/AGLILC/questions/6789fca786383bb338ed8fd8>

## ■ Linear Transformations

Transform data from one measurement scale to another.

Examples:

Convert length measured in  $X$  feet to measurement in  $Y$  meters, i.e.

$$Y = 0.3048 X$$

Convert temperature in Fahrenheit to Centigrade:

$$C = 0.5556 F - 17.778$$

- Linear Transformations

Which of the following are linear transformations?



<https://app.wooclap.com/events/AGLILC/questions/6789f629d8e0bc0325748f99>

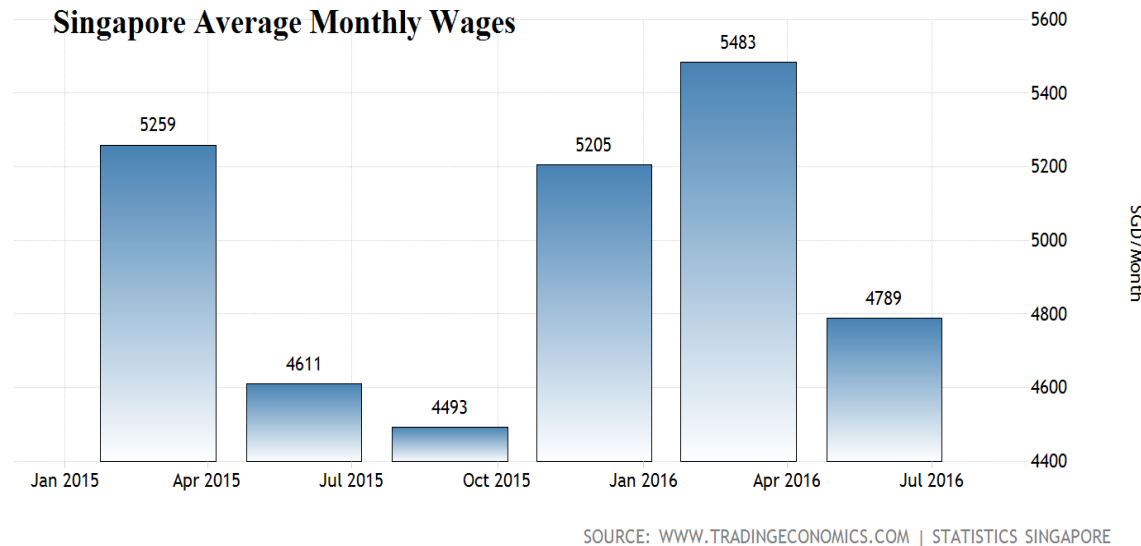
## Ch 2. Presenting Data

- Frequency tables and Charts
- Bar Charts
- Stem and Leaf Displays
- Histograms
- Box Plots
- others

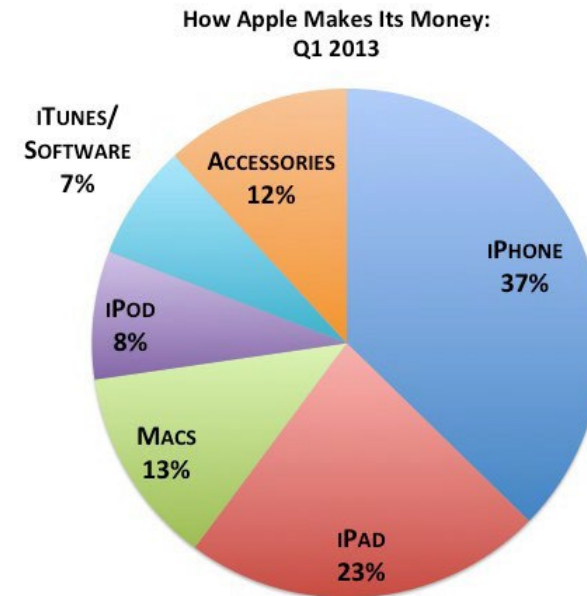
- Presenting **qualitative** or **discrete** data:

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Frequency table



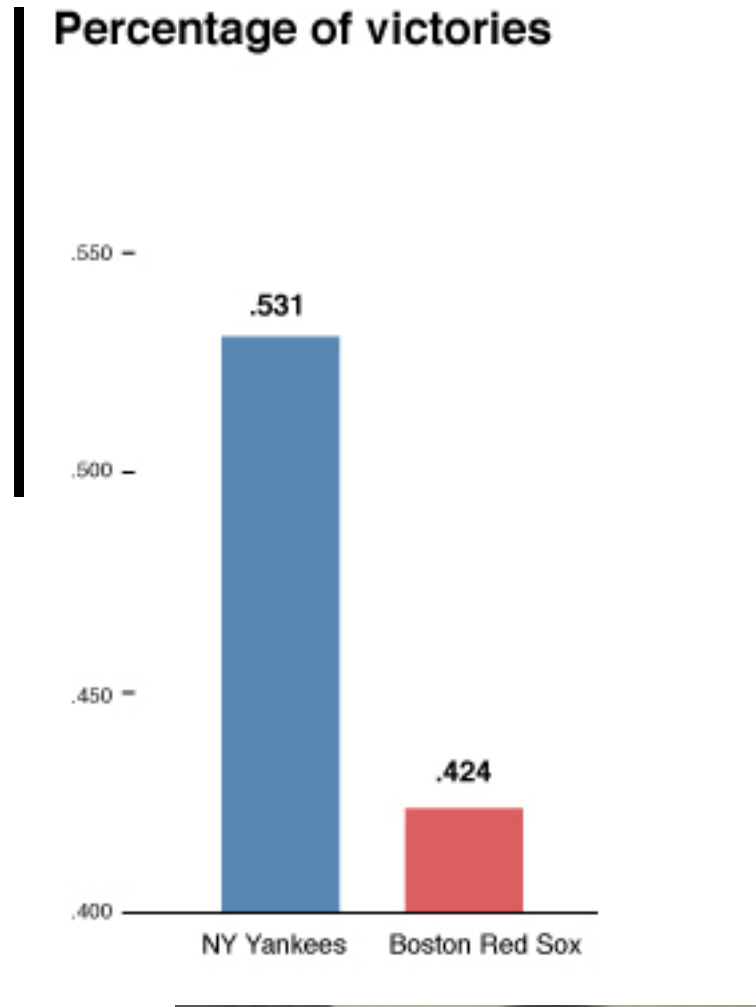
Bar chart



Pie chart



## Examples of misleading presentations:

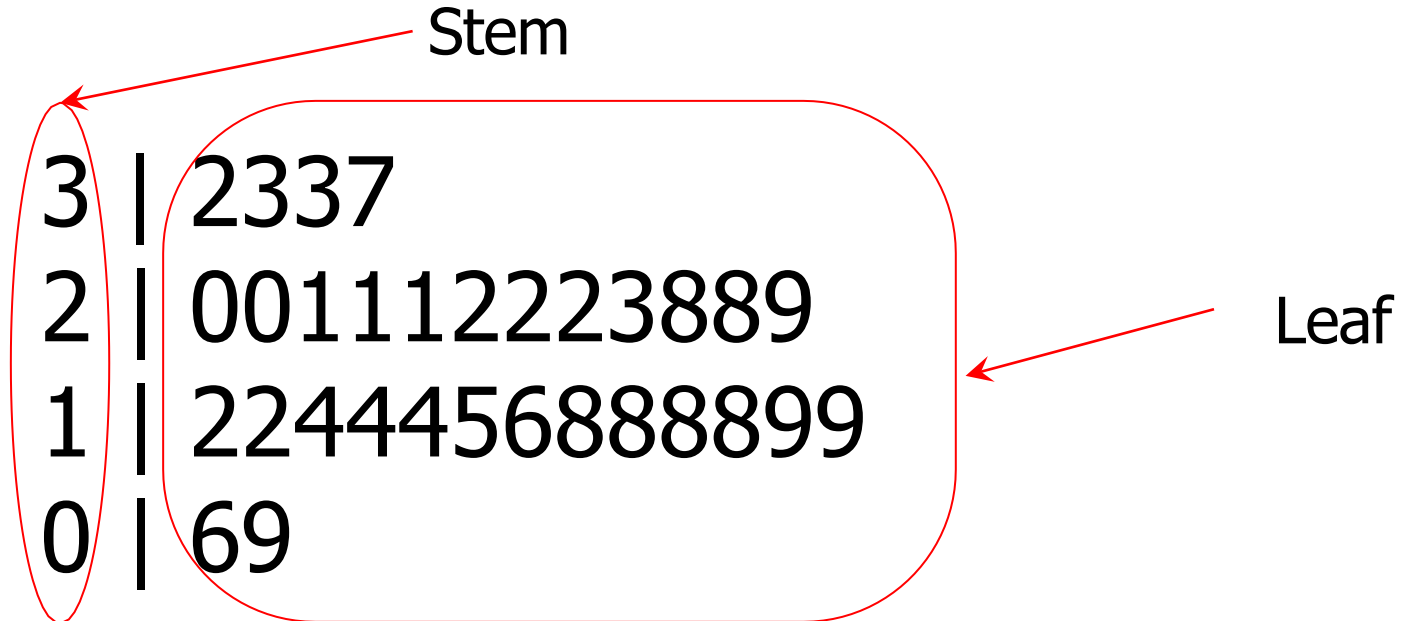


Is there any  
misleading  
presentation?

- Stem-and-leaf Presentation
- Useful when data are not too numerous

Eg: No. of touchdown passes by each of the 31 football teams.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20,  
20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6



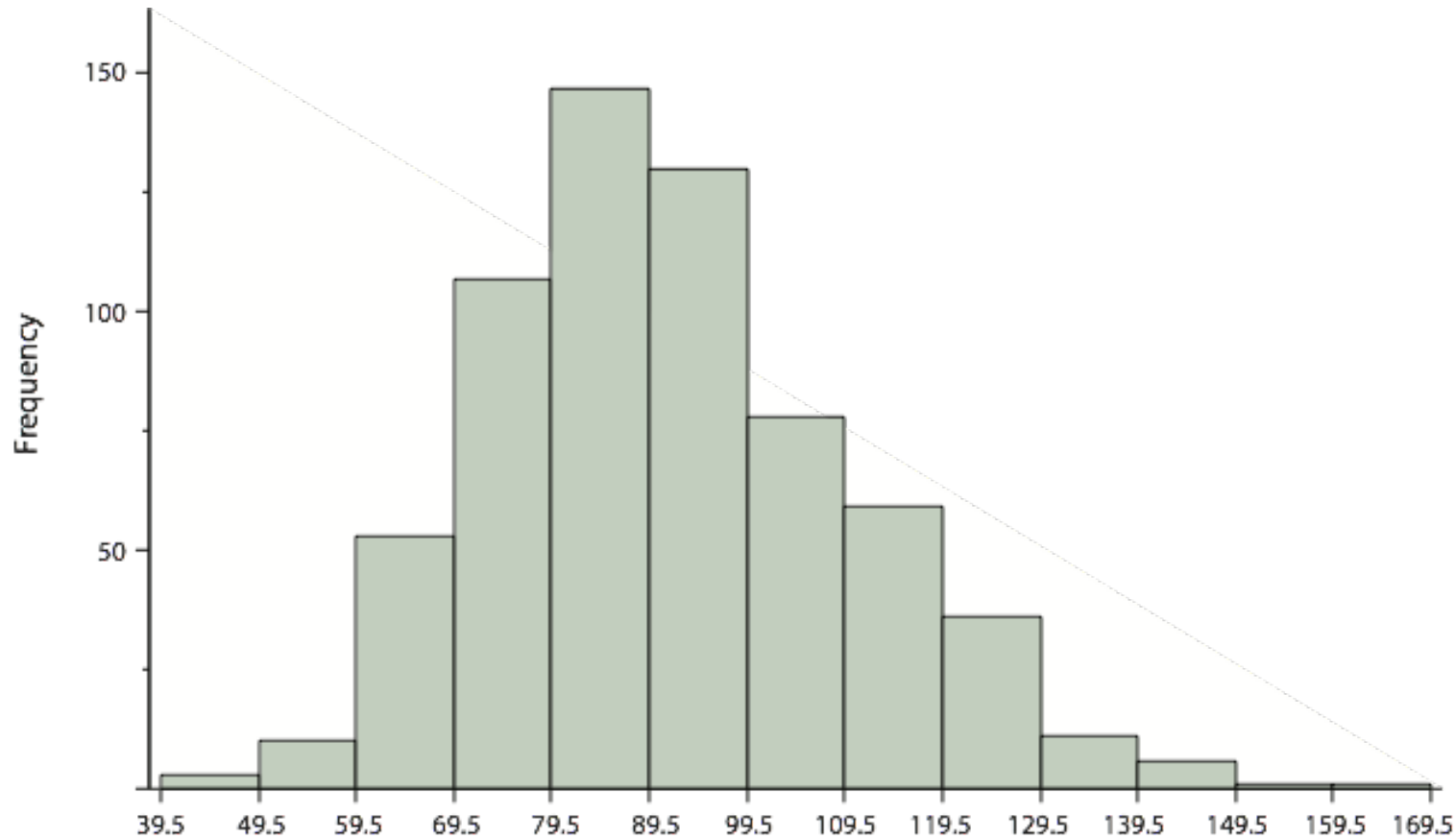
- Presenting **continuous** data:

Data range is divided in class intervals or bins

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

Grouped frequency distribution of certain test scores

- Presenting **continuous** data:  
Data range is divided in class intervals or bins



Histogram of the test scores

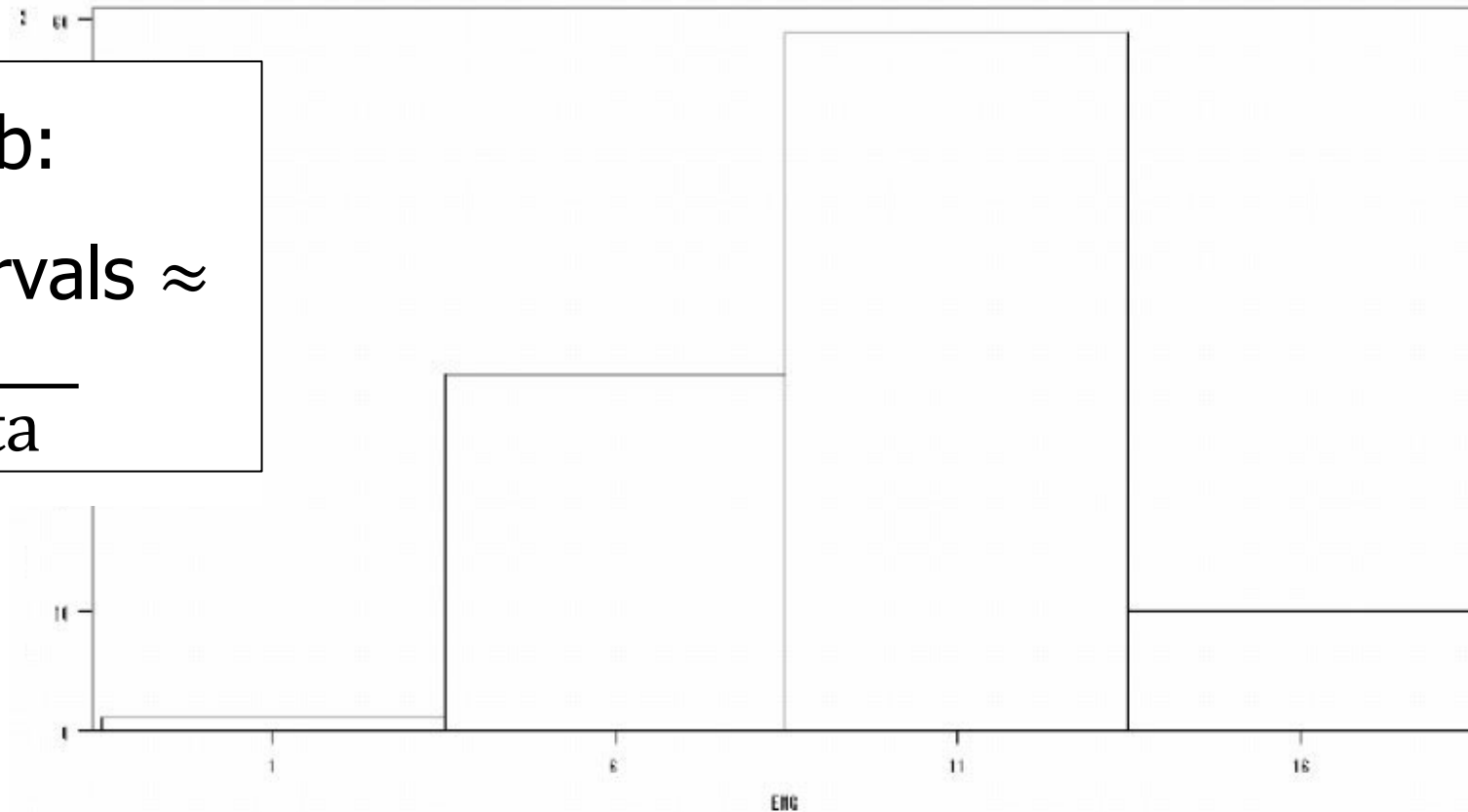
# Presenting continuous data - Histogram:

Examples: Class interval size affects the visual presentation

A rule of thumb:

# of class intervals  $\approx$

$$\sqrt{\text{\# of data}}$$



Histogram when class interval is 5

Question: You have to decide between displaying data with a histogram or a stem-and-leaf display. What factor would affect your choice?

With more data, a histogram can be very useful since it shows the overall shape of the distribution.

Stem-and-leaf display is better for smaller sets of data.

Question:

Suppose you are constructing a histogram for describing the distribution of salaries for CTOs.

- (a) What is on the Y-axis? (b) What is on the X-axis?
- (c) What would be the probable shape of the salary distribution? Explain why.

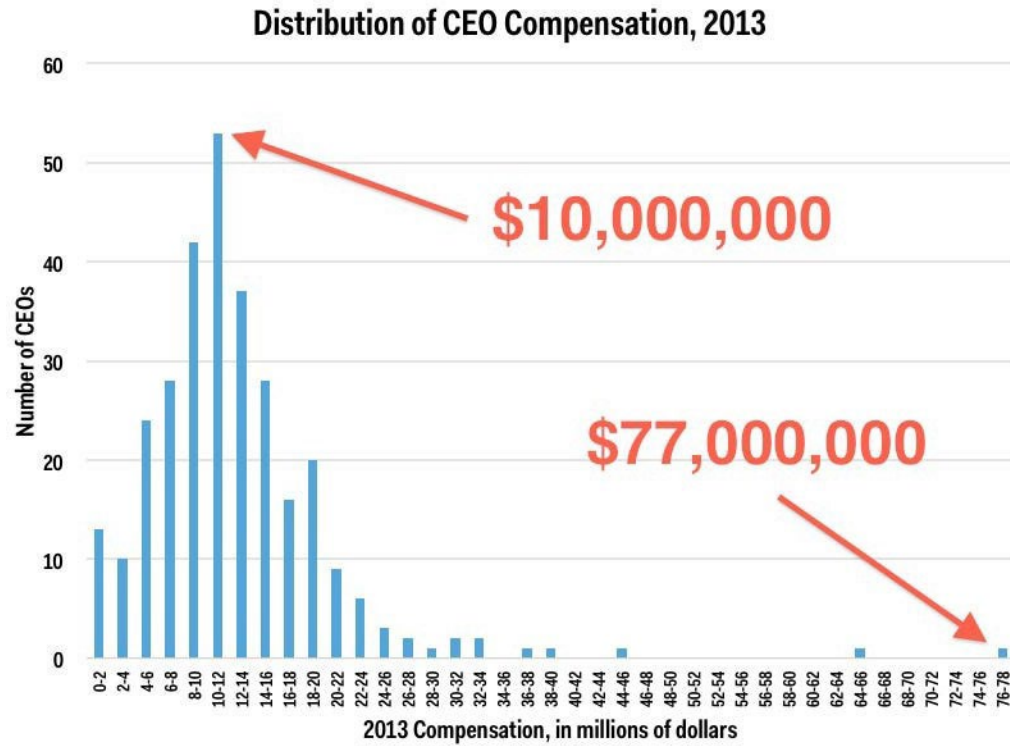
Question:

Suppose you are constructing a histogram for describing the distribution of salaries for CTOs.

(a) What is on the Y-axis? (b) What is on the X-axis?  
(c) What would be the probable shape of the salary distribution? Explain why.

- a) The Y-axis would be the frequency of individuals.
- b) Salary would be on the X-axis because this is the variable whose distribution is of interest.
- c) The distribution is expected to be positively skewed with a some earning well above the average salary.





Tuesday, July 30, 2024

### Top-earning jobs

Most of the top-earning jobs in 2023 were related to finance, technology, maritime or media, according to the latest Ministry of Manpower occupational wage data released in June 2024.

Rank	Job title	Number of workers in sample	Median monthly gross wage (\$)*
1	Oil and bunker trader	87	14,911
2	Chief information officer/chief technology officer/chief security officer	727	13,840
3	Enterprise/solution architect	392	13,682
4	Business valuer	97	13,649
5	University lecturer	2,950	13,108
6	Insurance services manager	272	12,407
7	Strategic planning manager	557	12,312
8	Chief operating officer/general manager	2,159	12,137
9	Risk management manager	432	11,558
10	Digital forensics specialist	42	11,298
11	Director (stage, film, television, game, commercial, video and radio)	156	11,252
12	Infocomm technology sales and services professional	237	11,250
13	Financial/investment adviser (e.g. relationship manager)	1,366	11,129
14	Financial risk manager	380	10,817
15	Specialist medical practitioner (medical)	422	10,693
16	Regional sales manager	563	10,597
17	Financial services manager	1,531	10,580
18	Editor (news and periodicals)	410	10,577
19	Marine superintendent (engineer)	151	10,310
20	Research and development manager	1,813	10,262

NOTE: \*As at June 2023. It is the sum of the basic wage, overtime payments, commissions, allowances and other regular cash payments. It is before deduction of employee CPF contributions and personal income tax, and excludes employer CPF contributions, bonuses, stock options, other lump sum payments and payments in kind.

Source: MINISTRY OF MANPOWER STRAITS TIMES GRAPHICS

GRAPHIC BY: THE STRAITS TIMES

A fun site on salaries:  
<https://www.straitstimes.com/multimedia/graphics/2024/07/salary-guide-2024/index.html>

- Box Plot:

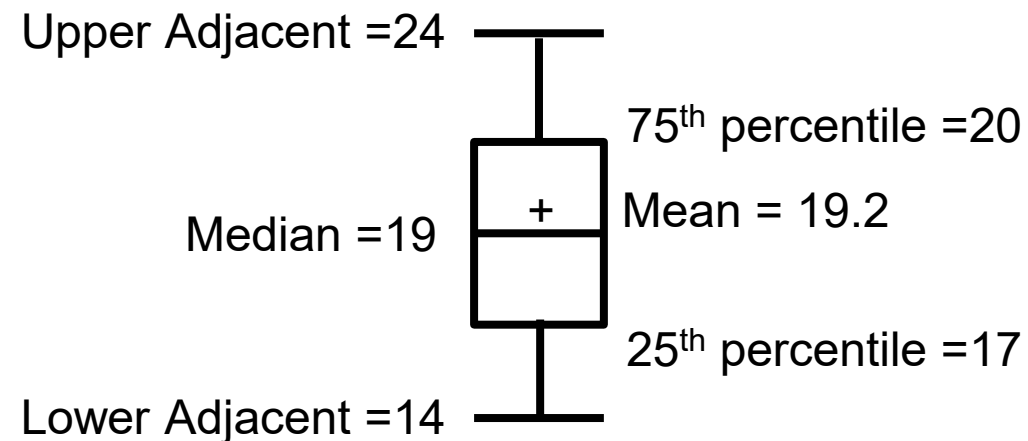
Draw the Box Plot for the following data set (31 values).

14	15	16	16	17	17	17	17	17	18	
18	18	18	18	18	19	19	19	20	20	
20	20	20	20	21	21	22	23	24	24	29

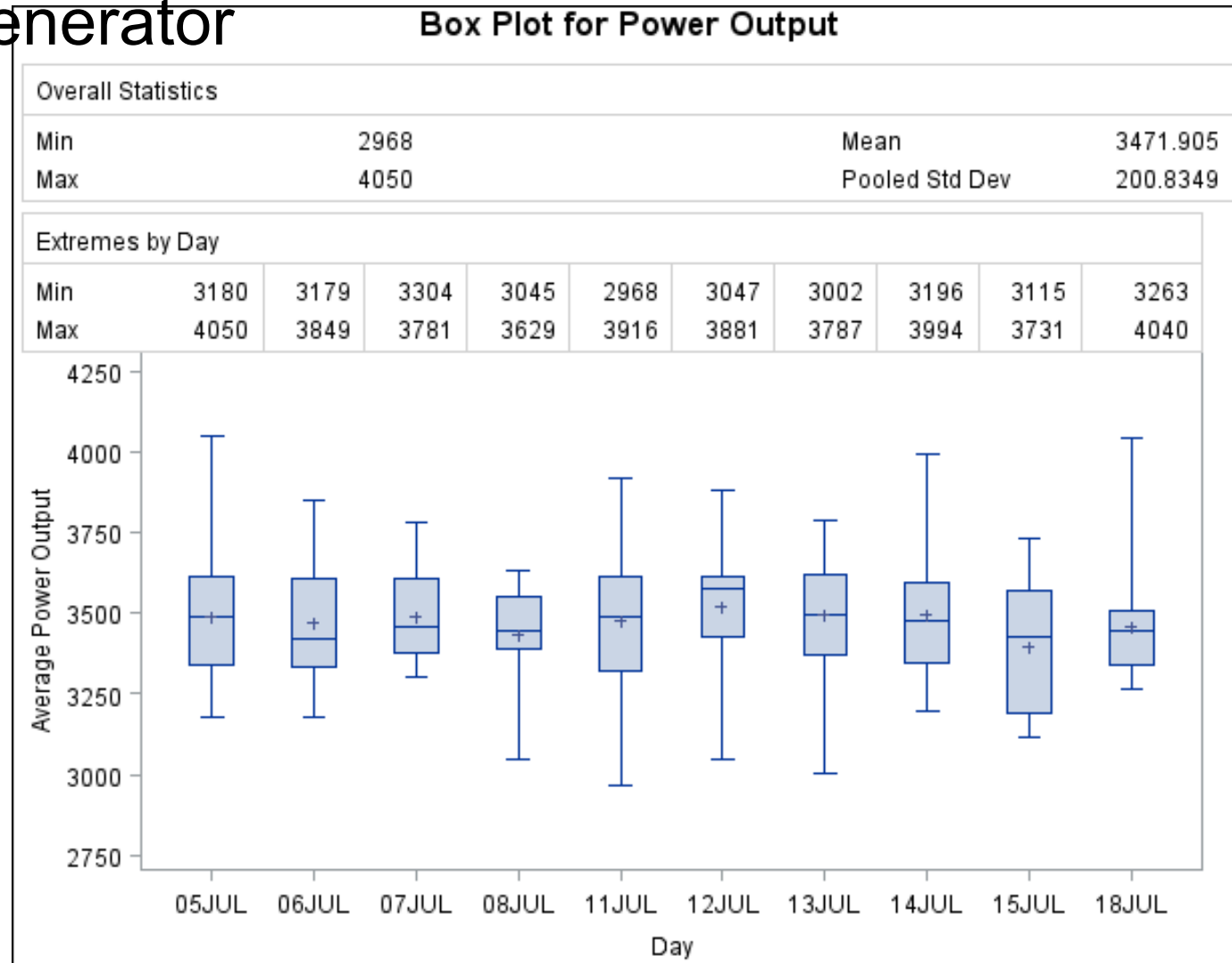
Upper Hinge = 75th Percentile (3 <sup>rd</sup> quartile)	20
Lower Hinge = 25th Percentile (1 <sup>st</sup> quartile)	17
H-Spread = Upper Hinge - Lower Hinge (IQR)	3
Step = 1.5 x H-Spread	4.5
Upper Inner Fence = Upper Hinge + 1 Step	24.5
Lower Inner Fence = Lower Hinge - 1 Step	12.5

Upper Outer Fence = Upper Hinge + 2 Steps	29
Lower Outer Fence = Lower Hinge - 2 Steps	8
Upper Adjacent = Largest value $\leq$ Upper Inner Fence	24
Lower Adjacent = Smallest value $\geq$ Lower Inner Fence	14
Outside Value	<div> <div>Outlier =</div> <div>Value &gt; upper adjacent or &lt; lower adjacent</div> </div>
Far Out Value	
	29
	none

○ Outlier = 29



# Practical example of Box Plot: Daily Power Output of an Electric Generator



# Another practical example of Box Plot: Daily share prices

OCBC Bank  
■ O39.SI - Daily

