Statistics for Research

Assoc. Prof. Anan Phonphoem, Ph.D.
(รศ.ดร. อนันต์ ผลเพิ่ม)

Computer Engineering Department,
Kasetsart University

Outline

- Qualitative V.S. Quantitative
- Descriptive Statistics
  - Definition
  - Data representation
  - Sample statistics

Research method categories

- Qualitative (เชิงคุณภาพ)
- Quantitative (เชิงปริมาณ)

Qualitative Research

- Concentrates on Collecting and analyzing
  - Subjective data
  - Non-numerical data
- (Usually) People perceptions
  - Verbal data rather than measurements
- Emphasis on
  - Interpretative manner, subjective, impressionistic
  - Knowledge

Quantitative Research

- Concentrates on Collecting and analyzing
  - Objective data
  - Measurable data
- Usually involves some form of mathematics
  - Statistical, Calculus, Discrete

Causality Problem

- Data taken from 20 year study of 2438 middle-aged Welsh men's (e.g. UK. Ireland) shaving habits discovered
- that the unshaven are;
  - Less likely to be married
  - More likely to be blue-collar
  - Had a 45% higher death rate
  - Had a 70% higher risk of stroke (โรคหลอดเลือดสมอง)
  - Were shorter
  - More likely to suffer from Angina (โรคเจ็บที่หัวใจ, เจ็บหน้าอก)

- **Conclusion:** Not shaving causes these problems?

From "Research Methods: Quantitative and Qualitative Research Methods" by Computer Science, Ryerson University

Quality V.S. Quantity

- Quantitative
  - We have an hypothesis that monkeys will put bananas to their ears
  - We gave bananas to monkeys
  - If we see banana to ear == “Monkeycide”
  - We counted xx instances of Monkeycide over yy trials
  - Our hypothesis is accepted if xx > 0

From “Research Methods: Quantitative and Qualitative Research Methods” by Computer Science, Ryerson University

Quality V.S. Quantity

- Qualitative
  - We saw monkeys pick up bananas
  - We observed the monkeys placing bananas to their ears
  - From observation we have the concept: “Monkeycide”
  - Monkeys Jenny, Irene and Blake exhibited Monkeycide

From “Research Methods: Quantitative and Qualitative Research Methods” by Computer Science, Ryerson University

Qualitative V.S. Quantitative

	Qualitative	Quantitative
Goal of the Research	provide a complete, detailed description of the research topic	focuses more in •counting •classifying features •constructing statistical models
Usage	earlier phases of research projects	•latter phases of the research projects •clearer picture of what to expect

Summarized from http://www.experiment-resources.com/quantitative-and-qualitative-research.html

Qualitative V.S. Quantitative

	Qualitative	Quantitative
Data Gathering Instrument	researcher serves as instrument •Interviews • focus groups • narratives •participant observation	• questionnaires • surveys • equipment to collect numerical or measurable data
Type of Data	• words (from interviews) • images (videos) • objects (such as artifacts) • discussion are figures • in form of graphs	• tables containing data • numbers and statistics
Approach	• primarily subjective • understand behavior • reasons that govern such behavior	• objectively separated from the subject matter • seeks precise measurements and analysis

Summarized from http://www.experiment-resources.com/quantitative-and-qualitative-research.html

Which one should be used ?

- Quantitative Research
  - To inquiry through numerical evidence
- Qualitative Research
  - Explain the reason for particular event
  - Explain this particular phenomenon occurred
- Sometimes required to use both methods
  - Complement each other

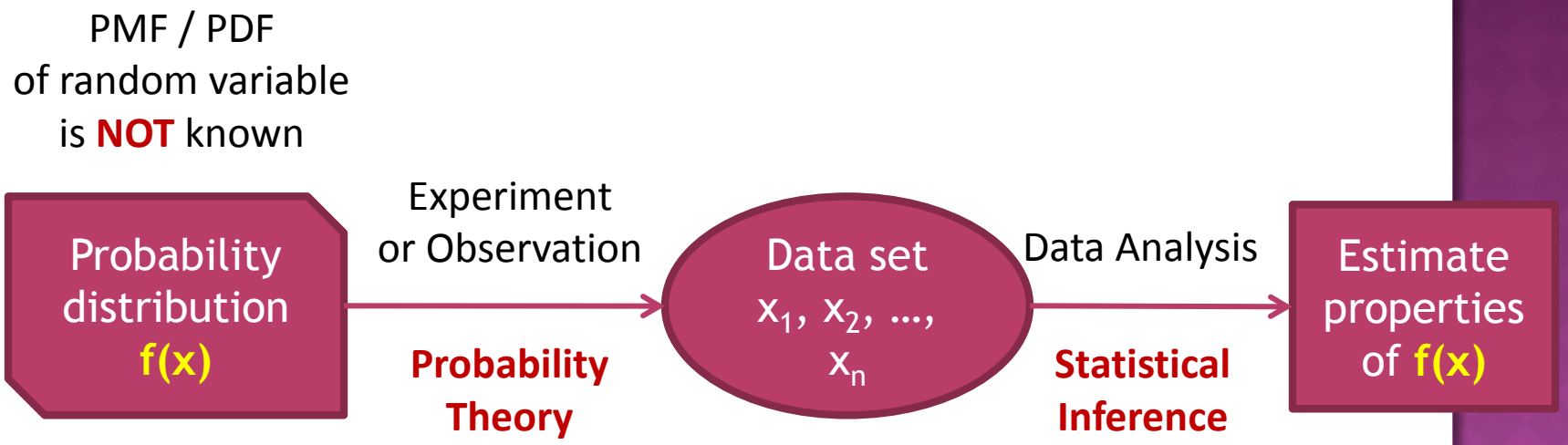
Outline

- Qualitative V.S. Quantitative
- Descriptive Statistics
  - Definition
  - Data representation
  - Sample statistics

Introduction

- Most applications
  - PMF/PDF of a random variable is NOT known
- Find out the probability distribution of the random variable
  - Perform an experiment → **Data set**
  - From data set → probability distribution called **Statistical Inference**

Probability V.S. Statistics



Definitions

- Population
  - All possible observations from probability distribution
- Sample
  - A subset of population
  - Use to investigate the unknown probability distribution
- Random sample
  - Elements of the sample are randomly chosen
  - To be the Representative

Example

- Observe computer network connection down in the University Network

Data Set

Caused by	Frequency in one year
Power down	15
DHCP server	25
DNS server	14
Switch/Router	5
Others	4

Quality of Representativeness

- What is the population?
  - Tough! The previous/next year data
- Is the data set the representative?
  - Factors that make data set cannot be
- Sample question
  - major cause for next year?
  - If we hire an experienced server administrator next year, major cause?
  - If we change new model switch/routers, major cause?

Data Set

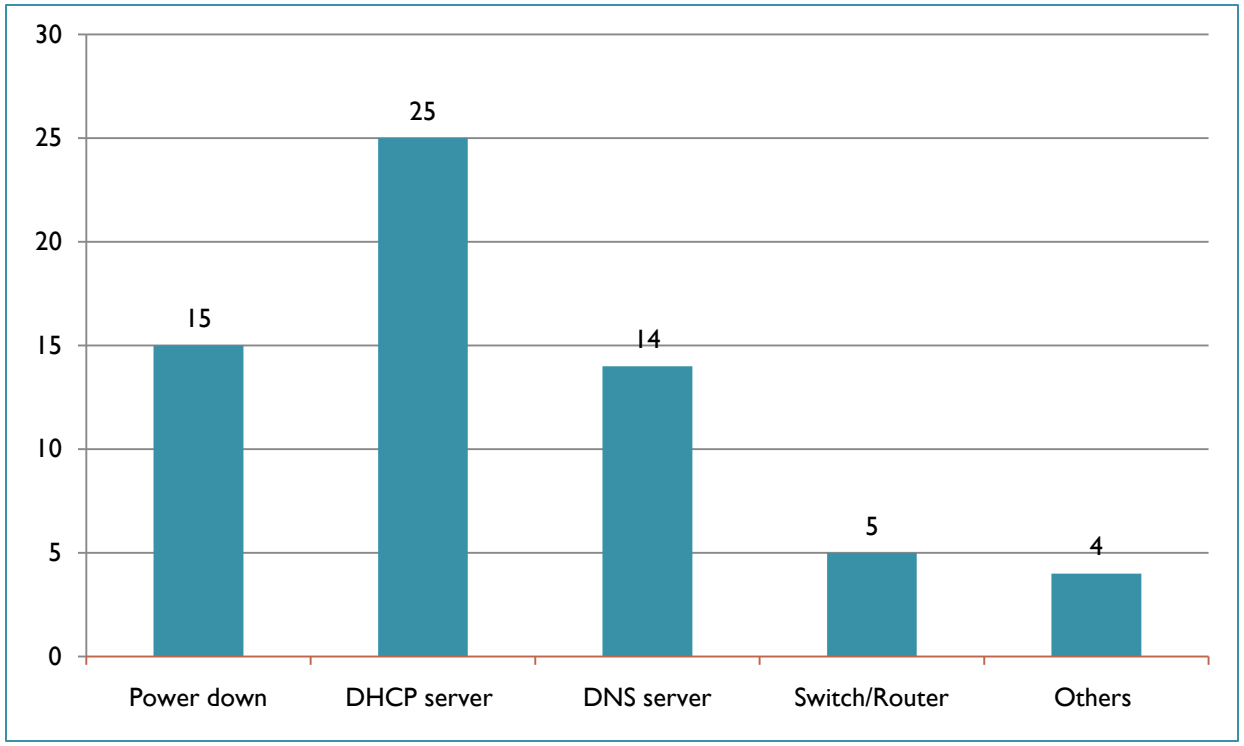
Caused by	Frequency in one year
Power down	15
DHCP server	25
DNS server	14
Switch/Router	5
Others	4

Data Presentation

- Tabular
  - Not very informative
- Graphical
  - Bar chart
  - Pareto chart
  - Pie chart
  - Histogram

Bar Chart

- Illustrate a **categorical** data set

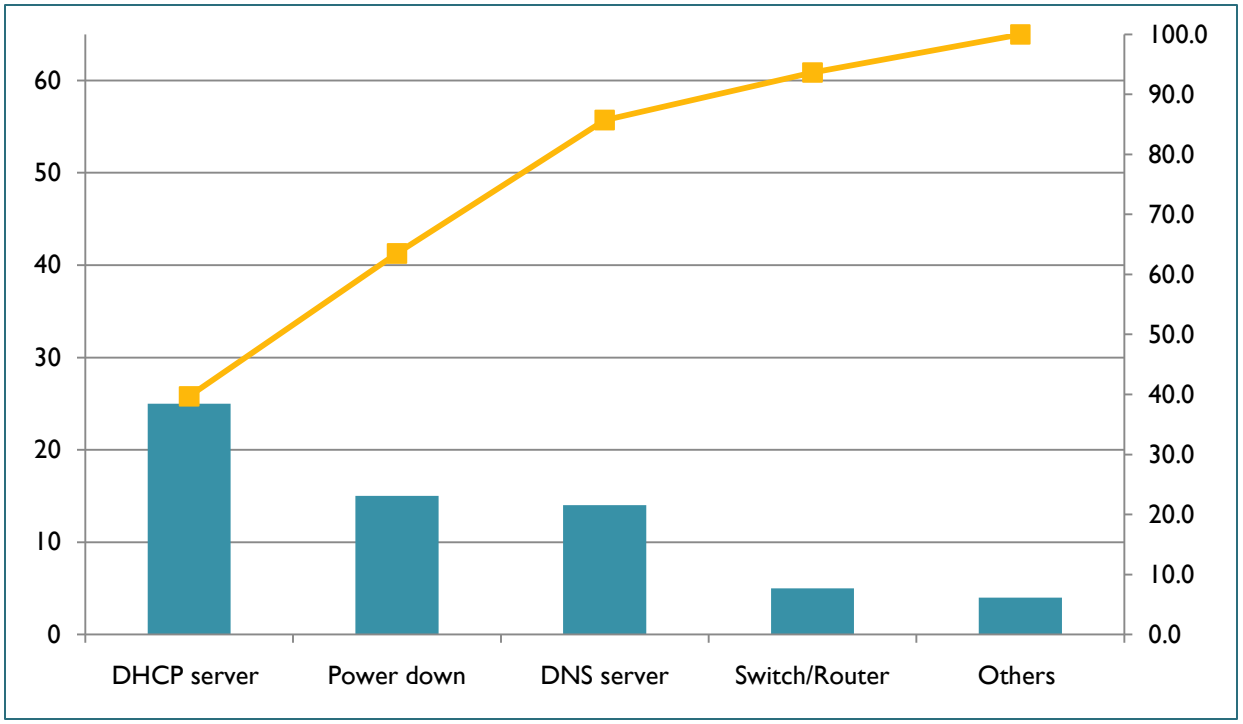


Data Set

Caused by	Frequency in one year
Power down	15
DHCP server	25
DNS server	14
Switch/Router	5
Others	4

Pareto Chart

- Categories in **decreasing frequency**
- Quality control / Failure

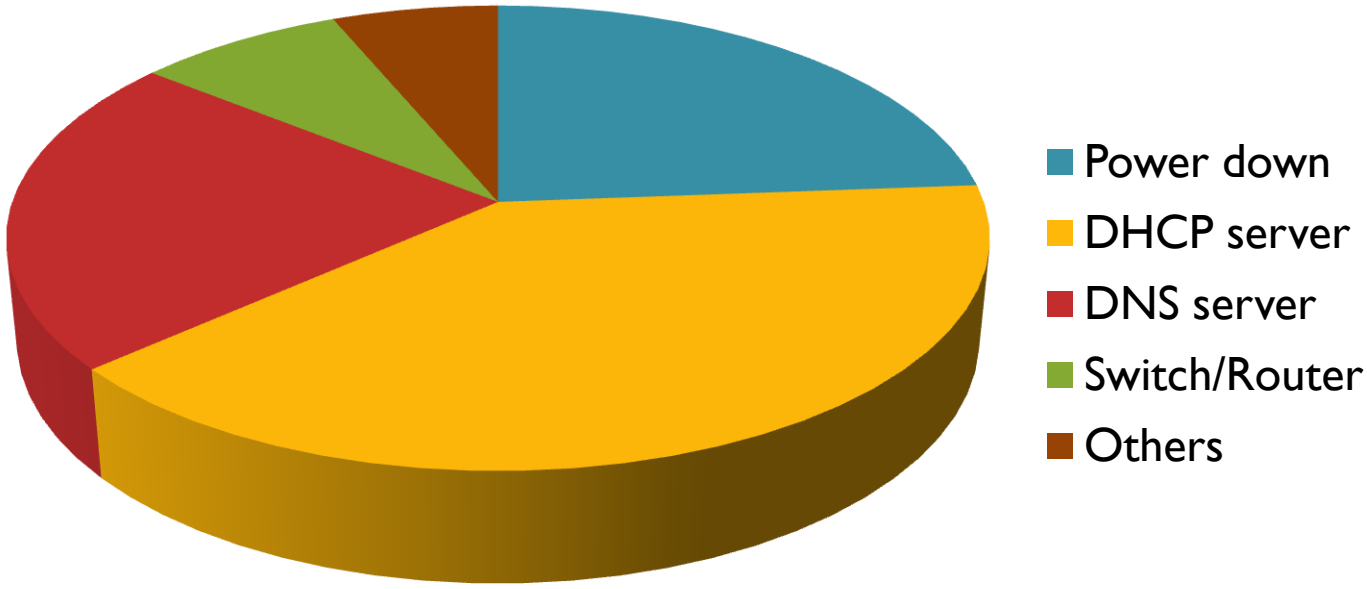


Data Set

Caused by	Frequency in one year
Power down	15
DHCP server	25
DNS server	14
Switch/Router	5
Others	4

Pie Chart

- Emphasize the **proportion** of data set

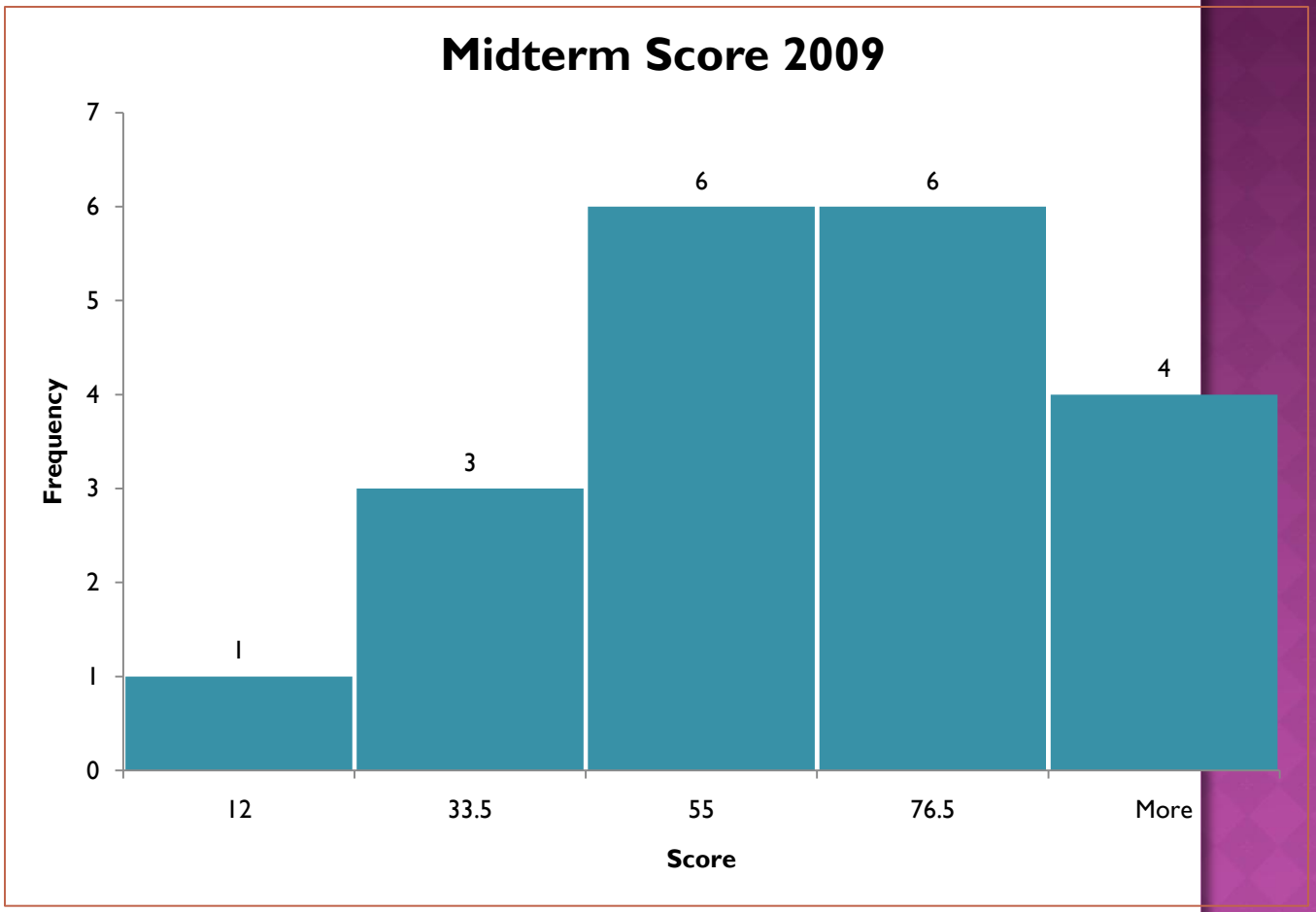


Histogram

- Look similar to Bar Chart
- Illustrate **numerical data** rather than categorical data
- Shape of histogram ←→PMF/PDF

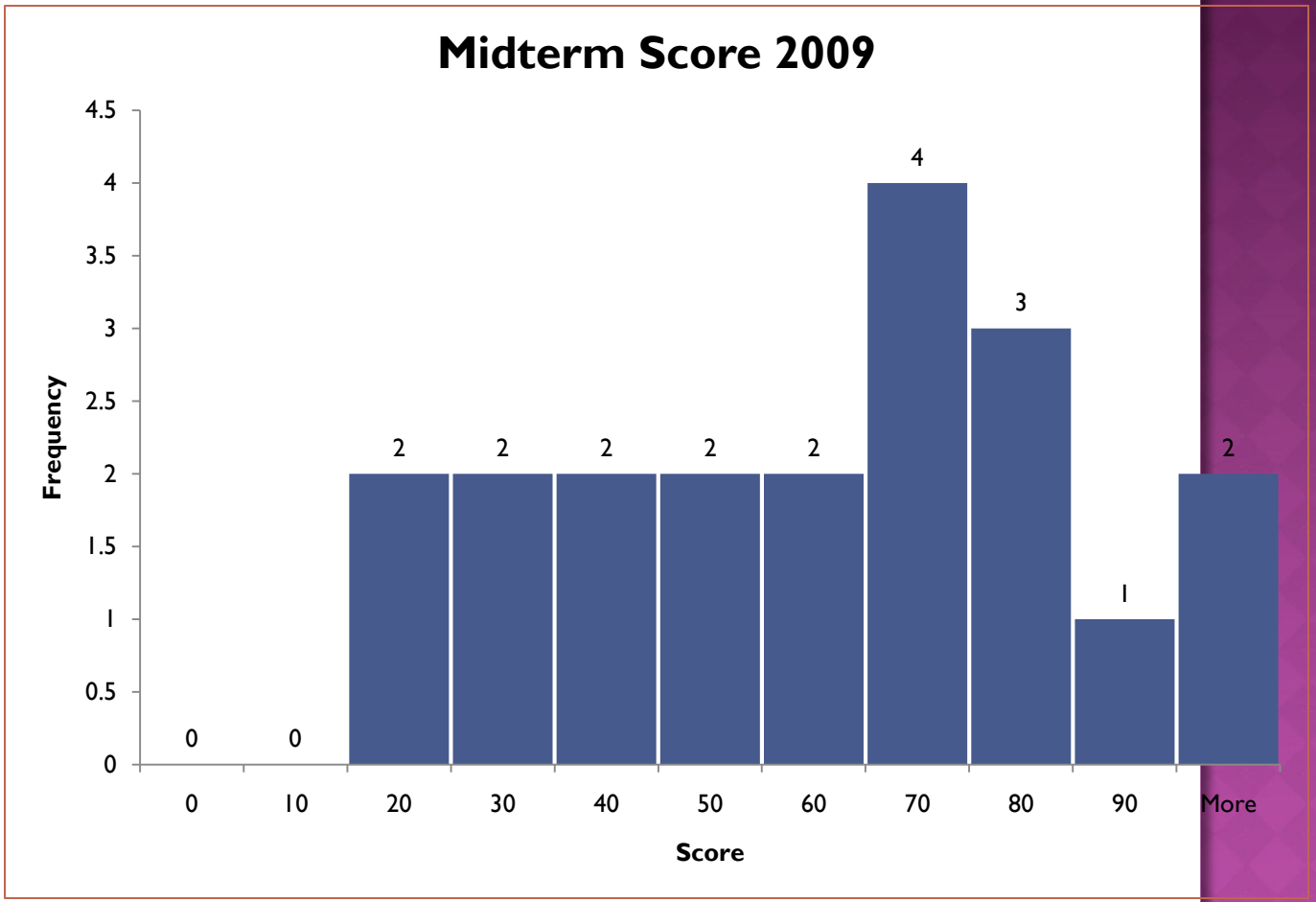
Histogram

No	Total
	100
1	82
2	61
3	49
4	52
5	71
6	68
7	36
8	61
9	50
10	95
11	55
12	26
13	73
14	12
15	24
16	17
17	77
18	98
19	62
20	39



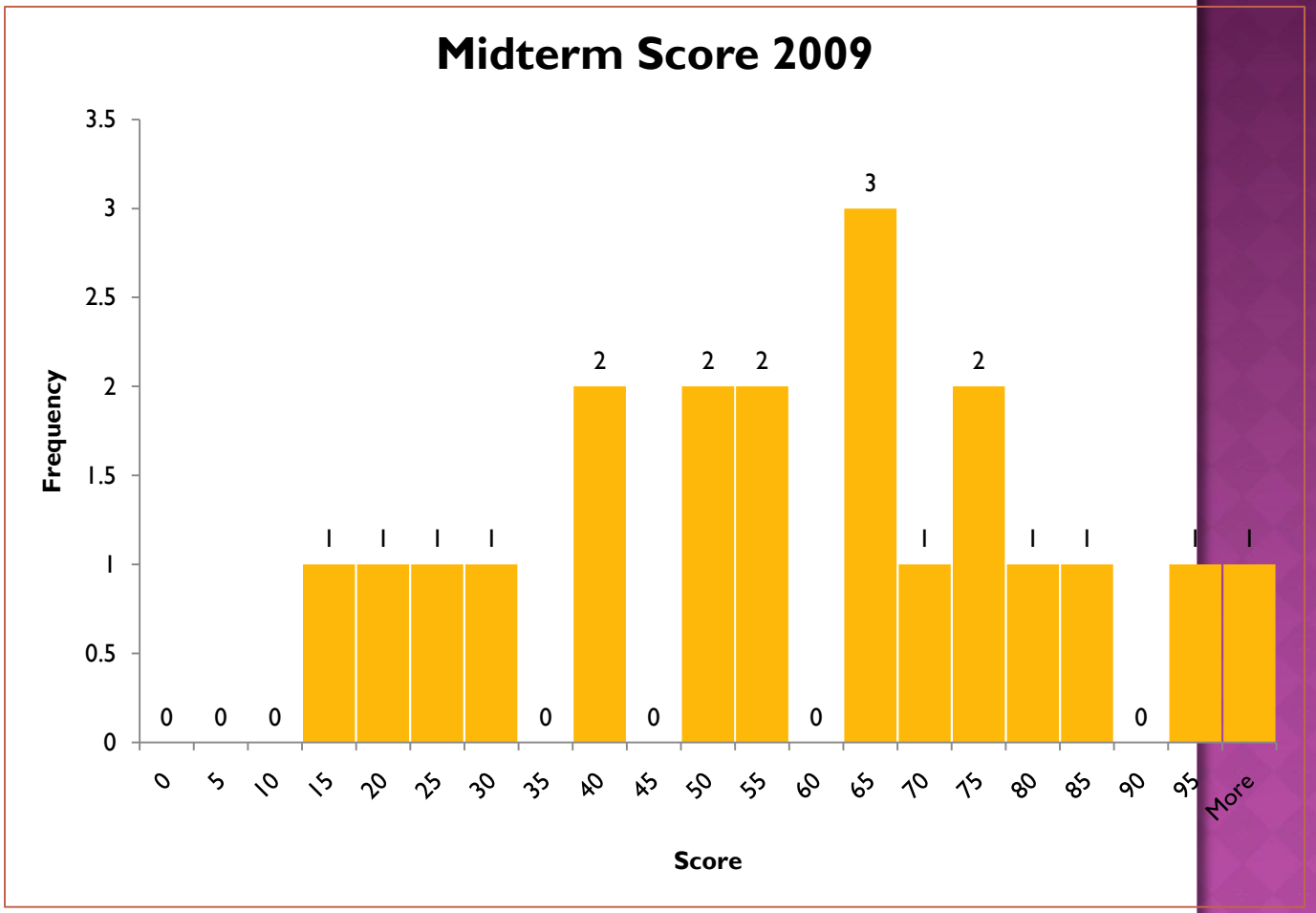
Histogram

No	Total
	100
1	82
2	61
3	49
4	52
5	71
6	68
7	36
8	61
9	50
10	95
11	55
12	26
13	73
14	12
15	24
16	17
17	77
18	98
19	62
20	39

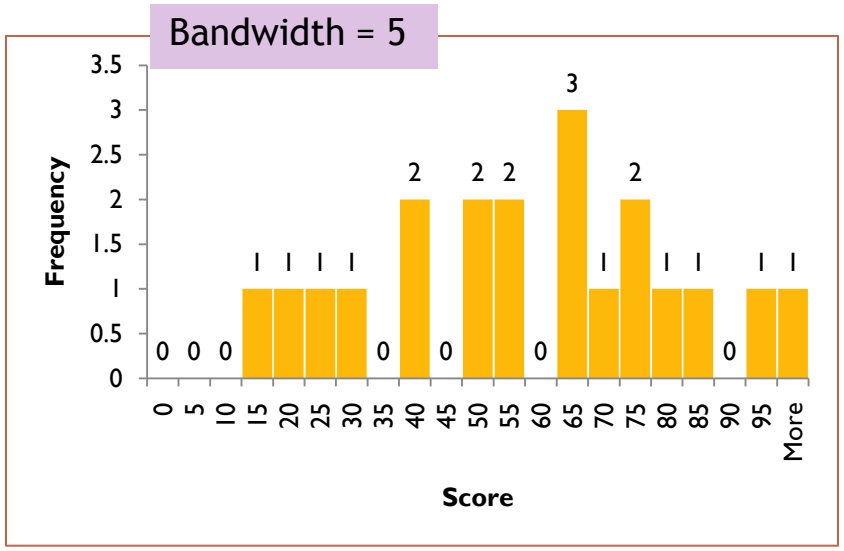


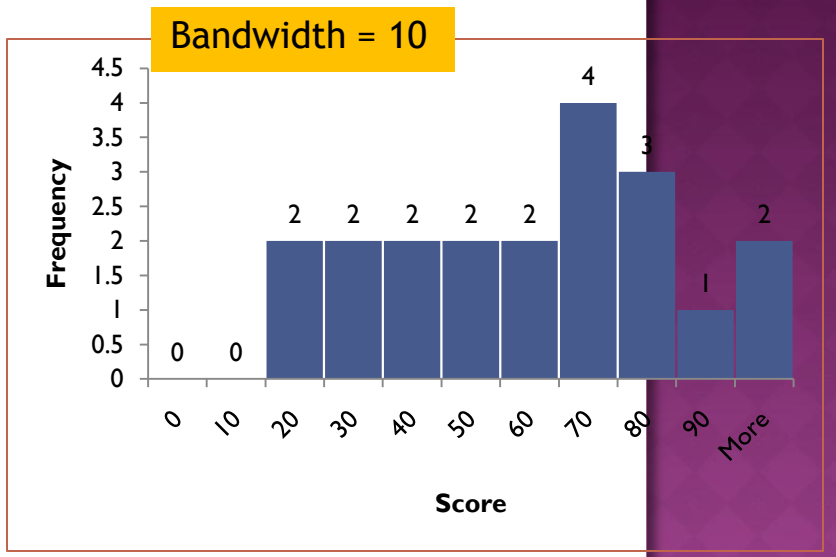
Histogram

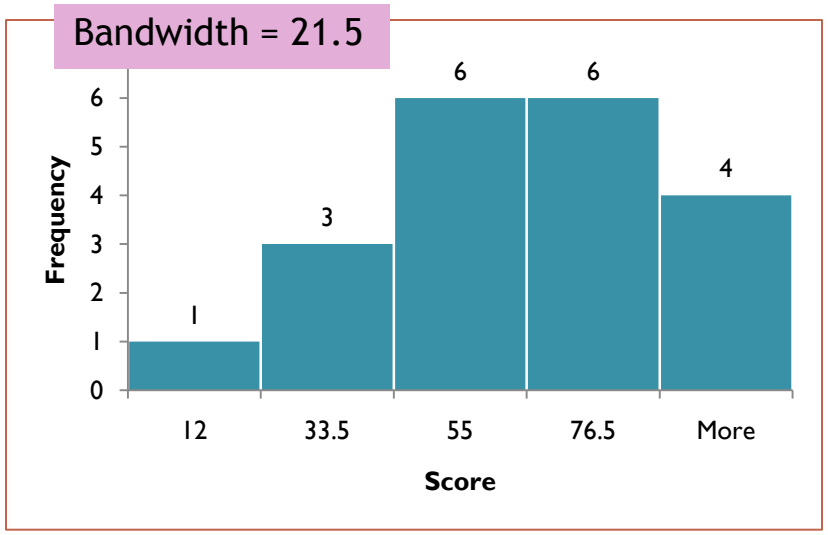
No	Total
	100
1	82
2	61
3	49
4	52
5	71
6	68
7	36
8	61
9	50
10	95
11	55
12	26
13	73
14	12
15	24
16	17
17	77
18	98
19	62
20	39



Which one is better?

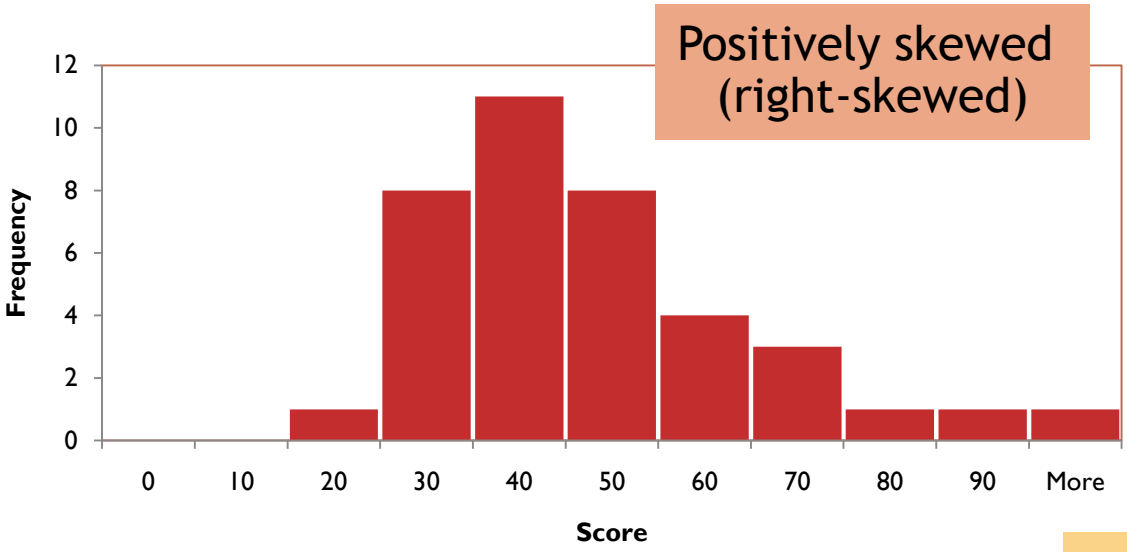


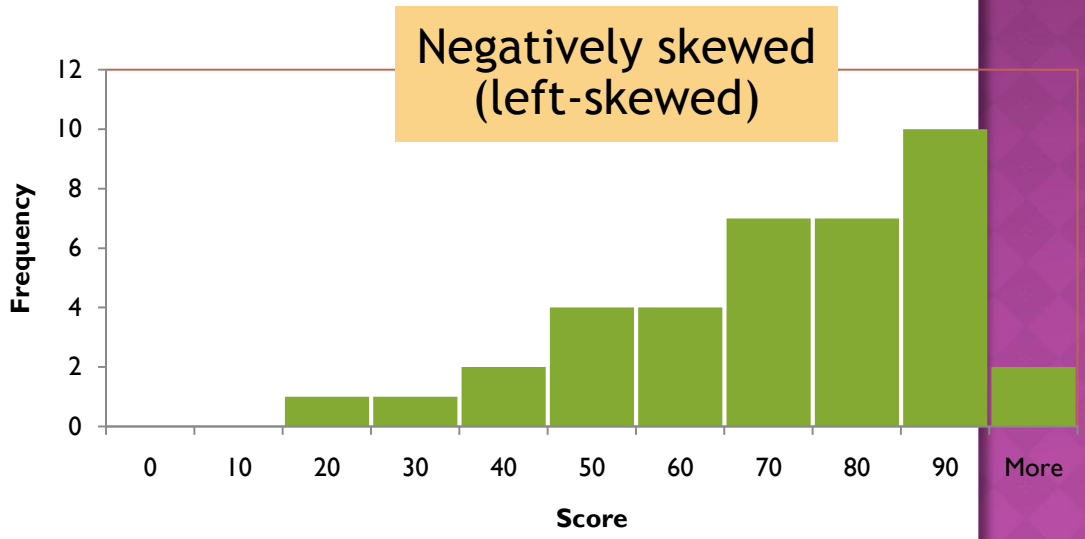




What is pdf ?

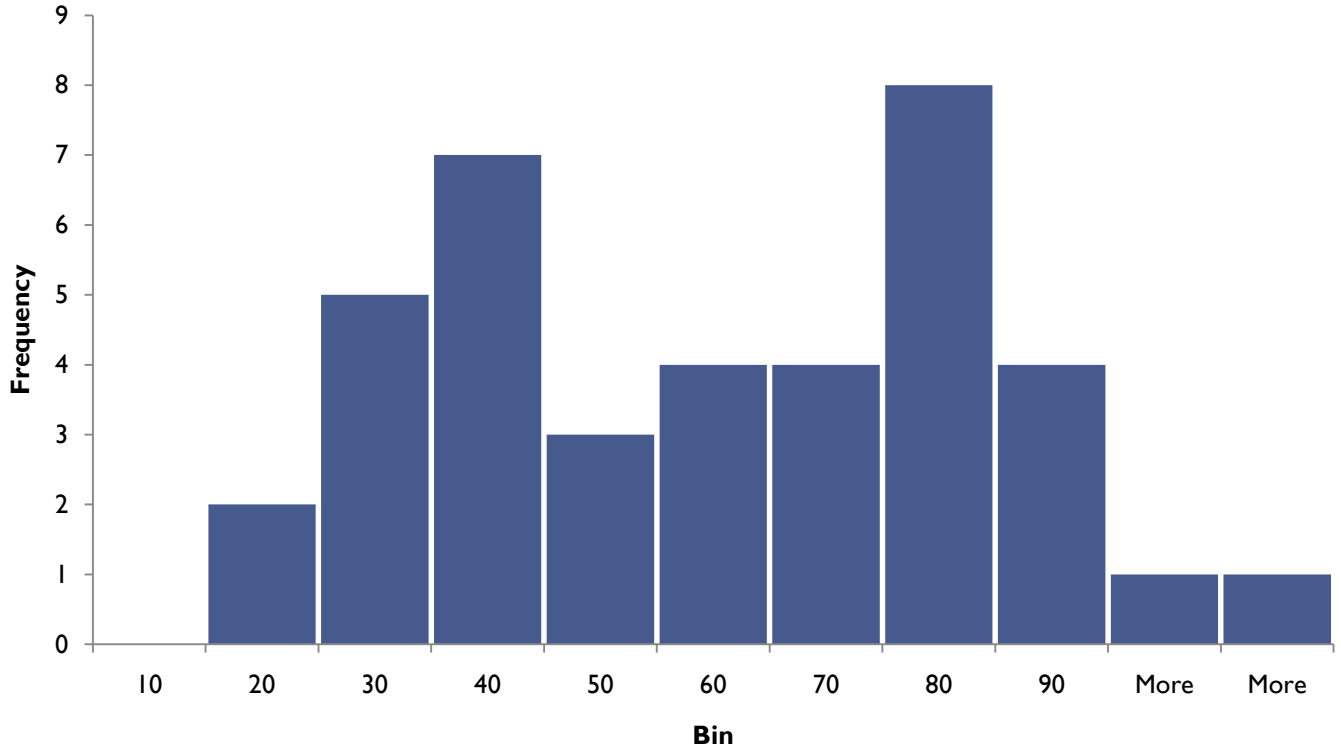
Skewness


Positively skewed
(right-skewed)
12
10
8
6
4
2
0
Frequency
0
10
20
30
40
50
60
70
80
90
More
Score


Negatively skewed
(left-skewed)
12
10
8
6
4
2
0
Frequency
0
10
20
30
40
50
60
70
80
90
More
Score

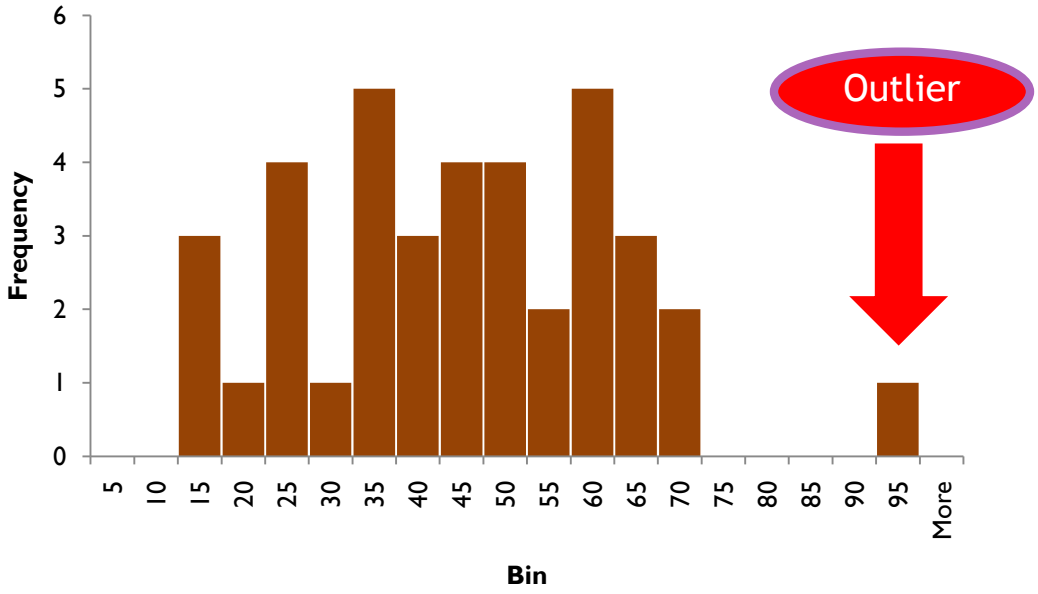
Bimodal Histogram

- Combination of two data sets
- Ex. Man / Woman



Outlier

- Strange data point, separate from the rest
- Should it be removed?
- Maybe misrecorded
- Special condition



Sample Statistics

- Sample Mean
- Sample Median
- Sample Trimmed Mean
- Sample Mode
- Sample Variance
- Sample Quantiles

Sample Mean

- A data set consists on n observations
  $x_1, x_2, x_3, \ldots, x_n$
- Middle Value → E[X]
- Estimate of expectation of unknown probability distribution

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Mean

12	17	24	26	36	39	49	50	52	55	61	61	62	68	71	73	77	82	95	98

$$\bar{x} = \frac{12 + 17 + 24 + \ldots + 98}{20} = 55.400$$

Sample Median

- **Middle** value of ordered data set

12	17	24	26	36	39	49	50	52	55	61	61	62	68	71	73	77	82	95	98

$$\frac{55 + 61}{2} = 58$$

- For a symmetric sample

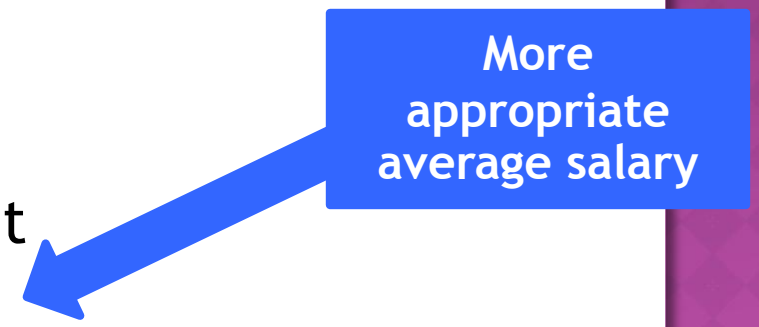
  sample mean ≈ sample median

- For positively skewed

  sample mean > sample median

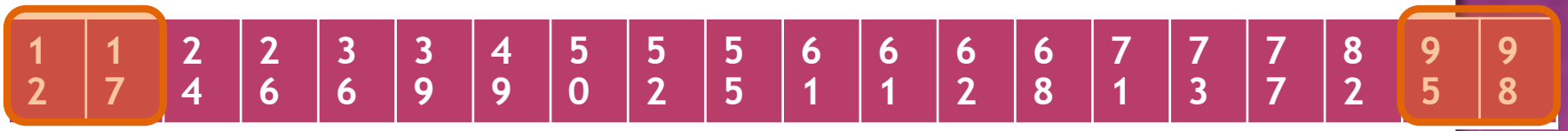
Sample Median

- Man power in a company
  - 100 people in the operator position
  - 25 people in the middle management
  - 4 people in the top executive
- What is the average salary?
  - Operator = 10,000 Baht
  - Middle management = 25,000 Baht
  - Top executive = 200,000 Baht
- Positively skewed
  - Sample mean = 15,746.45 Baht
  - Sample median = 10,000 Baht



Sample Trimmed Mean

- Deleting some largest/smallest data observations
- Taking mean of the remaining observations
- Usually 10 % trim
- Not so **sensitive** to the tail (especially, **outlier**)



$$\frac{24 + 26 + 36 + \ldots + 82}{16} = 55.375$$

Example: Olympics Diving

4.5	7.5	7.7	7.9	8.1	8.1	8.4	8.6	8.6	10

Sample Mean $\bar{x} = \frac{4.5 + 7.5 + 7.7 + \ldots + 10}{10} = 7.94$

Sample Median 8.1

Sample Trimmed Mean $\frac{7.5 + 7.7 + 7.9 + \ldots + 8.6}{8} = 8.11$

Data Analysis from Excel

4.5	7.5	7.7	7.9	8.1	8.1	8.4	8.6	8.6	10

Column1	
Mean	7.94
Standard Error	0.440505
Median	8.1
Mode	8.1
Standard Deviation	1.392998
Sample Variance	1.940444
Kurtosis	4.848537
Skewness	-1.61261
Range	5.5
Minimum	4.5
Maximum	10
Sum	79.4
Count	10

Data Analysis from Excel

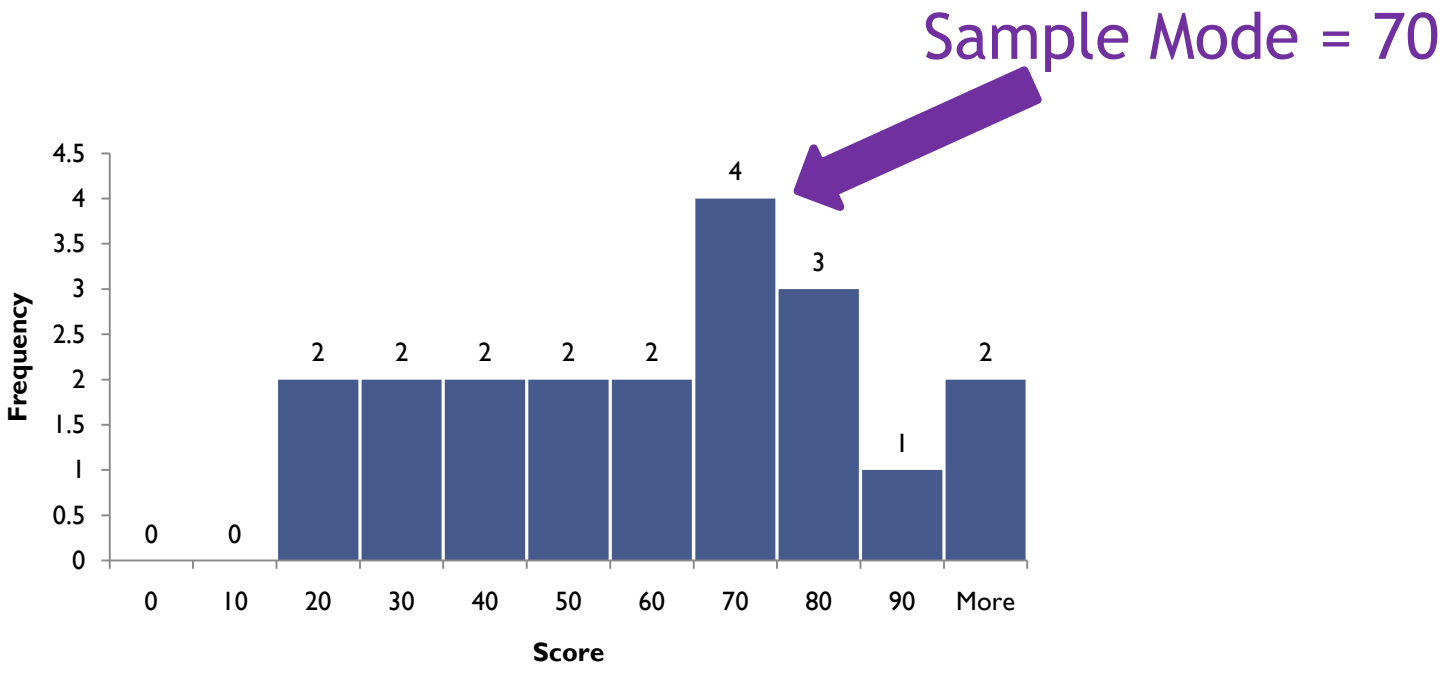
Midterm Exam Score

12	17	24	26	36	39	49	50	52	55	61	61	62	68	71	73	77	82	95	98

Column1	
Mean	55.4
Standard Error	5.455755
Median	58
Mode	61
Standard Deviation	24.398878
Sample Variance	595.30526
Kurtosis	-0.6216068
Skewness	-0.1075727
Range	86
Minimum	12
Maximum	98
Sum	1108
Count	20

Sample Mode

- To denote category or data value contains **largest number** of observations



Sample Variance

- A data set consists on n observations x_1, x_2, x_3, …, x_n
- Sample **standard deviation** = s
  - How deviate from sample mean
- Sample **variance** = s^2
- Denominator "n-1" → not "n"

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

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Sample Variance

Midterm Exam Score

12	17	24	26	36	39	49	50	52	55	61	61	62	68	71	73	77	82	95	98

Column1	
Mean	55.4
Standard Error	5.455755
Median	58
Mode	61
Standard Deviation	24.398878
Sample Variance	595.30526
Kurtosis	-0.6216068
Skewness	-0.1075727
Range	86
Minimum	12
Maximum	98
Sum	1108
Count	20

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

$\bar{x} = 55.4$ $\quad \sum_{i=1}^{20} x_i^2 = 12^2 + 17^2 + \ldots + 98^2 = 72,694$

$$s^2 = \frac{72694 - (20 * 55.4^2)}{19} = 595.30526$$

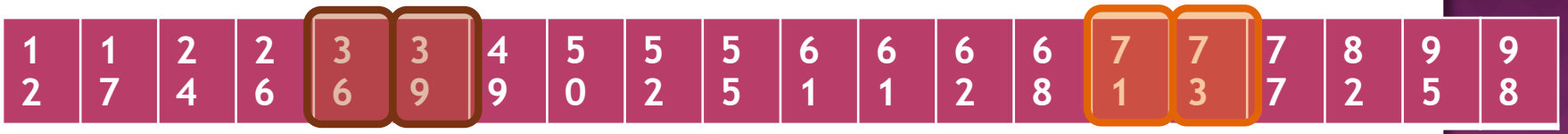
$$s = 24.3988$$

Sample Quantiles

- A value that has
  - a proportion **p** of sample taking smaller values
  - And **(1-p)** of sample taking larger values
- Sample Percentile
  - Usually take a value between two data observations (weighted average)
  - Upper sample quartiles (75th percentile)
  - Lower sample quartiles (25th percentile)
  - Sample interquartile range

Sample Quantiles

Midterm Exam Score



$$\left(\frac{3}{4} \times 36\right) + \left(\frac{1}{4} \times 39\right) = 36.75$$

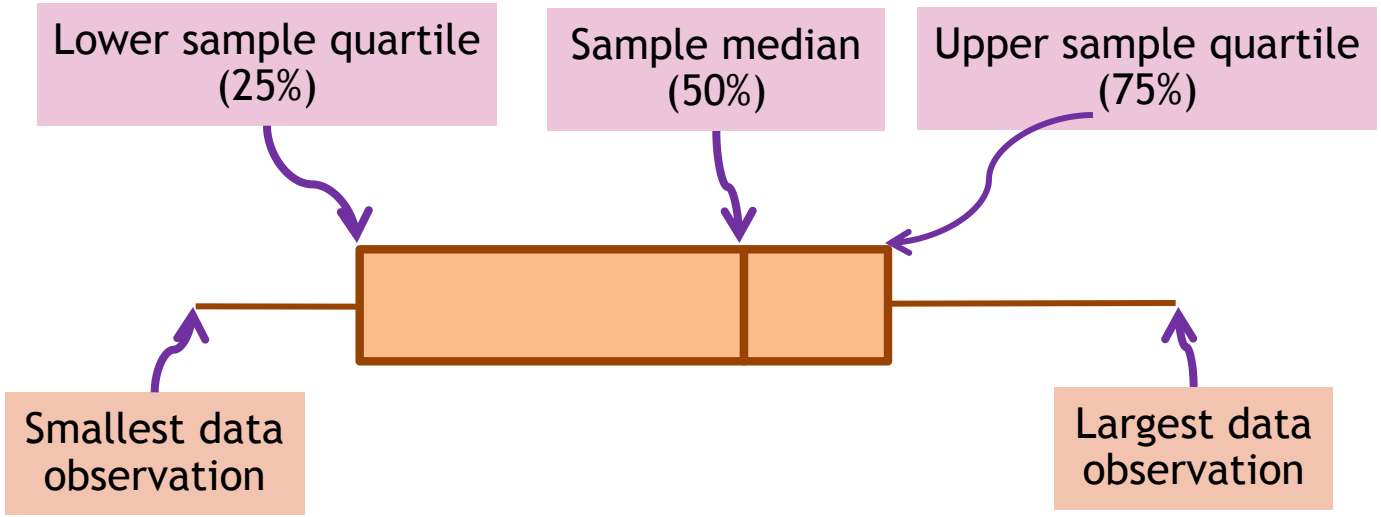
Lower sample quartile

$$\left(\frac{1}{4} \times 71\right) + \left(\frac{3}{4} \times 73\right) = 72.5$$

Upper sample quartile

Boxplots

- Schematic presentation of
  - sample median
  - upper/smaller quartiles
  - Largest/smallest data observations



Boxplots

Midterm Exam Score

12	17	24	26	36	39	49	50	52	55	61	61	62	68	71	73	77	82	95	98

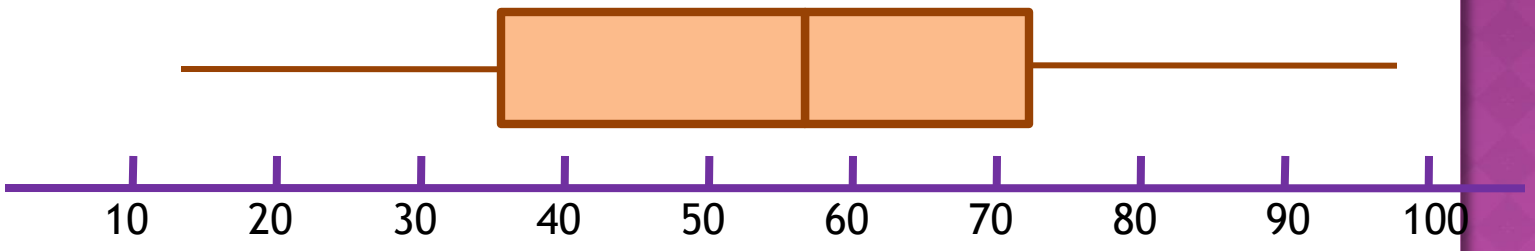
$$\left(\frac{3}{4} \times 36\right) + \left(\frac{1}{4} \times 39\right) = 36.75$$

Lower sample quartile

$$\left(\frac{1}{4} \times 71\right) + \left(\frac{3}{4} \times 73\right) = 72.5$$

Upper sample quartile

Sample median = 58



Coefficient of Variation

- Sample mean ($\bar{x}$) → middle values
- Standard deviation (s) → spread of data set

- Coefficient of variation

  → spread of the data **relative** to the middle value

$$CV = \frac{s}{\bar{x}}$$

- **Comparison** between different data sets
  - In terms of variation relative to sample average

Coefficient of Variation

- Zoologist interests in weight of animals
- Elephants
  - Sample average = 4,500 Kg
  - Sample standard deviation = 150 Kg
- Mice
  - Sample average = 30 g
  - Sample standard deviation = 1.67 g

$$CV_e = \frac{s}{\bar{x}} = \frac{150}{4550} = 0.033$$

$$CV_m = \frac{s}{\bar{x}} = \frac{1.67}{30} = 0.056$$

Mice have more variability in weights than elephants

References

- "*Research Methods: Quantitative and Qualitative Research Methods*" by Dept. of Computer Science, Ryerson University
- "*COMPARING QUANTITATIVE AND QUALITATIVE RESEARCH*", http://www.experiment-resources.com/quantitative-and-qualitative-research.html
- Probability and statistics for engineers and scientists," Anthony Hayter, 3rd Edition, Thomson Brooks/Cole, 2007, ISBN: 0-495-10878-2
- "Probability and Statistics" by Anan Phonphoem, Dept.of Computer Engineering, Kasetsart University

- how error bars relate to significance
- Confidence Intervals
  - very few studies actually measure an entire population

Standard deviation

- how much ***variation*** or "dispersion" there is from the average (mean, or expected value)
- Low standard deviation indicates
  - data points tend to be very close to the mean
- High standard deviation indicates
  - data are spread out over a large range of values

Standard error

- Measurement or ***estimation*** of the standard deviation of the sampling distribution associated with the estimation method
- Refer to an estimate of that standard deviation,
  - derived from a particular sample used to compute the estimate

For example

- Sample mean
  - Estimator of a population mean
  - Different samples drawn from that same population would in general have different values of the sample mean
- Standard error of the mean
  - Using sample mean as a method of estimating the population mean
  - standard deviation of those sample means over all possible samples (of a given size) drawn from the population
  - standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.