



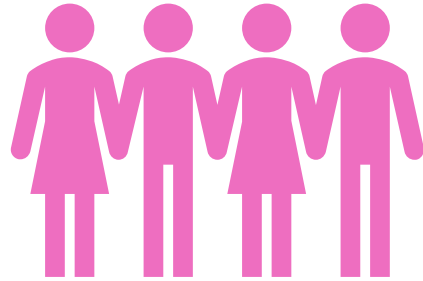
Descriptive Statistics 1

CSGE602013 –STATISTICS AND PROBABILITY
FACULTY OF COMPUTER SCIENCE UNIVERSITAS INDONESIA

References

- Introduction to Probability and Statistics for Engineers & Scientists, 4th ed. Sheldon M. Ross, Elsevier, 2009.
- Applied Statistics for the Behavioral Sciences, 5th Edition. Hinkle., Wiersma., Jurs., Houghton Mifflin Company, New York, 2003.
- Statistics for the Behavioral Sciences, 9th Edition. Frederick J. Gravetter, Larry B. Wallnau, Cengage Learning, 2012
- Elementary Statistics A Step-by-step Approach, 8th ed. Allan G. Bluman, Mc Graw Hill, 2012.

Recall



Population & Sample

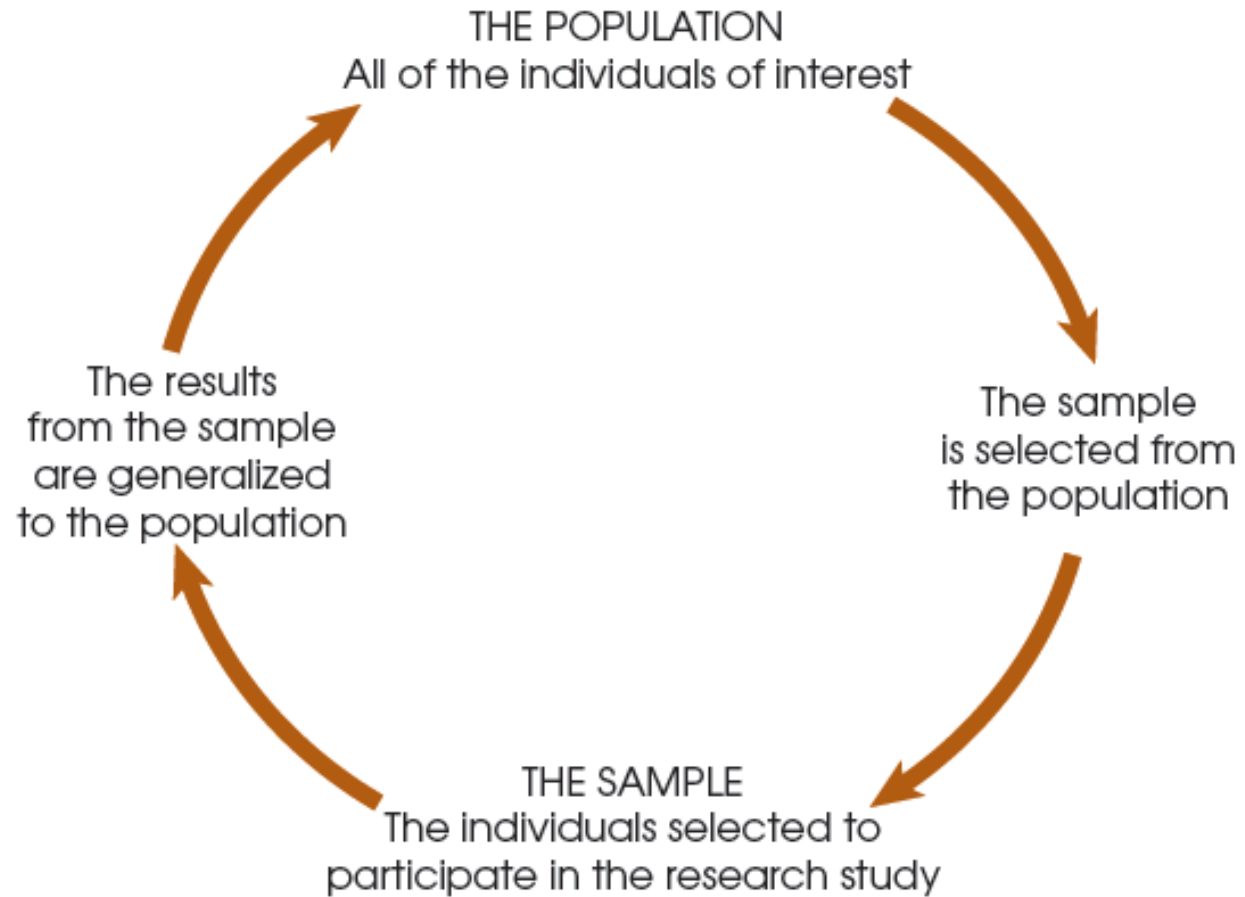


Parameters & Statistics

Population vs Sample

FIGURE 1.1

The relationship between a population and a sample.



Parameters & Statistics

- A parameter is to a population as a statistic is to a sample
- Parameters: Descriptive measures of a population
 - Ex: The mean of all UI students' GPA.
- Statistics: Descriptive measures of a sample
 - Ex: The mean of 100 UI students' GPA.

To estimate the parameter of the underlying (population's) probability distribution, we need to perform **statistical inference**.

Descriptive Statistics

- Using statistical inference, we may start deducing parameters of an underlying probability distribution from data.
- **But**, before we perform statistical inference, we usually need to describe and summarize our data set.

This is descriptive statistics !

Descriptive Statistics



Describing
data sets



Summarizing
data sets



Standard
Scores



Chebyshev's
Inequality

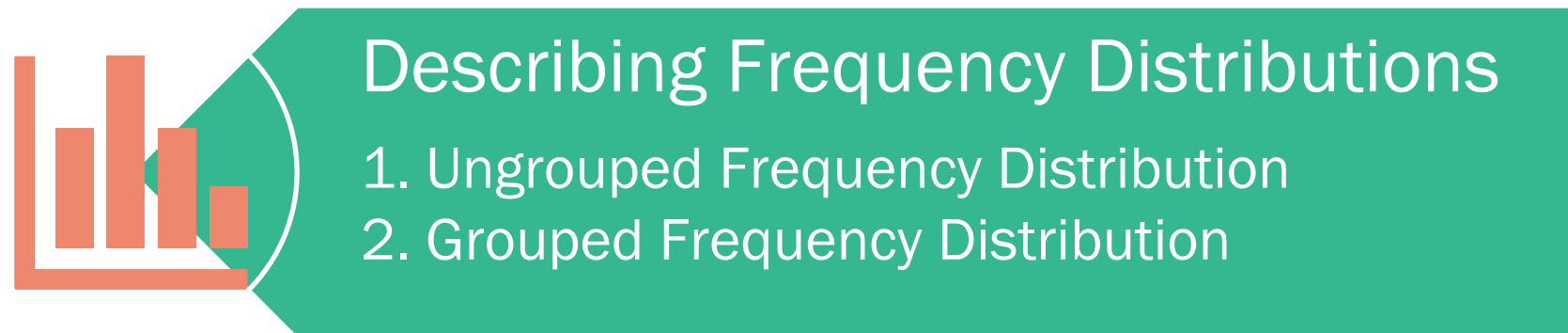
Describing Data Sets

- How would you like your steak?
 - Well-done or rare?
- How would you like your data?
 - Clear and Concise
 - Clear : make sure our meaning is **clear**
 - Concise : using the fewest words possible to convey an idea
 - So we can understand easily
 - It is recommended to present it visually

Describing Data Sets (2)

- The observed data should be presented **clearly** and **concisely**, so that observer can quickly **obtain a feel** for the **essential characteristics** of the data.
- **Tables & graphs** are particularly useful and powerful ways of presenting data.

Describing Data Sets (3)



STEM AND LEAF PLOT

Stem and Leaf Plot

- An efficient way of organizing a small- to moderate-sized data set.
 - Not for large data set !
- A plot is obtained by first dividing each data value into two parts – its stem & its leaf.

Stem and Leaf Plot (2)

- If data are all two-digit numbers, we could make the first digit as its stem, and the second digit as its leaf
- Expression for 62
 - Stem 6
 - Leaf 2

7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2	9.0, 9.8

51.0, 51.3, 52.0, ...



Stem and Leaf Plot Sample

DESCRIBING FREQUENCY DISTRIBUTIONS

1. UNGROUPED FREQUENCY DISTRIBUTION
2. GROUPED FREQUENCY DISTRIBUTION

Frequency Distribution

- Frequency distribution is a tabulation/summary that describes the **number of times** an individual score OR a group of scores occurs.
- Usual ways of presenting frequency distribution:

Frequency
table

Line
graph

Bar graph

Frequency
polygon

Histogram

Etc..

1. UNGROUPED FREQUENCY DISTRIBUTION

- The ungrouped frequency distribution is usually used for data that can be placed in specific categories (categorical data),
 - such as nominal- or ordinal-level data. [Bluman, 2012]
 - or there are relatively small number of distinct values

Frequency Table

- Suppose you purchased a bag of Skittles, You found that there are 55 candies inside.
- The distribution of Skittles color frequencies:

Color	Frequency
Purple	2
Red	18
Yellow	7
Green	7
Blue	17
Orange	4

Frequency Table (2)

- Starting yearly salaries of 42 recently graduated students
- When is the frequency table convenient?
 - When we have a relatively small number of distinct values!

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

In \$1,000

Visual Representations

- Other than the table, what types of visualizations of frequency do you know?

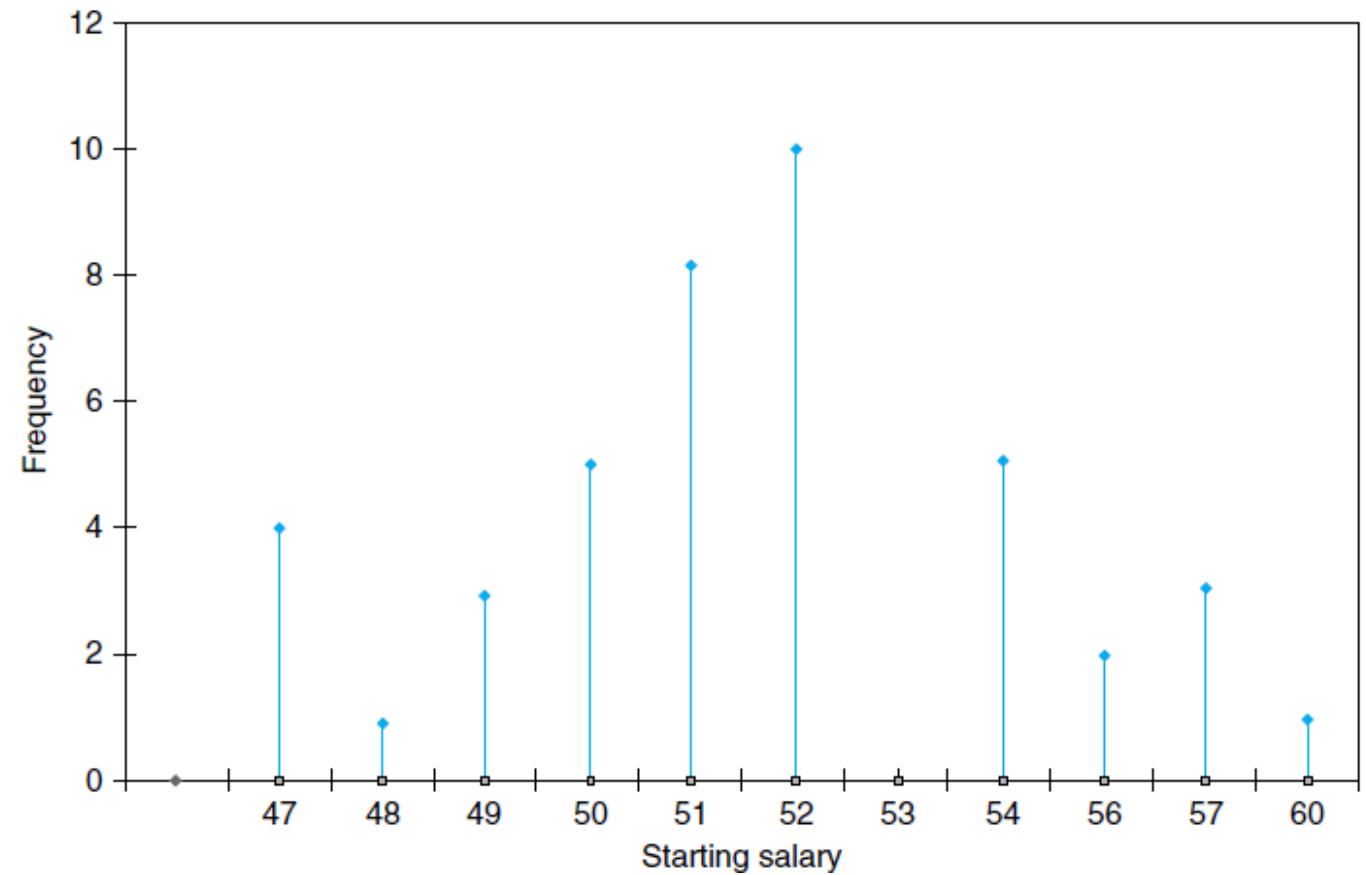
Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

In \$1,000

Line Graph

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

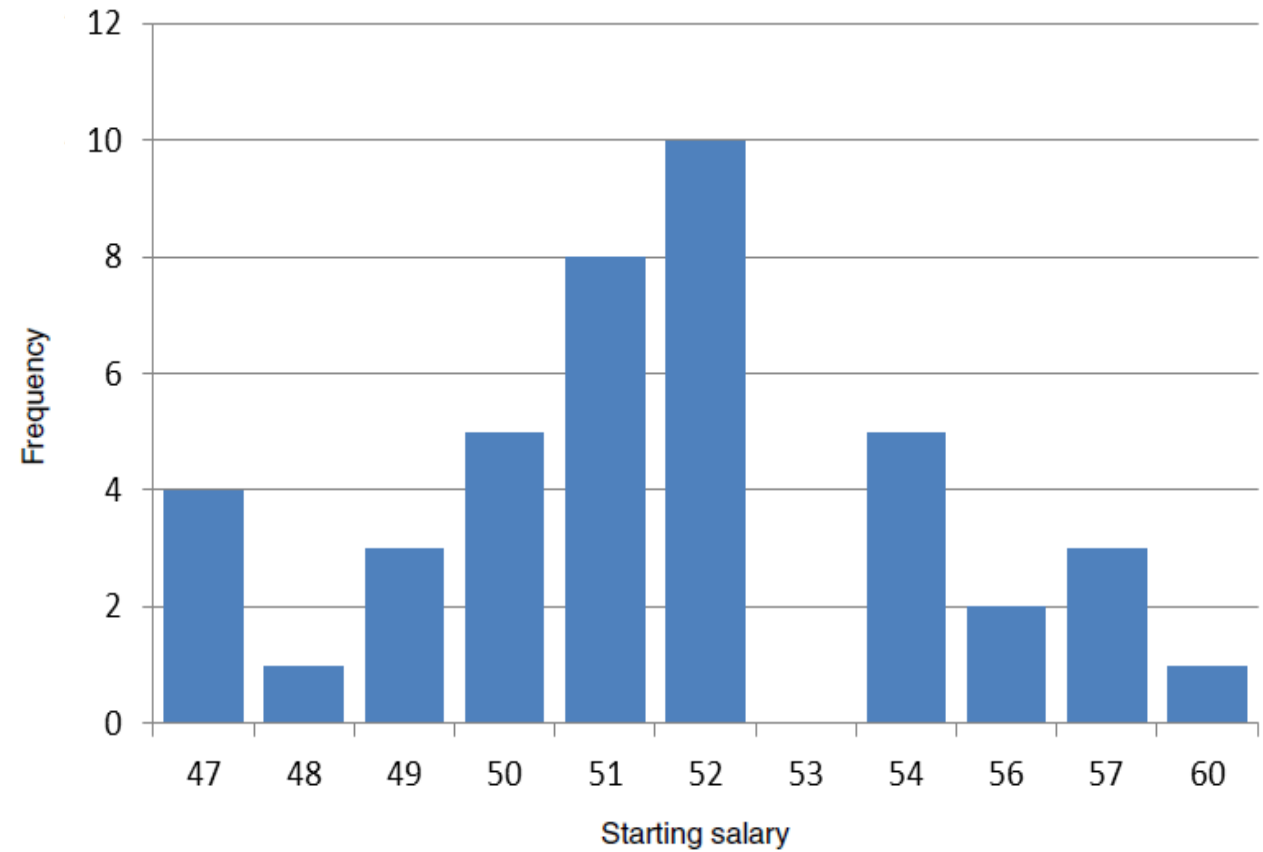
In \$1,000



Bar Graph

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

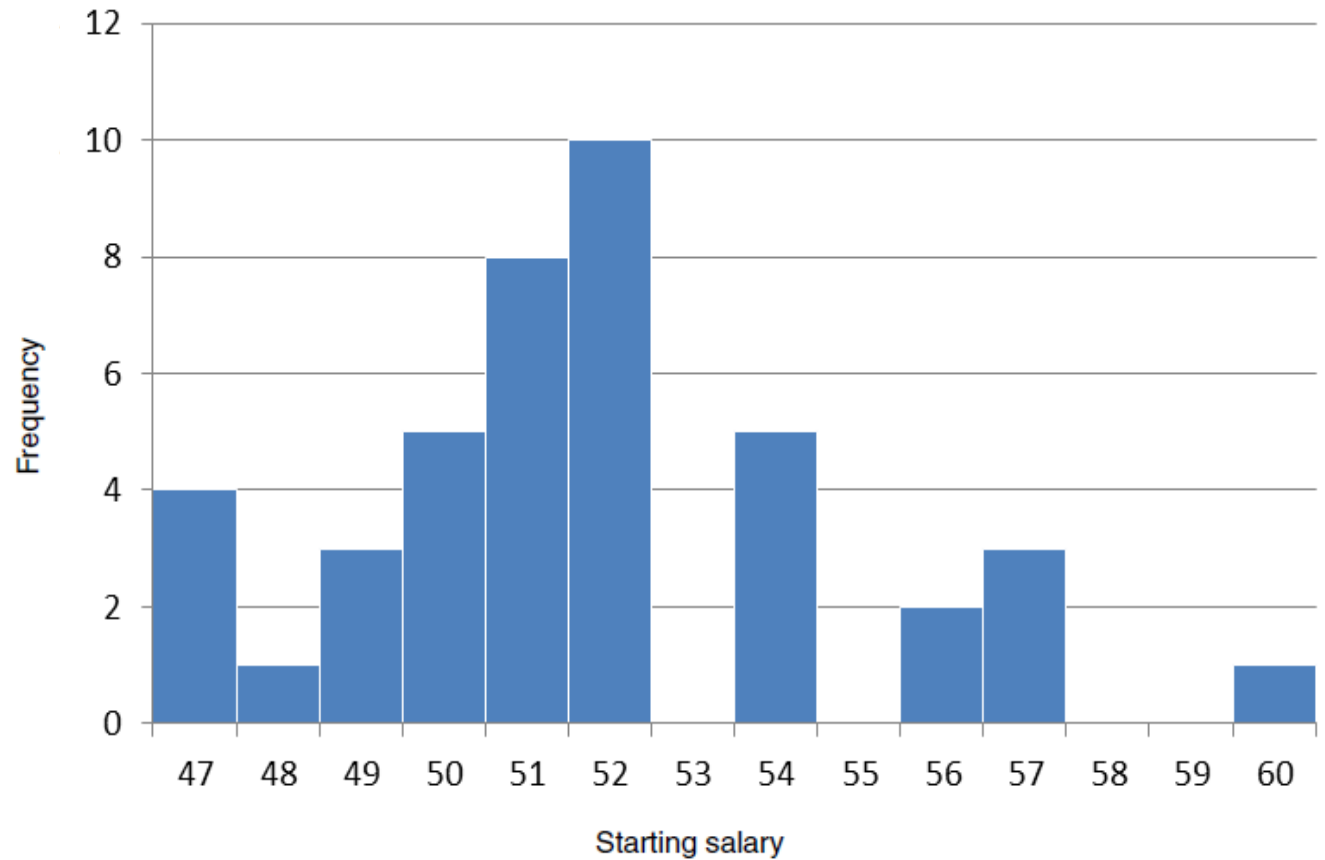
In \$1,000



Histogram

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

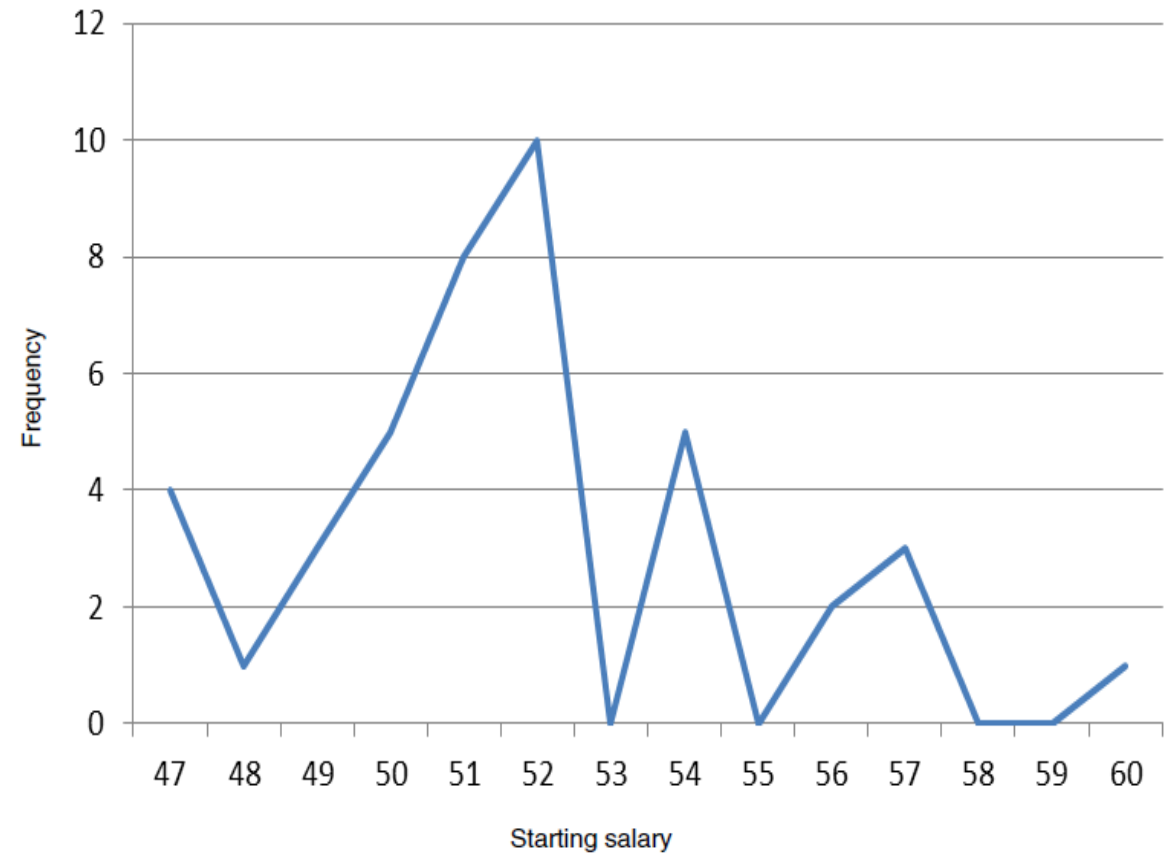
In \$1,000



Frequency Polygon

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
55	2
56	3
57	1
60	1

In \$1,000



Relative Frequency Distribution

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

In \$1,000

- For a starting salary of 47, there are 4 graduates.
- How many is 4?
- In context: 4 is out of 42

Relative Frequency Distribution (2)

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

In \$1,000

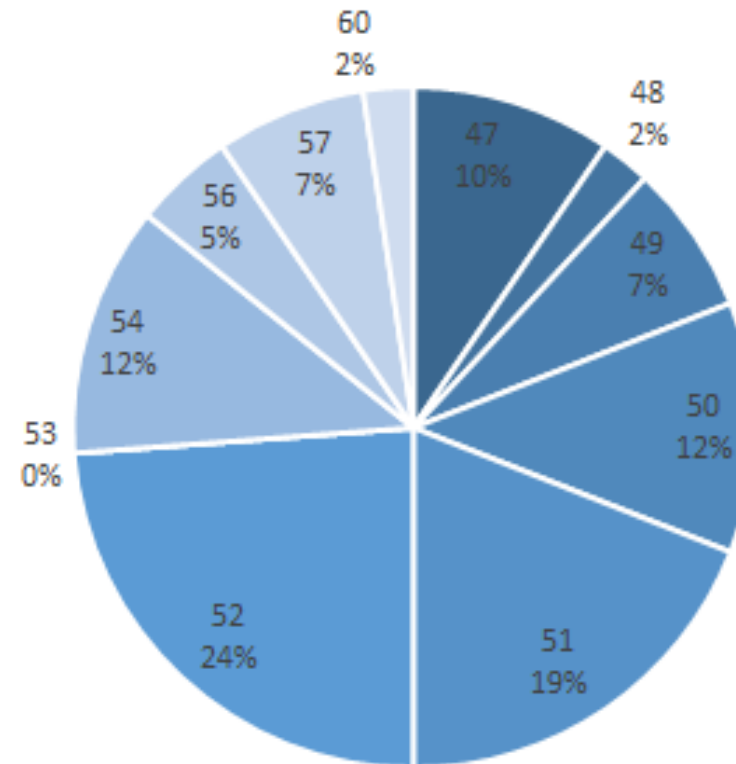
- Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio f/n is called its relative frequency.
- A relative frequency distribution presents the corresponding proportions of observations within the classes.
- Presenting relative frequency distribution:
 - Relative frequency table
 - Pie chart

Relative Frequency Table

Starting Salary	Frequency	Relative Frequency
47	4	$4/42 = .0952$
48	1	$1/42 = .0238$
49	3	$3/42$
50	5	$5/42$
51	8	$8/42$
52	10	$10/42$
53	0	0
54	5	$5/42$
56	2	$2/42$
57	3	$3/42$
60	1	$1/42$

In \$1,000

Starting Salary of 42 Recent Graduates



2. GROUPED FREQUENCY DISTRIBUTION

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

In \$1,000

Only 11
discrete
values

- Previously on ungrouped frequency distribution: we had an ordered listing of all values of a variable and their frequencies (or relative frequencies).
- How many values did we have?

Grouped Frequency Distribution

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

In \$1,000

- What if the number of distinct values is too large ?
- What if the variable is continuous?
- Group the Data

 - divide the values into groupings, or class intervals
 - and then, plot the number of data values falling in each class interval.

Grouped Data

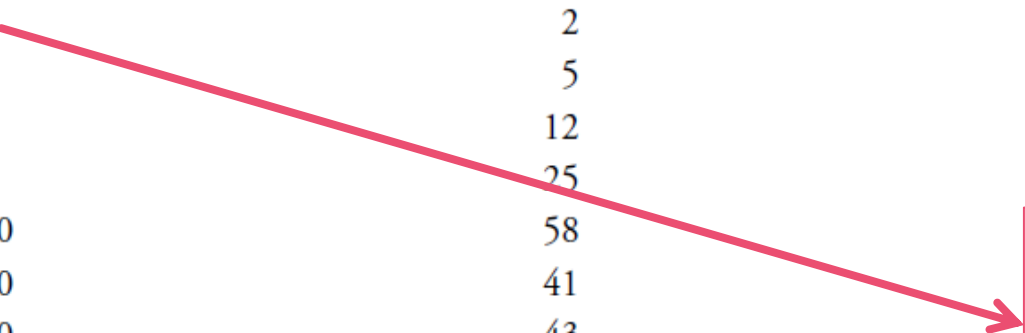
- Life (in Hours) of 200 Incandescent Lamps
- We will introduce two ways of presenting grouped frequency distribution.
 - A. Exclusive class interval with left-end inclusion convention [Ross, 2009]
 - B. Inclusive class interval [Hinkle, et al., 2003]

1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Grouped Data – Exclusive class interval with left-end inclusion convention [Ross, 2009]

- The endpoints of class interval (class boundaries) w/ **left-end inclusion!**

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

- 
- greater than or **equal** to 500
 - and less than 600
 - 600 is not included

Life in Hours of 200 Incandescent Lamps

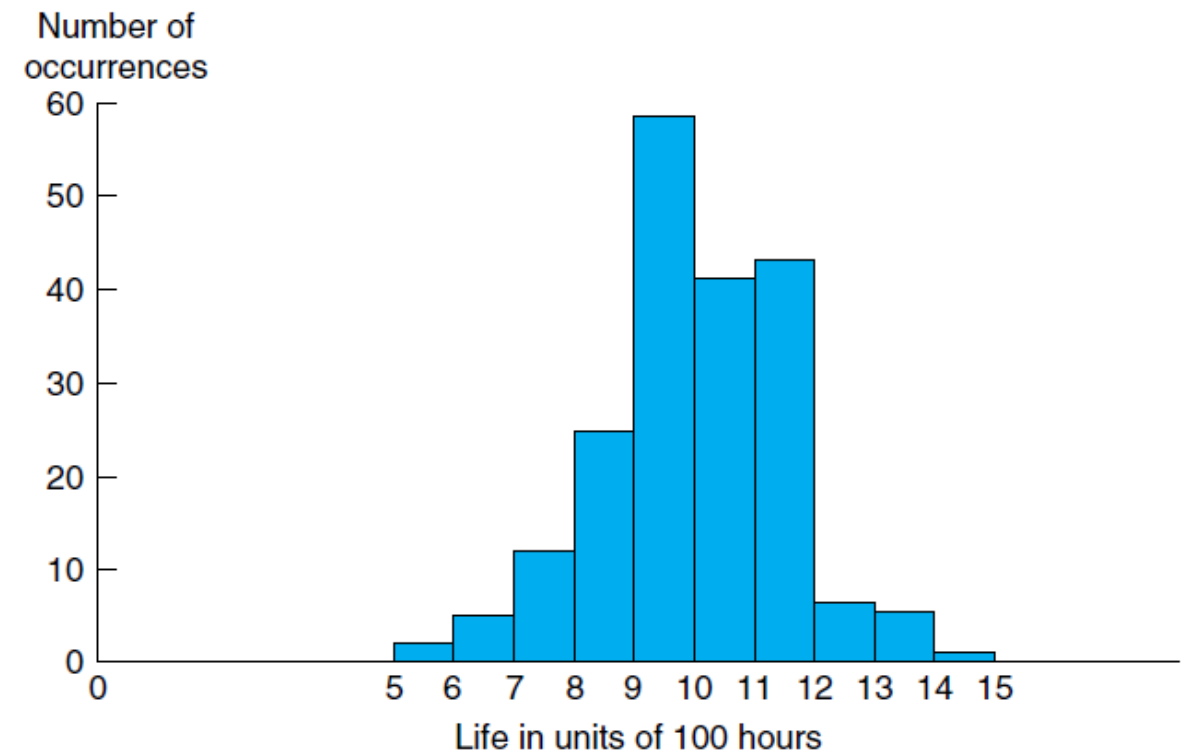
Grouped Data – Exclusive Class Interval [Ross, 2009]

Histogram

- A histogram plot of class data, with the bars placed adjacent to each other.

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

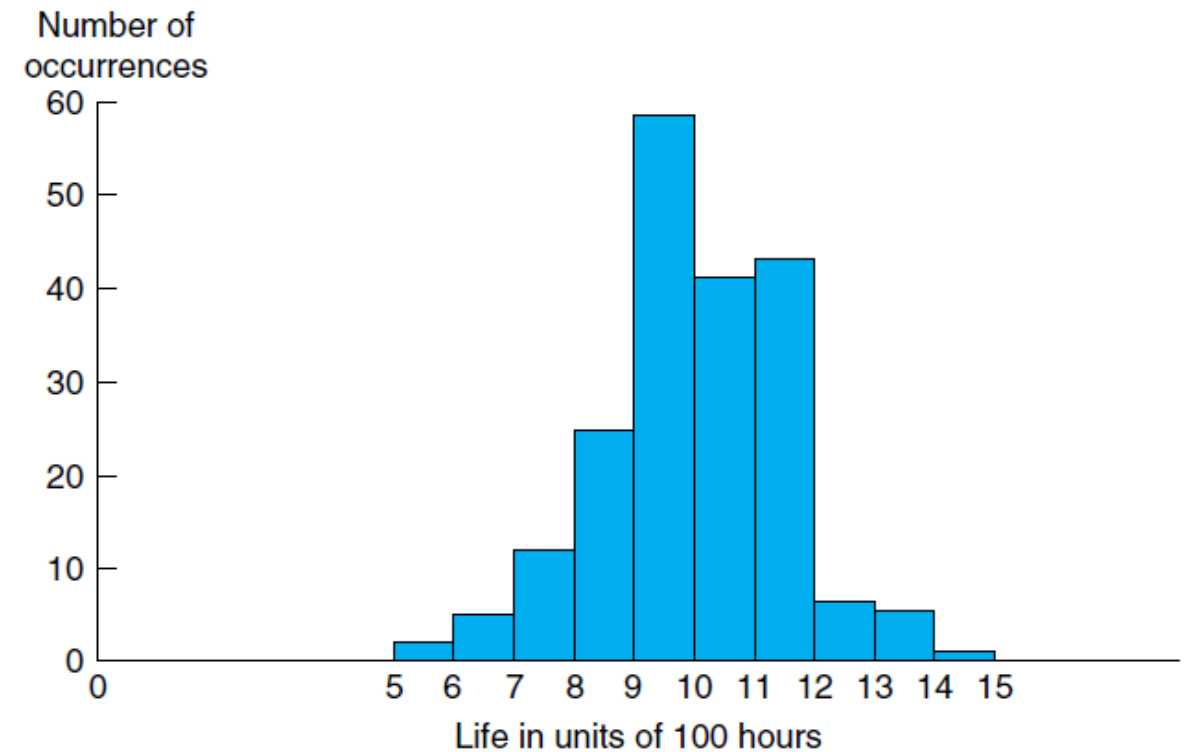
Life in Hours of 200 Incandescent Lamps



Grouped Data – Exclusive Class Interval [Ross, 2009]

Histogram (2)

- **X Axis** represents a possible data value.
- **Y Axis** gives the number (or proportion) of the data whose values are less than or equal to it.
 - **Number of data:** Frequency Histogram
 - **Proportion of data:** Relative Frequency Histogram



Grouped Data – Exclusive Class Interval [Ross, 2009]

Exercise

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

Life in Hours of 200 Incandescent Lamps

- Recall other visualization methods
 - Plot the frequency with a **polygon**
 - Plot the frequency **cumulatively**

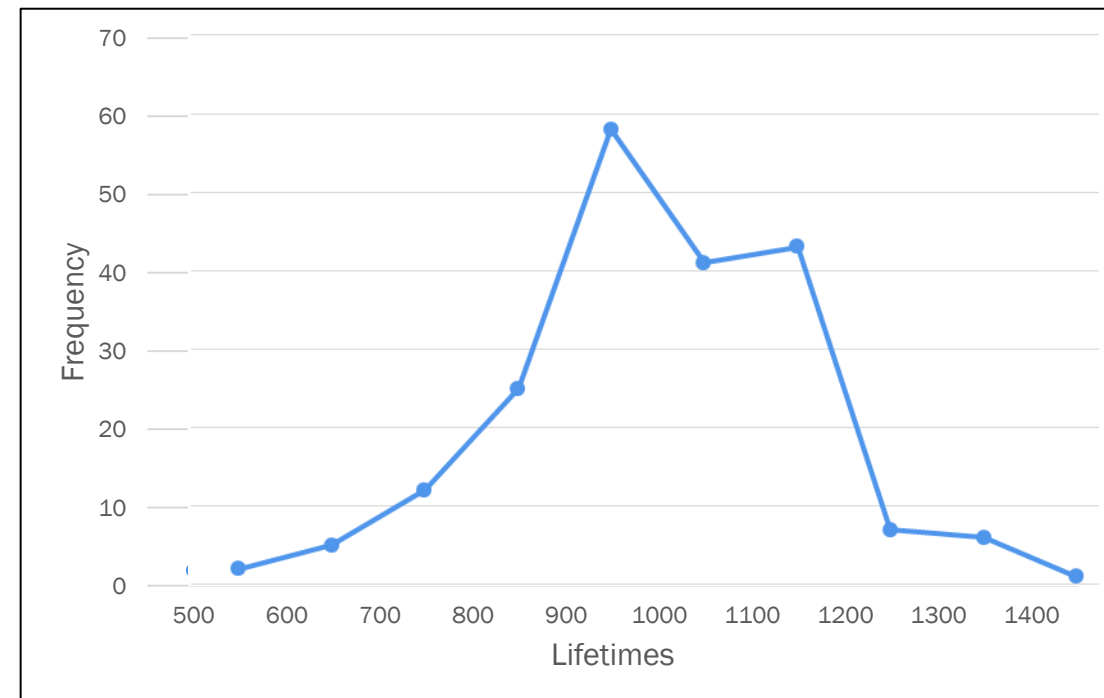
Grouped Data – Exclusive Class Interval [Ross, 2009]

Frequency Polygon

- Display the frequency of each class interval

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

Life in Hours of 200 Incandescent Lamps

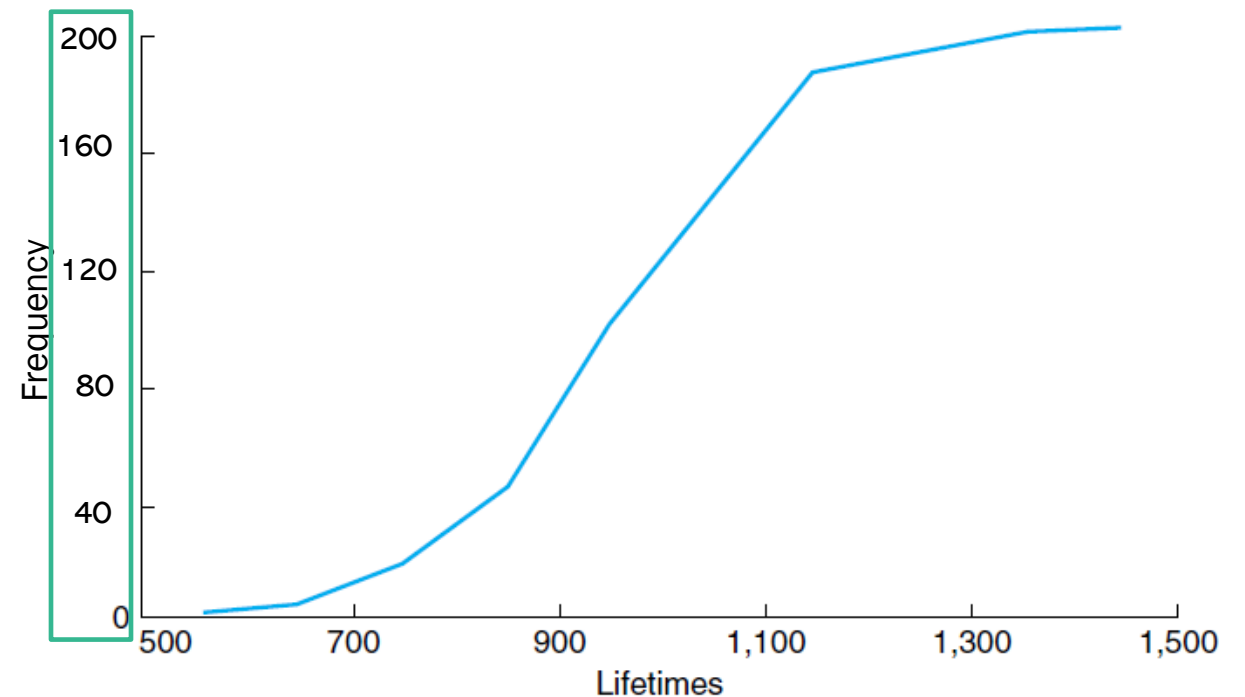


Cumulative Frequency

- The **axis** represents a possible data value. The **ordinate** gives the number of the data whose values are less than or equal to it.

Class Interval	Frequency	Cumulative Frequency
500–600	2	2
600–700	5	7
700–800	12	19
800–900	25	44
900–1000	58	102
1000–1100	41	143
1100–1200	43	186
1200–1300	7	193
1300–1400	6	199
1400–1500	1	200

Life in Hours of 200 Incandescent Lamps



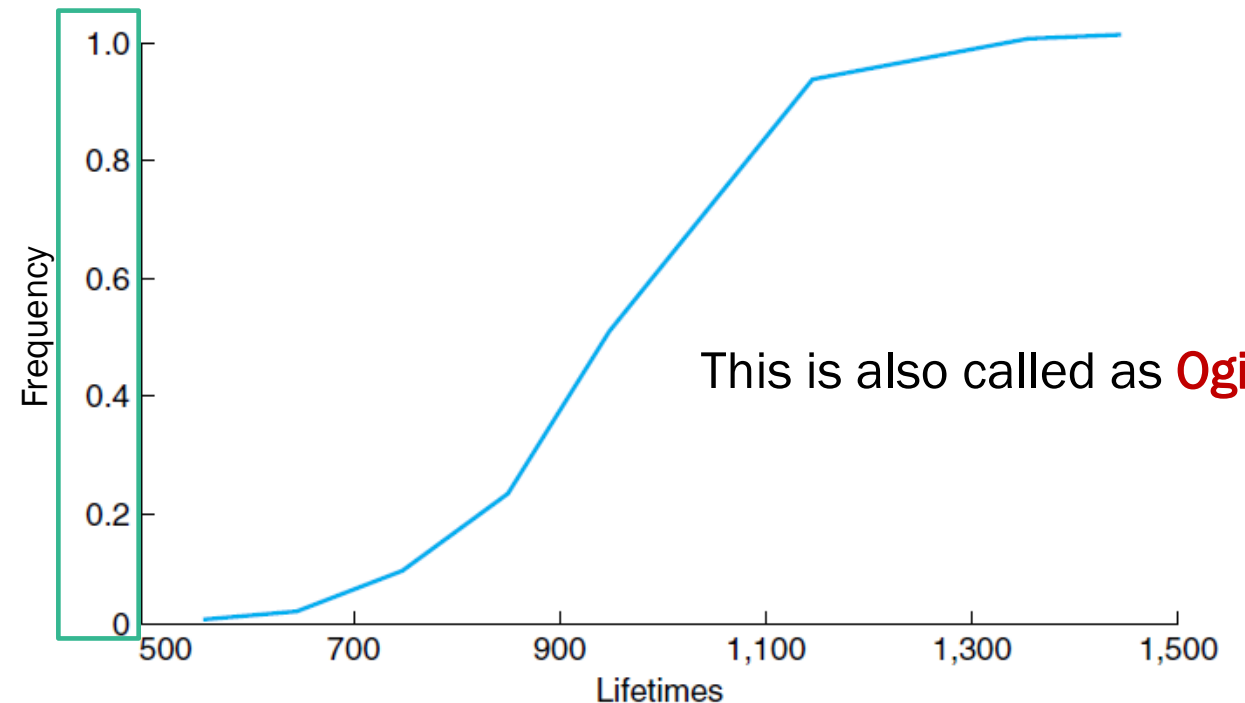
Grouped Data – Exclusive Class Interval [Ross, 2009]

Relative Cumulative Frequency

- **Axis** represents a possible data value. The **ordinate** gives the proportion of the data whose values are less than or equal to it.

Class Interval	Frequency	Relative Frequency	Relative Cumulative Frequency
500–600	2	0,010	0,010
600–700	5	0,025	0,035
700–800	12	0,060	0,095
800–900	25	0,125	0,220
900–1000	58	0,290	0,510
1000–1100	41	0,205	0,715
1100–1200	43	0,215	0,930
1200–1300	7	0,035	0,965
1300–1400	6	0,030	0,995
1400–1500	1	0,005	0,100

Life in Hours of 200 Incandescent Lamps



Grouped Data - [Hinkle, 2003]

- Final examination scores for freshman psychology students
 - How is it different from [Ross, 2009]?

Class interval of width 5

Midpoint: the point halfway through the interval

Class Interval	Midpoint	Frequency	Class Interval	Midpoint	Frequency
65-69	67	6	40-44	42	22
60-64	62	15	35-39	37	18
55-59	57	37	30-34	32	7
50-54	52	30	25-29	27	2
45-49	47	42	20-24	22	1

All scores between 20 and 24 inclusive! 

Grouped Data - [Hinkle, 2003] (2)

■ Disadvantage:

- This table no longer specifies the exact number of students & scores.
- It tells us only that there are six scores in the interval 65 – 69.

Class Interval	Midpoint	Frequency	Class Interval	Midpoint	Frequency
65-69	67	6	40-44	42	22
60-64	62	15	35-39	37	18
55-59	57	37	30-34	32	7
50-54	52	30	25-29	27	2
45-49	47	42	20-24	22	1

Grouped Data - [Hinkle, 2003] (3)

- We use the notion of exact limits to make continuous class interval :
 - Exact limits of a score extend from one-half unit below to one-half unit above the recorded score → exclusive class interval
 - E.g., a score of 53 represents a score somewhere between 52.5 and 53.5
- The score within any class interval are assumed to be **uniformly distributed** throughout the interval
- Represented by the **midpoint**.

Grouped Data - [Hinkle, 2003] (4)

TABLE 3.2

Frequency Distribution of Final Examination Scores, Including Cumulative Frequencies and Cumulative Percents

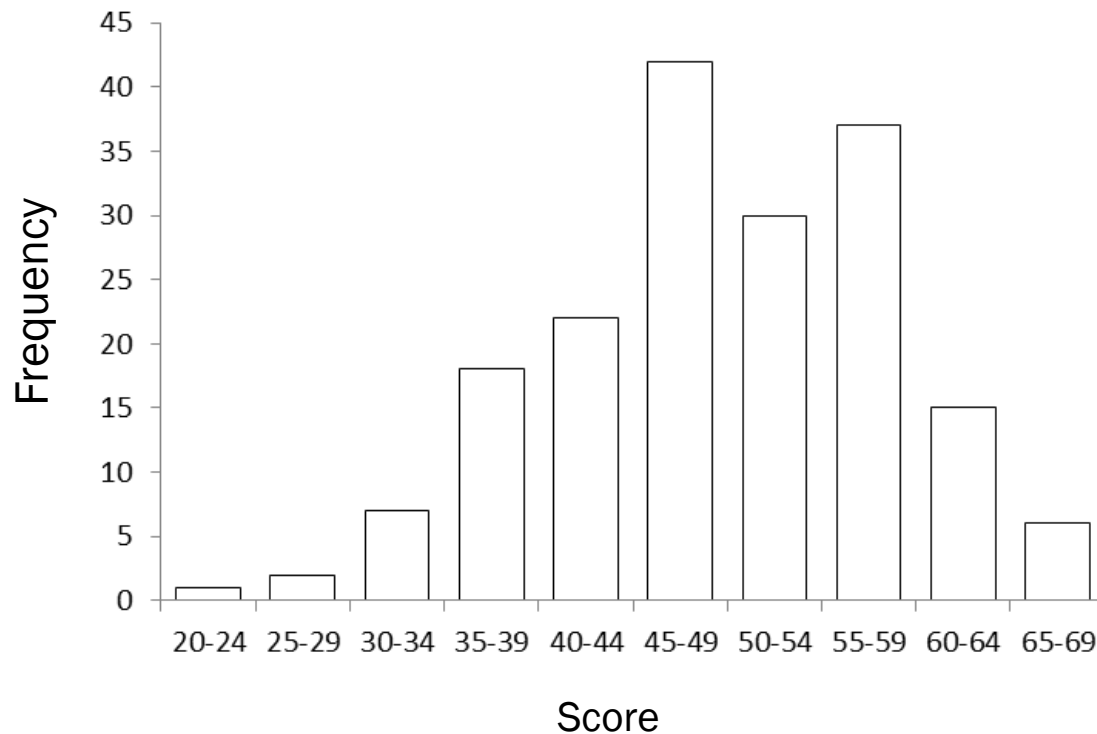
<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>	<i>%</i>	<i>c%</i>
65–69	64.5–69.5	67	6	180	3.33	100.00
60–64	59.5–64.5	62	15	174	8.33	96.67
55–59	54.5–59.5	57	37	159	20.56	88.34
50–54	49.5–54.5	52	30	122	16.67	67.78
45–49	44.5–49.5	47	42	92	23.33	51.11
40–44	39.5–44.5	42	22	50	12.22	27.78
35–39	34.5–39.5	37	18	28	10.00	15.56
30–34	29.5–34.5	32	7	10	3.89	5.56
25–29	24.5–29.5	27	2	3	1.11	1.67
20–24	19.5–24.5	22	1	1	0.56	0.56

Score Limits

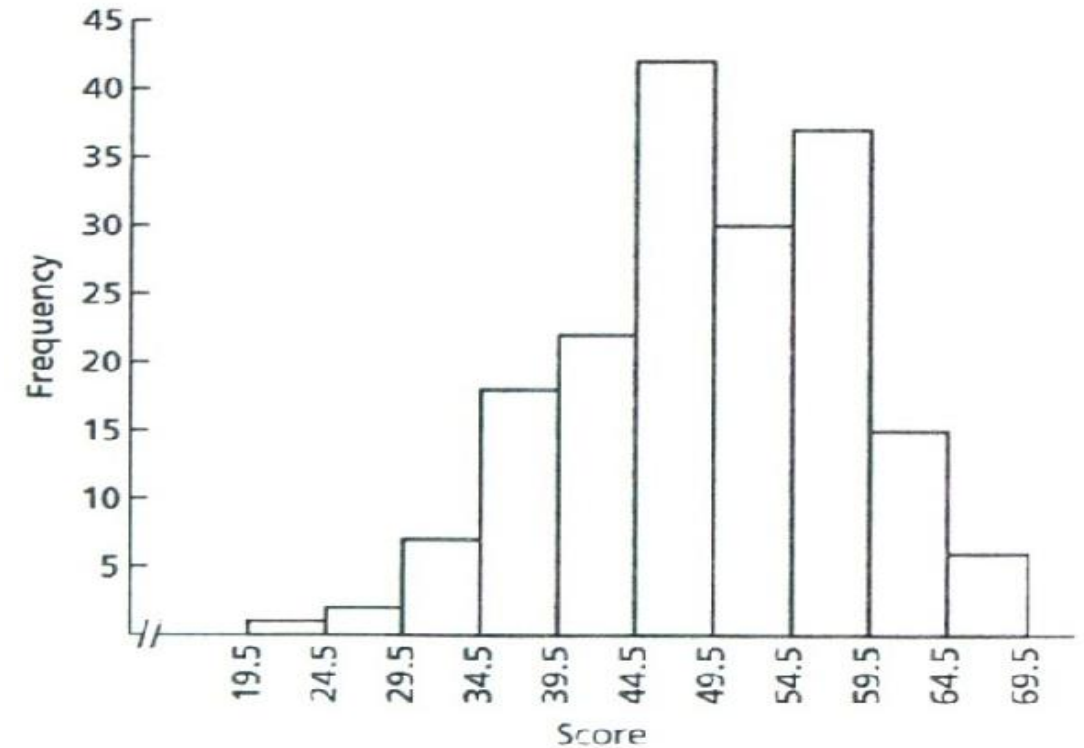


Grouped Data – Frequency Distribution

■ Bar Graph

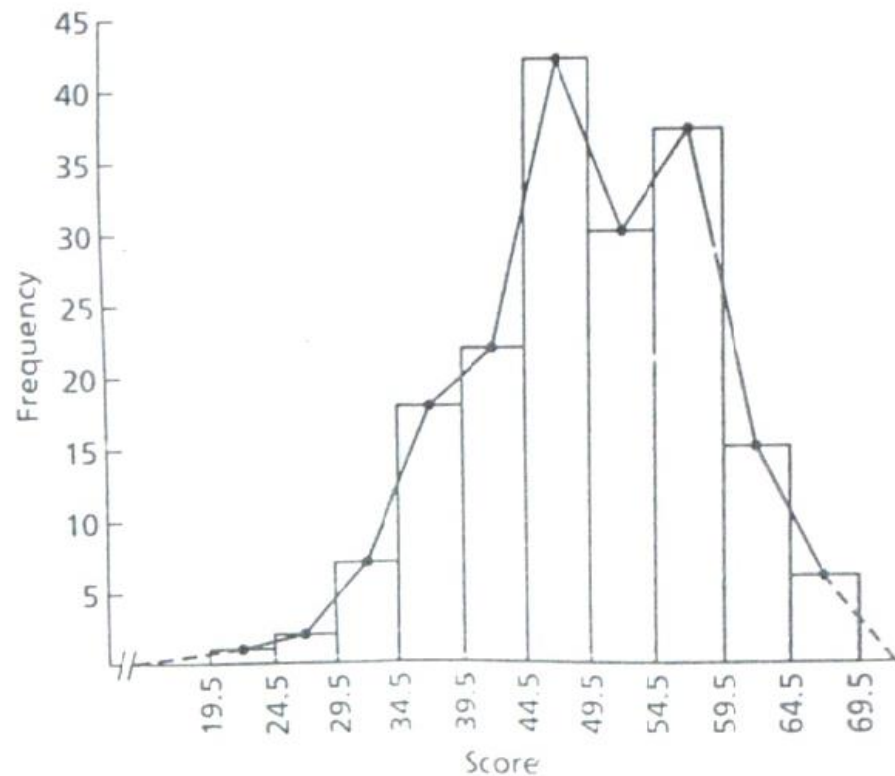


■ Histogram

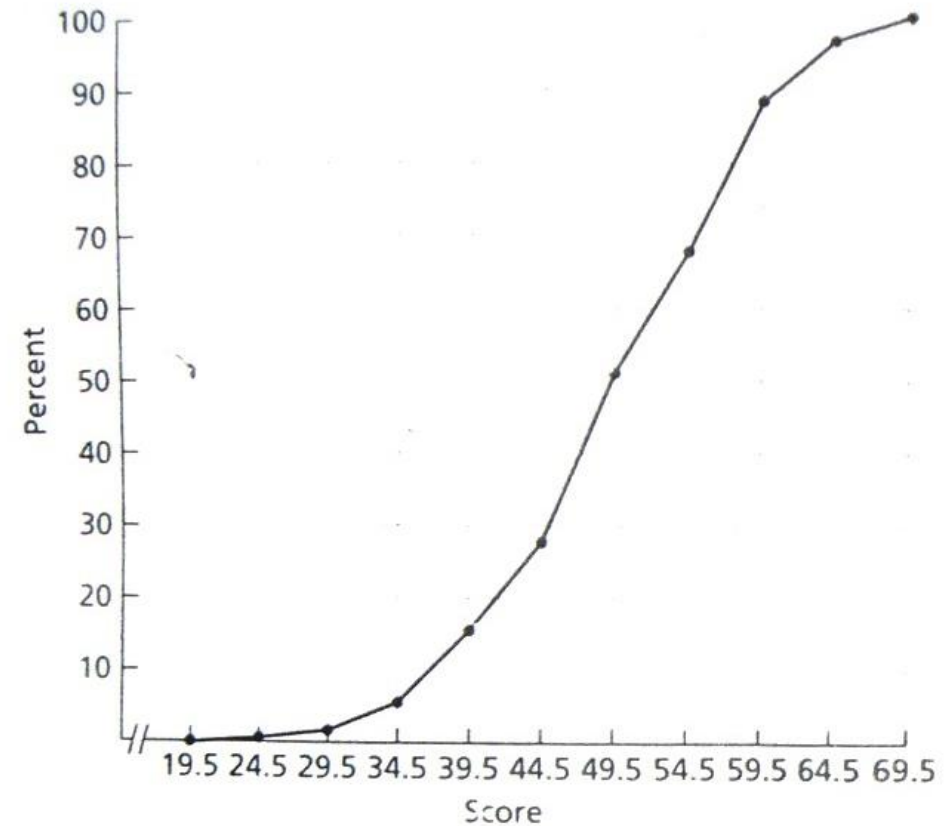


Grouped Data – plot using exact limits

■ Histogram and frequency polygon



■ Ogive



Discuss

- How should we choose the **number of class intervals**?



Too few

losing too much information about the actual data values in a class.



Too many

each class's frequency is too small; losing the data pattern.

Two General Rules for Class Intervals

- Number of intervals
 - For large data sets (>100 observations) with a wide range of scores, 10 to 20 intervals are common.
 - For smaller data sets, 6 to 12 intervals work well.
- The width of the class interval should be an odd number, whenever possible.
 - So, the midpoint of the interval will be a whole number.
 - Midpoint is the point halfway through the interval.
 - This rule makes computation easier (no need to compute midpoint)

Discuss

- How should we choose the **type** of class intervals?



Discrete Variable

We can use both **exclusive** and **inclusive** class interval.



Continuous Variable

Use **exclusive** class interval.

Shapes of Frequency Distributions

Uniform
frequency
distribution

Normal
frequency
distribution

Skewed
frequency
distribution

Bimodal
frequency
distribution

Symmetric
frequency
distribution

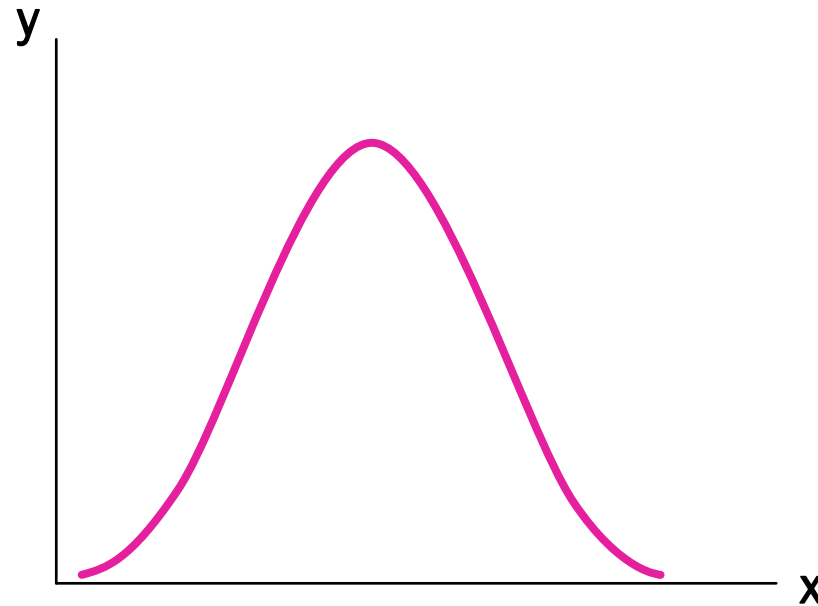
Uniform Frequency Distribution

- The scores are uniformly distributed between an interval



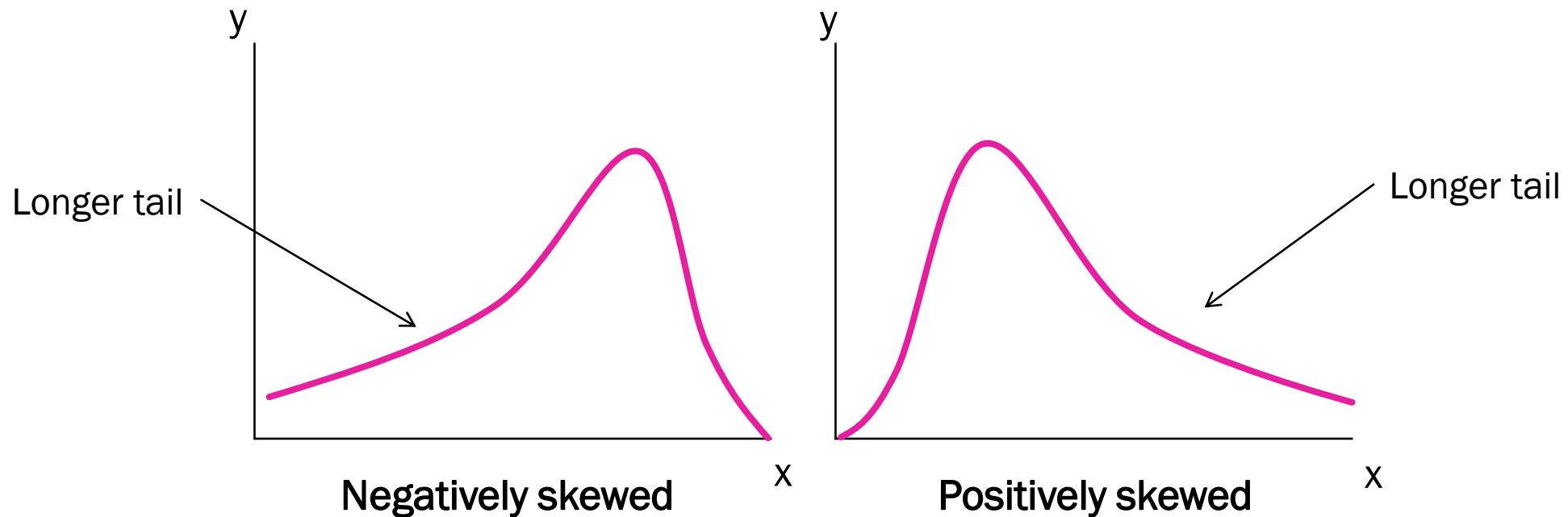
Normal Frequency Distribution

- This distribution often reaches their peak at the **median**.
- Bell-shaped symmetric fashion, with one peak (**unimodal**)



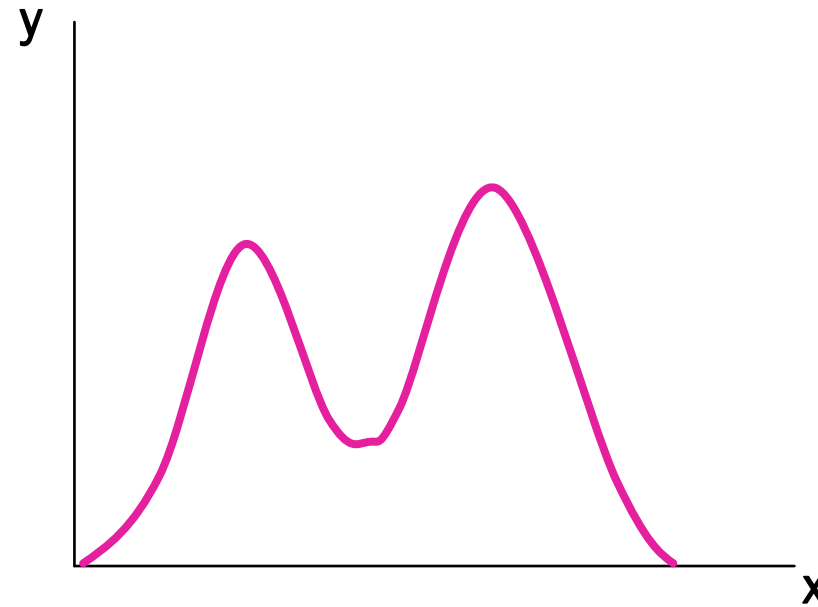
Skewed Frequency Distribution

- Most likely, there are many more **outliers** on the longer tail area
- Skewed to the right = positive skew, skewed to the left = negative skew



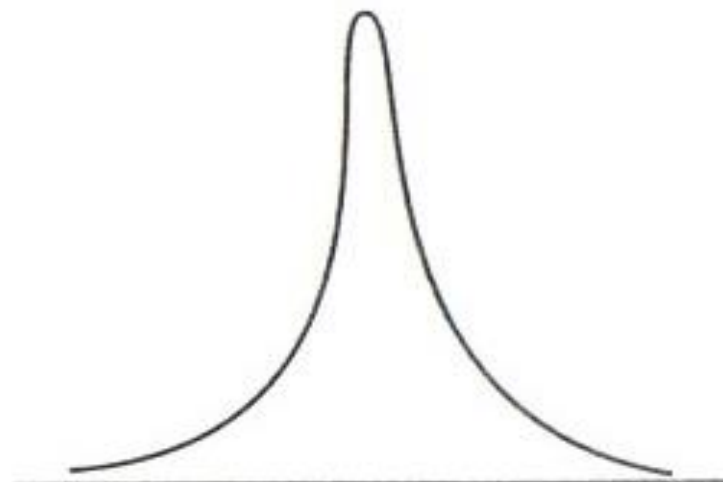
Bimodal Frequency Distribution

- It has two peaks
- Indicates two separate sub-populations in the study with different characteristics.

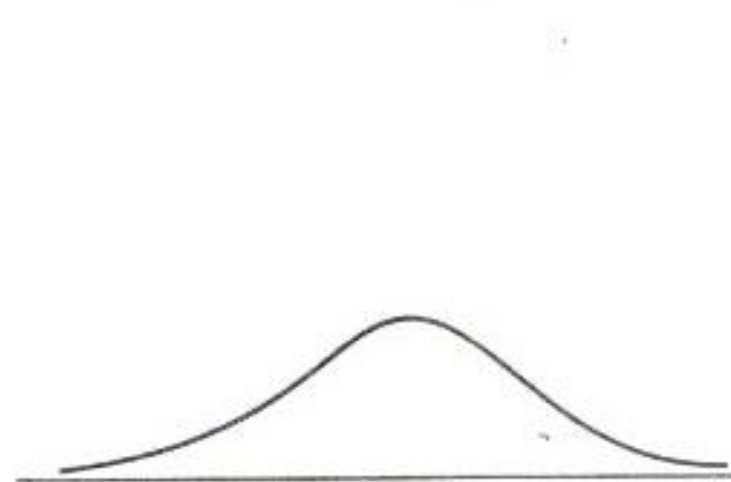


Symmetric Frequency Distributions

- Symmetric distributions such as the normal distribution may vary in **kurtosis** – degree of peakedness.
 - **Leptokurtic**: if the vast majority of the scores tend to be located at the center.
 - **Platykurtic**: if scores are distributed more uniformly, yet many scores still cluster at the center.



E. Leptokurtic distribution



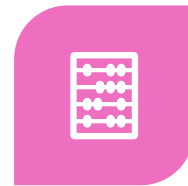
F. Platykurtic distribution

SUMMARIZING DATA SETS

1. MEASURES OF CENTRAL TENDENCY
2. MEASURES OF VARIATION
3. MEASURES OF POSITION

1. MEASURES OF CENTRAL TENDENCY

- Where are the majority of scores concentrated?
- Statistics that are used for describing the center of a set of data values:



MEAN



MEDIAN



MODE

Mean

- Mean is the arithmetic average of the scores in distribution.
- Population Mean

$$\mu = \frac{\sum x_i}{N}$$

- μ is for mean of population
- N is size of the population

- Sample Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

- \bar{x} is for mean of sample
- n is size of the sample

Sample Mean

- Let $x_1, x_2, x_3, \dots, x_n$ are n numerical values of our data set, then the sample mean, denoted by \bar{x} , is defined by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Sample Mean (2)

$$y_i = ax_i + b$$

$$\bar{y} = \sum_{i=1}^n \frac{ax_i + b}{n}$$

$$\bar{y} = \sum_{i=1}^n \frac{ax_i}{n} + \sum_{i=1}^n \frac{b}{n}$$

$$\bar{y} = a\bar{x} + b$$

$$\bar{x} = \frac{\bar{y} - b}{a}$$

- Modified data; multiply with a **constant a** and add with a **constant b** .
- The constants, a and b , will impact the *mean* of the modified data.
- **Relatively simplify the calculation of the mean.**

Example

- Find the sample mean of the following scores (The winning scores in the U.S. Masters golf tournament 1999-2008).

$$\{280, 278, 272, 276, 281, 279, 276, 281, 289, 280\}$$

- It is easy to first subtract 280 from these values, $y_i = x_i - 280$

$$\{0, -2, -8, -4, 1, -1, -4, 1, 9, 0\}$$

- It is easy to determine the mean of y_i 's, i.e $\bar{y} = -0.8$
- So, the mean of original data is,

$$\bar{x} = \bar{y} + 280 = 279.2$$

Exercise

- Find the sample mean of the following Statprob Scores.

$\{90, 87, 85, 92, 90, 86, 98, 95, 91, 81\}$

Exercise

- Find the sample mean of the following Statprob Scores.

$\{90, 87, 85, 92, 90, 86, 98, 95, 91, 81\}$

- Mean = 89.5

Mean in Class Intervals

- Mean for data distribution that are grouped into class intervals (in **grouped frequency table**).

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

- m_i = mid-point of i^{th} interval.
- f_i = frequency of i^{th} interval.

Properties of the Mean

1. The sum of deviations of all scores from the mean is zero.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

- Prove it using the Statprob scores:
 - {90, 87, 85, 92, 90, 86, 98, 95, 91, 81}
 - Mean $\bar{x} = 89.5$

Properties of the Mean (2)

2. The **sum of squares** of the deviation from the mean is **smaller than** the sum of squares of the deviation from any other value in the distribution.

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - m)^2, m \in R$$

- Prove it using the Statprob scores:
 - {90, 87, 85, 92, 90, 86, 98, 95, 91, 81}
 - Mean = 89.5

Properties of the Mean (3)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \overline{95})^2$
90	0.5	0.25	25
87	-2.5	6.25	64
85	-4.5	20.25	100
92	2.5	6.25	9
90	0.5	0.25	25
86	-3.5	12.25	81
98	8.5	72.25	9
95	5.5	30.25	0
91	1.5	2.25	16
81	-8.5	72.25	196
Σ	0	222.5	525

Sample Median

- Order the values of a data set of size n from smallest to largest.
 - If n is odd, the sample median is the value in position $(n + 1)/2$
 - if n is even, the sample median is the average of the values in positions $n/2$ and $n/2 + 1$.
- **Median is actually second quartile.**
- Example
 - $\{3, 6, 12, 18, 19, 21, 23\} \rightarrow \text{median} = 4\text{th datum} = 18.$
 - $\{3, 6, 12, 18, 19, 21, 23, 25\} \rightarrow \text{median} = (18 + 19) / 2 = 18.5$

Sample Median (2)

- For grouped frequency table

$$Mdn = ll + \left(\frac{n(0.50) - cf}{f_i} \right) (w)$$

- ll : lower exact limit of the interval containing the $n(0.50)$ score
- n : total number of score
- cf : cumulative freq. of scores below the interval containing the $n(0.50)$ score
- f_i : freq. of scores in the interval containing the $n(0.50)$ score
- w : width of class interval

For left-end-inclusion [Ross, 2009] case, lower limit of an interval is the left-interval-bound

Exercise: Sample Median

<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>
65–69	64.5–69.5	67	6	180
60–64	59.5–64.5	62	15	174
55–59	54.5–59.5	57	37	159
50–54	49.5–54.5	52	30	122
45–49	44.5–49.5	47	42	92
40–44	39.5–44.5	42	22	50
35–39	34.5–39.5	37	18	28
30–34	29.5–34.5	32	7	10
25–29	24.5–29.5	27	2	3
20–24	19.5–24.5	22	1	1

$$Mdn = ll + \left(\frac{n(0.50) - cf}{f_i} \right) (w)$$

Exercise: Sample Median

<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>
65–69	64.5–69.5	67	6	180
60–64	59.5–64.5	62	15	174
55–59	54.5–59.5	57	37	159
50–54	49.5–54.5	52	30	122
45–49	44.5–49.5	47	42	92
40–44	39.5–44.5	42	22	50
35–39	34.5–39.5	37	18	28
30–34	29.5–34.5	32	7	10
25–29	24.5–29.5	27	2	3
20–24	19.5–24.5	22	1	1

$$Mdn = ll + \left(\frac{n(0.50) - cf}{f_i} \right) (w)$$

$$Med = 44.5 + \left(\frac{90 - 50}{42} \right) (5) = 49.26$$

Mean vs Median

- Mean is highly sensitive to outliers !
- Suppose we have a data set consisting 4 persons' weight:

{60, 70, 80, 990}

- The mean of this sample is

$$\frac{(60 + 70 + 80 + 990)}{4} = 300$$

- The mean 300 fails to present a realistic picture of the major part of the data. 990 seems to be an **outlier** !

Mean vs Median (2)

- We need another statistic → the median.
- For the data set consisting 4 persons' weight:

{60, 70, 80, 990}

- The median of this sample is

$$\frac{(70 + 80)}{2} = 75$$

- In this case, 3 observations out of 4 lie between 60-80, so the median is a good statistic here.

Sample Mode

- Mode is the **most frequent score** in a distribution.
- Calculate frequency distribution

Score	f
783	6
785	4
786	2
788	2
789	2
790	2
791	3
792	2

783 is the most frequent score (6 times)
Mode of the data is 783

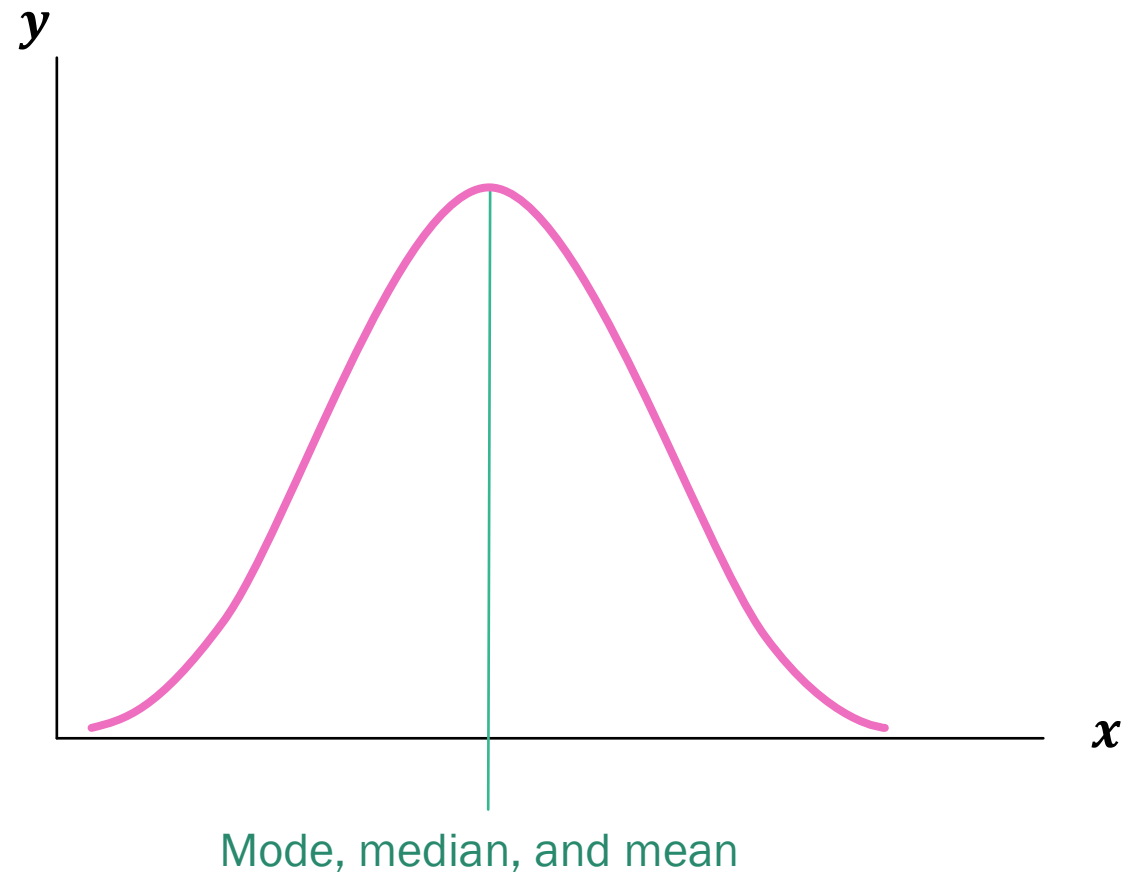
Sample Mode

- When data are grouped into class intervals [Hinkle, 2003], the mode is a modal interval. And the midpoint of this interval is considered the mode.

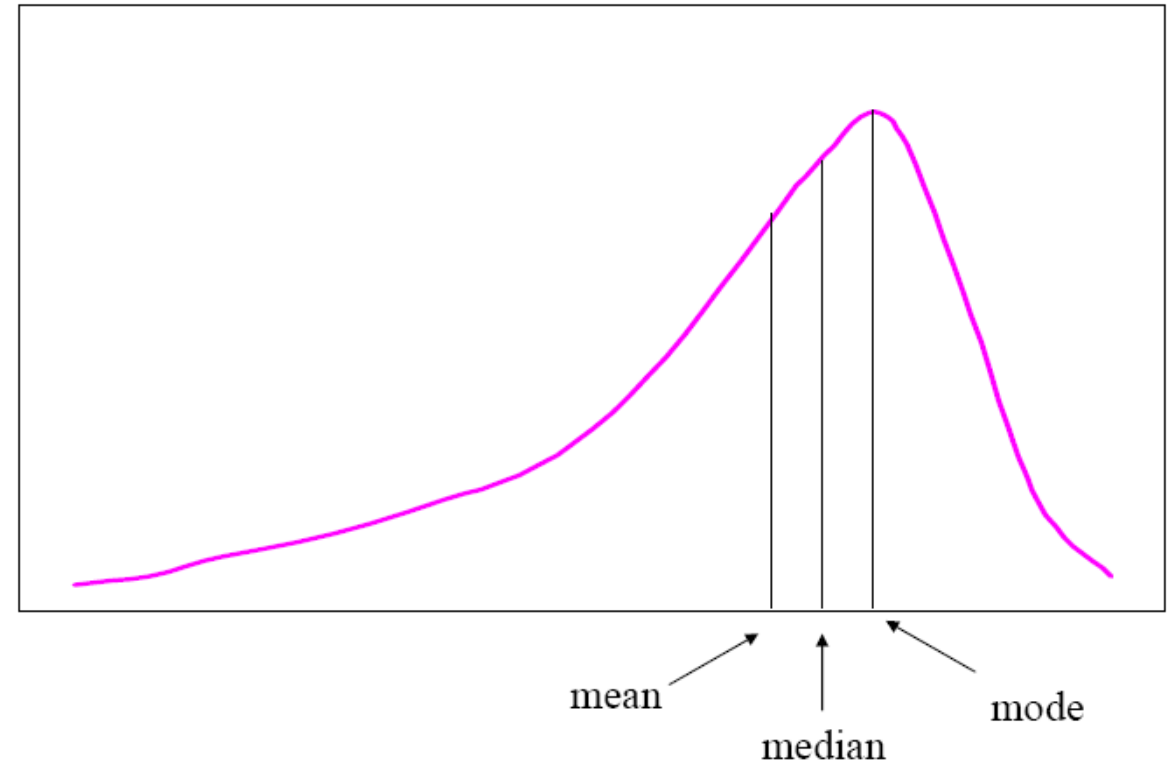
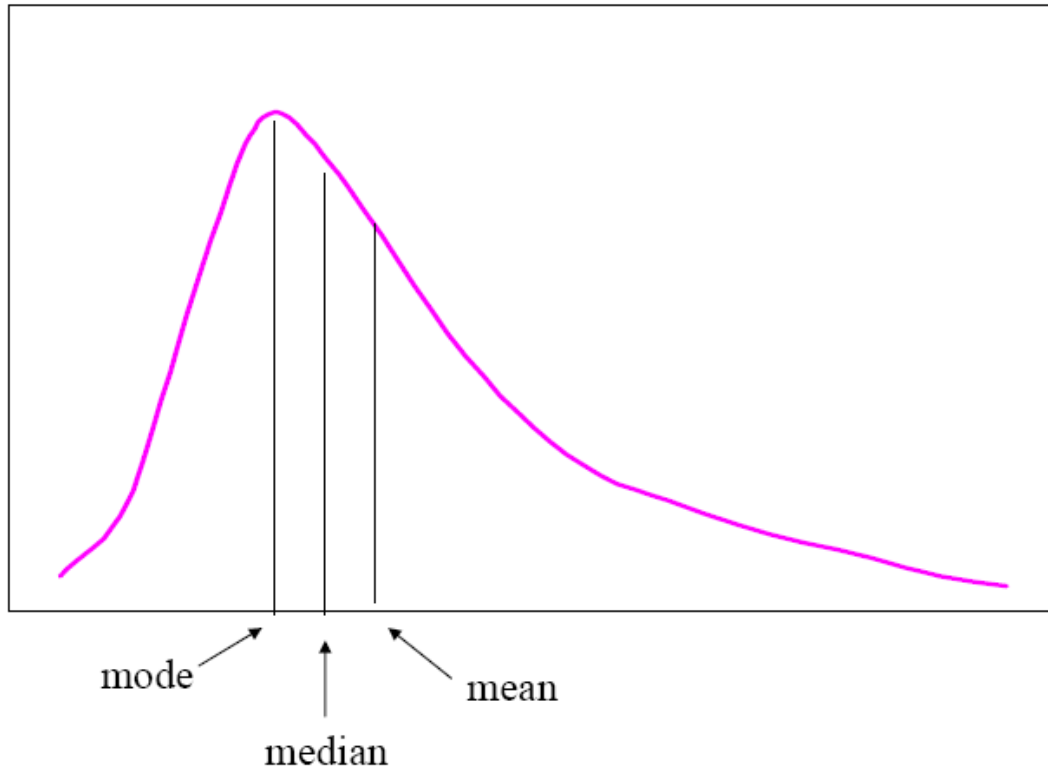
<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>
65-69	64.5-69.5	67	6	180
60-64	59.5-64.5	62	15	174
55-59	54.5-59.5	57	37	159
50-54	49.5-54.5	52	30	122
45-49	44.5-49.5	47	42	92
40-44	39.5-44.5	42	22	50
35-39	34.5-39.5	37	18	28
30-34	29.5-34.5	32	7	10
25-29	24.5-29.5	27	2	3
20-24	19.5-24.5	22	1	1

Modal interval is interval 45-49.
Mode of the data is 47

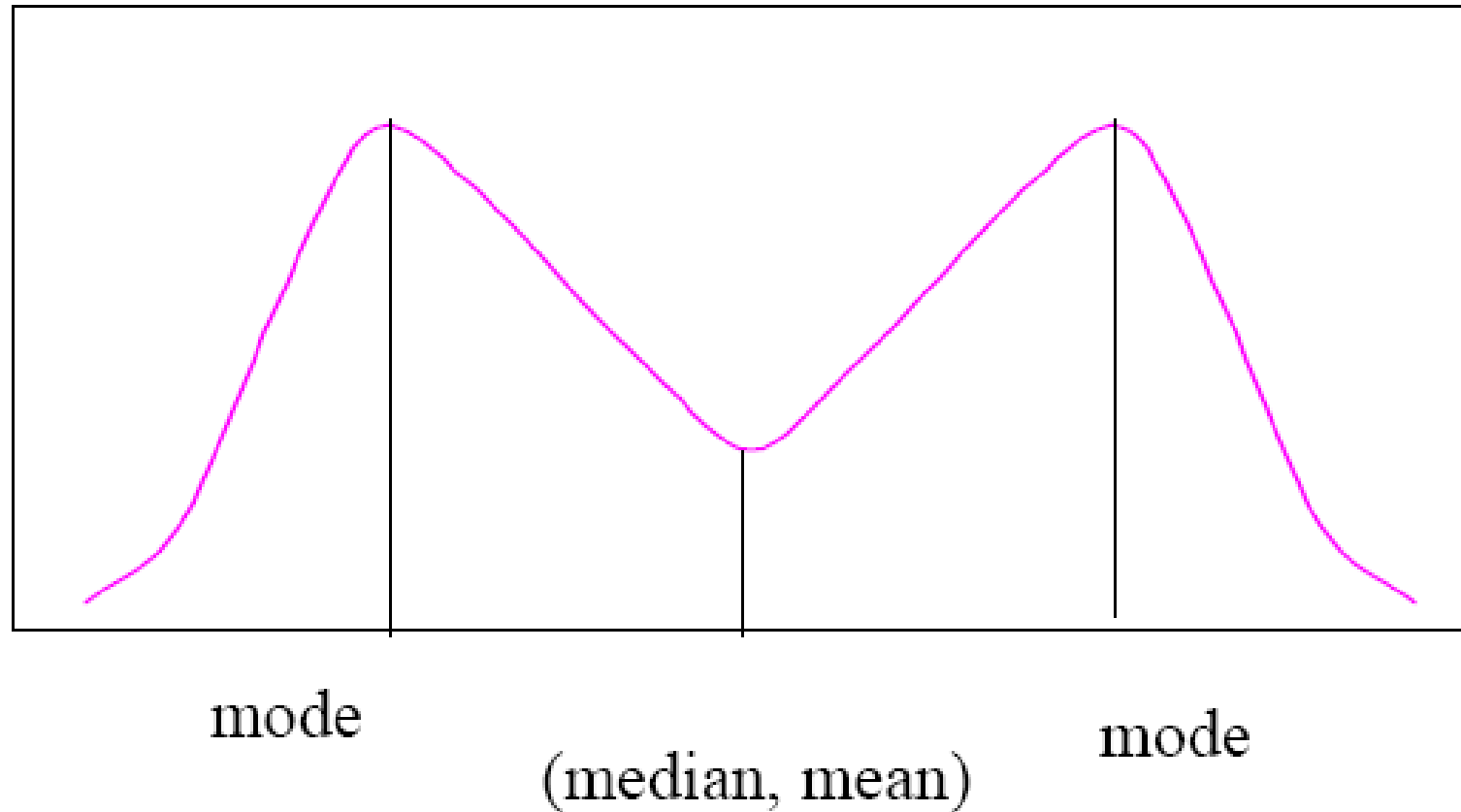
Mode, Median, and Mean in Normal Distribution



Mode, Median, and Mean in Skewed Distribution



Mode, Median, and Mean in Bimodal Distribution



2. MEASURES OF VARIABILITY

- How widely are scores spread throughout the distribution ?
- These statistics measure how much our variables vary from the mean.
- The measures of variation to be discussed:



RANGE



MEAN
DEVIATION



VARIANCE



STANDARD
DEVIATION

Range

- Range is the number of units on the scale of measurement that include the highest and lowest values.

$$\text{Range} = (\text{highest score} - \text{lowest score}) + 1 \text{ unit}$$

- Sample: ordered data:

Distribution 1	11	16	18	...	31	37
Distribution 2	18	19	21	...	26	29

- Compute the range of both distributions!

Range

- Range is the number of units on the scale of measurement that include the highest and lowest values.

$$\text{Range} = (\text{highest score} - \text{lowest score}) + 1 \text{ unit}$$

- Sample: ordered data:

Distribution 1	11	16	18	...	31	37
Distribution 2	18	19	21	...	26	29

- Distribution 1 = $37 - 11 + 1 = 27$
- Distribution 2 = $29 - 18 + 1 = 12$

Range

- Grouped frequency distribution, e.g. inclusive class interval, uses the class interval limits as the highest and lowest scores.

<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>
65–69	64.5–69.5	67	6	180
60–64	59.5–64.5	62	15	174
55–59	54.5–59.5	57	37	159
50–54	49.5–54.5	52	30	122
45–49	44.5–49.5	47	42	92
40–44	39.5–44.5	42	22	50
35–39	34.5–39.5	37	18	28
30–34	29.5–34.5	32	7	10
25–29	24.5–29.5	27	2	3
20–24	19.5–24.5	22	1	1

$$\begin{aligned}
 \text{Range} &= (\text{highest score} - \text{lowest score}) + 1 \text{ unit} \\
 &= (69 - 20) + 1 = 50
 \end{aligned}$$

Range

- However, for exclusive class interval with left-end inclusion, the range is computed as follows

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

Life in Hours of 200 Incandescent Lamps

$$\begin{aligned}
 \text{Range} &= \text{highest score} - \text{lowest score} \\
 &= 1500 - 500 = 1000
 \end{aligned}$$

- The same can be done for the inclusive class interval to obtain range as long as we use the exact limits

Range

$$\text{Range} = 69.5 - 19.5 = 50$$

<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>
65–69	64.5–69.5	67	6	180
60–64	59.5–64.5	62	15	174
55–59	54.5–59.5	57	37	159
50–54	49.5–54.5	52	30	122
45–49	44.5–49.5	47	42	92
40–44	39.5–44.5	42	22	50
35–39	34.5–39.5	37	18	28
30–34	29.5–34.5	32	7	10
25–29	24.5–29.5	27	2	3
20–24	19.5–24.5	22	1	1

- Note that we still consider 20 as the lowest score and 69 as the highest score for this dataset because the class intervals represent the true data, unlike the exact limits.

Mean Deviation

- **Deviation score** is the difference between the given score and the mean.

$$DS_i = (x_i - \bar{x})$$

- **Mean deviation (MD)** is the average of the absolute values of the deviation scores.

$$MD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^n |DS_i|}{n}$$

- A larger MD shows a greater variation

Variance

- Using square instead of absolute.
- Variance is the **average** of the **sum of squared deviations** around the **mean**.
- Population variance σ^2

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

SS: sum of square

- Sample Variance s^2

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

Sample Variance

- The sample variance, call it s^2 , of the data set $x_1, x_2, x_3, \dots, x_n$ is defined by

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

Sample Variance (2)

$$y_i = ax_i + b$$

$$s_y^2 = \frac{\sum_{i=1}^n ((ax_i + b) - (a\bar{x} + b))^2}{n - 1}$$

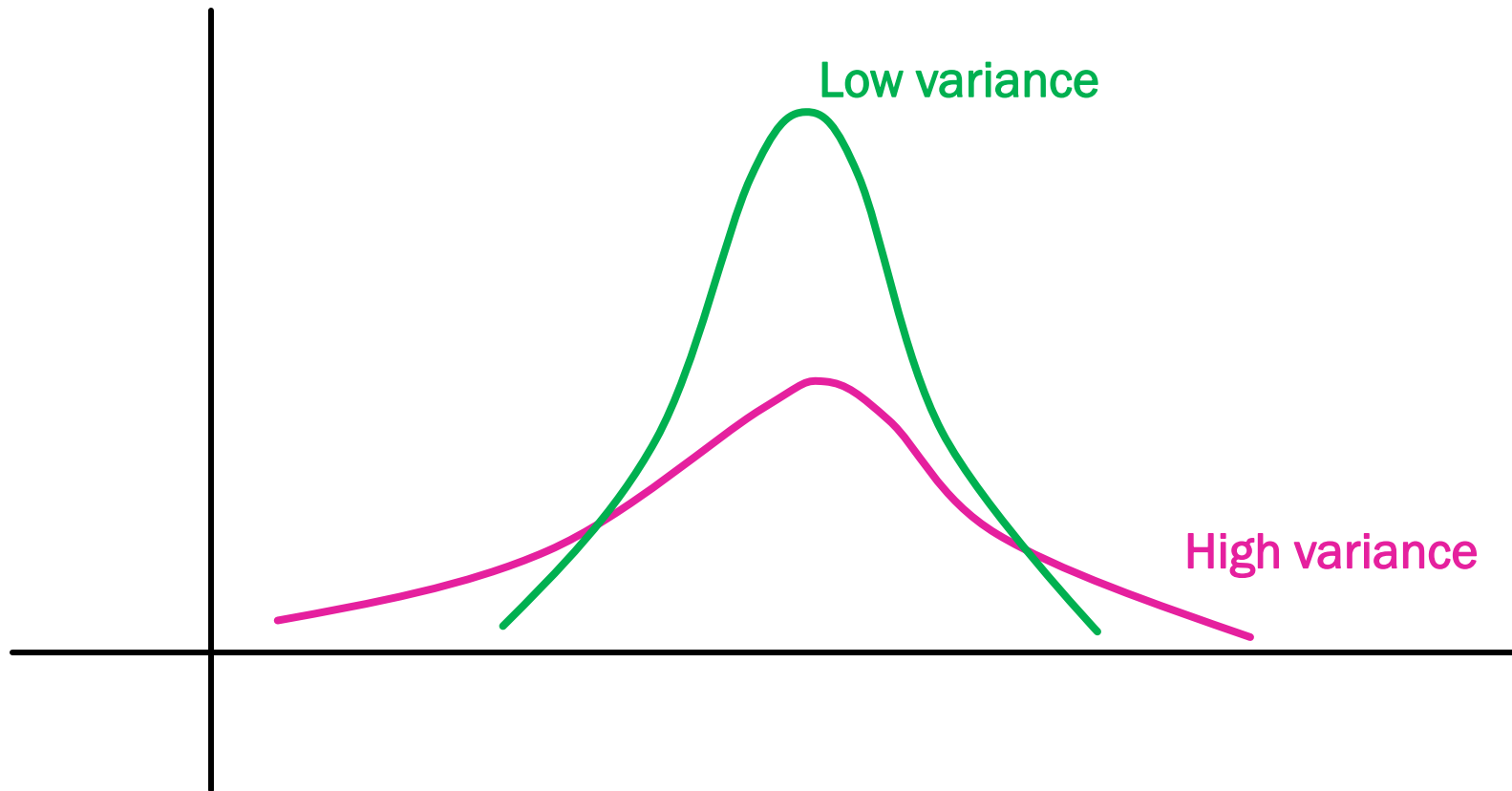
$$s_y^2 = \frac{\sum_{i=1}^n (ax_i - a\bar{x})^2}{n - 1}$$

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

$$s_y^2 = a^2 s_x^2$$

- Modified data; multiply with a **constant a** and add with a **constant b** .
- The constants, a and b , will impact the *mean* of the modified data.
- **Relatively simplify the calculation of the mean.**

Sample Variance (3)



Sample Variance for Grouped Data

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

- f_i : frequency of the i^{th} interval
- m_i : midpoint of the i^{th} interval
- k : number of interval

Exercise

- Compute the sample variance

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

<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>
65–69	64.5–69.5	67	6	180
60–64	59.5–64.5	62	15	174
55–59	54.5–59.5	57	37	159
50–54	49.5–54.5	52	30	122
45–49	44.5–49.5	47	42	92
40–44	39.5–44.5	42	22	50
35–39	34.5–39.5	37	18	28
30–34	29.5–34.5	32	7	10
25–29	24.5–29.5	27	2	3
20–24	19.5–24.5	22	1	1

Standard Deviation

- Standard deviation is the square root of the variance.
- Symbols:
 - Standard deviation of population σ

$$\sigma = \sqrt{\sigma^2}$$

- Standard deviation of sample s

$$s = \sqrt{s^2}$$

Example

- Compute the standard deviation

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	9	3	9
2	12	6	36
3	7	1	1
4	5	-1	1
5	2	-4	16
6	3	-3	9
7	4	-2	4
Σ	42	0	76

- $n = 7$

- Total $\sum x_i = 42$

- Mean $\bar{x} = \frac{42}{7} = 6$

- $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{76}{6} = 12.67$

$$s = \sqrt{s^2} = \sqrt{12.67} = 3.56$$

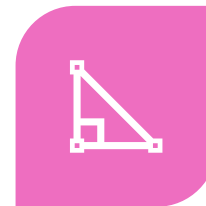
Exercise

- Compute **mean**, **mean deviation**, and **variance** of the following grouped frequency table!

Class Interval	Frequency
0-2	3
3-5	6
6-8	6
9-11	4
12-14	1

3. MEASURES OF POSITION

- Where can you find a certain score in reference to the other scores?
- Statistics to give context or frame of reference, i.e., **relative position of a score among other scores.**
- Some statistics for this problem:



PERCENTILE



PERCENTILE
RANK

Sample Percentile

- The sample **100p percentile** is that data value such that:
 - 100p percent of the data are less than or equal to it
 - 100(1 - p) percent are greater than or equal to it
- If two data values satisfy this condition, then the sample 100p percentile is the average of these two values.
- Writing convention
 - Sample 100p percentile = P_{100p}
 - Sample 25 percentile = P_{25}

Sample Percentile (2)

- To determine the sample $100p$ percentile of a data set of size n , we need to determine the data values such that:
 - At least np of the values are **less than or equal** to it.
 - At least $n(1 - p)$ of the values are **greater than or equal** to it.
- First, You need to arrange the data in increasing order !

Sample Percentile for Non-Grouped Data

- To determine the sample $100p$ percentile of data of size n :
 1. Arrange the data in order (lowest to highest)
 2. Compute np
 3. Test:
 - a. If np is not whole number, round up to the next whole number !
 - b. If np is whole number, compute the average of values in the position np and $np + 1$.

Sample Percentile for Non-Grouped Data (2)

- If $n = 22$, determine the position of 80 percentile !
- What we can conclude
 - $np = 22(0.8) = 17.6$ of the values are less than or equal to it.
 - $n(1 - p) = 22(0.2) = 4.4$ of the values are greater than or equal to it.
 - So the **18th smallest value** satisfies both conditions.
 - This is the sample 80 percentile, where $P_{80} = 18^{th}$ value.
- If np is an integer (e.g 18) , then both values in positions np and $np + 1$ satisfy both conditions, and so the sample 100p percentile is the average of these values, where

$$\frac{(x_{18} + x_{19})}{2}$$

Sample Percentile for Non-Grouped Data

- First quartile (Q_1) : the sample 25 percentile.
- Second quartile (Q_2) : the sample 50 percentile → Sample median.
- Third quartile (Q_3) : the sample 75 percentile
- Interquartile Range (IQR) : $Q_3 - Q_1$.

Example

- Determine first, second, and third quartile, as well as P70 of the following data set !

$\{17.11, 6.6, 6.59, 11.06, 2.78, 6.96, 3.79, 4.3\}$

- Ordered data set:

$\{2.78, 3.79, 4.3, 6.59, 6.6, 6.96, 11.06, 17.11\}$

Example

- Ordered data set:

$\{2.78, 3.79, 4.3, 6.59, 6.6, 6.96, 11.06, 17.11\}$

P_{25}	$np = 8(0.25) = 2$	$P_{25} = \frac{(3.79 + 4.3)}{2} = 4.045$
P_{50}	$np = 8(0.50) = 4$	$P_{50} = \frac{(6.59 + 6.6)}{2} = 6.595$
P_{75}	$np = 8(0.75) = 6$	$P_{75} = \frac{(6.96 + 11.06)}{2} = 9.01$
P_{70}	$np = 8(0.70) = 5.6$	$P_{70} = 6.96$
Interquartile Range (IQR) = $Q_3 - Q_1 = 9.01 - 4.045 = 4.965$		

Sample Percentile for Grouped Data

- For grouped frequency table

$$X^{th} \text{ percentile} = P_x = ll + \left(\frac{n.p - cf}{f_i} \right) (w)$$

- ll : lower exact limit of the interval containing the $n(p)$ score
- n : total number of scores
- p : proportion corresponding to the desired percentile
- cf : cumulative freq. of scores below the interval containing the $n(p)$ score
- f_i : freq. of scores in the interval containing the $n(p)$ score
- w : width of class interval

For left-end-inclusion [Ross, 2009] case, lower limit of an interval is the left-interval-bound

Example

- Find the 34th percentile!

$$X^{th} \text{ percentile} = P_x = ll + \left(\frac{n.p - cf}{f_i} \right)(w)$$

$$P_{34} = 44.5 + \left(\frac{180(0.34) - 50}{42} \right)(5) = 45.83$$

Class Interval	Exact Limits	Midpoint	f	cf	%	c%
65-69	64.5-69.5	67	6	180	3.33	100.00
60-64	59.5-64.5	62	15	174	8.33	96.67
55-59	54.5-59.5	57	37	159	20.56	88.34
50-54	49.5-54.5	52	30	122	16.67	67.78
45-49	44.5-49.5	47	42	92	23.33	51.11
40-44	39.5-44.5	42	22	50	12.22	27.78
35-39	34.5-39.5	37	18	28	10.00	15.56
30-34	29.5-34.5	32	7	10	3.89	5.56
25-29	24.5-29.5	27	2	3	1.11	1.67
20-24	19.5-24.5	22	1	1	0.56	0.56

Percentile Rank

- Percentile rank of a score is the **percent** of scores less than or equal to that score.
- Suppose you got 65 on the final exam of this course. You want to know what percent of students scored lower.
- Writing convention
 - Percentile rank of score 65 = PR_{65}

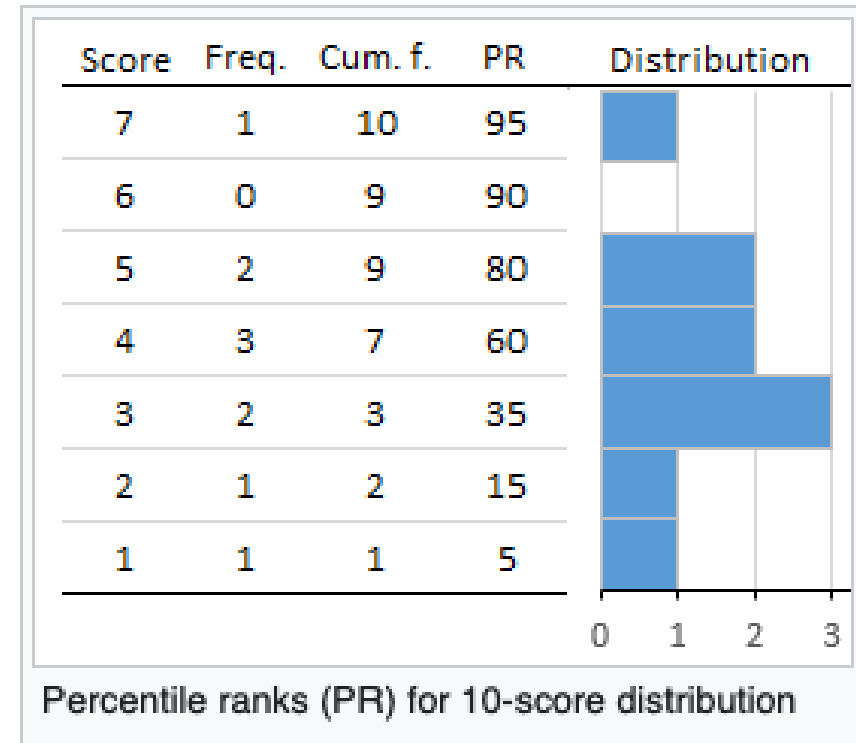
Percentile Rank for Non-Grouped Data

$$PR = \frac{CF' + (0.5 \times F)}{N} \times 100.$$

CF' = the count of all scores less than the score of interest

F = frequency of the score

N = number of scores



Percentile Rank for Non-Grouped Data (2)

- Find percentile rank of a score of 12 from the following dataset

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

- Ordered data set:

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

- Percentile rank:

$$PR_{12} = \frac{6 + (0.5)(1)}{10} \times 100 = 65$$

Percentile Rank for Grouped Data

- Percentile rank in grouped data

$$PR_X = \left(\frac{cf + \frac{X - ll}{w} f_i}{n} \right) (100)$$

- PR_X = percentile rank of score X
- cf = cumulative frequency of scores below the interval containing percentile point
- ll = exact lower limit of the interval containing percentile point
- w = width of class interval
- f_i = frequency of scores in the interval containing percentile point
- n = total number of scores

For left-end-inclusion [Ross, 2009] case, lower limit of an interval is the left-interval-bound

Percentile Rank

- Find percentile rank of score 61 !

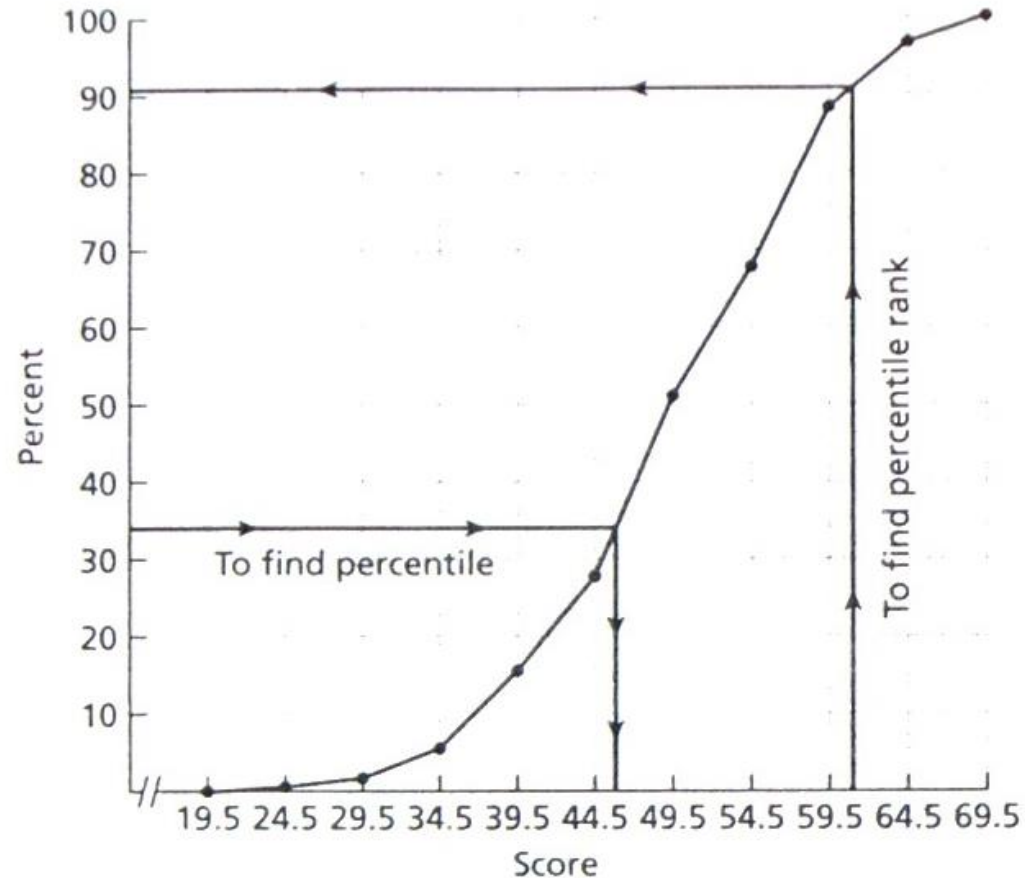
$$PR_x = \left(\frac{cf + \frac{X - l}{w} f_i}{n} \right) (100)$$

$$PR_{61} = \left(\frac{159 + \frac{61 - 59.5}{5} 15}{180} \right) (100) = 90.83$$

<i>Class Interval</i>	<i>Exact Limits</i>	<i>Midpoint</i>	<i>f</i>	<i>cf</i>	<i>%</i>	<i>c%</i>
65-69	64.5-69.5	67	6	180	3.33	100.00
60-64	59.5-64.5	62	15	174	8.33	96.67
55-59	54.5-59.5	57	37	159	20.56	88.34
50-54	49.5-54.5	52	30	122	16.67	67.78
45-49	44.5-49.5	47	42	92	23.33	51.11
40-44	39.5-44.5	42	22	50	12.22	27.78
35-39	34.5-39.5	37	18	28	10.00	15.56
30-34	29.5-34.5	32	7	10	3.89	5.56
25-29	24.5-29.5	27	2	3	1.11	1.67
20-24	19.5-24.5	22	1	1	0.56	0.56

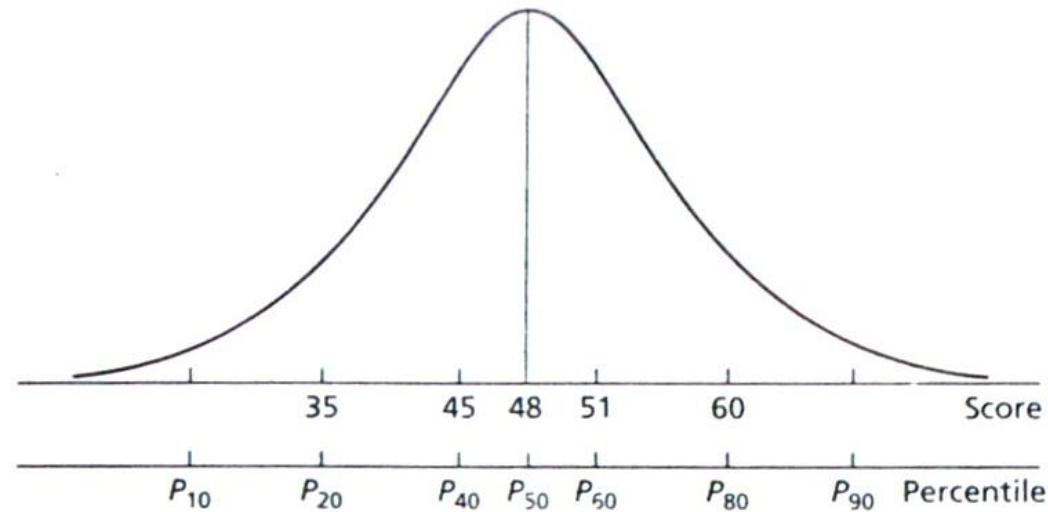
Ogive & Percentile

- Ogive can be used to find percentile & percentile rank



Percentile Rank as an Ordinal Scale

- Position of percentile for normal distribution



- In the middle, a difference of 6 raw score (45-51) is equivalent to a difference of 20 percentile points
- In the tails, the opposite phenomenon occurs
- Percentile Rank is an ordinal scale

Percentile Rank as an Ordinal Scale

- Recall the criteria of an ordinal scale
 - Mutually exclusive
 - Some logical order

→ yes
- The difference between $P_{50} - P_{40}$ and $P_{20} - P_{10}$ may not be the same → ordinal
- Percentile should be used only for describing points in a distribution (relative position/rank in a distribution), **NOT** for making comparisons accross distribution.

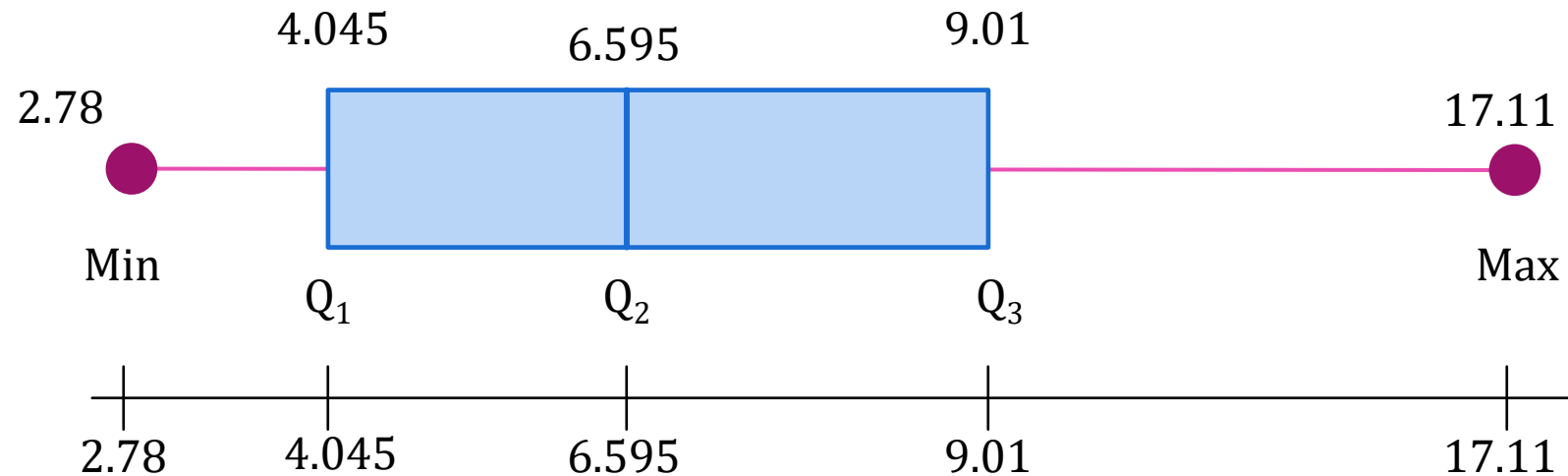
Box Plot

- A straight line segment stretching from the smallest to the largest data value. It contains information about first to the third quartile on the “box” part.

P_{25}	$np = 8(0.25) = 2$	$P_{25} = \frac{(3.79 + 4.3)}{2} = 4.045$
P_{50}	$np = 8(0.50) = 4$	$P_{50} = \frac{(6.59 + 6.6)}{2} = 6.595$
P_{75}	$np = 8(0.75) = 6$	$P_{75} = \frac{(6.96 + 11.06)}{2} = 9.01$
P_{70}	$np = 8(0.70) = 5.6$	$P_{70} = 6.96$
Interquartile Range (IQR) = $Q_3 - Q_1 = 9.01 - 4.045 = 4.965$		

Box Plot (2)

- A straight line segment stretching from the smallest to the largest data value. It contains information about first to the third quartile on the “box” part.



Outliers

- An **outlier** is an unusual score in a distribution that may warrant special consideration.
- Outliers can arise because of a **measurement or recording error** or because of equipment failure during an experiment, etc.
- An outlier might be **indicative of a sub-population**, e.g. an abnormally low or high value in a medical test could indicate presence of an illness in the patient.

Outliers & Box Plot

- 5 important numbers on the box plot:

- RUB (reasonable upper boundary)

$$RUB = Q_3 + 1,5 (IQR)$$

- Q_3 (third quartile)

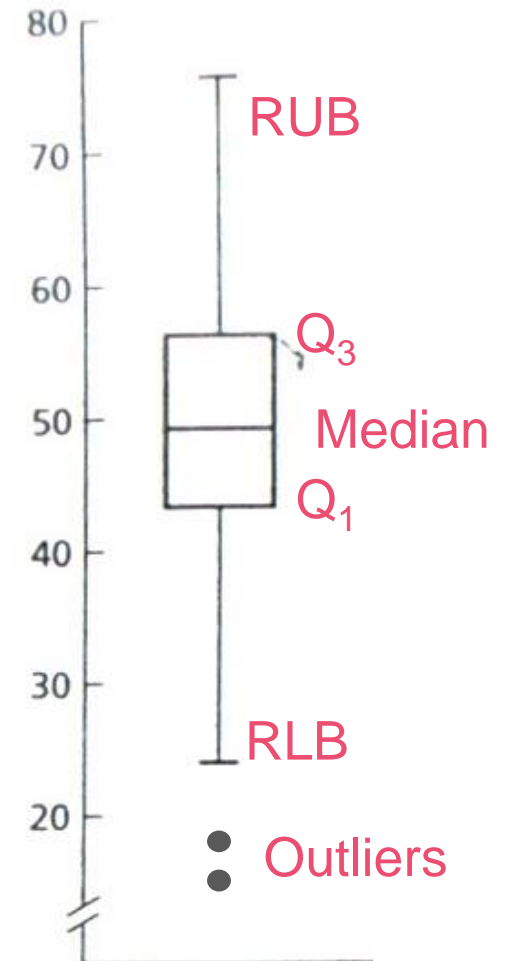
- Median (Q_2)

- Q_1 (first quartile)

- RLB (reasonable lower boundary)

$$RLB = Q_1 - 1,5(IQR)$$

Outliers are all scores above the RUB or below the RLB



STANDARD SCORES

Standard Scores

- Suppose a student has the scores in 3 classes: 68 in math, 77 in physics, and 83 in history.
- In which class did the student perform best?

We can not use raw scores

- Those 3 distributions may have different mean & variance
- Those 3 scores may have different scale of measurement

We can not use percentiles

- Percentile is an ordinal scale

Standard Scores (2)

- One way is to transform the scores into scores on an **equal interval scale**.
- Standard scores do this by using standard deviation as the unit of measure.
- Standard score or z-score is computed as follow:

$$z = \frac{x - \bar{X}}{s}$$

Properties of Z-score

- Retains the shape of the distribution of the original scores
- mean = 0
- variance = 1,
- standard deviation = 1

$$z = \frac{x - \bar{X}}{s}$$

- Z-score indicates the number of standard deviations a corresponding raw score is above or below the mean.

Standard Scores

- Suppose a student has the scores in 3 classes: 68 in math, 77 in physics, and 83 in history.
- In which class did the student perform best?
- Suppose we know the mean and standard deviation of those 3 distributions. So we can compute **z-scores**:

Subject	x	\bar{X}	s	z
Math	68	65	6	0.50
Physics	77	77	9	0.00
History	83	89	8	-0.75

Relative to the other students, this student performed best on the math exam.

Kandidat	Nilai Asli				Dalam z score			
	IPK	TOEFL	Paper		IPK	TOEFL	Paper	Rata-rata
A	3.3	570	80		-1.46	0.01	-0.14	-0.53
B	3.5	600	75		-0.31	0.85	-0.86	-0.11
C	3.6	550	85		0.26	-0.55	0.57	0.09
D	3.7	610	80		0.83	1.13	-0.14	0.61
E	3.75	580	80		1.11	0.29	-0.14	0.42
F	3.9	585	75		1.97	0.43	-0.86	0.51
G	3.51	563	90		-0.26	-0.19	1.29	0.28
H	3.48	475	70		-0.43	-2.65	-1.57	-1.55
I	3.46	573	80		-0.54	0.10	-0.14	-0.20
J	3.35	590	95		-1.17	0.57	2.00	0.47
mean	3.555	569.6	81		0.00	0.00	0.00	
standar deviasi	0.18	35.65	7.00		1.00	1.00	1.00	

Transformed Standard Scores

- We need to transform those z-scores into a different distribution of scores, so that people are easy to interpret those value.

$$X' = (s')(z) + \overline{X'}$$

X' = the transformed score

s' = the desired standard deviation

$\overline{X'}$ = the desired mean

Dalam z score					Transformasi	
IPK	TOEFL	Paper	Rata-rata		rata2 85, std 10	
-1.46	0.01	-0.14	-0.53		79.71	
-0.31	0.85	-0.86	-0.11		83.94	
0.26	-0.55	0.57	0.09		85.93	
0.83	1.13	-0.14	0.61		91.06	
1.11	0.29	-0.14	0.42		89.21	
1.97	0.43	-0.86	0.51		90.15	
-0.26	-0.19	1.29	0.28		87.81	
-0.43	-2.65	-1.57	-1.55		69.49	
-0.54	0.10	-0.14	-0.20		83.03	
-1.17	0.57	2.00	0.47		89.67	
0.00	0.00	0.00				
1.00	1.00	1.00				

Weighted Averages

- How to develop a composite score from two or more individual scores ?
- For example: we want to compute composite scores of two technical test and one personality test to evaluate a job seeker.

$$Weighted_score_j = \frac{\sum_i^n w_i z_{ij}}{\sum_i^n w_i}$$

- $n = \text{number of test}$
- $w_i = \text{weight of each test}$
- $z_{ij} = \text{standard score for person } j \text{ on test } i$

Dalam z score				Transformasi		Nilai Berbobot		IPK	3
IPK	TOEFL	Paper	Rata-rata	rata2 85, std 10		weighed z-score		TOEFL	3
								Paper	4
-1.46	0.01	-0.14	-0.53	79.71		-0.49	80.10		
-0.31	0.85	-0.86	-0.11	83.94		-0.18	83.19		
0.26	-0.55	0.57	0.09	85.93		0.14	86.41		
0.83	1.13	-0.14	0.61	91.06		0.53	90.31		
1.11	0.29	-0.14	0.42	89.21		0.36	88.64		
1.97	0.43	-0.86	0.51	90.15		0.38	88.78		
-0.26	-0.19	1.29	0.28	87.81		0.38	88.82		
-0.43	-2.65	-1.57	-1.55	69.49		-1.55	69.47		
-0.54	0.10	-0.14	-0.20	83.03		-0.19	83.09		
-1.17	0.57	2.00	0.47	89.67		0.62	91.21		
0.00	0.00	0.00							
1.00	1.00	1.00							

Chebyshev's Inequality

Chebyshev's Inequality (2)

- Let \bar{x} and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, x_2, x_3, \dots, x_n$, where $s > 0$. Let

$$\begin{aligned} S_k &= \left\{ i, 1 \leq i \leq n : |x_i - \bar{x}| < ks \right\} \\ &= \left\{ i, 1 \leq i \leq n : \bar{x} - ks < x_i < \bar{x} + ks \right\} \end{aligned}$$

- And let $N(S_k)$ be the number of elements in the set S_k . Then, for any $k \geq 1$,

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

Chebyshev's Inequality (4)

$$\frac{N(S_k)}{n} > 1 - \frac{1}{k^2}$$

- Which means, more than **(at least) $100(1 - \frac{1}{k^2})$** percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$.
- We only need to know the **standard deviation & mean !**

Example

- 10 top-selling passenger cars in the U.S in 2008.
- Compute the **standard deviation & mean** !

$$\bar{x} = 31879.4 \quad s = 7514.7$$

1.	Ford F Series	44,813
2.	Toyota Camry	40,016
3.	Chevrolet Silverado.....	37,231
4.	Honda Accord Hybrid.....	35,075
5.	Toyota Corolla Matrix	32,535
6.	Honda Civic Hybrid	31,710
7.	Chevrolet Impala.....	26,728
8.	Dodge Ram	24,206
9.	Ford Focus	23,850
10.	Nissan Altima Hybrid	22,630

Example

- How large is the percentage of car sales lie between $\bar{x} - 1.5s$ to $\bar{x} + 1.5s$?

$$\bar{x} = 31,879.4 \quad s = 7514.7$$

- Chebyshev's Inequality

$$(\bar{x} - 1.5s, \bar{x} + 1.5s) = (20,607.35, 43,151.45)$$

- Chebyshev's Inequality shows that greater than $100 \left(\frac{5}{9} \right) = 55.55\%$ of the data from any data set lies between $\bar{x} - 1.5s$ to $\bar{x} + 1.5s$. Here, we know that $k = \frac{3}{2}$.

1.	Ford F Series	44,813
2.	Toyota Camry	40,016
3.	Chevrolet Silverado.....	37,231
4.	Honda Accord Hybrid.....	35,075
5.	Toyota Corolla Matrix	32,535
6.	Honda Civic Hybrid	31,710
7.	Chevrolet Impala.....	26,728
8.	Dodge Ram	24,206
9.	Ford Focus	23,850
10.	Nissan Altima Hybrid	22,630