Descriptive statistics

CSGE602013 - Statistika dan Probabilitas

Fakultas Ilmu Komputer

Universitas Indonesia

Semester Genap 2017/2018

References

- Introduction to Probability and Statistics for Engineers & Scientists, 4th ed.,
  - Sheldon M. Ross, Elsevier, 2009.

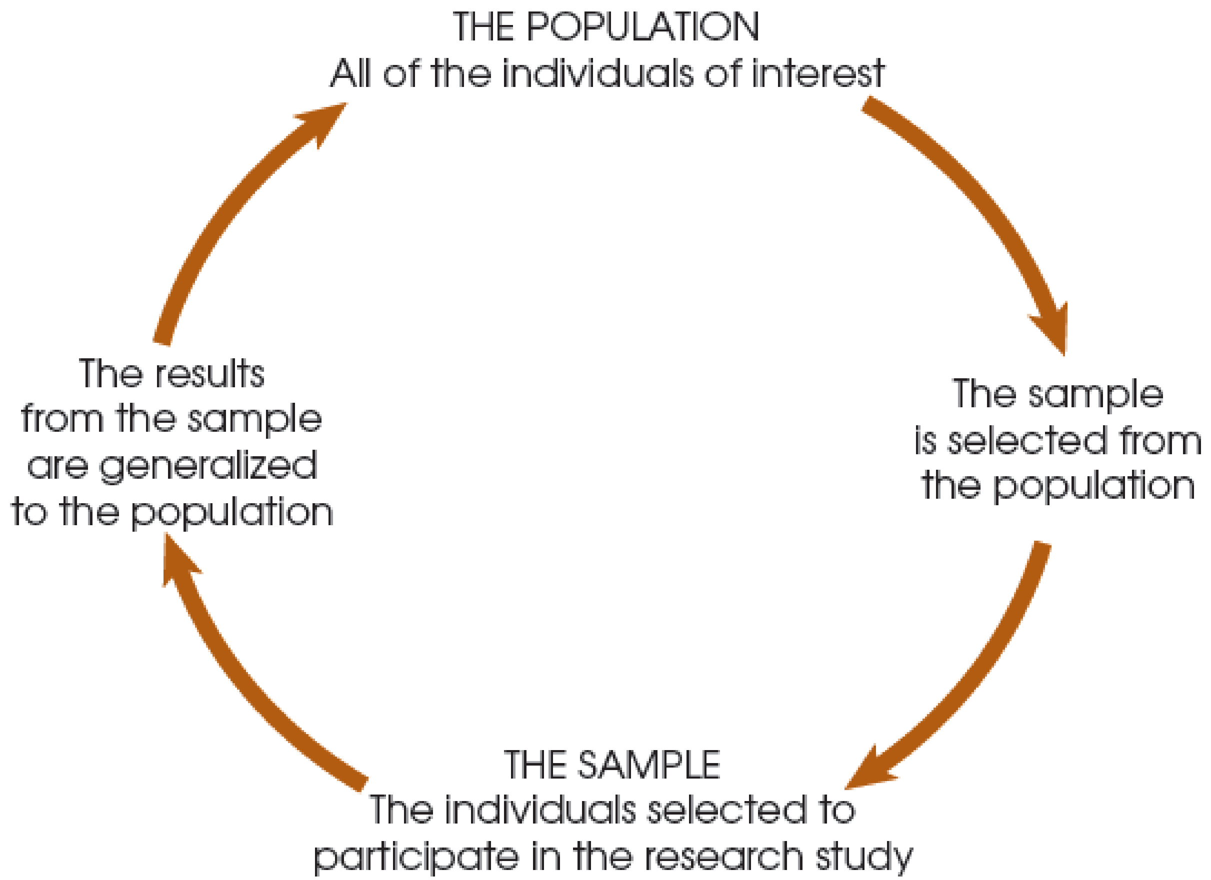
- Applied Statistics for the Behavioral Sciences, 5th Edition,
  - Hinkle., Wiersma., Jurs., Houghton Mifflin Company, New York, 2003.

- 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.

FIGURE 1.1

The relationship between a population and a sample.



[Gravetter & Wallnau, 2012]

Introduction

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

[Recall] **statistical inference**:

- The science of deducing properties (parameters) of an **underlying probability distribution** from **data**.

Before we perform statistical inference, we usually need to **describe** and **summarize** our **data set**.

- This is **descriptive statistics** !

Outline

- Describing data sets (presentation)
  - Stem-and-Leaf Display
  - Ungrouped Frequency Distribution
  - Grouped Frequency Distribution
- Summarizing data sets
  - Measures of Central Tendency
  - Measures of Variations
  - Measures of Position
- Chebyshev’s Inequality
- Normal Data Set

Describing data sets

- Stem-and-Leaf Display
- Ungrouped Frequency Distribution
- Grouped Frequency Distribution

Describing data sets

The **observed data** should be **presented** clearly, concisely so that **observer** can quickly **obtain a feel** for the **essential characteristics** of the data.

Over the years, **tables** & **graphs** are particularly useful and powerful ways of presenting data.

We will learn some common graphical and tabular ways of presenting data.

Stem and Leaf Plot (1)

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**.

If data are all two-digit numbers, we could let

- First digit as its stem
- Second digit as its leaf

Expression for 62

Stem Leaf

6 2

Stem and Leaf Plot (2)

Two values 62 and 67 can be represented as

Stem Leaf

6 2, 7

Ex: Annual average daily minimum temperature (35 points)

51.0, 51.3, 52.0, …

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

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..

Two types:

- Ungrouped Frequency Distribution
- Grouped Frequency Distribution

I. Ungrouped Freq. Distribution

The **ungrouped frequency distribution** is usually used for data that can be placed in specific **categories (Categorical)**, such as **nominal**- or **ordinal**-level data. [Bluman, 2012]

… or there are **relatively small number of distinct values**

Ex : political affiliation, religious affiliation, etc.

Frequency Table (1)

Suppose you purchased a bag of **M&M's chocolate candies** ! You found that there are **55** candies inside.

The distribution of M&M color frequencies:

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

http://onlinestatbook.com/2/introduction/distributions.html

Frequency Table (2)

Starting yearly salaries of 42 recently graduated students

In $1,000 unit

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

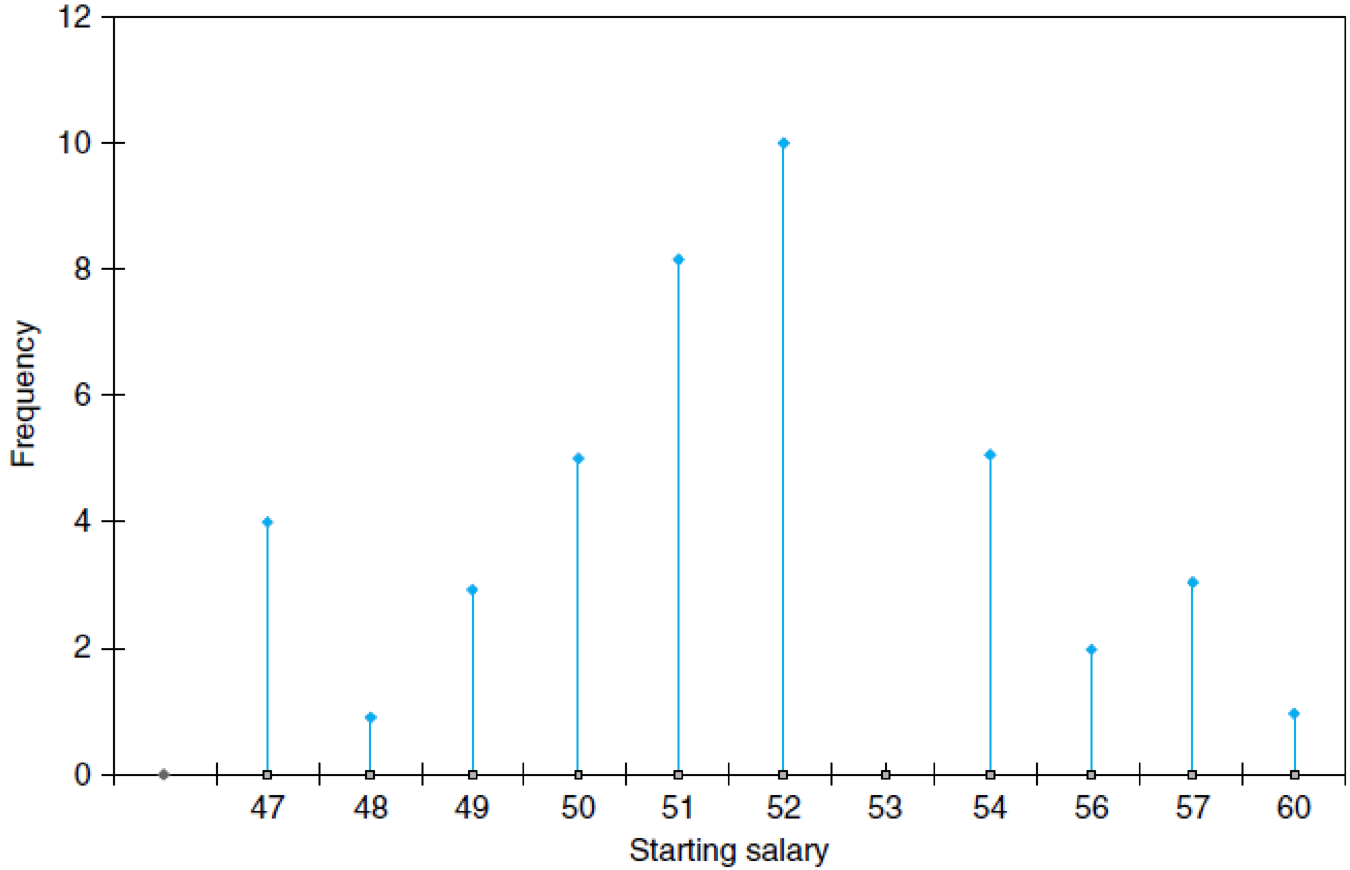
relatively small number of distinct values !

That's why we can use this simple frequency table.

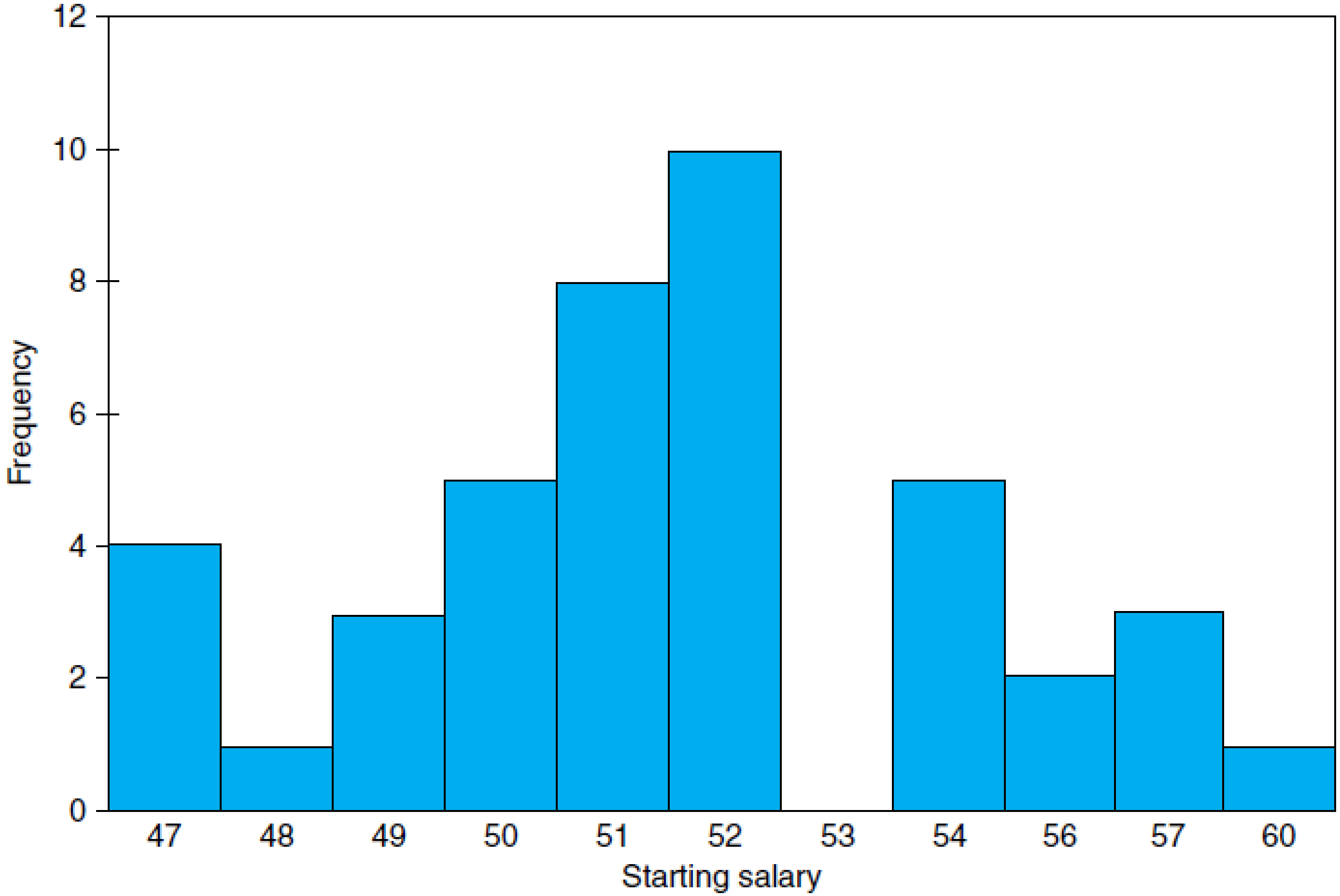
Line Graph

We present the frequency distribution of “starting salary”

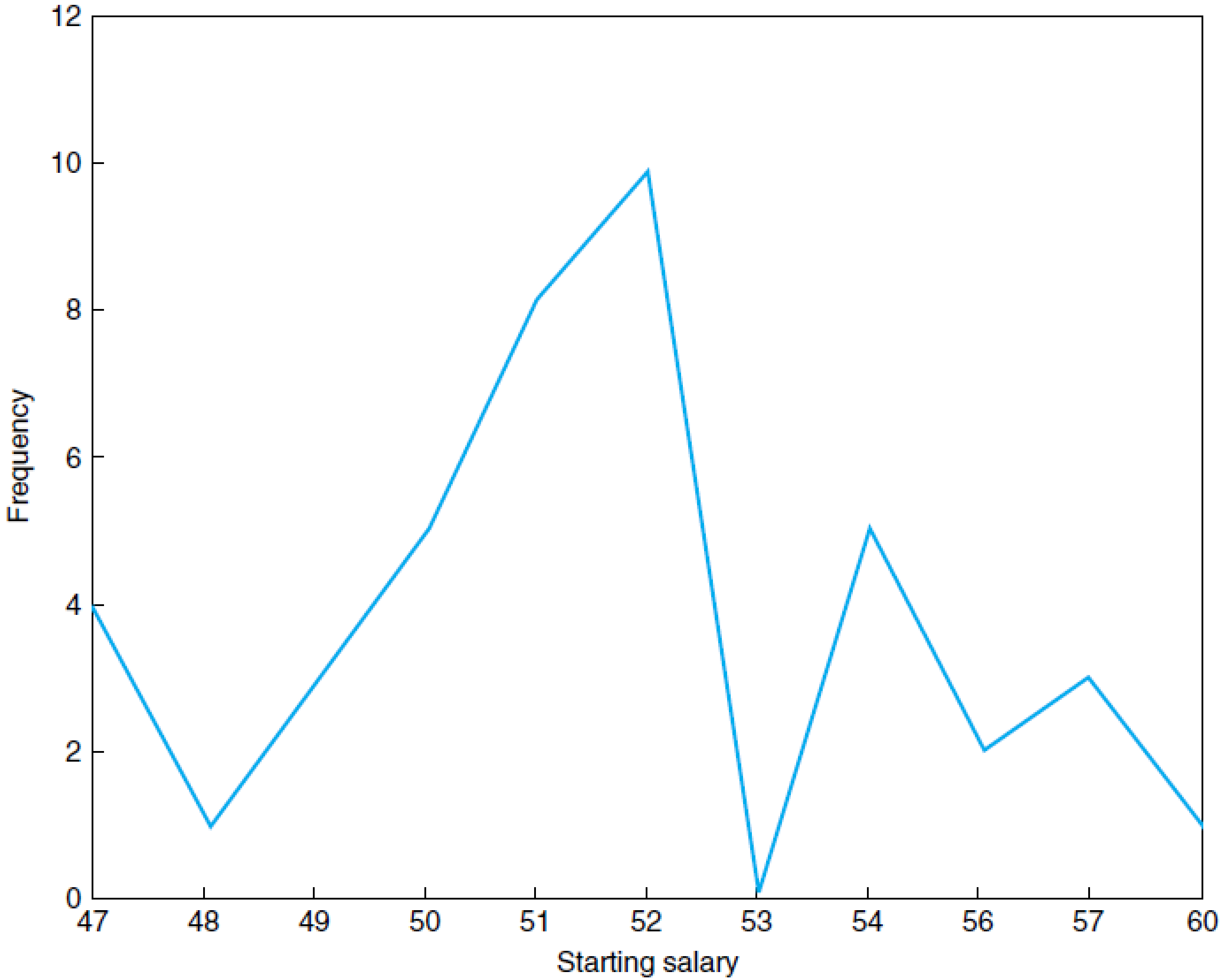
u



Bar Graph


12
10
8
6
4
2
0
Frequency
47
48
49
50
51
52
53
54
56
57
60
Starting salary

Frequency Polygon



Relative Frequency Distribution

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.

Usual Ways of presenting relative frequency distribution:

- Relative frequency table
- Pie chart

Relative Frequency Table

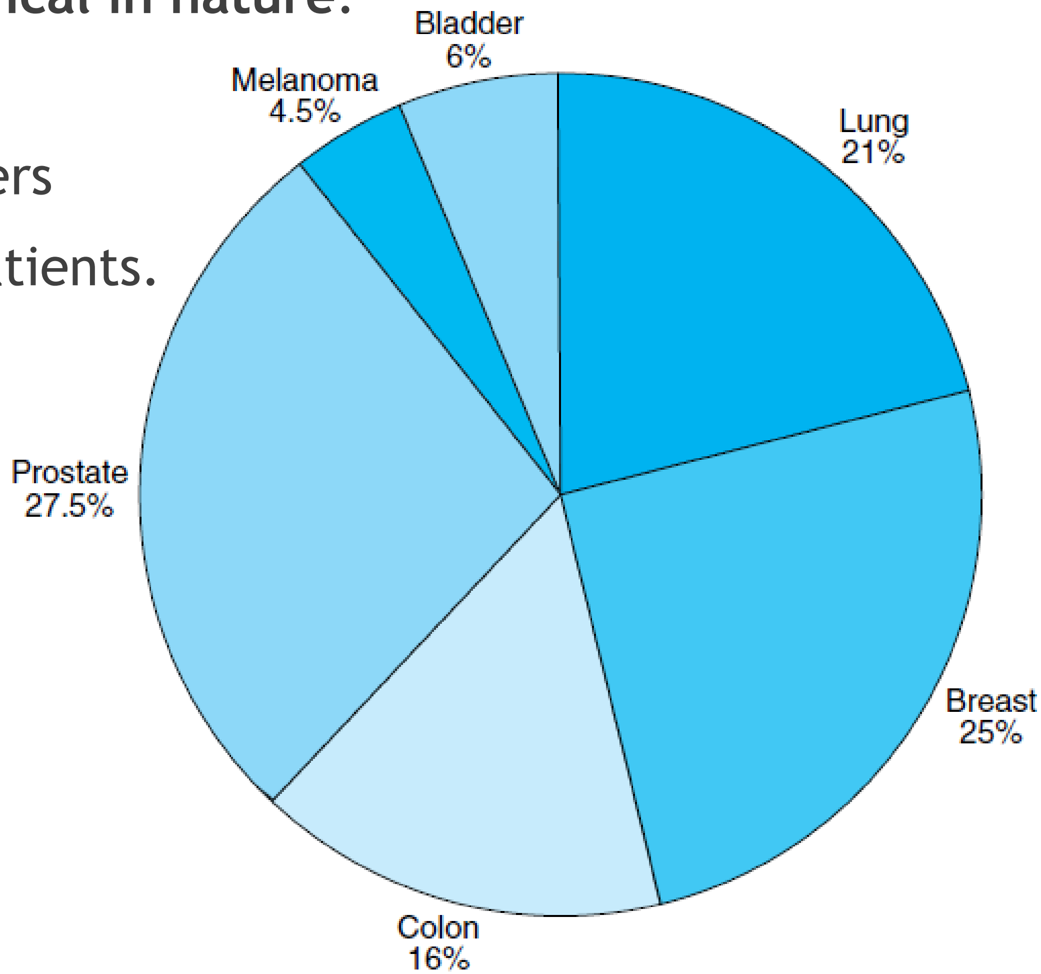
Relative frequency distribution for “starting salary”

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

Pie Chart

Pie chart is **often used** to indicate relative frequencies when the data are **not numerical in nature.**

Different types of cancers affecting 200 sample patients.



II. Grouped Freq. Distribution

Grouped Data

Previously, you learned about **Ungrouped Frequency Distribution** .

- An ordered listing of all values of a variable and their frequencies (or relative frequencies).

When a set of data over a **wide range of values**, it is unreasonable to list all the individual scores in a TFD table.

Solution: **Grouped Frequency Distribution (GFD) !**

http://www.chsbs.cmich.edu/k_han/psy511/table.htm

Grouped Data

- What if the **number of distinct values** of data sets **is too large** ?
- What if the variable is **continuous** ?

In such cases, it is useful to

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

Grouped Data

Life in Hours of 200 Incandescent Lamps

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

We will introduce two ways of presenting grouped frequency distribution.

- Left-end inclusion convention [Ross, 2009]
- [Hinkle, et al., 2003]

Versi yang sering digunakan.
Kita akan lebih banyak menggunakan versi ini

Grouped Data [Ross, 2009]

The endpoints of class interval: ***class boundaries***.

[Ross, 2009] adopts the **left-end inclusion convention** !

Class frequency table

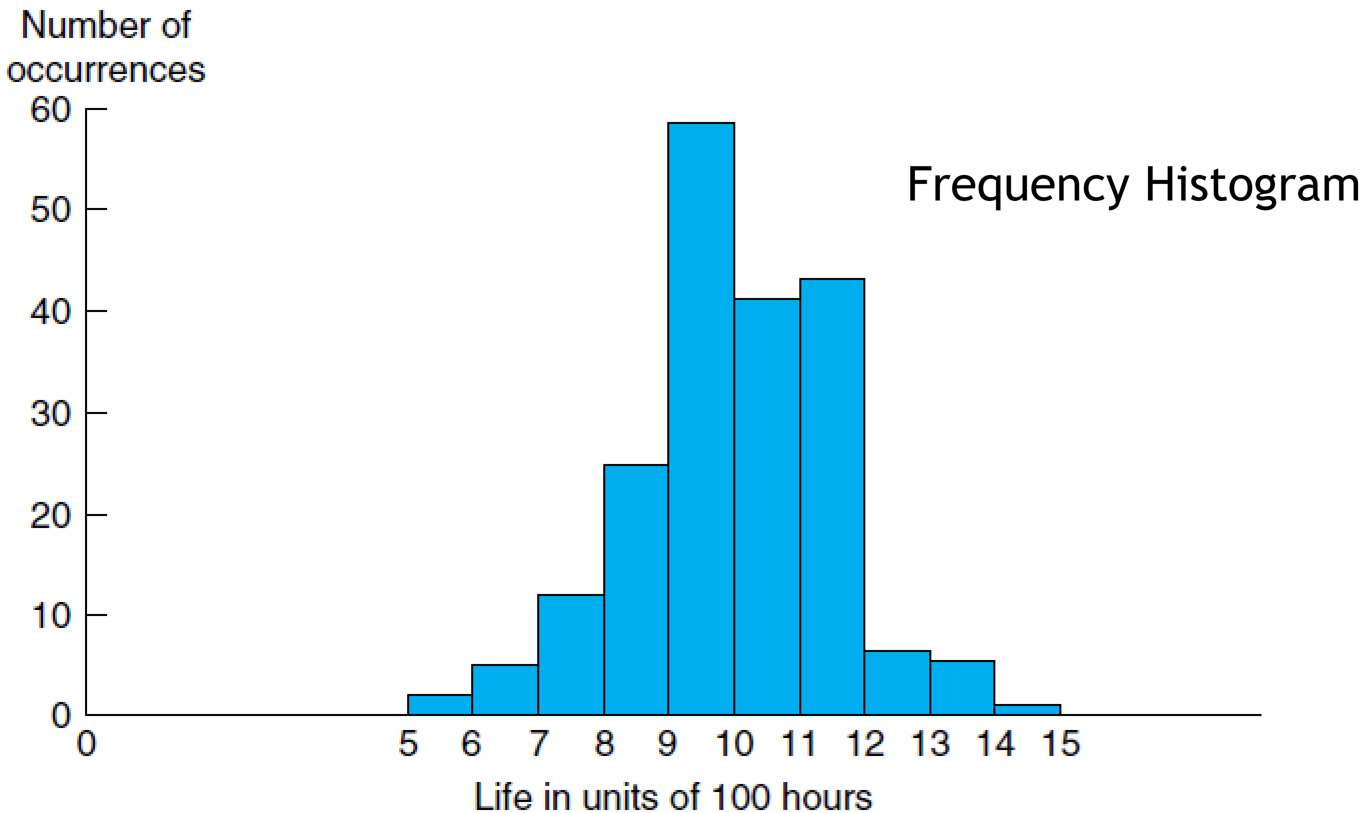
Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

Life in Hours of 200 Incandescent Lamps

greater than or equal to 500 and less than 600

Grouped Data [Ross, 2009]

Histogram: A bar graph plot of class data, with the bars placed adjacent to each other.

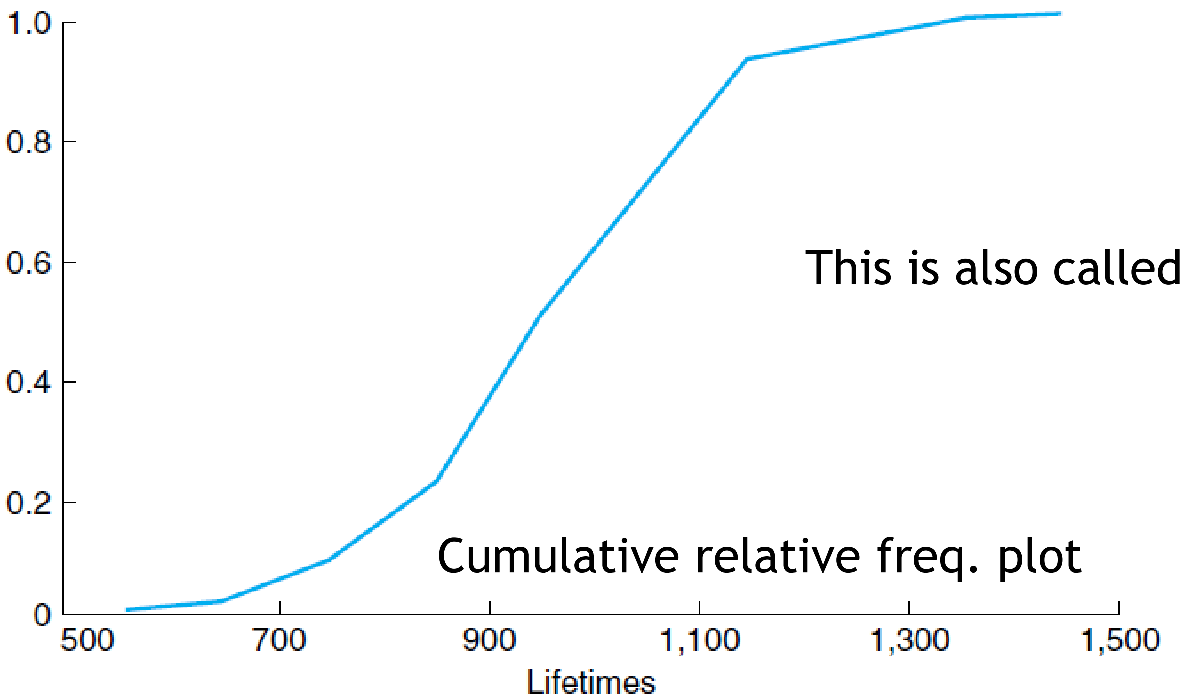


Life in Hours of 200 Incandescent Lamps

Cumulative (or Cumulative Relative) Frequency [Ross, 2009]

Axis represents a possible data value.

Ordinate gives the **number (or proportion)** of the data whose values are **less than or equal to it**.



This is also called as ***Ogive***.

Cumulative relative freq. plot

Life in Hours of 200 Incandescent Lamps

Grouped Data [Hinkle, 2003]

Suppose we have **final examination scores** for freshman psychology students.

68	52	69	51	43	36	44	...
55	54	54	53	33	48	32	...
65	57	64	49	51	56	50	...
42	49	41	48	50	24	49	...
64	63	63	64	54	45	53	...
45	54	44	55	63	55	62	...
56	38	55	37	68	46	67	...
59	46	58	47	57	58	56	...
...	...	...	...	...	...	...	...

Grouped Data [Hinkle, 2003]

Class Interval	Midpoint	Freq.	Class Interval	Midpoint	Freq.
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

Class interval of width 5

All scores between 20 and 24 **inclusive** !

It's different from **[Ross, 2009]** !

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

Midpoint: the point halfway through the interval

Grouped Data [Hinkle, 2003]

Previously, we considered final examination scores as **discrete values** !

Now, we assume that final examination score as a **continuous variable,** although we may record a score as a whole number !

For example:

A score of 53 represents a score somewhere between 52.5 and 53.5

Here, 52.5 and 53.5 represents the **exact limits** of the score 53.

Grouped Data [Hinkle, 2003]

We need to use the notion of **exact limits** :

- Exact limits of a score extend from **one-half unit below to one-half unit above** the recorded score.
- 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, and all are assumed to be adequately represented by the **midpoint**.

If the measurement is more precise …

- If the score limits of a class interval are **17.3** and **18.7**, then the **exact limits** are **17.25** and **18.75**.

Grouped Data [Hinkle, 2003]

When we use class interval, we need to remember two assumptions:

- The score within any class interval are assumed to be **uniformly distributed** throughout the interval.
- Since the table no longer specifies the exact number of data, each class interval is represented by its **midpoint**.

Grouped Data [Hinkle,

TABLE 3.2
Frequency Distribution of Final Examination Scores, Including Cumulative Frequencies and Cumulative Percents

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

This is also called **Score limits**

Grouped Data [Hinkle, 2003]

Choosing the number of class intervals (trade-off)

- **Too few** : losing too much information about the actual data values in a class.
- **Too many** : each class's frequency is too small; loosing the data pattern.

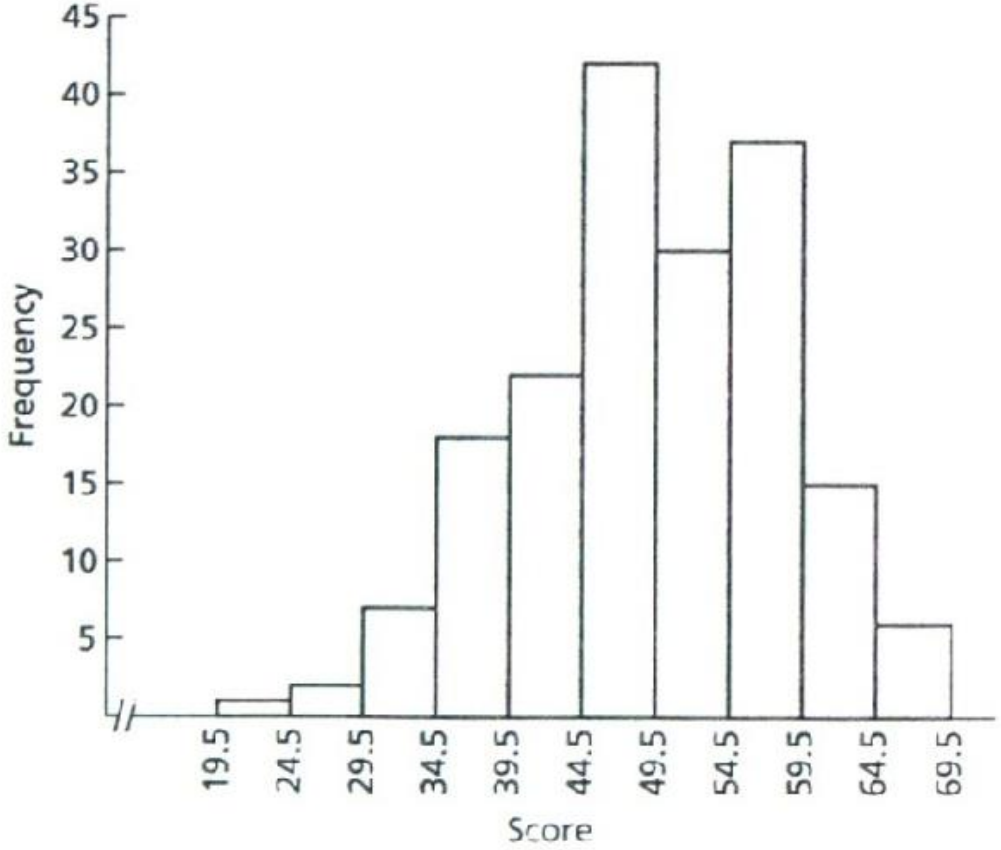
Grouped Data [Hinkle, 2003]

Two General Rules for Class Interval

- 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, 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)

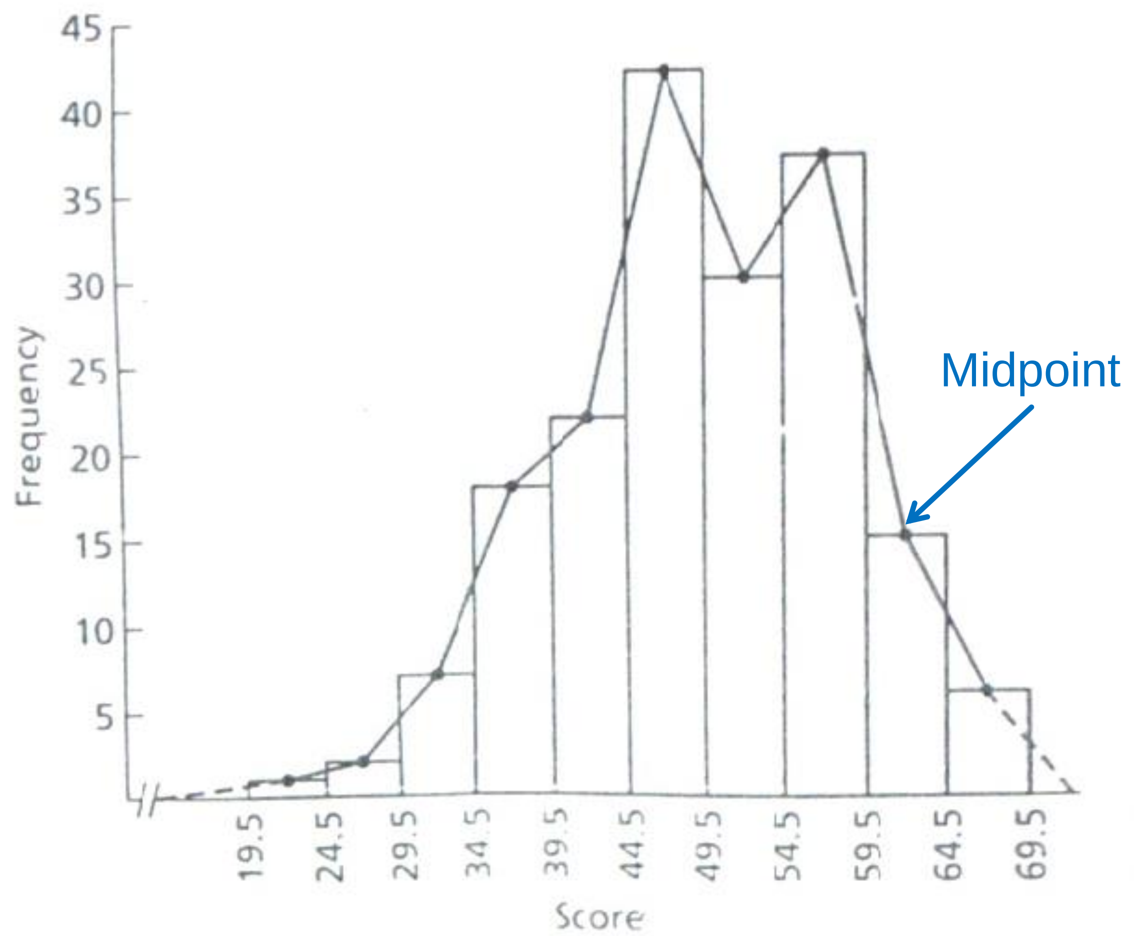
Grouped Data [Hinkle, 2003]

Histogram



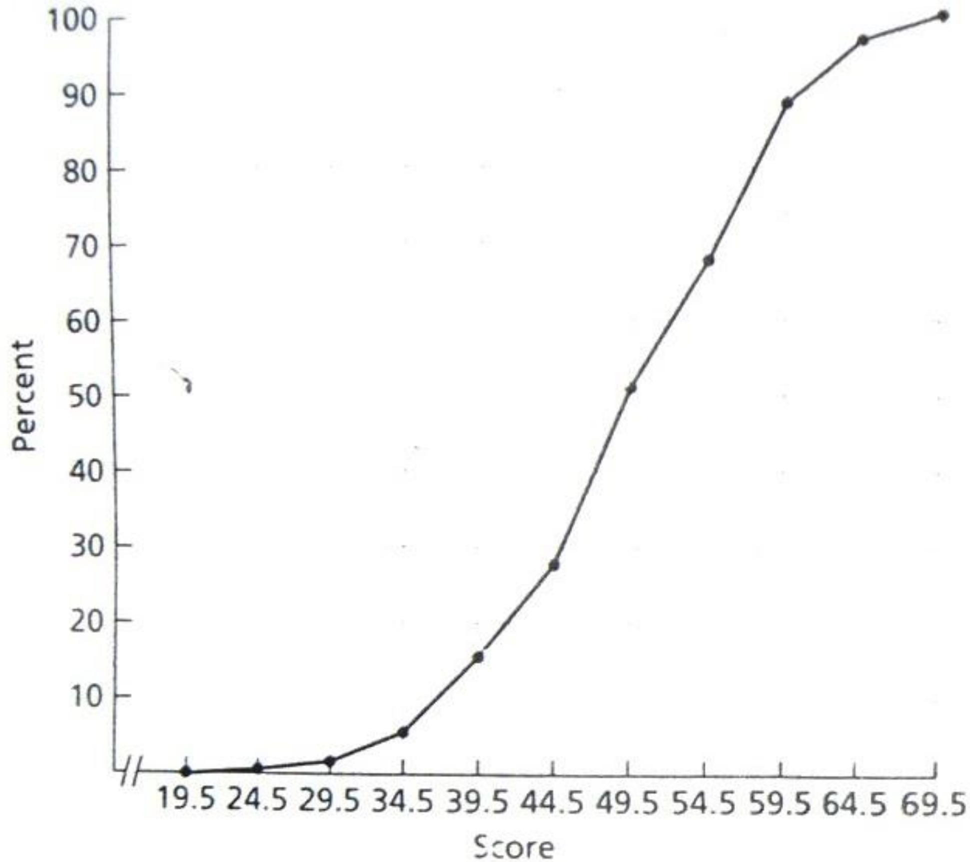
Grouped Data [Hinkle, 2003]

Histogram and **frequency polygon**

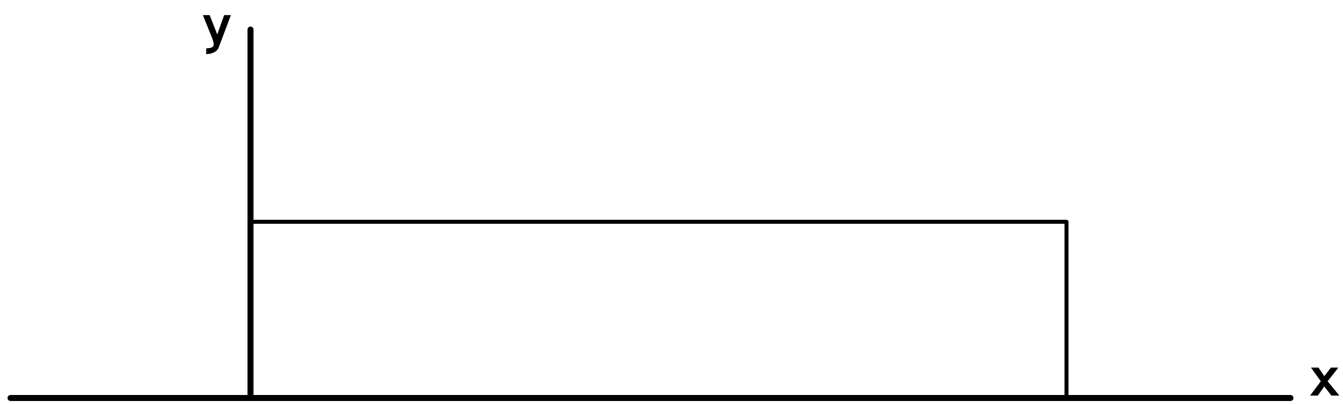


Grouped Data [Hinkle, 2003]

Ogive = cumulative frequency **percentage** distribution



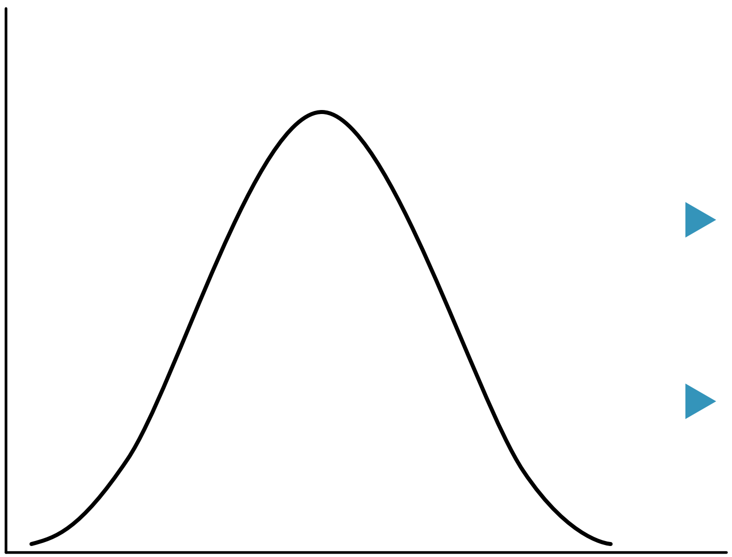
Shapes of Frequency Distribution (1)
(shape of frequency polygon)



Uniform frequency distribution

The scores are uniformly distributed between an interval.

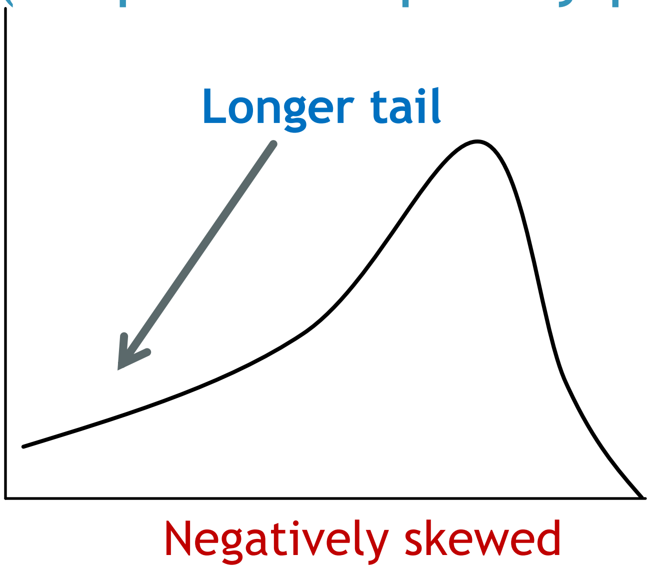
Shapes of Frequency Distribution (2)
(shape of frequency polygon)

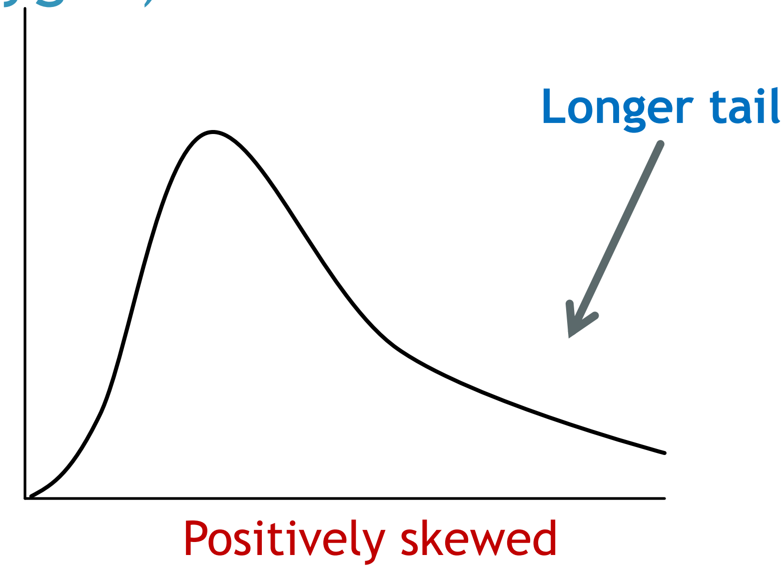


Normal Frequency Distribution

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

Shapes of Frequency Distribution (3)
(shape of frequency polygon)





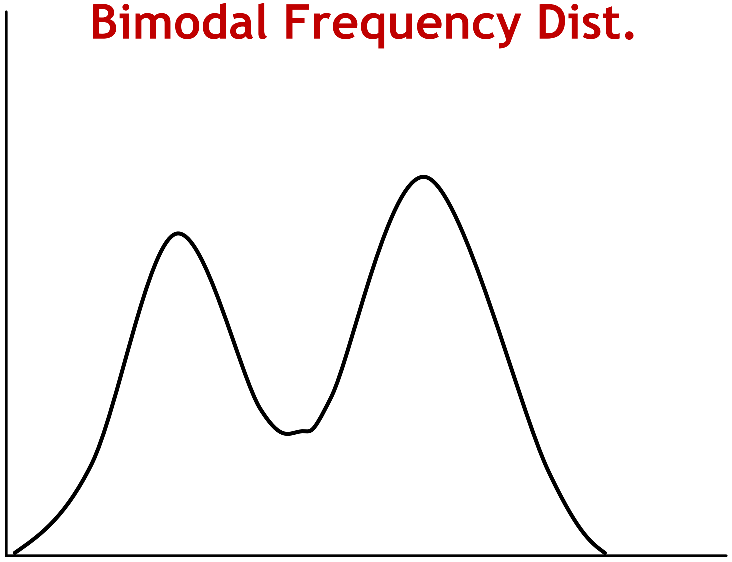
Skewed to the right = positive skew

Skewed to the left = negative skew

Why ? Most likely, there are many more **outliers** on the longer tail area.

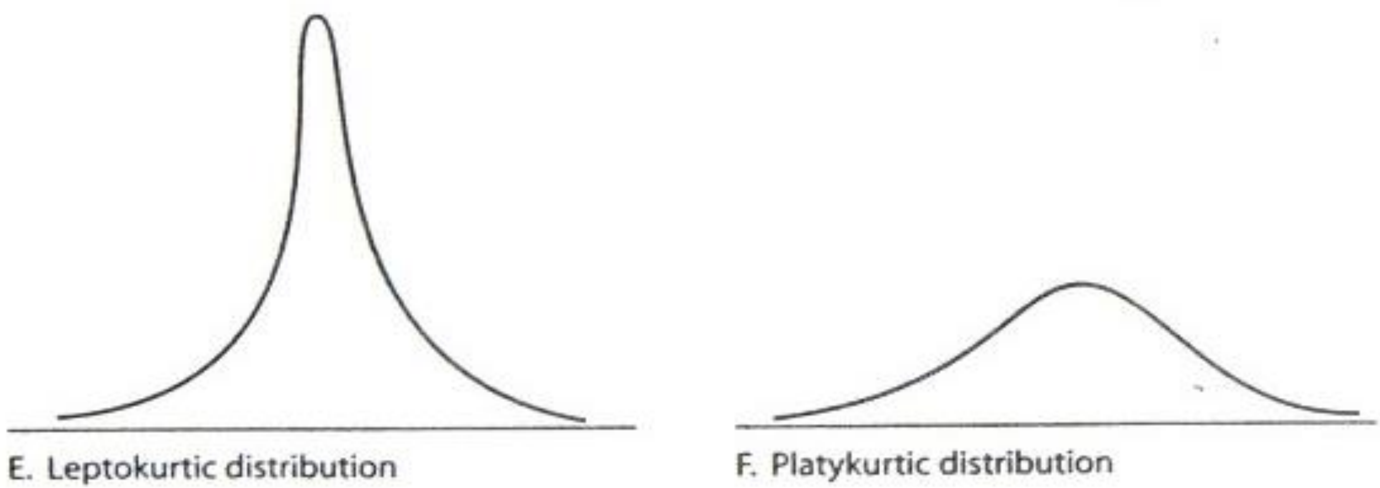
Shapes of Frequency Distribution (4)
(shape of frequency polygon)

Bimodal Frequency Dist.



- It has two peaks
- One possibility is that there are two separate sub-populations in the study. They have different characteristics.

Shapes of Frequency Distribution (5)
(shape of frequency polygon)



Symmetric distributions like normal distribution may vary in **kurtosis** – degree of peakedness.

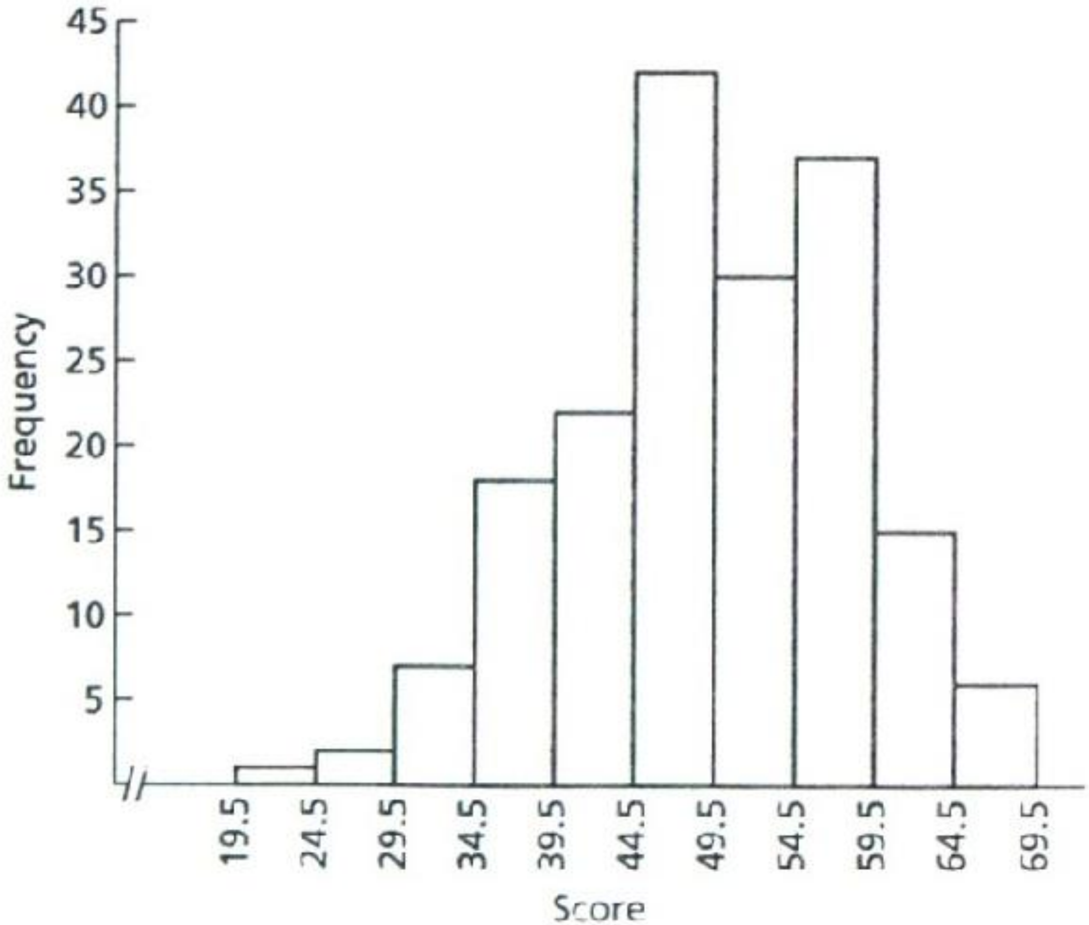
Leptokurtic: if the vast majority of the scores tend to be located at the center.

Platykurtic: if scores are distributed more uniformly, yet many scores still cluster at the center.

Summarizing data sets

Summarizing data sets

How to describe the data set entirely?



Summarizing data sets

- Measures of Central Tendency
- Measures of Variations
- Measures of Position

Summarizing data sets

[Recall] operational definition

Parameter

- Characteristics or measures by using **all** the data values from a **population**.

Statistic

- Characteristics or measures obtained by using the data values from a **sample**.

Summarizing data sets

Part I: Measures of central tendency

Measures of Central Tendency

Information about the **concentration** of scores in a distribution.

Some statistics that are used for describing the **center** of a set of data values:

- Mean
- Median
- Mode

Measures of Central Tendency

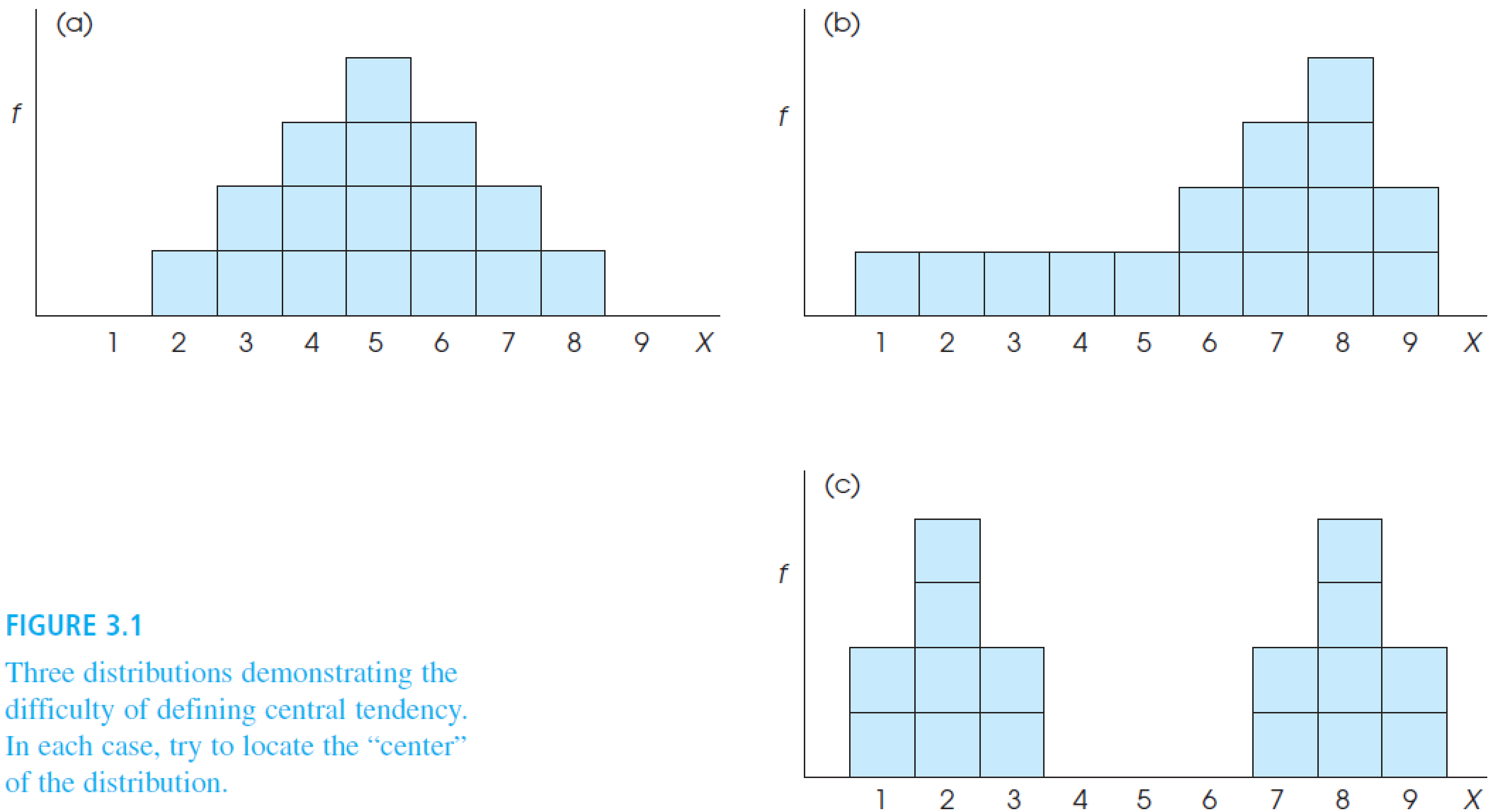


FIGURE 3.1

Three distributions demonstrating the difficulty of defining central tendency. In each case, try to locate the "center" of the distribution.

Mean

Mean is the arithmetic average of the scores in distribution.

Symbol:

μ is for mean of population.

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

Population Mean !

N is size of the population.

$\bar{x}$ is for mean of sample.

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

Sample Mean !

n is size of the sample.

Sample Mean

Definition

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

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

Sample Mean

$$y_i = ax_i + b$$

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

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

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

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

- 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.**

Sample 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, $\mathbf{y_i = x_i - 280}$.

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

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

So, the mean of original data is, $\bar{x} = \bar{y} + 280 = 279.2$

Sample Mean

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

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

- Mean in Class Intervals
  - 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 ! Hint: using definition of the mean.

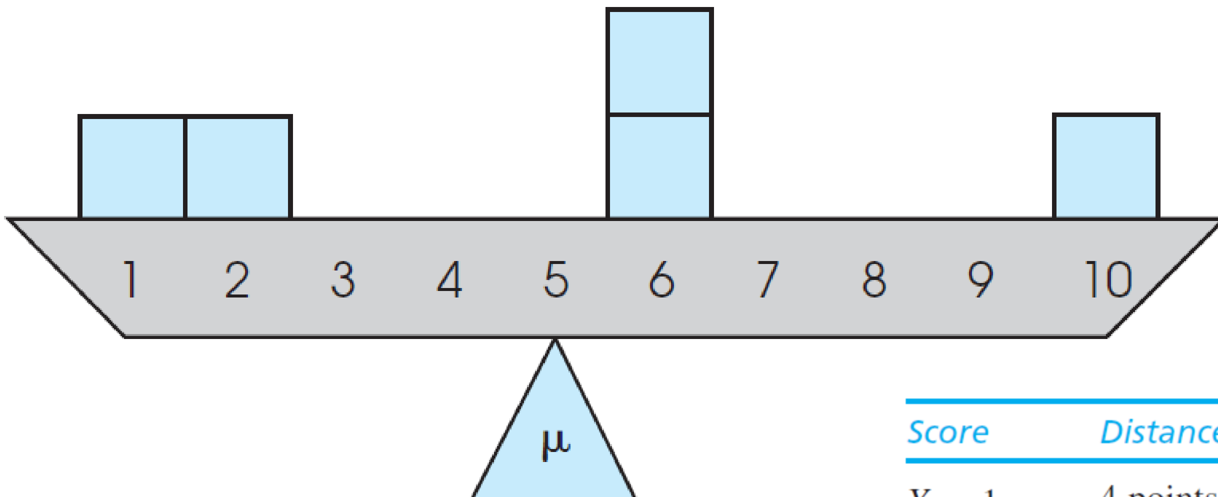
2. The **sum of squares** of the deviation from the mean is **smaller than** the sum of squares of the deviation from any other value in the distribution.

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

Properties of the mean

Idea of Mean

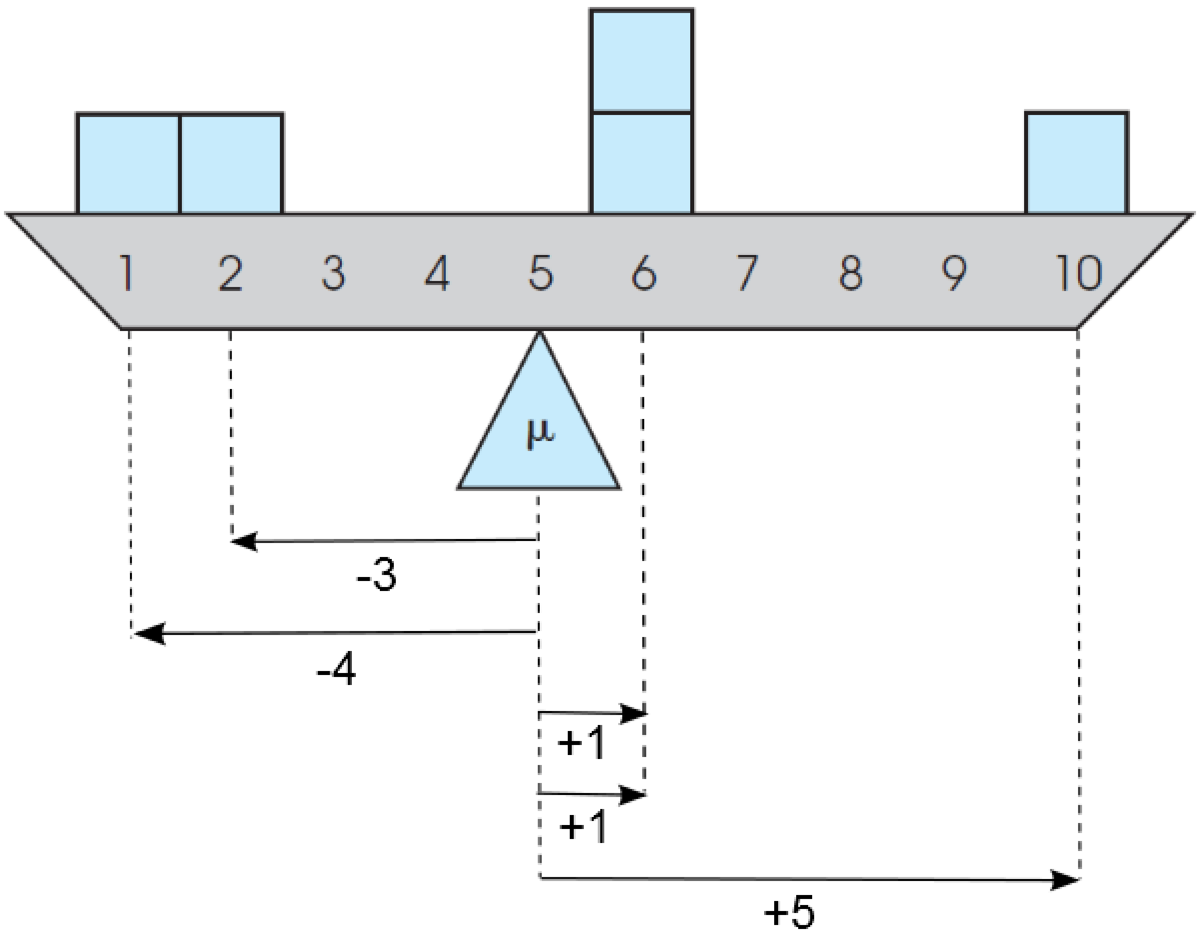
{1, 2, 6, 6, 10}



The frequency distribution shown as a seesaw balanced at the mean.

Score	Distance from the Mean
$X = 1$	4 points below the mean
$X = 2$	3 points below the mean
$X = 6$	1 point above the mean
$X = 6$	1 point above the mean
$X = 10$	5 points above the mean

Properties of the mean



Properties of the mean

$\boldsymbol{x_i}$	$(x_i - \bar{x})$	$(x - \bar{x})^2$	$(x - 8)^2$
9	3	9	1
12	6	36	16
7	1	1	1
5	-1	1	9
2	-4	16	36
3	-3	9	25
4	-2	4	16
Sum	**0**	**76**	**104**

Sample Median

Definition

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, it is the average of the values in positions **n/2** and **n/2 + 1**.

Median is actually **second quartile**.

{3, 6, 12, 18, 19, 21, 23} -> median = 4th datum = **18**.

{3, 6, 12, 18, 19, 21, 23, 25} -> median = (18 + 19) / 2 = **18.5**

Sample Median

For **grouped frequency table** **[Hinkle, 2003]**...

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

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 case, lower limit of an interval is the **left-interval-bound**

Sample Median

Example

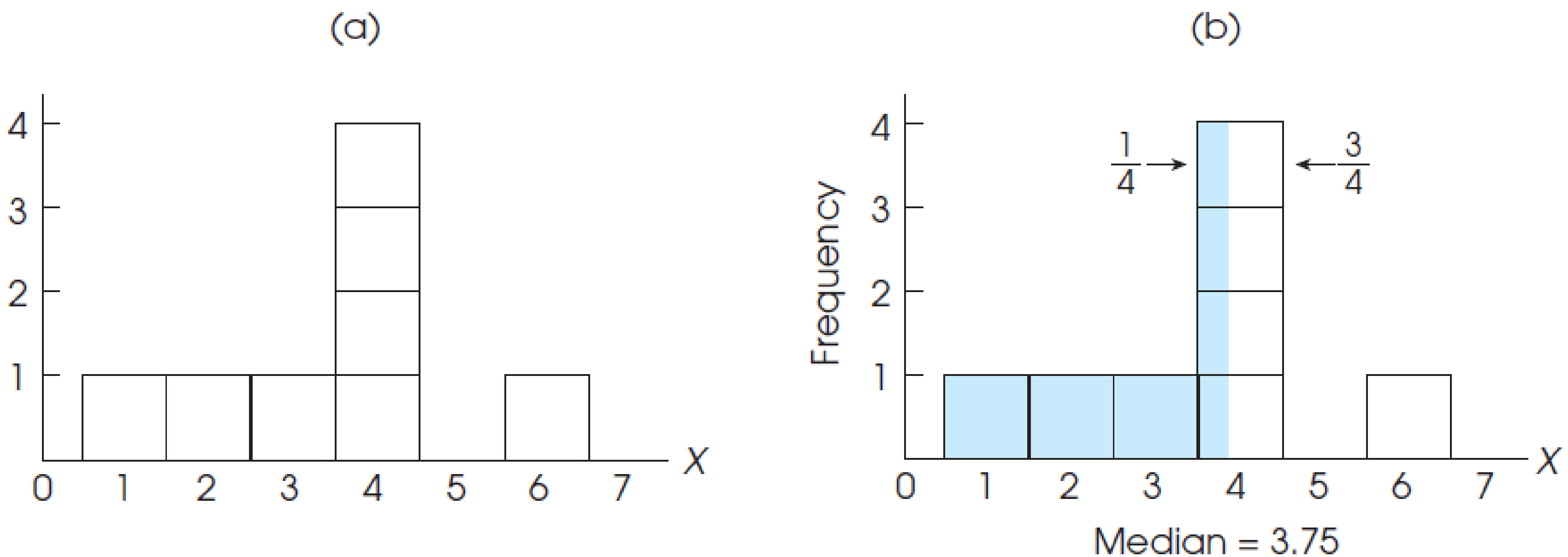
Class Interval	Exact Limits	Midpoint	f	cf
65–69	64.5–69.5	67	6	180
60–64	59.5–64.5	62	15	174
55–59	54.5–59.5	57	37	159
50–54	49.5–54.5	52	30	122
45–49	44.5–49.5	47	42	92
40–44	39.5–44.5	42	22	50
35–39	34.5–39.5	37	18	28
30–34	29.5–34.5	32	7	10
25–29	24.5–29.5	27	2	3
20–24	19.5–24.5	22	1	1

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

Sample Median

Be careful! → Continuous variables

{1., 2., 3., 4., 4., 4., 4., 6.}



Being "mean" is a problem
(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 (60 + 70 + 80 + 990)/4 = **300 ??**

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

Solution: we need another statistic -> **median**.

Median is (70+80)/2 = **75**. 3 observations out of 4 lie between 60-80, **Median** is a good statistic here ☺

Sample Mode

Mode is the most frequent score in a distribution.

Score	f		
783	**6**	←	783 is the most frequent score (6 times)
785	4		
786	2		Mode of the data is 783
788	2		
789	2		
790	2		
791	3		
792	2		

Sample Mode

Multiple Modes are possible: **bimodal or multimodal**

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

- 783 and 803 are the most frequent.
- The data has **dual mode** 783 and 789.
- If no single value occurs, all values that occur as the highest frequency are called **modal values**.

Sample Mode

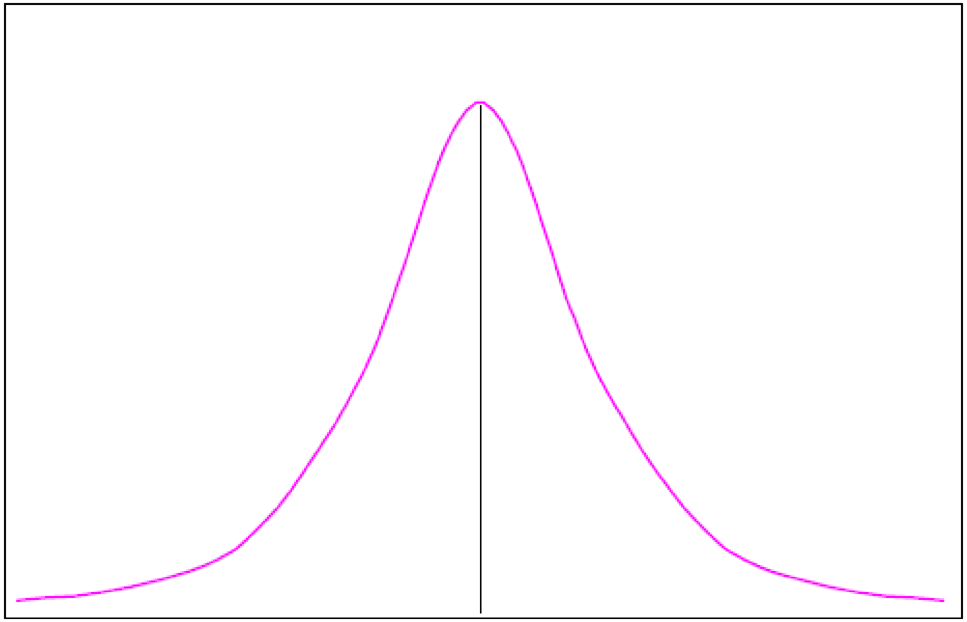
When data are grouped into class intervals (using [Hinkle, 2003]), the mode is a **modal interval**. And the **midpoint** of this interval is considered the mode.

Class Interval	*Exact Limits*	*Midpoint*	*f*	*cf*
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. Hence, the mode is **47**.

Comparisons of the mode, median, and mean in several distributions

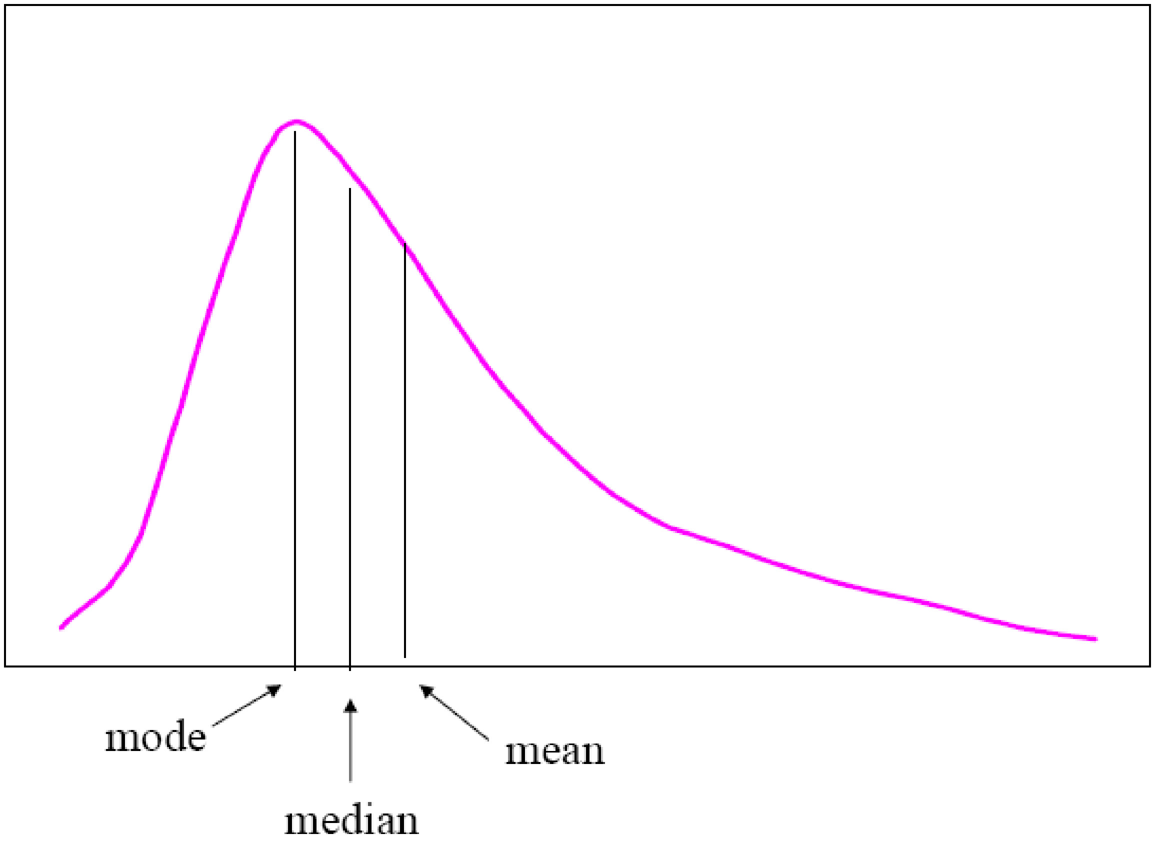
- Normal Distribution



mean, median, mode

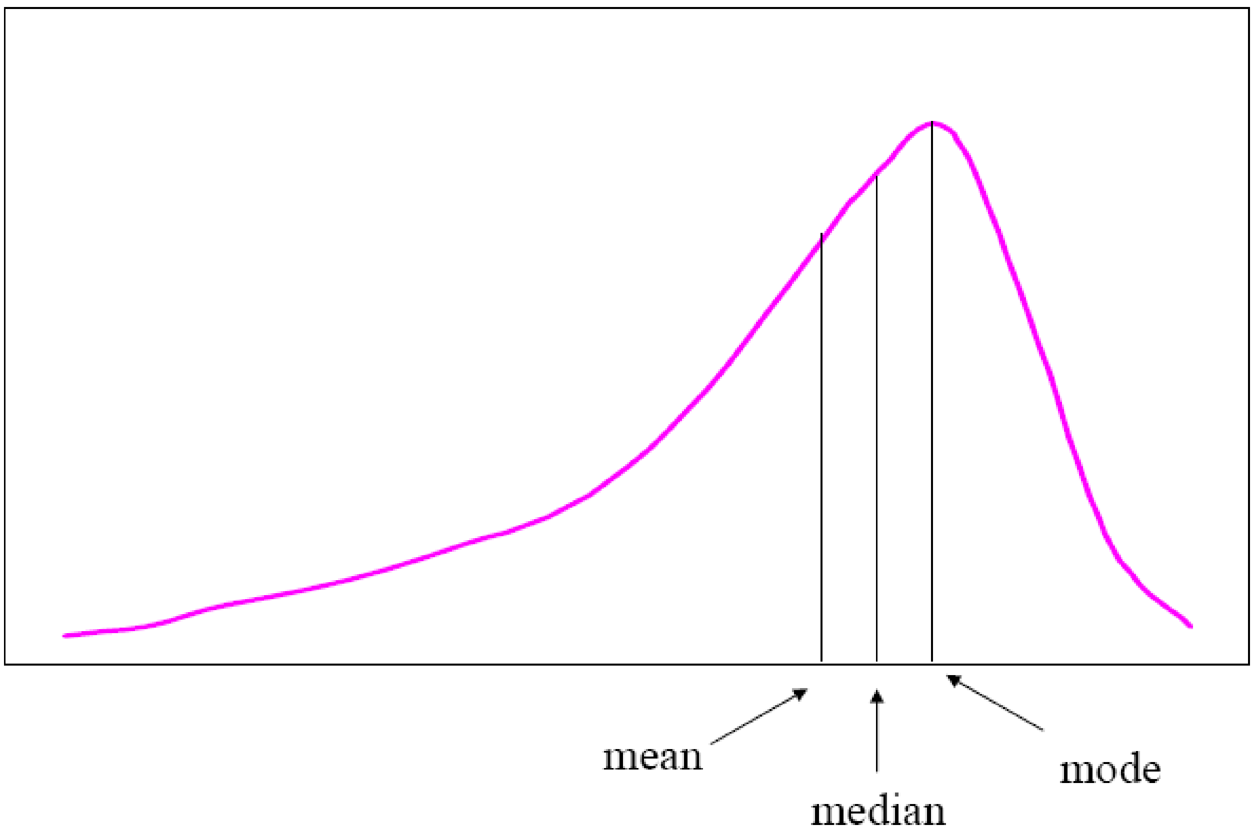
Comparisons of the mode, median, and mean in several distributions

- Positivelv Skewed



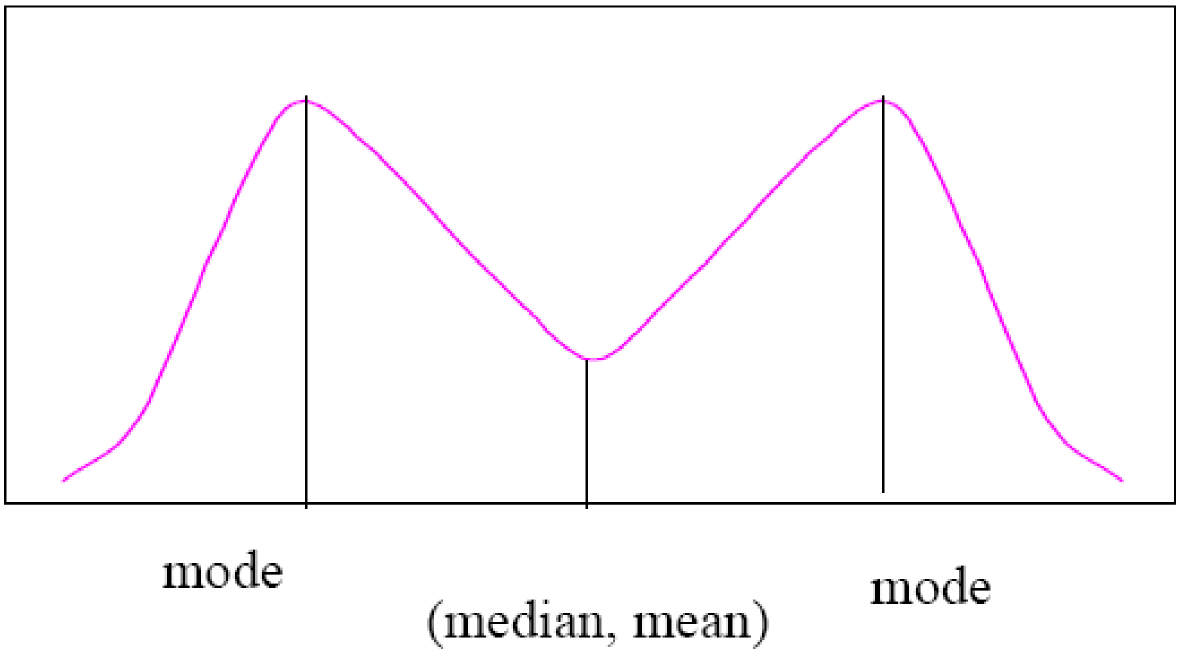
Comparisons of the mode, median, and mean in several distributions

- Negatively Skewed



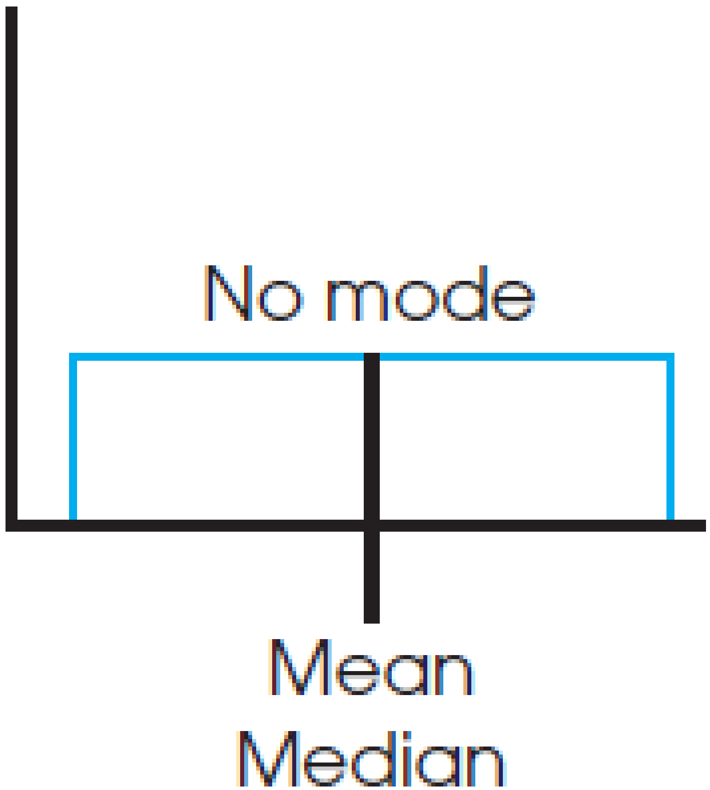
Comparisons of the mode, median, and mean in several distributions

- Dual Mode (bimodal)



Comparisons of the mode, median, and mean in several distributions

- Rectangular



Summarizing data sets

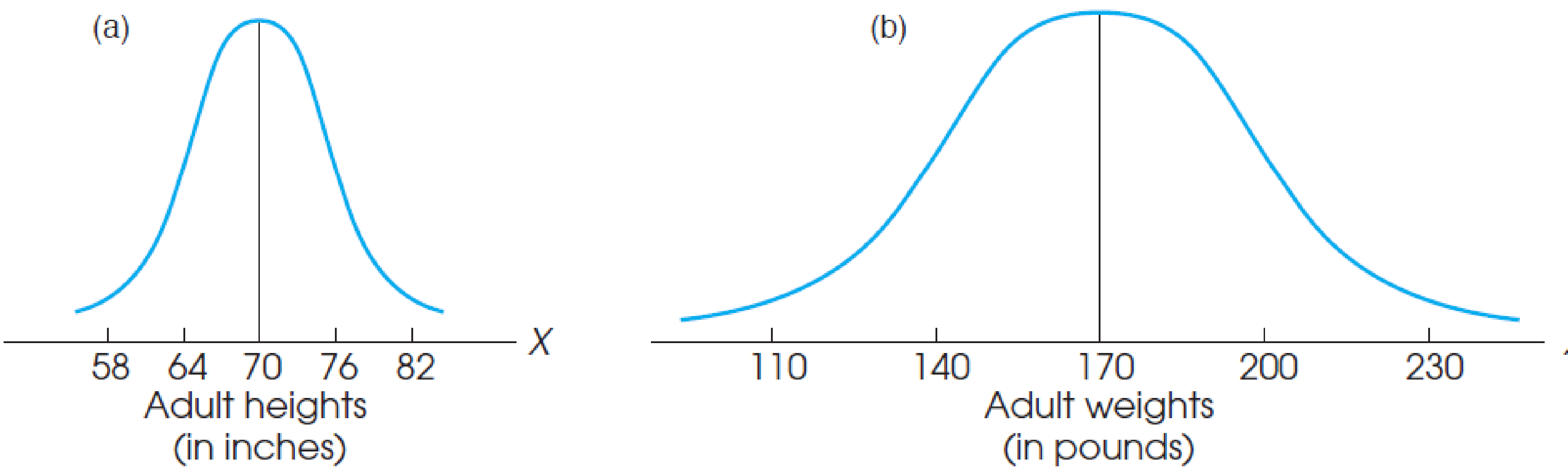
part II: measures of variability (dispersion)

Measures of variability (dispersion)

These statistics measure the **amount of scatter** in a data set.

Basic question: how widely scores are spread throughout the distribution ?

Measures of dispersion give us information about how much our variables vary from the mean.



Population distribution of adult heights and adult weights.

Measures of variability (dispersion)

The measures of variation to be discussed:

- range
- mean deviation
- variance
- standard deviation

Range

Range [Hinkle, 2003]

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

Range = (highest score - lowest score) + 1 unit

Sample: ordered data:

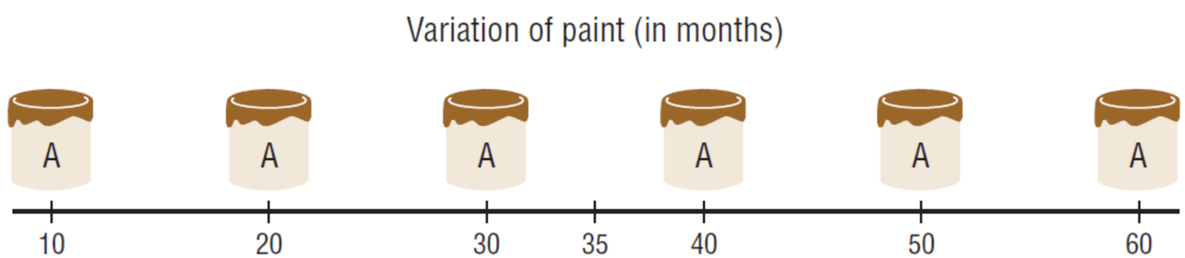
Dist. 1: 11 16 18 ... 31 37

Dist. 2: 18 19 21 ... 26 29

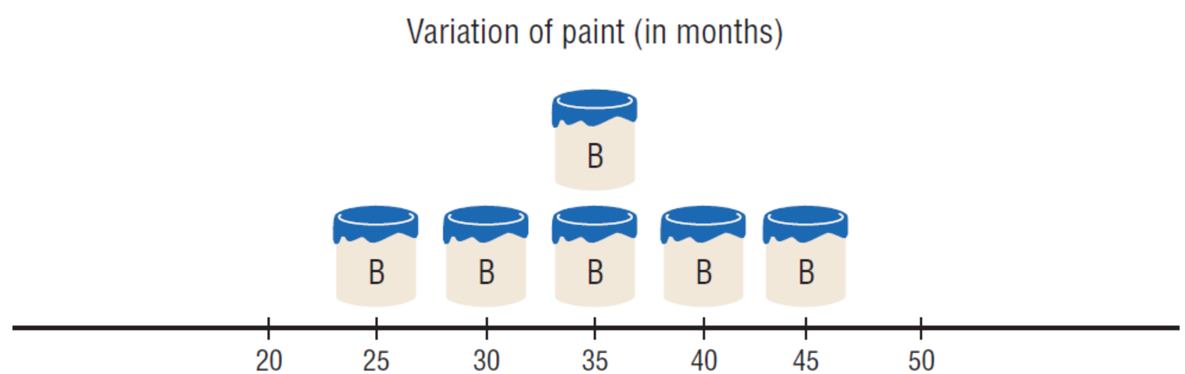
Range(Dist.1) = 37-11+1=27, Range(Dist.2) = 29-18+1=12

Dist.1 are more “varied” !

Range



(a) Brand A



(b) Brand B

Mean Deviation

Deviation score is the difference between 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.**

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

Larger MD have greater variations !

Variance

- Using square instead of absolute.
- Variance is the **average** of the **sum of squared deviations** around the **mean**.

Symbol:

σ^2 is the variance of a population

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

Population variance

SS: sum of square

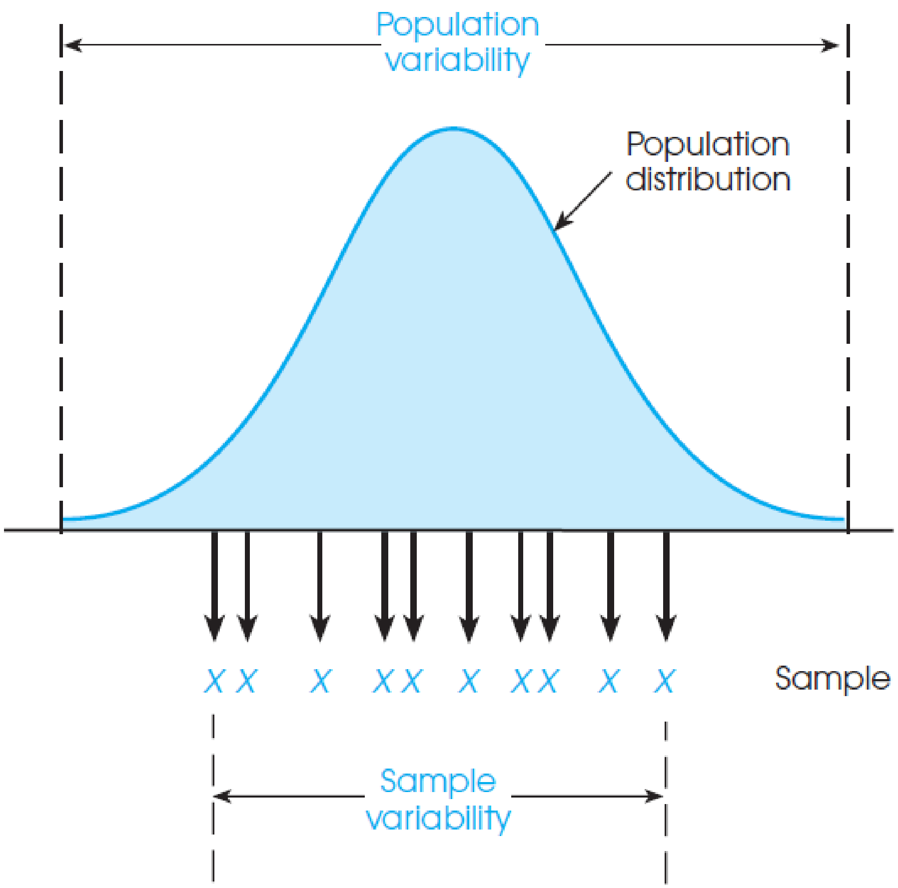
s^2 is the variance of a sample

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

Sample variance

Variance

Why *n*-1?



Population variance

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

Sample variance

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

The population of adult heights forms a normal distribution. If you select a sample from this population, you are most likely to obtain individuals who are near average in height. As a result, the scores in the sample will be less variable (spread out) than the scores in the population.

Sample Variance

Definition

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

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

Sample Variance

$$y_i = ax_i + b$$

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

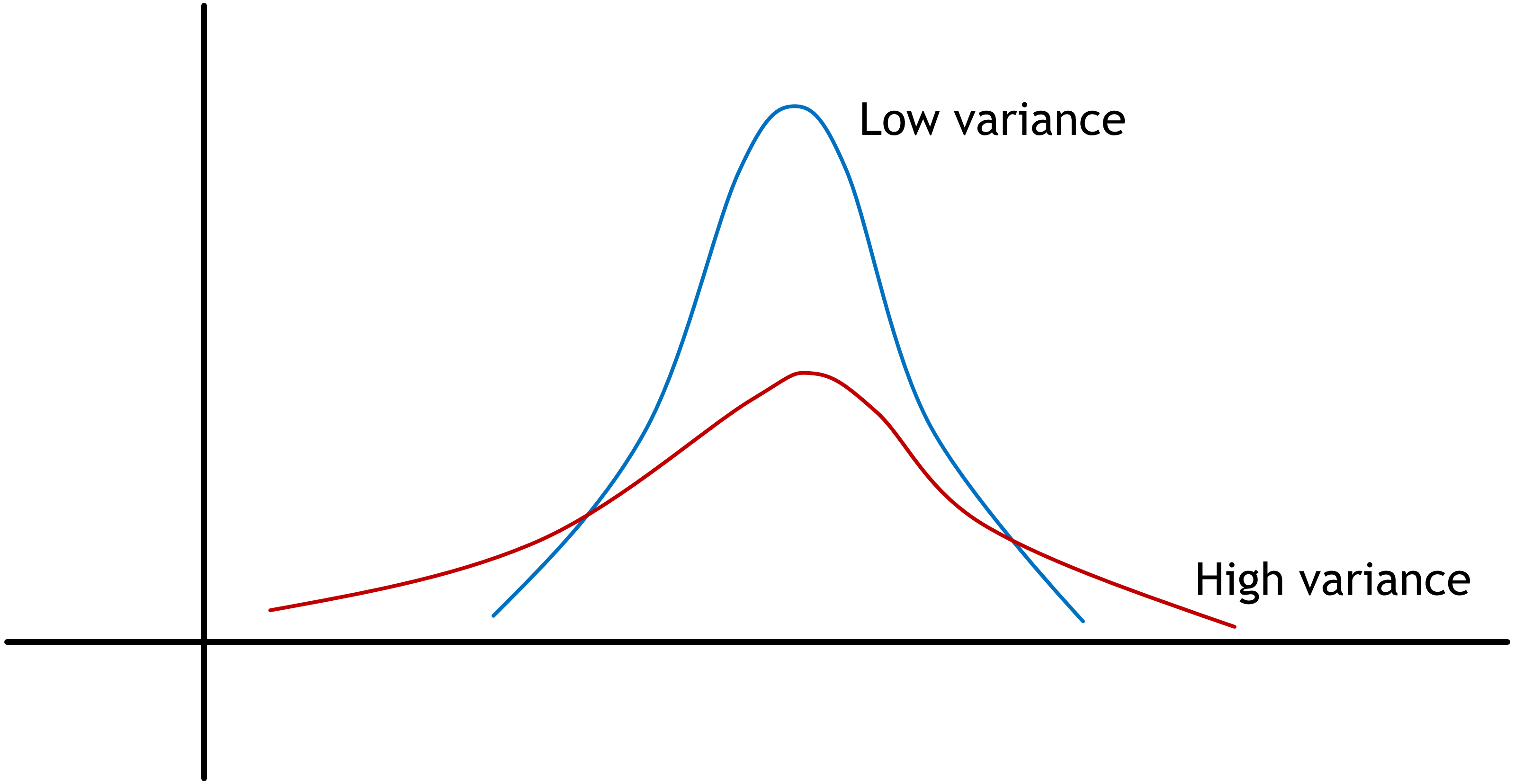
$$s_y{}^2 = \frac{\sum_{i=1}^{n}(ax_i - a\bar{x})^2}{n-1}$$

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

$$s_y{}^2 = a^2 s_x{}^2$$

- Modify the data; multiply with a **constant a** and add with a **constant b**.
- Only **constant a** affects the variance of the new data.
- This can be used to simplify our computation.

Sample Variance



Sample Variance

for **Grouped Data** **[Hinkle et al, 2003]**

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

***f*$_i$** : frequency of the ith interval
m$_i$: midpoint of the ith interval

$$n = \sum_{i=1}^{n} f_i$$

Algebraic Identity

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

Proof:

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

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

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

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

Standard Deviation

Standard deviation is the **square root** of the **variance**.

Symbol:

σ is the standard deviation of a population

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

s is the *sample standard deviation*

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

Example

A sample consisting of 7 elements

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
∑	42	0	76

n=7

total= 42

mean= 42/7= 6

$$s^2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{n-1} = \frac{76}{6} = 12.67$$

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

Problem

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

Summarizing data sets

part III: measures of position

Percentile & Percentile Rank

- Statistics for describing individual scores

- A score is actually meaningless without an adequate frame of reference, i.e., without an indication of the **relative position of a score in the total distribution of scores.**

- Some statistics for this problem
  - Percentile
  - Percentile Rank

Sample Percentile

Definition

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.**

Sample **100p** percentile = $\mathbf{P_{100p}}$

Sample **25** percentile = $\mathbf{P_{25}}$

Sample Percentile

To determine the sample ***100p*** percentile of a data set of size ***n***, we need to determine the data values such that:

1. At least ***np*** of the values are **less than or equal** to it.
2. 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** !

Example:

If **n = 22**, determine the position of **80 percentile** !

Sample Percentile

Example:

If **n = 22**, determine the position of **80 percentile !**

np = 22(0.8) = 17.6 of the values are less than or equal to it.

Clearly, only the **18th smallest value** satisfies both conditions! So, this is the sample 80 percentile ! **P_{80} = 18th value**.

If ***np* is integer**, then both values in positions **np** and **np+1** satisfy both conditions, and so the **sample 100p percentile is the average of these values**.

Sample Percentile

Summary, 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:
   1. If **np** is **not** whole number, round up to the **next** whole number !
   2. If **np** is whole number, compute the average of values in the position **np** and **np+1**.

Sample Percentile

Definition

First quartile (Q_1): the sample 25 percentile.

Second quartile (Q_2): the sample 50 percentile

Third quartile (Q_3): the sample 75 percentile

Second quartile is the sample **median**.

Interquartile Range (IQR) = $Q_3 - Q_1$.

Example:

Determine **first, second, and third quartile**, as well as P_{70} of the following data set !

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

Sample Percentile

Ordered data set:

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

$\mathbf{P_{25}}$ -> np = 8(0.25) = 2. $\mathbf{P_{25}}$ = (3.79 + 4.3)/2 = 4.045

$\mathbf{P_{50}}$ -> np = 8(0.50) = 4. $\mathbf{P_{50}}$ = (6.59 + 6.6)/2 = 6.595

$\mathbf{P_{75}}$ -> np = 8(0.75) = 6. $\mathbf{P_{75}}$ = (6.96 + 11.06)/2 = 9.01

$\mathbf{P_{70}}$ -> np = 8(0.70) = 5.6. $\mathbf{P_{70}}$ = 6.96

Interquartile Range (**IQR**) = $\mathbf{Q_3}$ **-** $\mathbf{Q_1}$ = 9.01 – 4.045 = 4.965

Sample Percentile

For **grouped frequency table [Hinkle, 2003]**...

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

ll: lower exact limit of the interval containing the percentile point

n: total number of scores

p: proportion corresponding to the desired percentile

cf: cumulative freq. of scores below the interval containing the percentile point

f_i: freq. of scores in the interval containing the percentile point

w: width of class interval

For left-end-inclusion case, lower limit of an interval is the **left-interval-bound**

Sample Percentile

Find **34th percentile** !

Class Interval	*Exact Limits*	*Midpoint*	*f*	*cf*
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

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

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.**

Find percentile rank of score 65 !

Notation : $\mathbf{PR_{65}}$

Determining percentile rank in **non-grouped data** is easy ! How ?

We will focus on how to determine percentile rank in **grouped data.**

Percentile Rank

For **non-grouped** data:

$$PR_X = \frac{\langle number\ of\ values\ below\ X \rangle + 0.5}{total\ number\ of\ values} \times 100$$

Find percentile rank of a score of 12 from the following data:

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

Percentile Rank

Percentile Rank in grouped data **[Hinkle, 2003]**

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

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 case, lower limit of an interval is the **left-interval-bound**

Percentile Rank

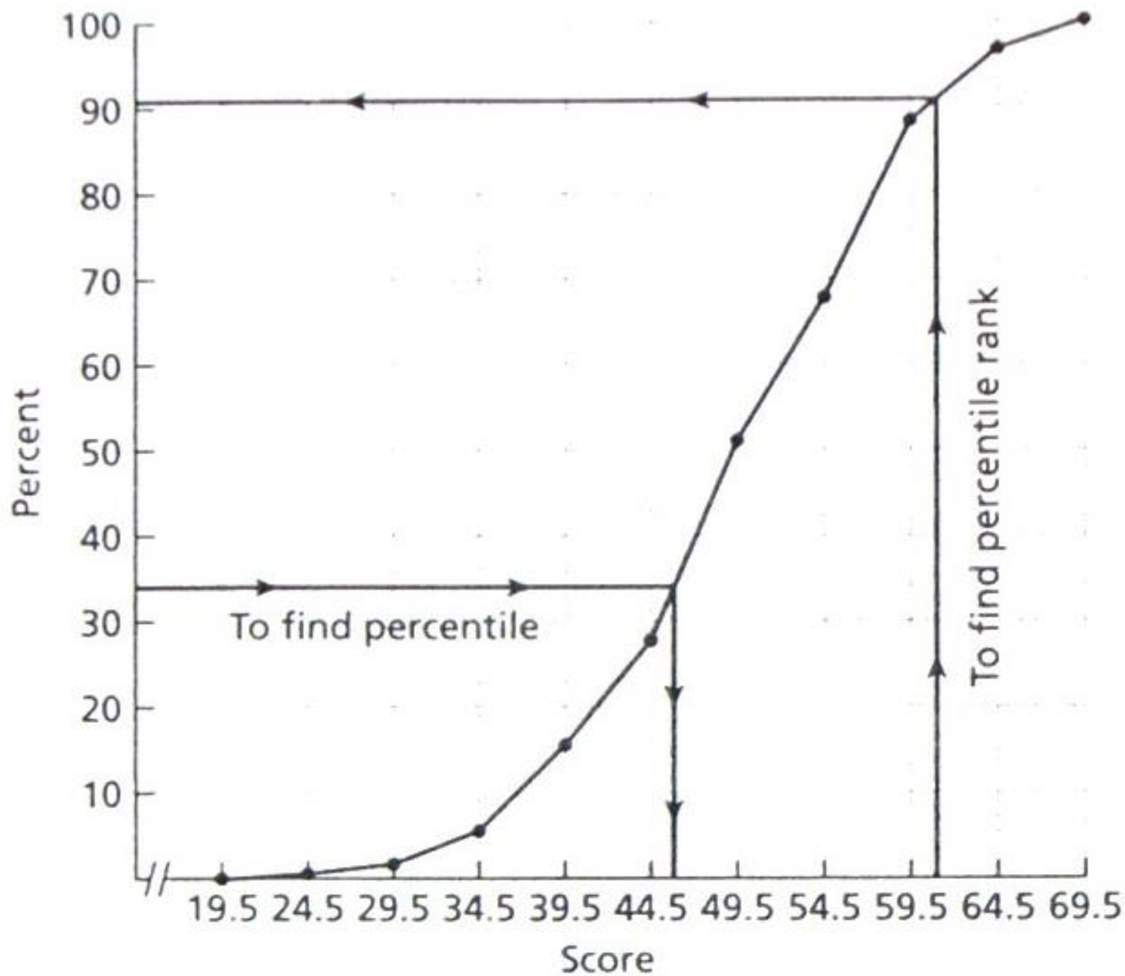
Find percentile rank of score 61 !

Class Interval	Exact Limits	Midpoint	f	cf
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

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

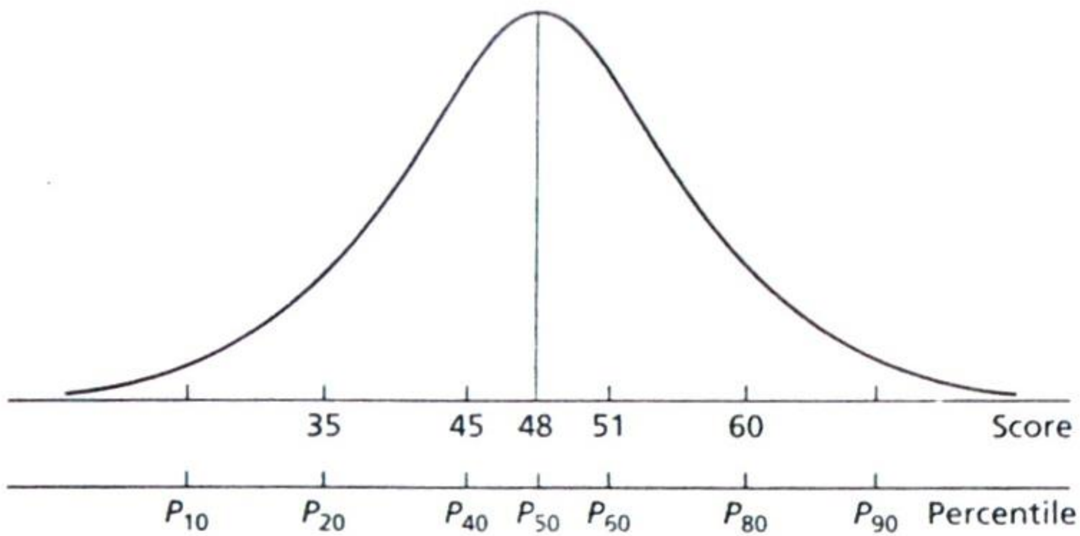
Ogive & Percentile

Ogive can be used to find percentile & percentile rank



Percentile Rank – 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 – an ordinal scale

The difference between $\mathbf{P_{50}}$-$\mathbf{P_{40}}$ and $\mathbf{P_{20}}$-$\mathbf{P_{10}}$ may **not** be the same ! -> **ordinal scale**

Suppose there are two distributions. Score A is from distribution 1, and score B is from distribution 2.

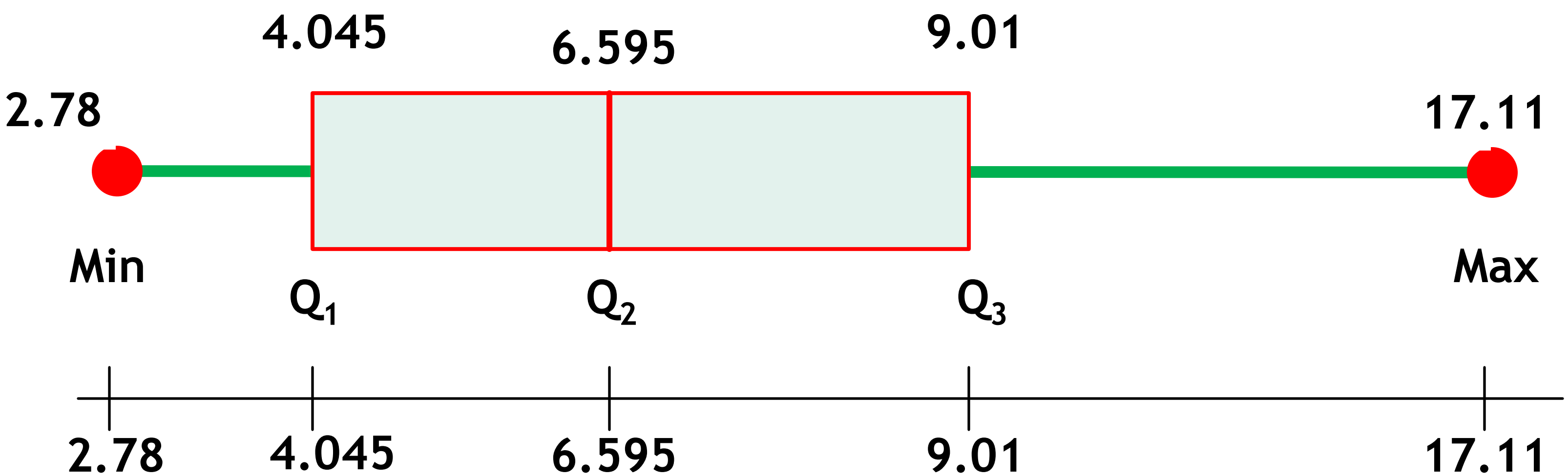
$PR_A = 50$, $PR_B = 48$

Even though the difference of Both PRs is **just 2 points**, we **don't** have any idea about |A – B|. It could be small or large.

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.



Box plot of previous example

Length of the box represents interquartile range.

Box Plot

Example: comparing sodium content of cheese

A dietitian is interested in comparing the sodium content of real cheese with the sodium content of a cheese substitute. The data for two random samples are shown. Compare the distributions, using boxplots.

Real cheese				Cheese substitute			
310	420	45	40	270	180	250	290
220	240	180	90	130	260	340	310

Find Q_1, MD, and Q_3 for the real cheese data.

$$\begin{array}{cccccccc} 40 & 45 & 90 & 180 & 220 & 240 & 310 & 420 \\ & \multicolumn{2}{c}{\uparrow} & \multicolumn{2}{c}{\uparrow} & \multicolumn{2}{c}{\uparrow} & \\ & \multicolumn{2}{c}{Q_1} & \multicolumn{2}{c}{\text{MD}} & \multicolumn{2}{c}{Q_3} & \end{array}$$

$$Q_1 = \frac{45 + 90}{2} = 67.5 \qquad \text{MD} = \frac{180 + 220}{2} = 200$$

$$Q_3 = \frac{240 + 310}{2} = 275$$

Find Q_1, MD, and Q_3 for the cheese substitute data.

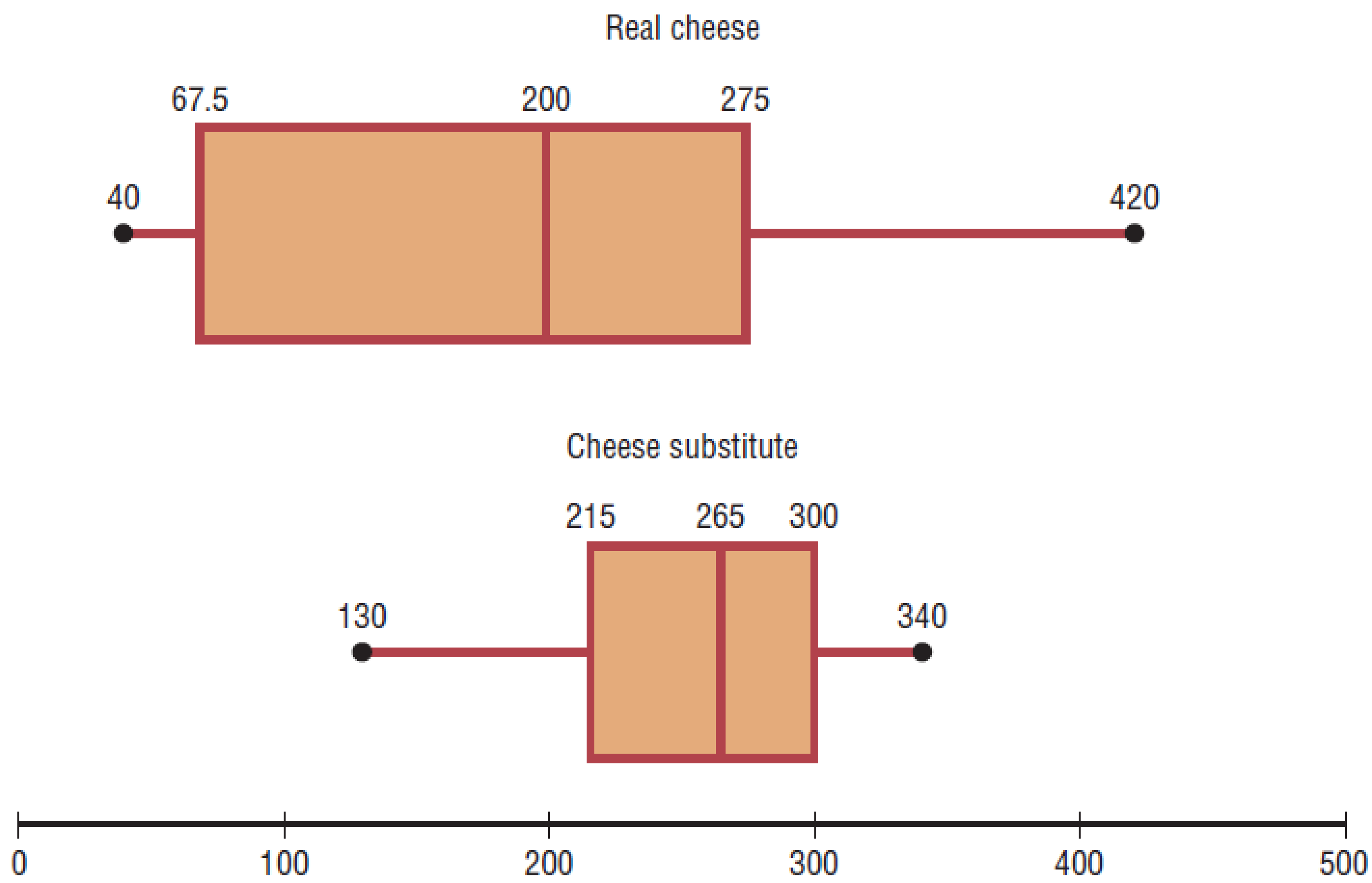
$$\begin{array}{cccccccc} 130 & 180 & 250 & 260 & 270 & 290 & 310 & 340 \\ & \multicolumn{2}{c}{\uparrow} & \multicolumn{2}{c}{\uparrow} & \multicolumn{2}{c}{\uparrow} & \\ & \multicolumn{2}{c}{Q_1} & \multicolumn{2}{c}{\text{MD}} & \multicolumn{2}{c}{Q_3} & \end{array}$$

$$Q_1 = \frac{180 + 250}{2} = 215 \qquad \text{MD} = \frac{260 + 270}{2} = 265$$

$$Q_3 = \frac{290 + 310}{2} = 300$$

Box Plot

Example: comparing sodium content of cheese



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.

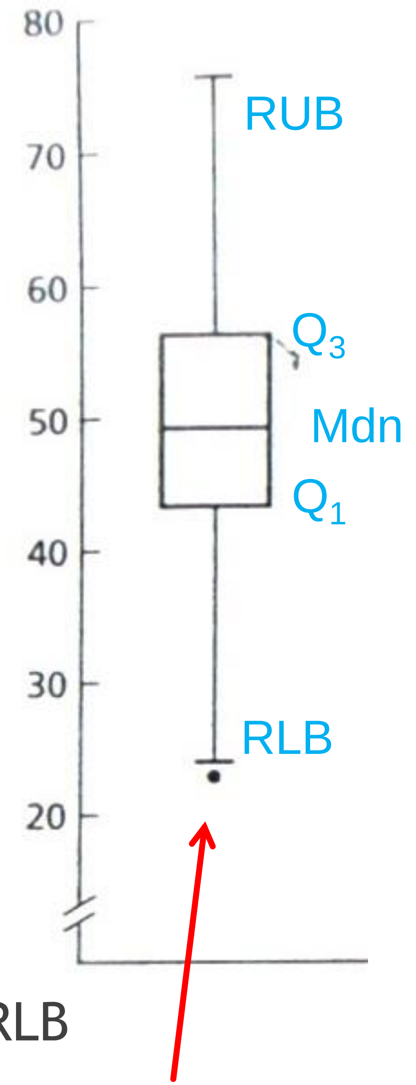
Outliers & Box Plot

- Modify box plot to use 5 numbers as follow:
  - RUB (reasonable upper boundary)

    RUB = Q_3 + 1,5 (IQR)
  - Q_3 (third quartile)
  - Median
  - 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

Modified Boxplot



This dot represents an **outlier** !

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 ?

How to answer this question ?

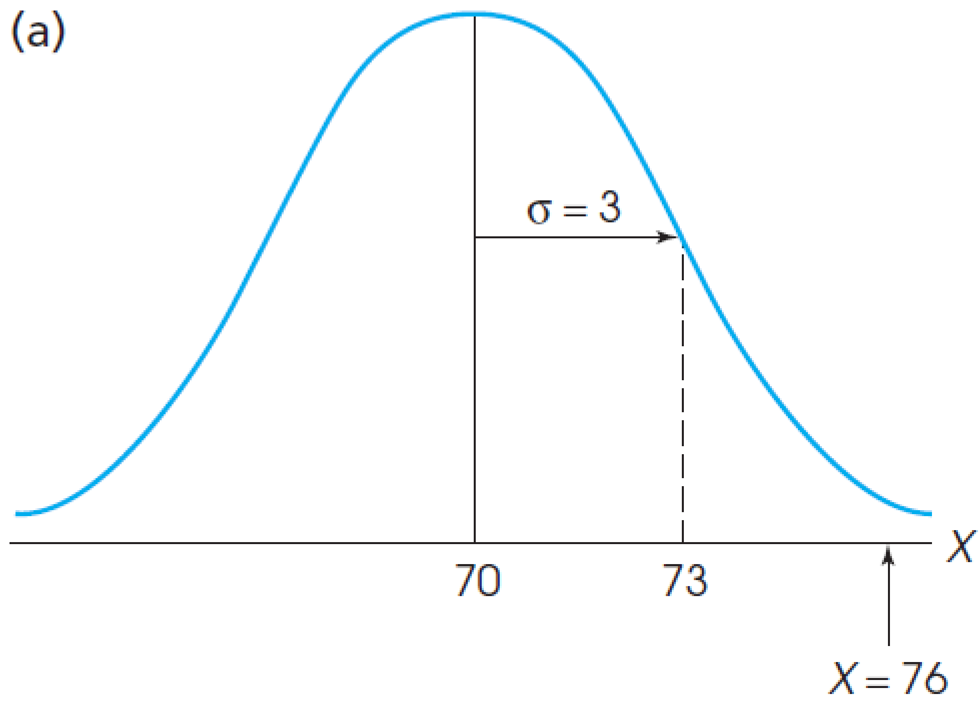
We can not use raw scores !

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

We can not use percentile ! Percentile is an ordinal scale !

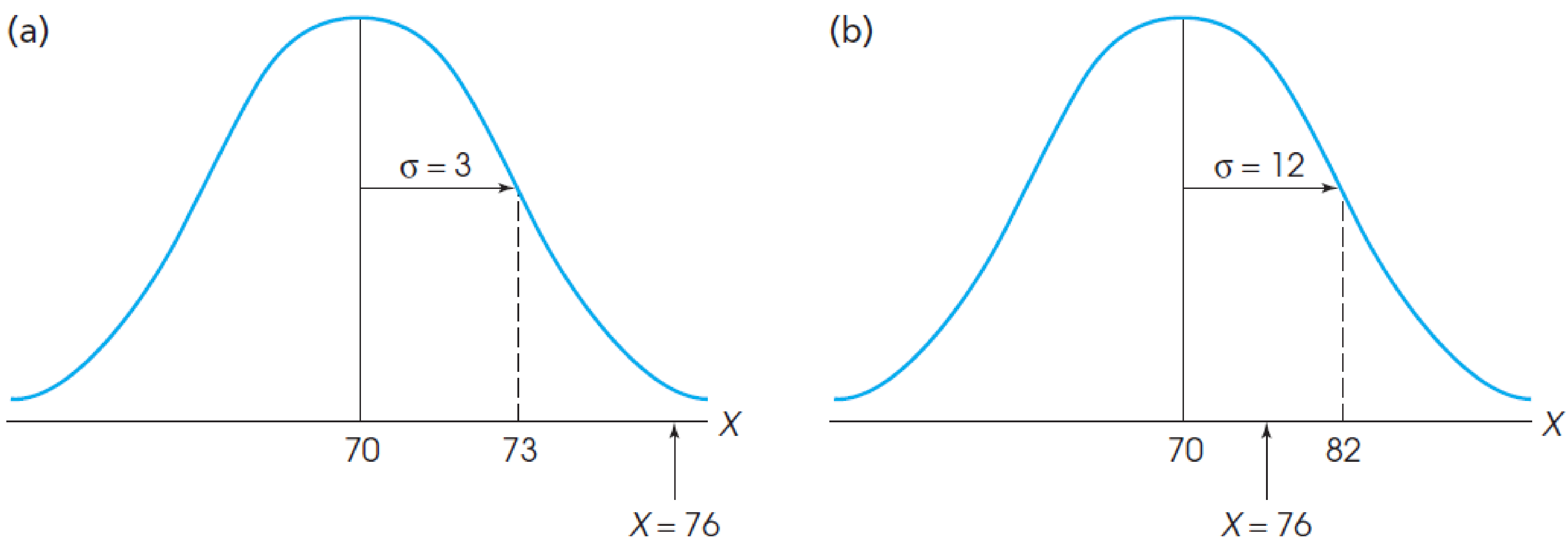
Standard Scores

Example: a score 76 in exam A.



Standard Scores

Compare the same scores in different exams.



Two distributions of exam scores. For both distributions, $\mu = 70$, but for one distribution, $\sigma = 3$, and for the other, $\sigma = 12$. The relative position of $X = 76$ is very different for the two distributions.

Standard Scores

One way is to transform the scores into scores on an **equal interval** scale.

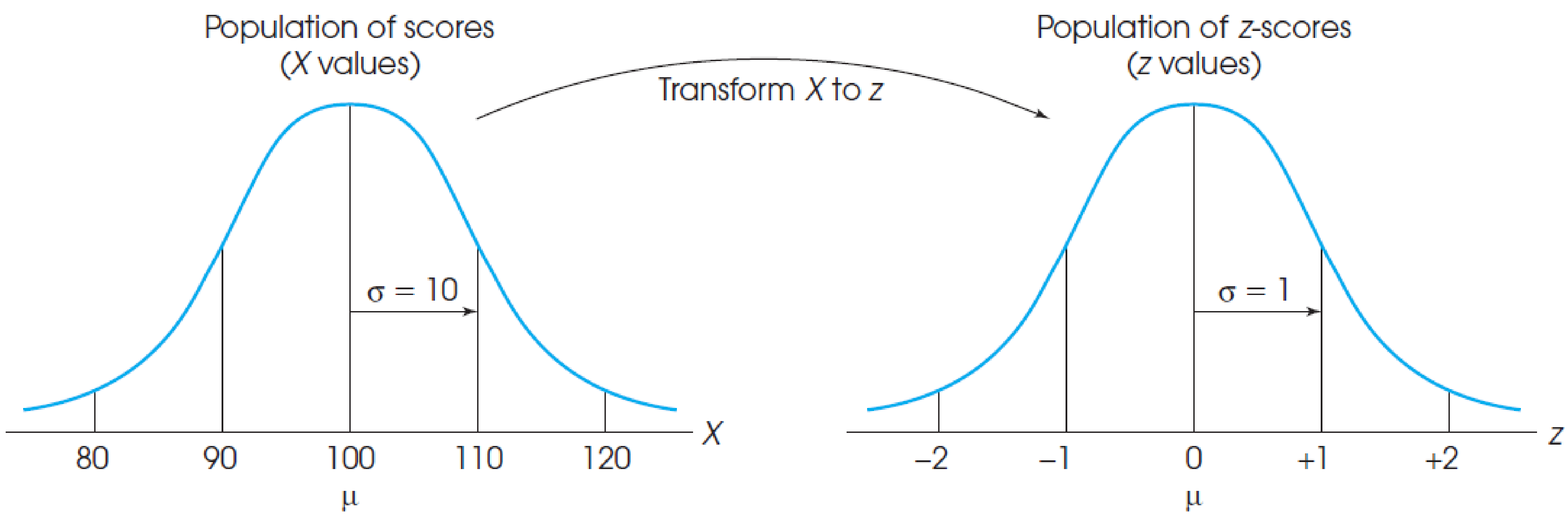
Standard scores do this by using standard deviation as the unit of measure.

Standard score or **z-score** is computed as follow:

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

Standard Scores

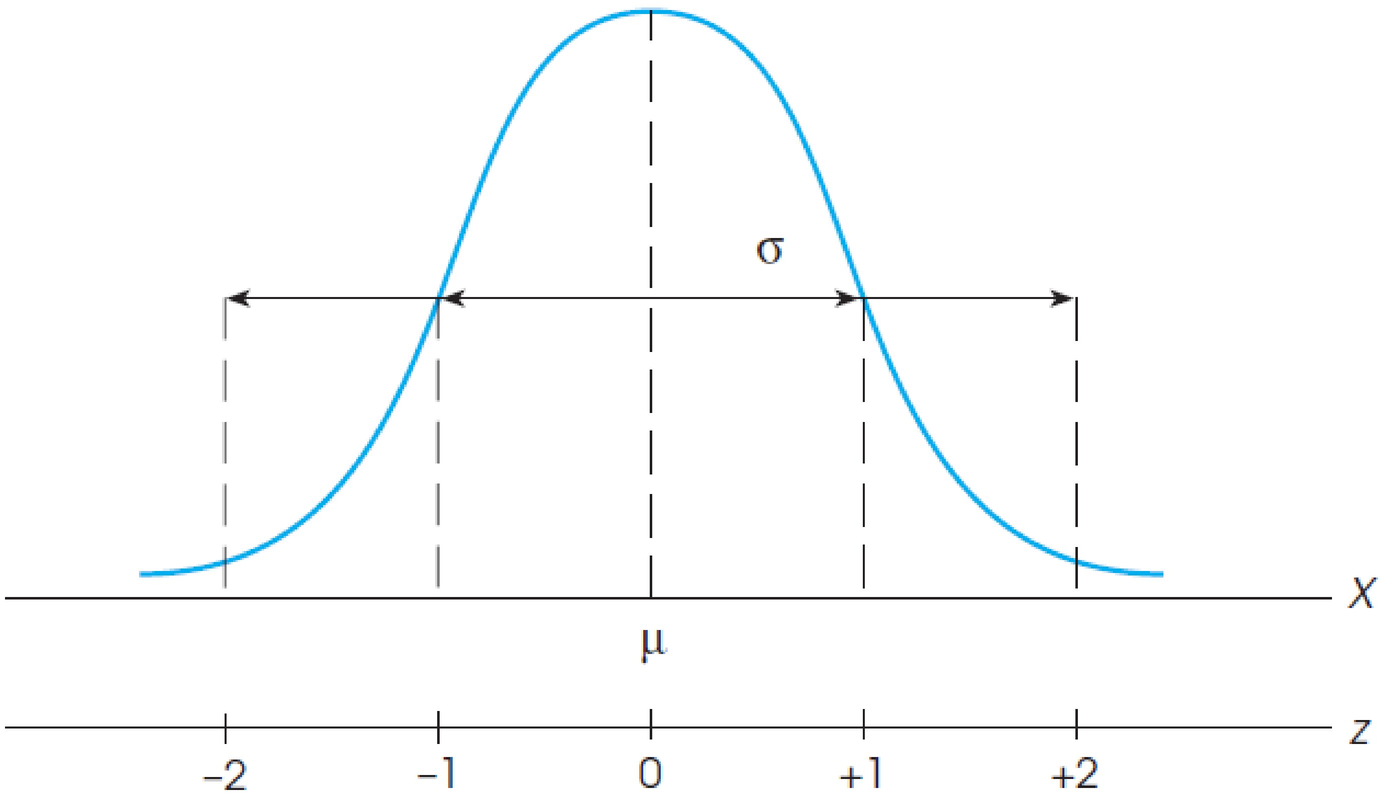
Transforming into z-scores



The transformation does not change shape of the population.

Standard Scores

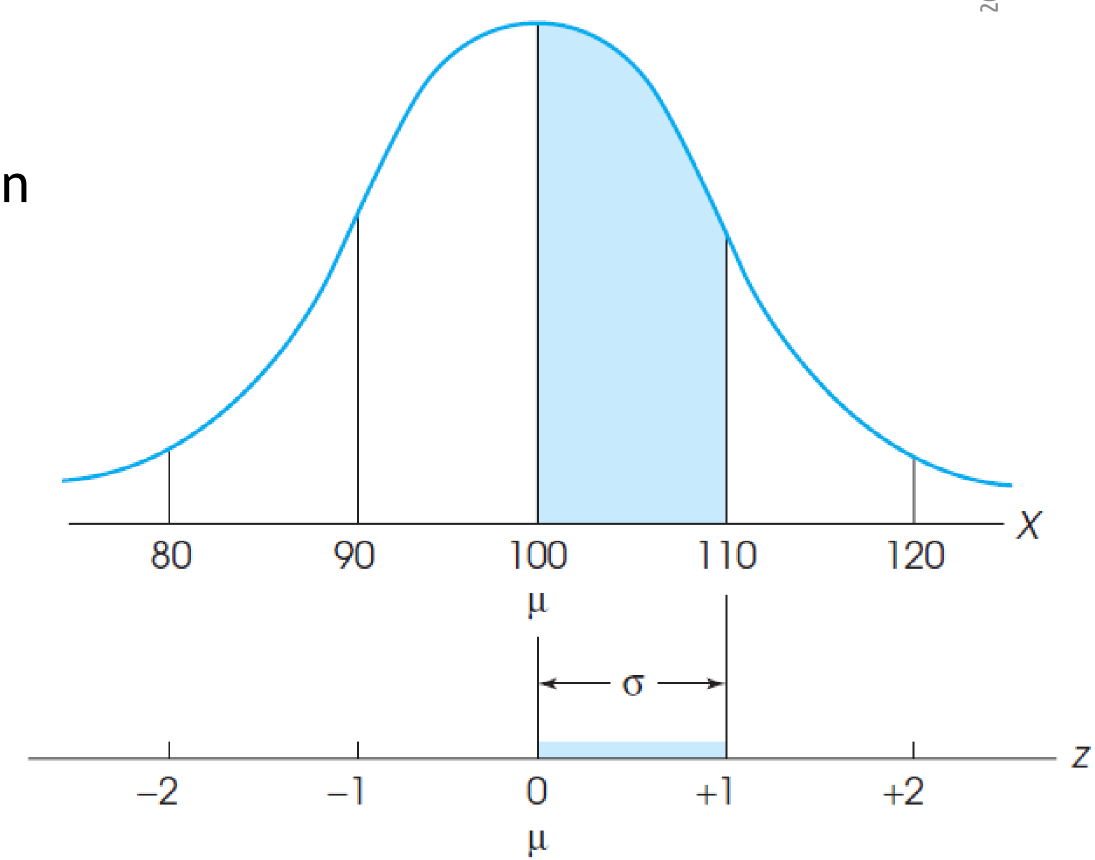
z-score values and locations in a distribution.



Standard Scores

Properties of z-score

- **retains** the shape distribution of the origin
- mean = 0
- variance = 1,
- standard deviation = 1



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

Standard Scores

Example

$z = (10-6)/3.18 = 1.26$

TABLE 3.6
Distribution of Raw Scores and z Scores

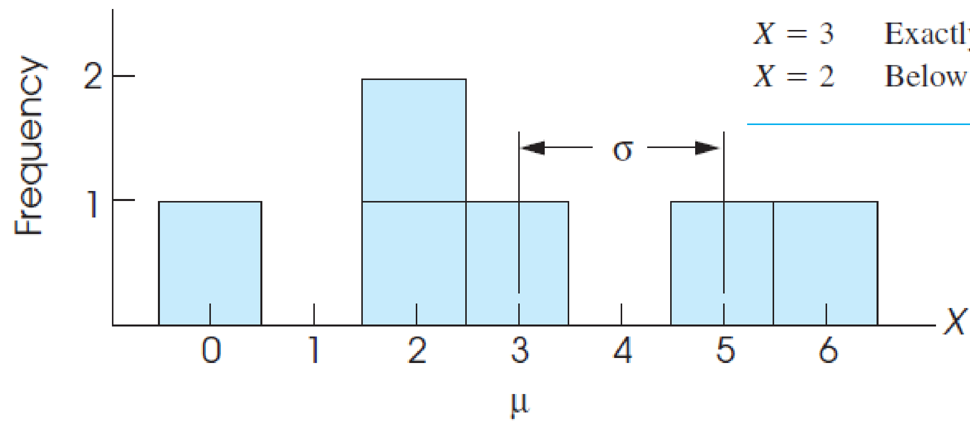
Subject	*Raw Score X*	*Standard Score z*
A	10	1.26
B	9	0.94
C	3	−0.94
D	10	1.26
E	9	0.94
F	2	−1.26
G	2	−1.26
H	10	1.26
I	5	−0.31
J	5	−0.31
K	1	−1.57
L	6	0.00
M	8	0.63
N	6	0.00
O	6	0.00
P	1	−1.57
Q	3	−0.94
R	6	0.00
S	10	1.26
T	8	0.63
$n = 20$		
Mean ($\bar{X}$)	6.0	0
Standard Deviation (s)	3.18	1.00

Standard Scores

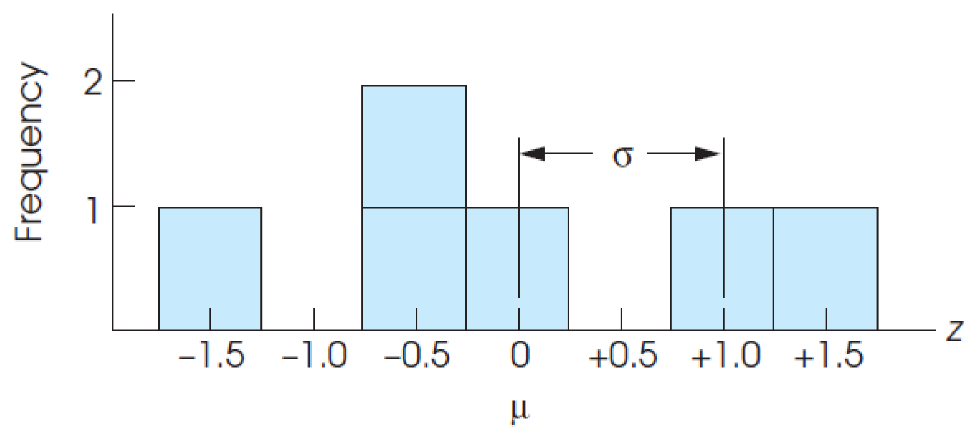
Example: {0, 6, 5, 2, 3, 2}

$X = 0$	Below the mean by $1\frac{1}{2}$ standard deviations	$z = -1.50$
$X = 6$	Above the mean by $1\frac{1}{2}$ standard deviations	$z = +1.50$
$X = 5$	Above the mean by 1 standard deviation	$z = +1.00$
$X = 2$	Below the mean by $\frac{1}{2}$ standard deviation	$z = -0.50$
$X = 3$	Exactly equal to the mean—zero deviation	$z = 0$
$X = 2$	Below the mean by $\frac{1}{2}$ standard deviation	$z = -0.50$

(a)



(b)



Standard Scores

Back to previous question....

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.**

Weighted Averages

How to develop a composite score from two or more individual score ?

For example: we want to compute composite scores of two technical test and one personality test to evaluate a job seeker.

$$\text{Weighted score}_j = \frac{\sum W_i z_{ij}}{W_i}$$

W_i = weight of each test

Z_{ij} = standard score for person j on test i

Why do we use standard score ?

Transformed Standard Scores

Z-scores are for purposes of **comparison.**

But, they can be misleading for people. In the previous table, the **score of math is 0.5** and **history is -0.75**. People might consider **"0.5"** as a bad score !

So, 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

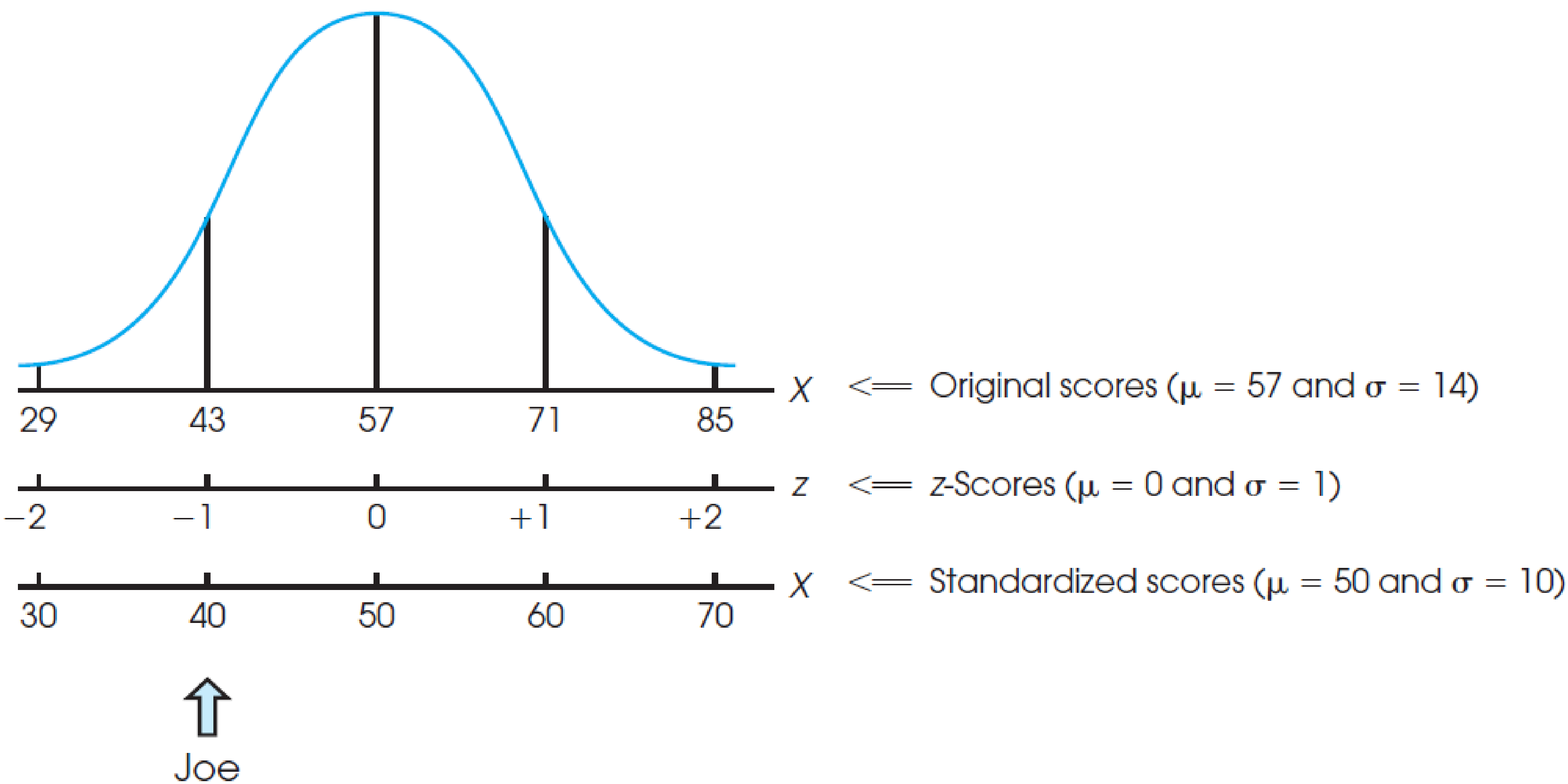
s' = the desired standard deviation

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

Transformed Standard Scores

Example

	Original Scores $\mu = 57$ and $\sigma = 14$		z-Score Location		Standardized Scores $\mu = 50$ and $\sigma = 10$
Maria	$X = 64$	$\rightarrow$	$z = +0.50$	$\rightarrow$	$X = 55$
Joe	$X = 43$	$\rightarrow$	$z = -1.00$	$\rightarrow$	$X = 40$

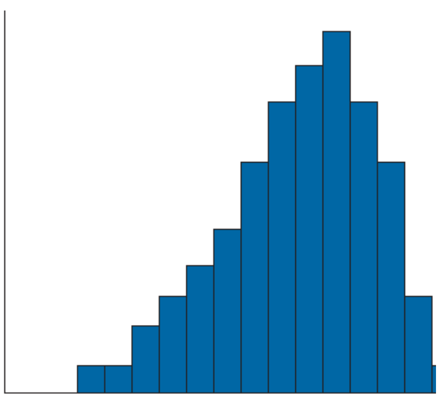


Normal Data Sets

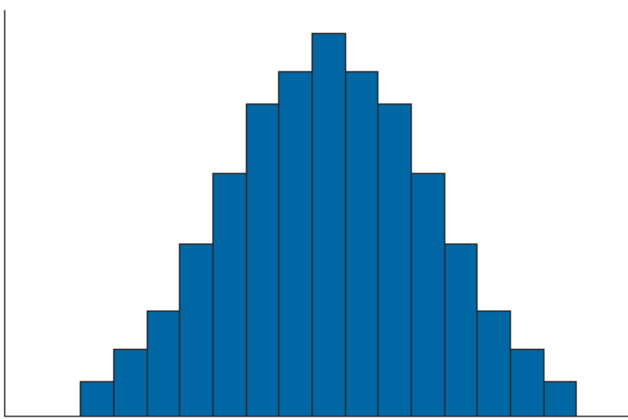
Normal Data Sets

Many of the large data sets observed **in practice** have histograms that are similar in shape.

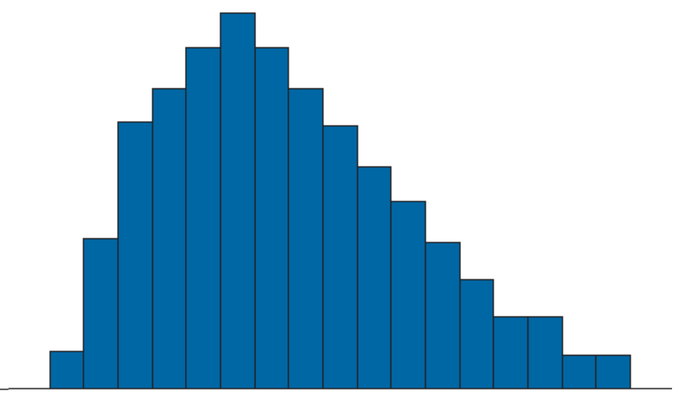
These histrograms often reach their peaks at the sample median and then decrease on both sides of this point in **bell-shaped symmetric fashion**.



Skewed left



Normal histogram



Skewed right

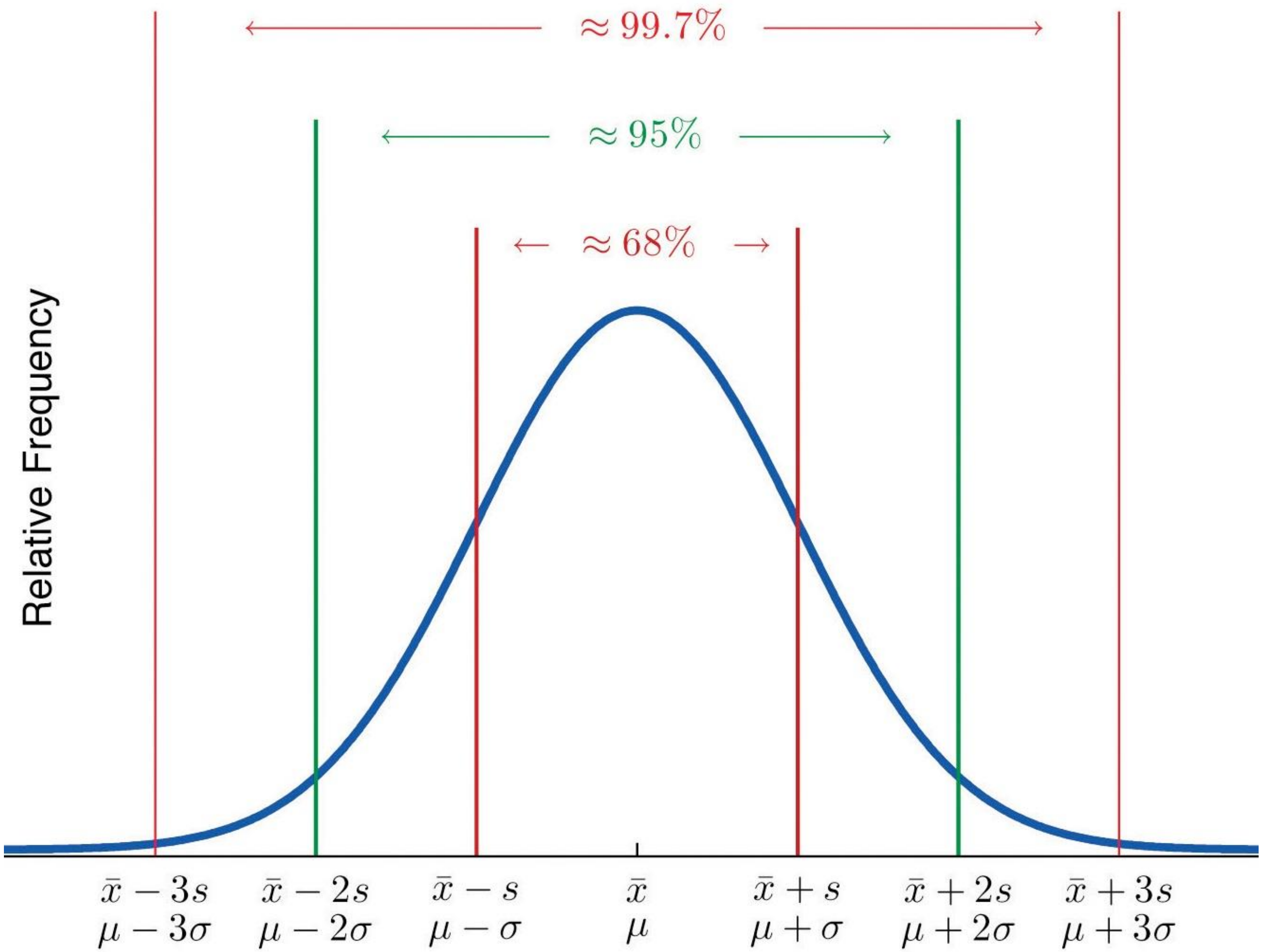
Normal Data Sets

If the histogram of a data set is close to being a normal histogram, then we say that the data set it ***approximately normal***.

If a data set is approximately normal with sample mean $\bar{x}$ and standard deviation s. The following statements are true:

- 68% of data lies within $\bar{x} \pm s$
- 95% of data lies within $\bar{x} \pm 2s$
- 99.7% lies within $\bar{x} \pm 3s$

Normal Data Sets



Chebyshev's Inequality

Chebyshev's Inequality

Idea: at least $1-\frac{1}{k^2}$ of data from a sample must fall within *k* standard deviations from the mean (within interval from $\bar{x} - ks$ to $\bar{x} + ks$), where *k* >= 1.

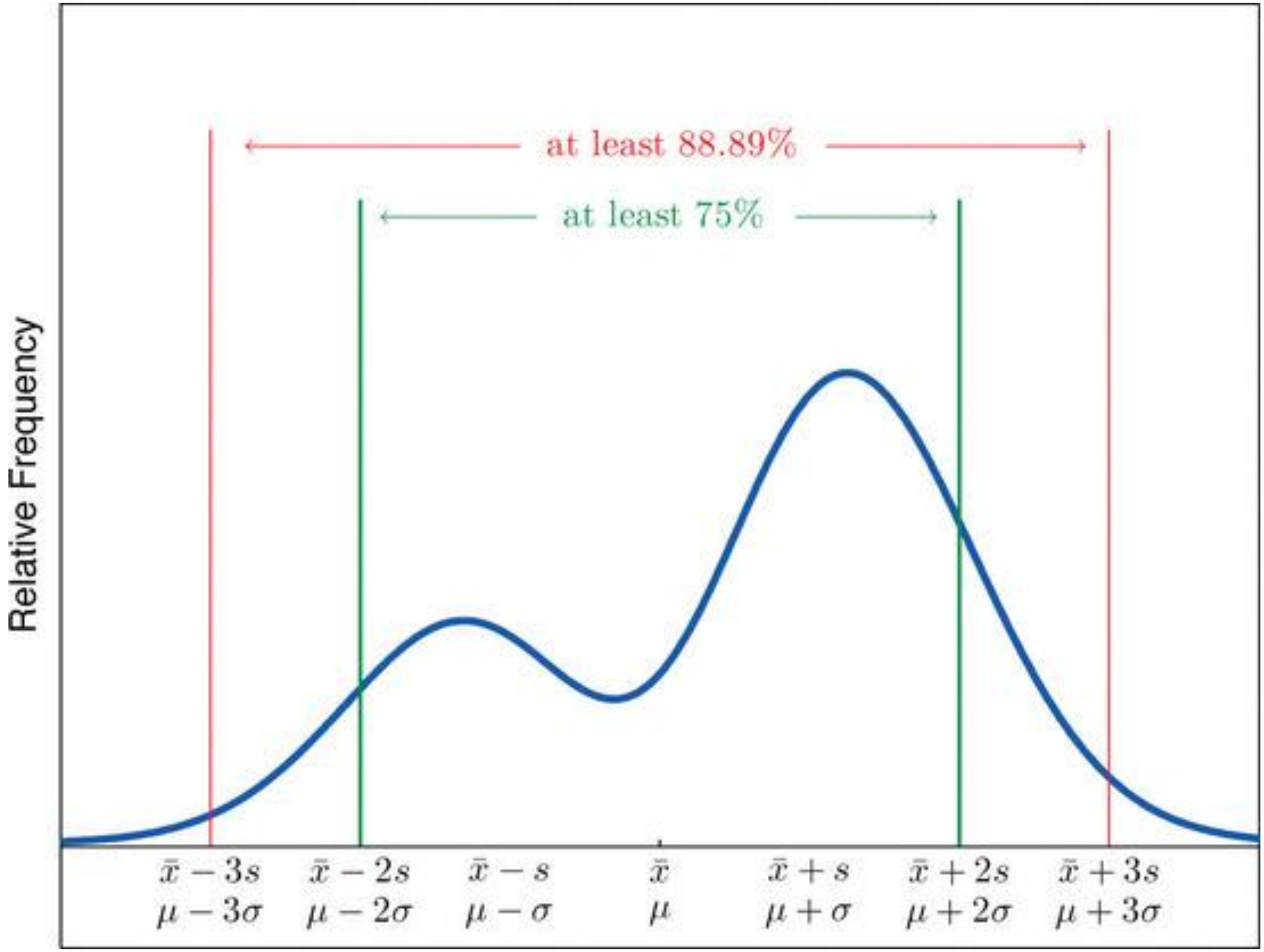
Able to deal with any data sets, not only the normal ones.

Proven mathematically (vs. emprical).

Example:

- *k* = 2 → 1 - ¼ = 75% (at least 75% of data...)
- *k* = 3 → 1 - 1/9 = 88.89% (at least 88.89% of data...)

Chebyshev's Inequality



Chebyshev's Inequality

A simple example

Weights of dogs in a local animal shelter have the mean of 20 pounds with standard deviation of 3 pounds.

Using *k*=2 → $(\bar{x} - ks,\ \bar{x} + ks)$ = (20-6, 20+6) = (14, 26).

It means, at least 75% of the dogs have weight within the range from 14 to 26 pounds.

Chebyshev's Inequality

Application of Chebyshevs In. to normal distribution

Interval	Chebyshev	Empirical rule
$\bar{x} \pm s$	0%	68.27%
$\bar{x} \pm 2s$	75%	95.45%
$\bar{x} \pm 3s$	88.89%	99.73%
$\bar{x} \pm 4s$	93.75%	99.99%
$\bar{x} \pm 5s$	96%	99.9999%

Chebyshev's Inequality

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

$$S_k = \left\{ i, 1 \le i \le n : \left| x_i - \bar{x} \right| < ks \right\}$$

And let **N(S$_k$) be the number of elements in the set S$_k$**. Then, for any **k ≥ 1**,

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

Chebyshev's Inequality

Proof:

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

By definition of sample SD.

$$\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i \in S_k}(x_i - \bar{x})^2 + \sum_{i \notin S_k}(x_i - \bar{x})^2 \\
&\ge \sum_{i \notin S_k}(x_i - \bar{x})^2 \\
&\ge \sum_{i \notin S_k} k^2 s^2 \\
&= k^2 s^2 (n - N(S_k))
\end{aligned}$$

By definition of S_k

$$\begin{aligned}
S_k &= \left\{\ i, 1 \le i \le n : \left|x_i - \bar{x}\right| < ks \right\} \\
S_k{}^C &= \left\{\ i, 1 \le i \le n : \left|x_i - \bar{x}\right| \ge ks \right\} \\
&= \left\{\ i, 1 \le i \le n : (x_i - \bar{x})^2 \ge k^2 s^2 \right\}
\end{aligned}$$

Chebyshev's Inequality

Proof (Cont'd):

$$(n-1)s^2 \geq k^2 s^2 (n - N(S_k))$$

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

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

$$= 1 - \frac{1 - {}^{1}\!/_{n}}{k^2}$$

$$> 1 - \frac{1}{k^2}$$

n > 0, sample size

Q.E.D

Chebyshev's Inequality

$$S_k = \{\, i, 1 \le i \le n : |x_i - \bar{x}| < ks \}$$
$$= \{\, i, 1 \le i \le n : \bar{x} - ks < x_i < \bar{x} + ks \}$$

proportions

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

It means, greater than (**at least**) **100(1 - 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** !

Chebyshev's Inequality

10 top-selling passenger cars in the U.S in 2008.

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

$\bar{x}$ **= 31,879.4**

s **= 7,514.7**

We obtain from Chebyshev's Inequality that greater than 100(5/9) = **55.55%** of the data from any data set lies between $\bar{x} - 1.5s$ to $\bar{x} + 1.5s$ (**k = 3/2**).→ Compare to 90% in actual data!

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