UMP OPEN COURSEWARE

Universiti Malaysia PAHANG

Engineering • Technology • Creativity

CHAPTER 4

Fundamentals of Statistics

Expected Outcomes

Know the difference between a variable and an attribute.
Perform mathematical calculations to the correct number of significant figures.
Construct histograms for simple and complex data.
Calculate and effectively use the different measures of central tendency, dispersion, and how related

Introduction

Definition of Statistics:

1. A collection of quantitative data pertaining to to a subject or group. Examples are blood pressure statistics etc.
2. The science that deals with the collection, tabulation, analysis, interpretation, and presentation of quantitative data

Collection of Data

Types of Data:

Attribute:

Discrete data. Data values can only be integers. Counted data or attribute data. Examples include:

- How many of the products are defective?
- How often are the machines repaired?
- How many people are absent each day?

Precision

Precision

description of a level of measurement that yields consistent results when repeated. It is associated with the concept of "random error", a form of observational error that leads to measurable values being inconsistent when repeated.

Accuracy

Accuracy

- The more common definition is that accuracy is a level of measurement with no inherent limitation
- The ISO definition is that accuracy is a level of measurement that yields true (no systematic errors) and consistent (no random errors) results.

Describing Data

Frequency Distribution:

- Three types--Categorical, Ungrouped, & Grouped
- Categorical frequency distributions
- Data that can be placed in specific categories, such as nominal- or ordinal-level data.

Categorical

EXAMPLE: Frequency distribution of injury type at a workplace

Injury Type	Frequency	Percent
Fall	14	30
Cut	8	17
Burn	3	6
Back injury	2	4
Other trauma	11	23
Injury not specified	9	19
TOTAL	47	100

Ungrouped

Ungrouped frequency distributions

- Ungrouped frequency distributions - can be used for data that can be enumerated and when the range of values in the data set is not large.

Ungrouped Population Mean	Ungrouped Sample Mean
10	10
17	17
12	12
10	10
14	14
9	9
8	8
3	3
14	14
16	16
$\sum X =$ 113	$\sum x =$ 113

Population	Sample
$\mu = \frac{\sum X}{N}$	$\bar{x} = \frac{\sum x}{n}$
$\mu = \frac{113}{10}$	$\bar{x} = \frac{113}{10}$
$\mu = 11.3$	$\bar{x} = 11.3$

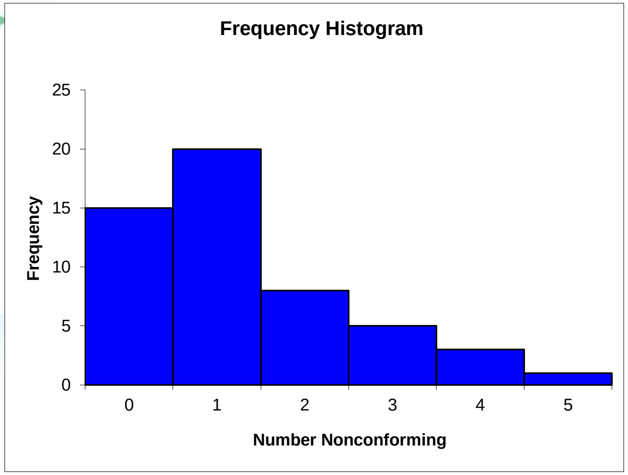
Grouped

- Grouped frequency distributions
- Can be used when the range of values in the data set is very large. The data must be grouped into classes that are more than one unit in width.

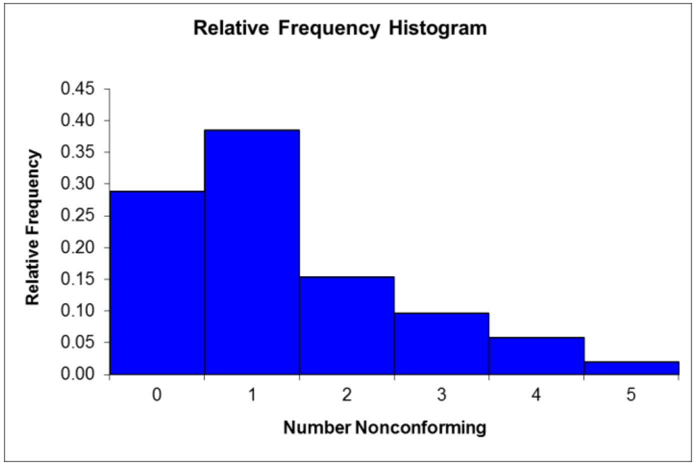
Height (in cm)	No of students
159-162	1
163-166	4
167-170	11
171-174	12
175-178	6
179-182	4
183-186	2

Frequency Distributions

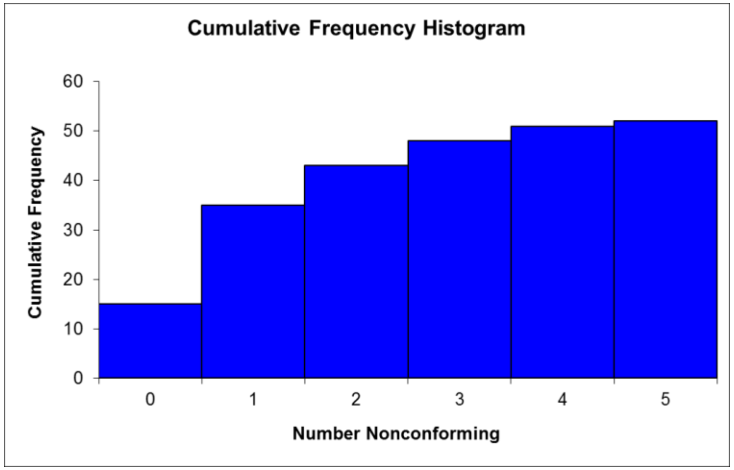
Number non conforming	Frequency	Relative Frequency	Cumulative Frequency	Relative Frequency
0	15	0.29	15	0.29
1	20	0.38	35	0.67
2	8	0.15	43	0.83
3	5	0.10	48	0.92
4	3	0.06	51	0.98
5	1	0.02	52	1.00



Frequency Histogram
25
20
15
10
5
0
Frequency
0
1
2
3
4
5
Number Nonconforming



Relative Frequency Histogram
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Relative Frequency
0
1
2
3
4
5
Number Nonconforming



Cumulative Frequency Histogram
60
50
40
30
20
10
0
Cumulative Frequency
0
1
2
3
4
5
Number Nonconforming

The Histogram

The histogram is the most important graphical tool for exploring the shape of data distributions.

Constructing a Histogram

Step 1: Find range of distribution, largest - smallest values

Step 2: Choose number of classes, 5 to 20

Step 3: Determine width of classes, one decimal place more than the data, class width = range/number of classes $\#classes = \sqrt{n}$

Step 4: Determine class boundaries

Step 5: Draw frequency histogram

Constructing a Histogram

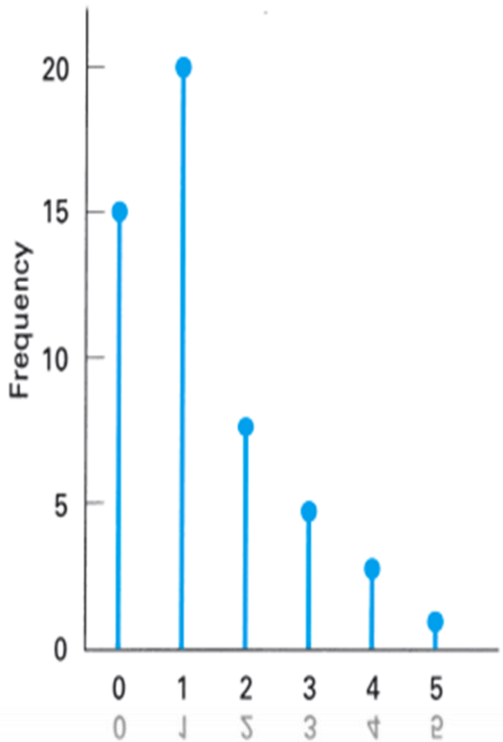
Number of groups or cells

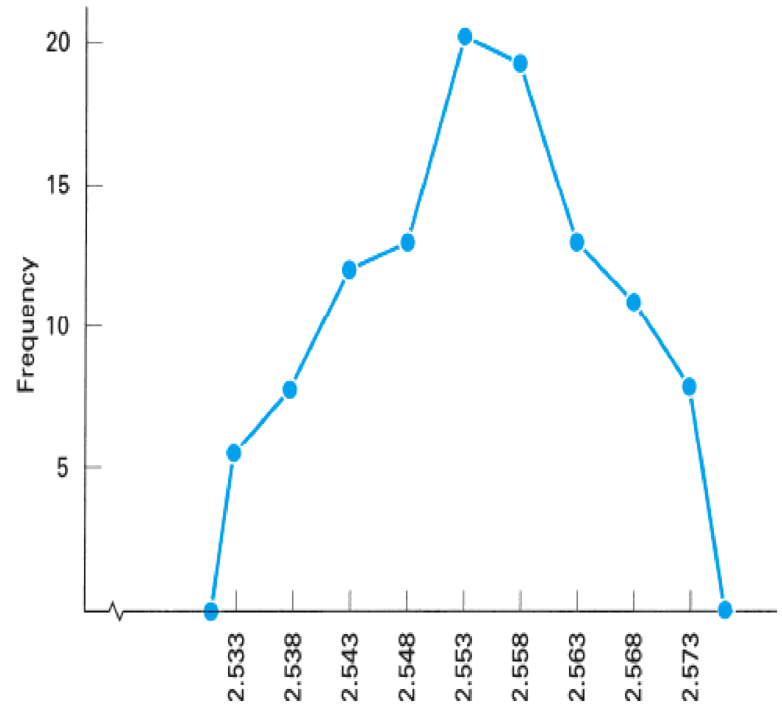
- If no. of observations < 100 – 5 to 9 cells
- Between 100-500 – 8 to 17 cells
- Greater than 500 – 15 to 20 cells

Other Types of Frequency Distribution Graphs

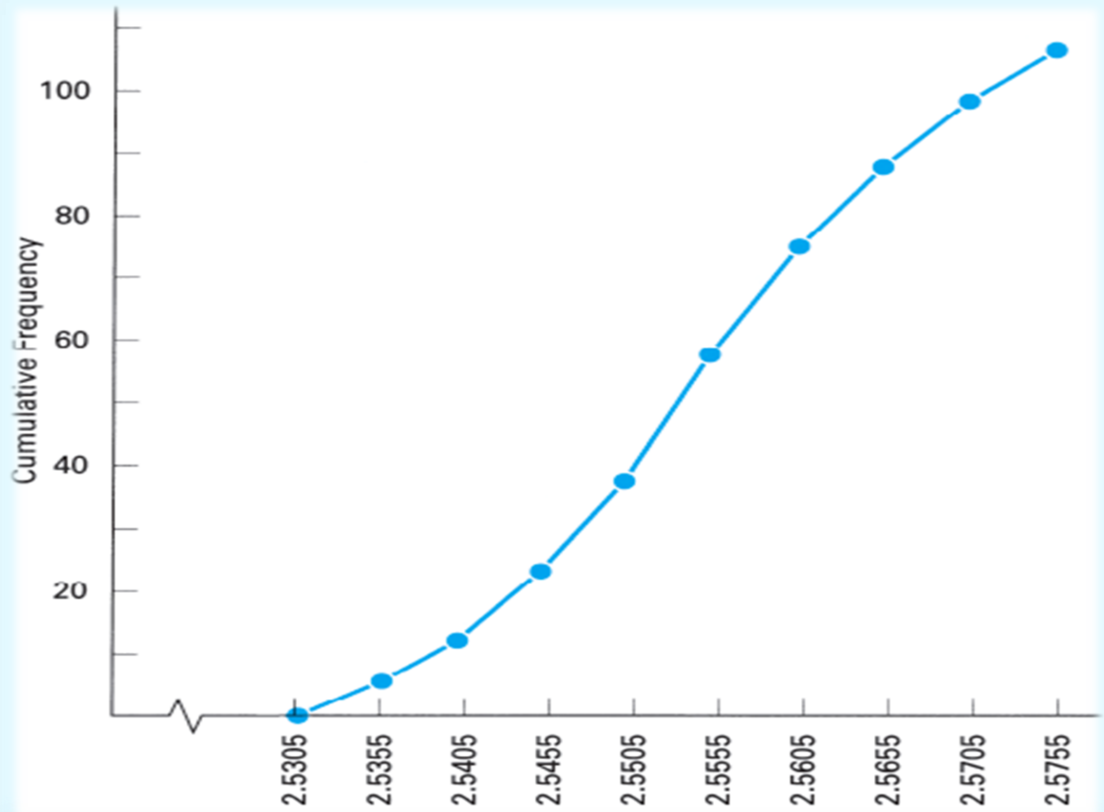
- ❑ Bar Graph
- ❑ Polygon of Data
- ❑ Cumulative Frequency Distribution or Ogive

Bar Graph and Polygon of Data

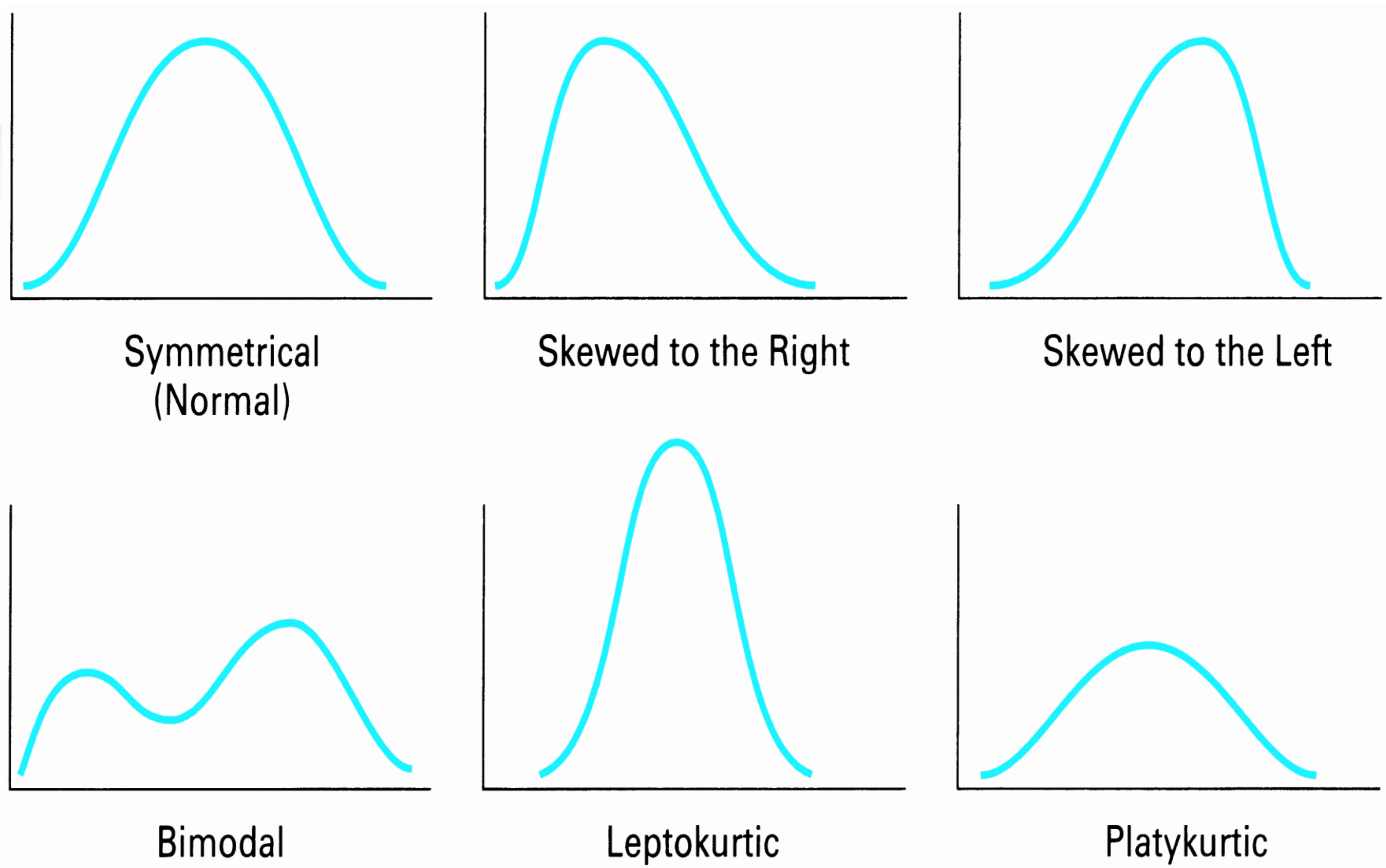




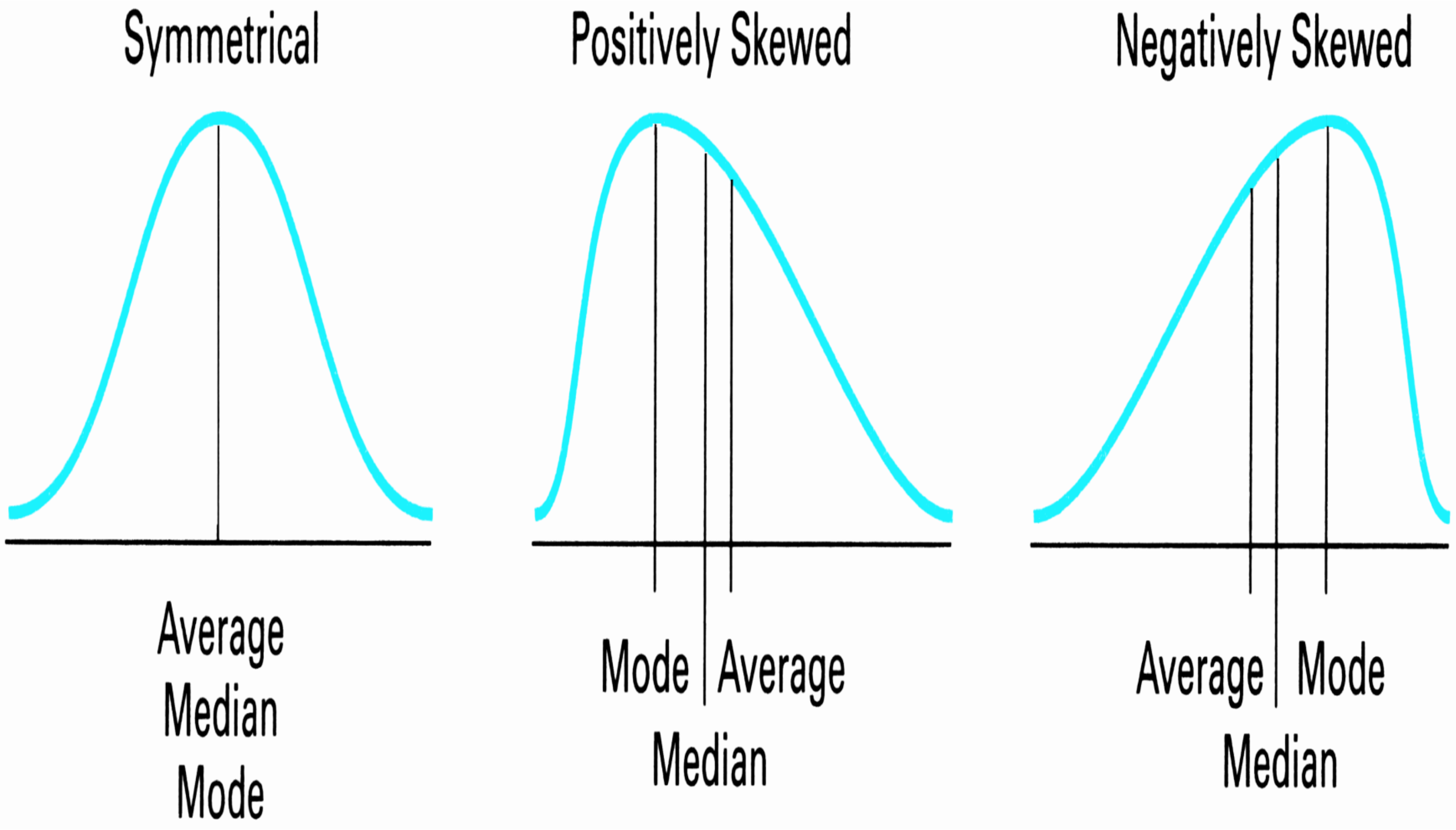
Cumulative Frequency



Characteristics of Frequency Distribution Graphs



Analysis of Histograms



Measures of Central Tendency

The three measures in common use are the:

- Average
- Median
- Mode

Average

There are three different techniques available for calculating the average three measures in common use are the:

- Ungrouped data
- Grouped data
- Weighted average

Average-Ungrouped Data

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

Average-Grouped Data

$$\overline{X} = \sum_{i=1}^{h} \frac{f_i X_i}{n}$$

$$= \frac{f_1 X_1 + f_2 X_2 \ldots + f_h X_h.}{f_1 + f_2 \ldots + f_h}$$

h = number of cells **fi=frequency**
Xi=midpoint

Average-Weighted Average

Used when a number of averages are combined with different frequencies

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i \overline{X}_i}{\sum_{i=1}^{n} w_i}$$

Median-Grouped Data

$$M_d = L_m + \left[\frac{\frac{n}{2} - cf_m}{f_m} \right] i$$

Lm=lower boundary of the cell with the median

N=total number of observations

Cfm=cumulative frequency of all cells below m

Fm=frequency of median cell

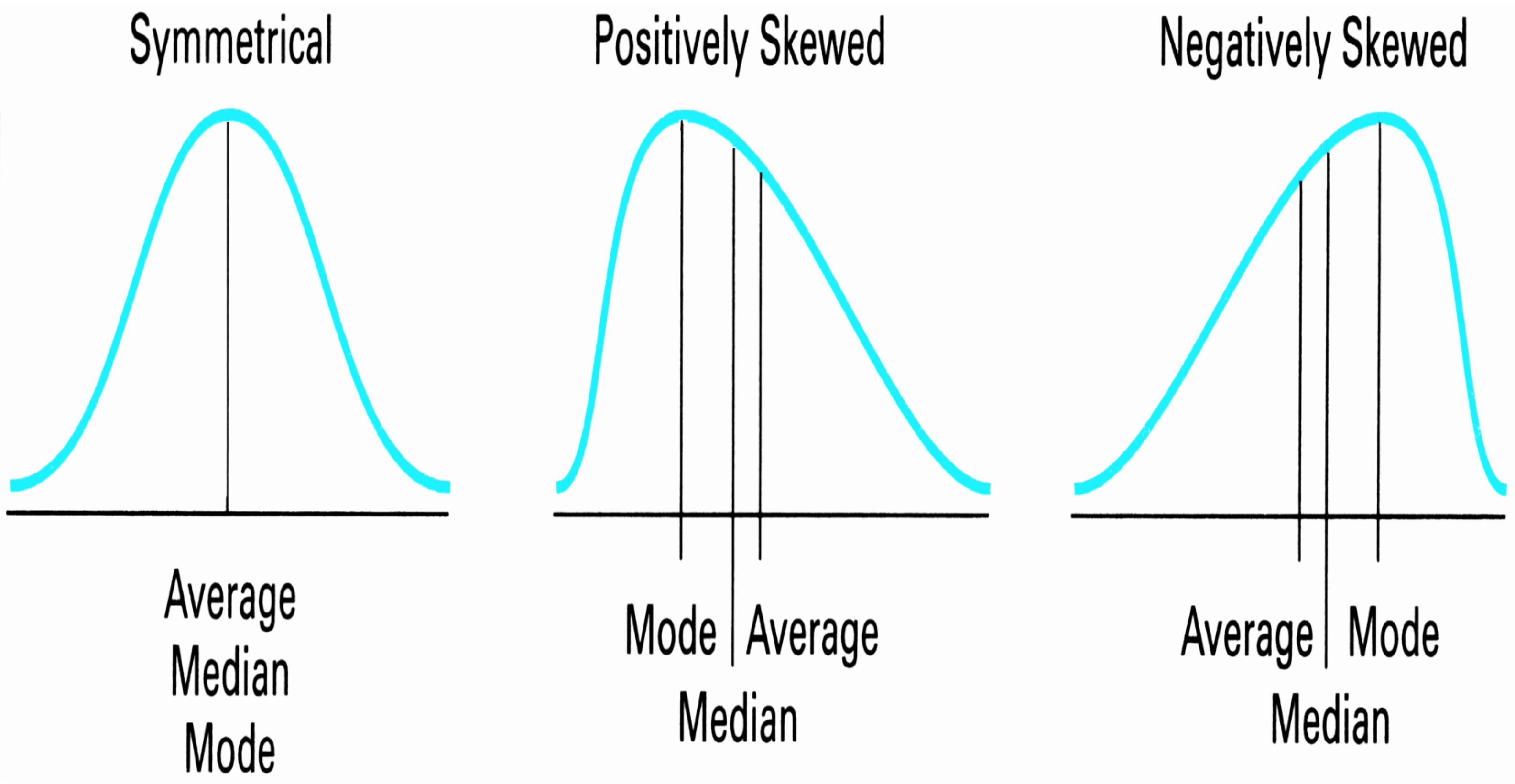
i=cell interval

Mode

The Mode is the value that occurs with the greatest frequency.

It is possible to have no modes in a series or numbers or to have more than one mode.

Relationship Among the Measures of Central Tendency



Measures of Dispersion

- Range
- Standard Deviation
- Variance

Measures of Dispersion-Range

The range is the simplest and easiest to calculate of the measures of dispersion.

Range = R = X_h - X_l

- Largest value - Smallest value in data set

Measures of Dispersion-Standard Deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n} (Xi - \overline{X})^2}{n-1}}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} Xi^2 - \left(\sum_{i=1}^{n} Xi\right)^2 / n}{n-1}}$$

Standard Deviation

$$S = \sqrt{\frac{n\sum_{i=1}^{n} Xi^2 - (\sum_{i=1}^{n} Xi)^2}{n(n-1)}}$$

Standard Deviation

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

Relationship Between the Measures of Dispersion

- As n increases, accuracy of R decreases
- Use R when there is small amount of data or data is too scattered
- If n> 10 use standard deviation
- A smaller standard deviation means better quality

Other Measures

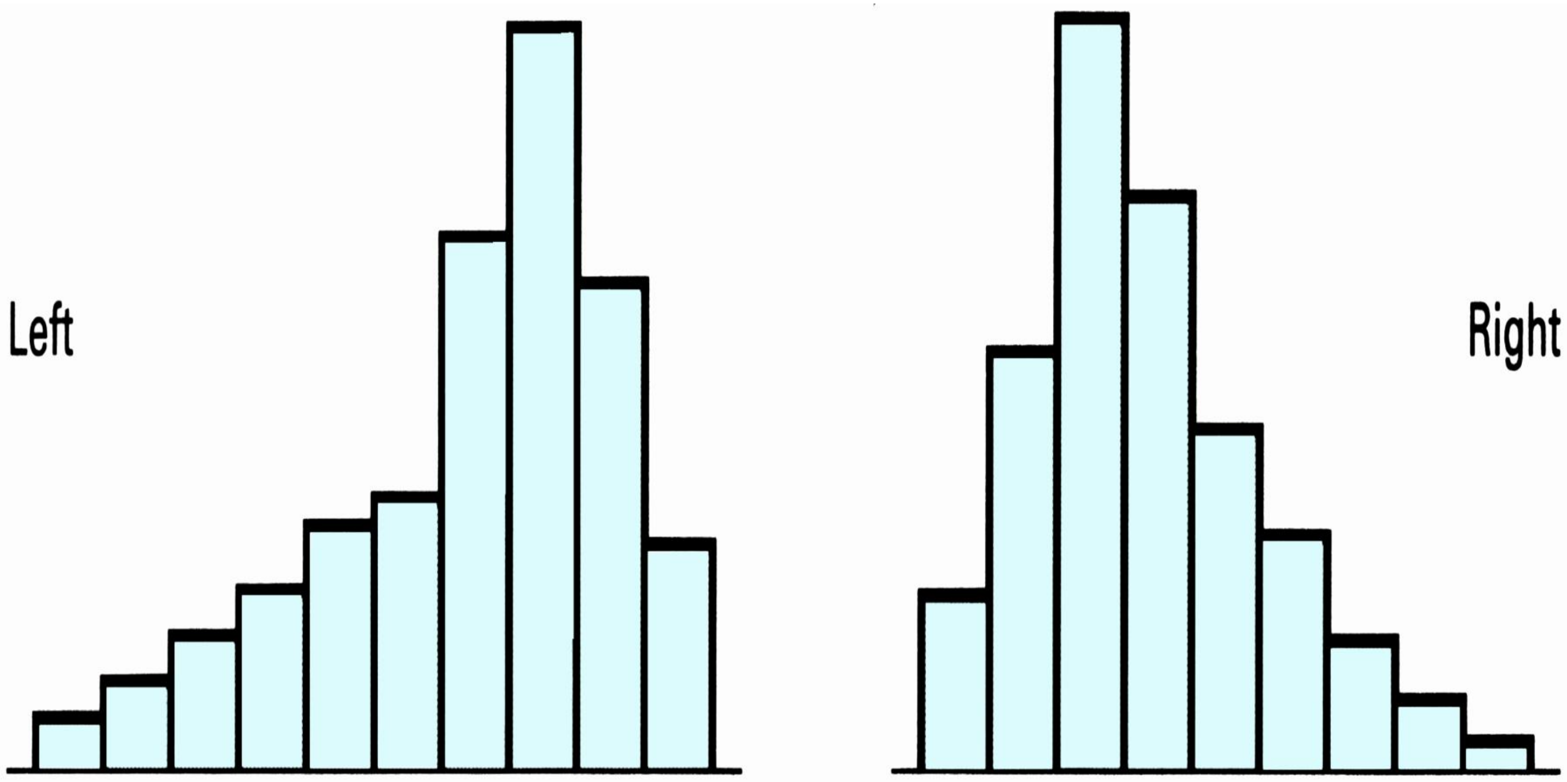
There are three other measures that are frequently used to analyze a collection of data:

- Skewness
- Kurtosis
- Coefficient of Variation

Skewness

$$a_3 = \frac{\sum_{i=1}^{h} f_i (X_i - \overline{X})^3 / n}{s^3}$$

Skewness



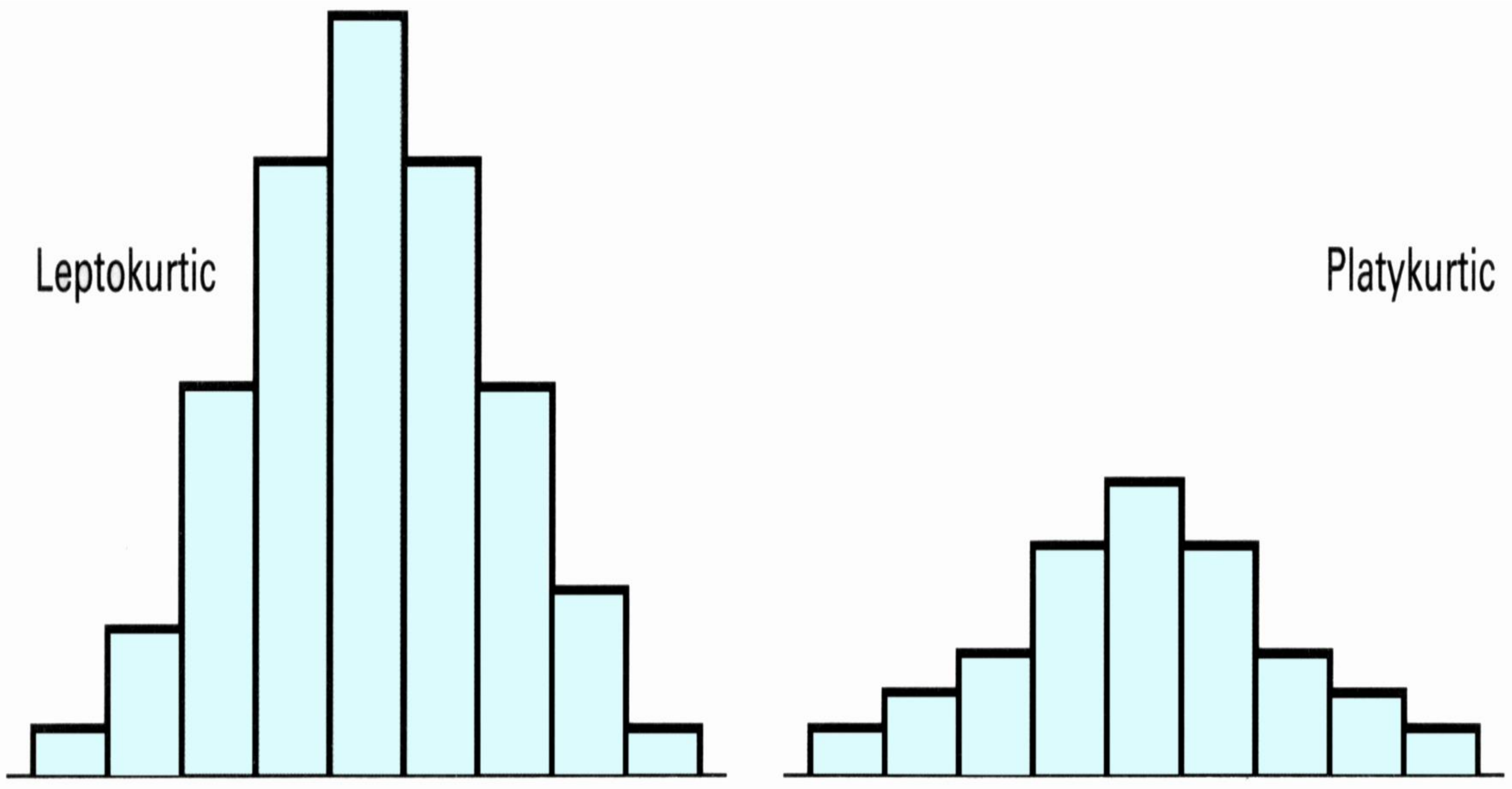
Kurtosis

Kurtosis provides information regrading the shape of the population distribution (the peakedness or heaviness of the tails of a distribution).

For grouped data:

$$a_4 = \frac{\sum_{i=1}^{h} f_i (X_i - \overline{X})^4 / n}{s^4}$$

Kurtosis



The Normal Curve

Characteristics of the normal curve:

- It is symmetrical -- Half the cases are to one side of the center; the other half is on the other side.
- The distribution is single peaked, not bimodal or multi-modal
- Also known as the Gaussian distribution
- Mean is best measure of central tendency

The Normal Curve

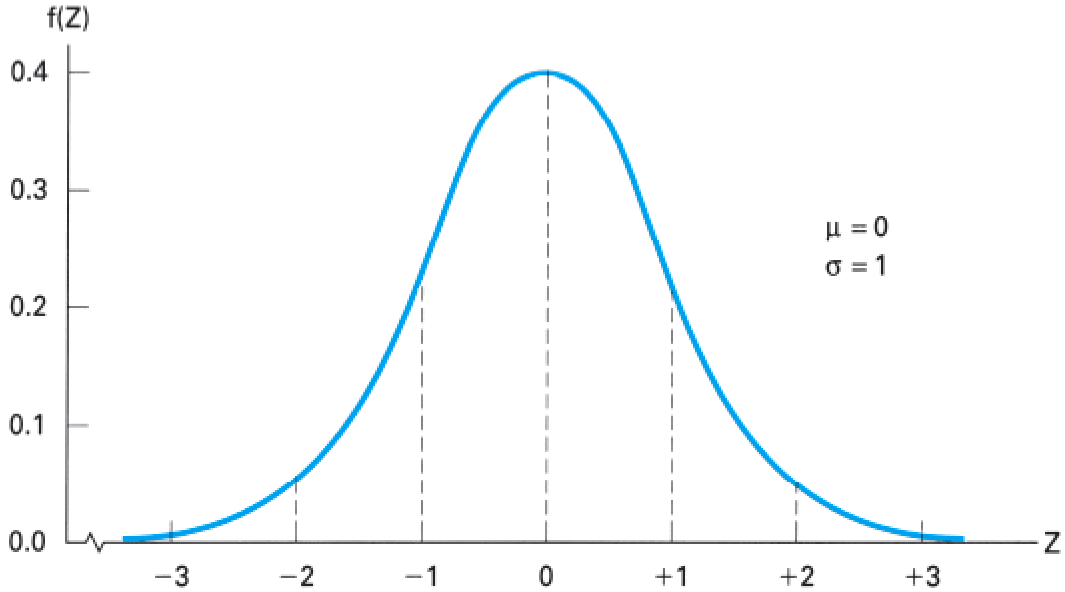
Characteristics:

Most of the cases will fall in the center portion of the curve and as values of the variable become more extreme they become less frequent, with "outliers" at the "tail" of the distribution few in number. It is one of many frequency distributions.

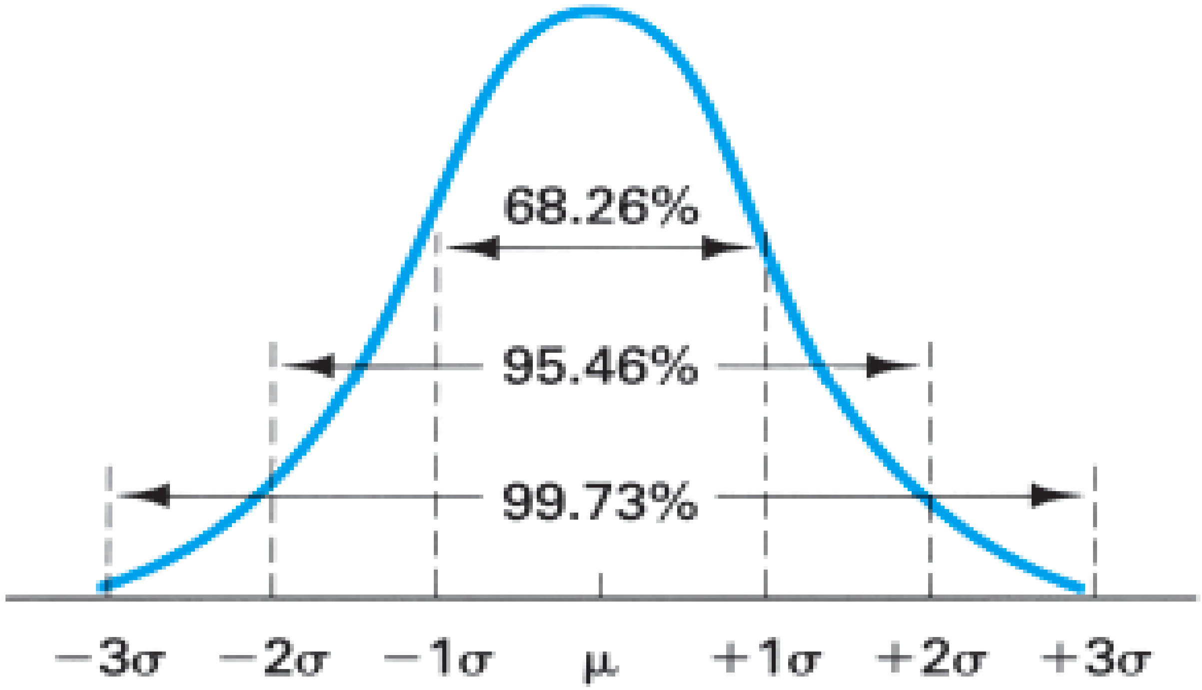
Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. Normal distributions can be transformed to standard normal distributions by the formula:

$$Z = \frac{X_{i-\mu}}{\sigma}$$

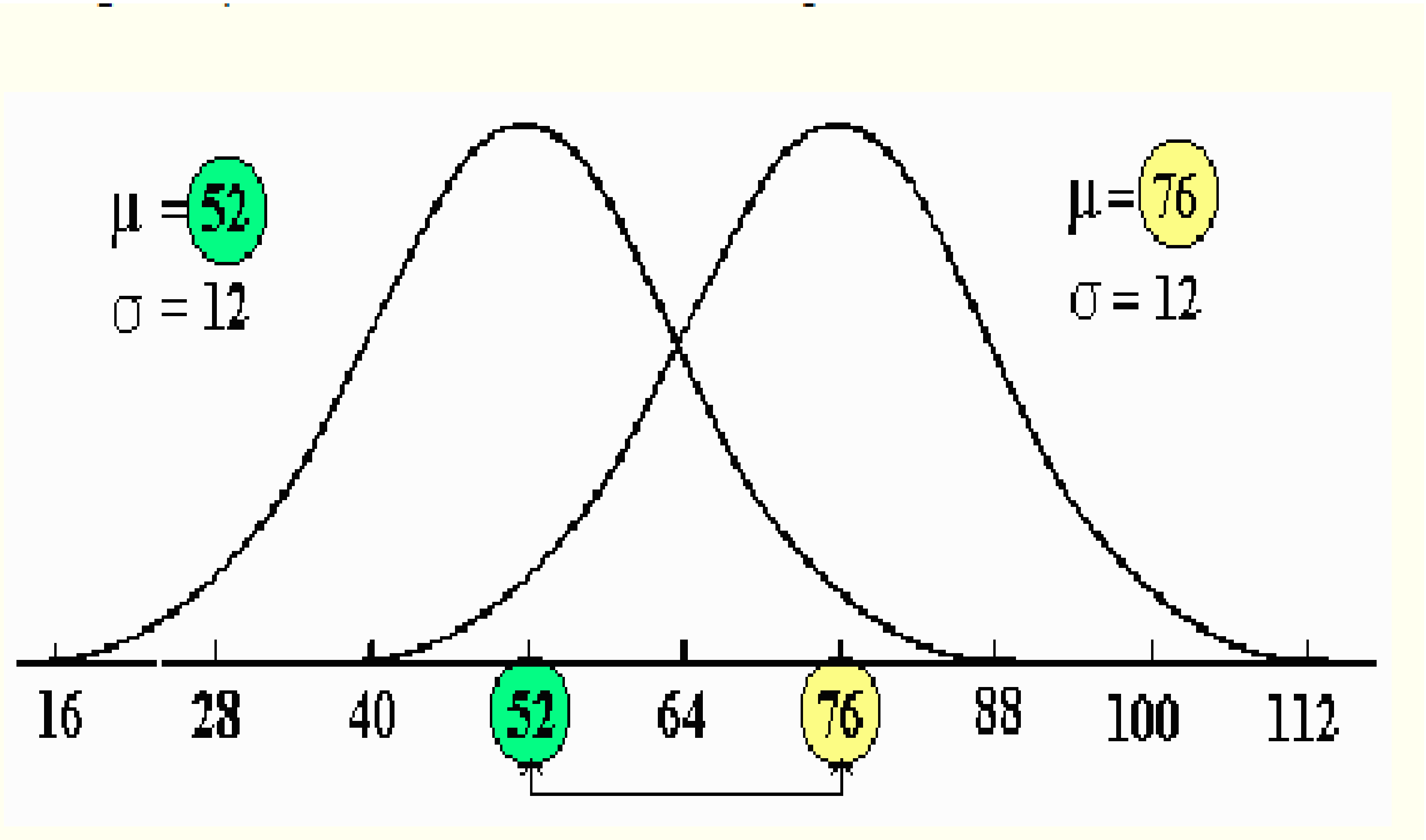


Standardized Normal Distribution with μ = 0 and σ = 1



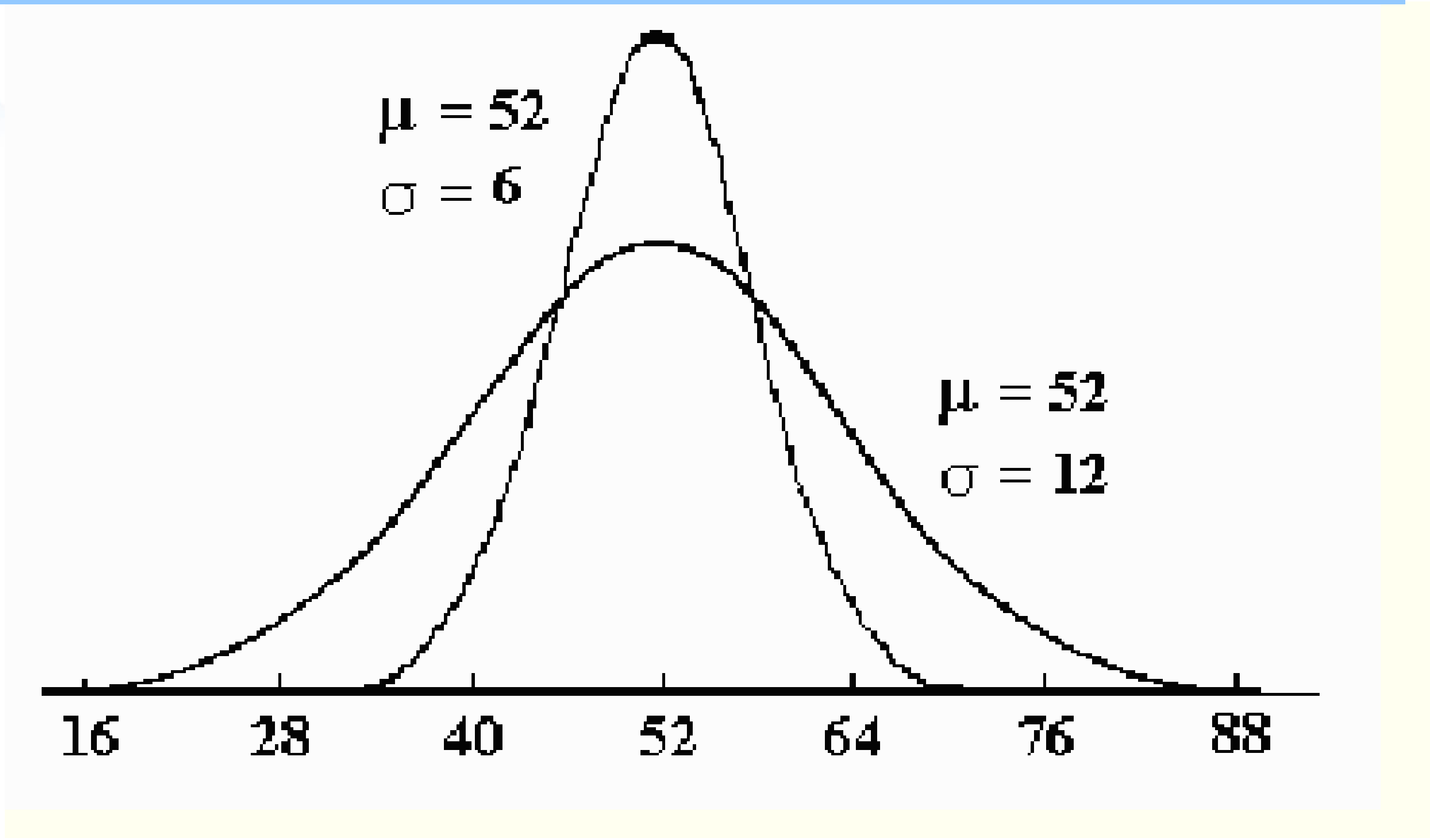
Percent of Items Included between certain values of the standard deviation

Relationship between the Mean and Standard Deviation



Mean and Standard Deviation

Same mean but different standard deviation



Tests for Normality

- Histogram
- Skewness
- Kurtosis

Tests for Normality

Histogram:

Shape

- Symmetrical

The larger the sampler size, the better the judgment of normality. A minimum sample size of 50 is recommended

Tests for Normality

Skewness (a3) and Kurtosis (a4)”

- Skewed to the left or to the right (a3=0 for a normal distribution)
- The data are peaked as the normal distribution (a4=3 for a normal distribution)
- The larger the sample size, the better the judgment of normality (sample size of 100 is recommended)

Tests for Normality

Probability Plots

- Order the data from the smallest to the largest
- Rank the observations (starting from 1 for the lowest observation)
- Calculate the plotting position

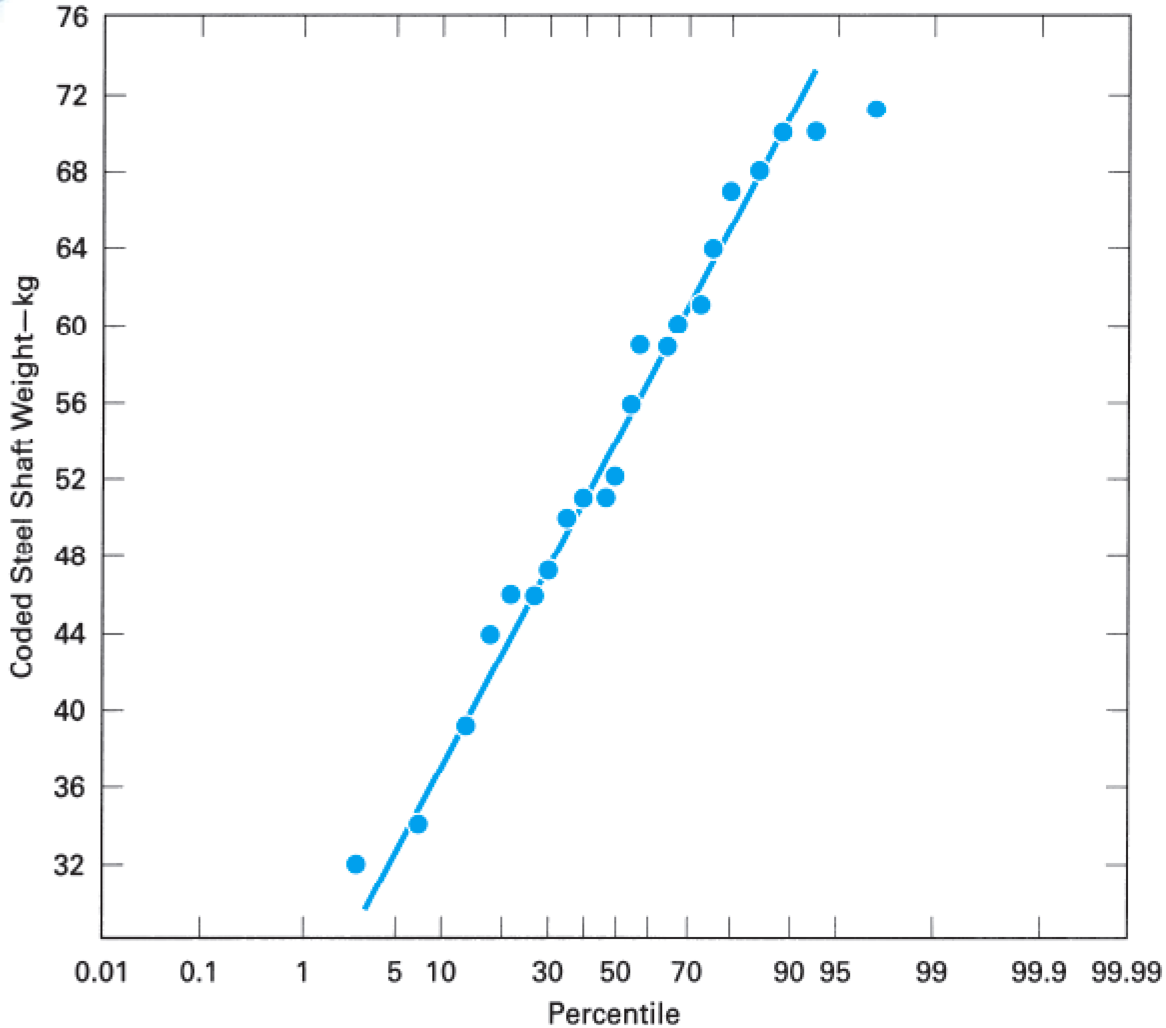
$$PP = \frac{100(i - 0.5)}{n}$$

Where i = rank PP=plotting position n=sample size

Probability Plots

Procedure cont'd:

- Order the data
- Rank the observations
- Calculate the plotting position
- Label the data scale
- Plot the points
- Attempt to fit by eye a "best line"
- Determine normality



Probability Plots