Introduction

Basic terminology

- Population vs. Sample
- Parameter vs. Statistic
- Descriptive vs. Inferential Statistics
- Types of variables

What is Statistics

- Statistics is the art of learning from data.

  ~Sheldon M. Ross

- Statistics is the science of learning from data.

  ~Moore, McCabe & Craig

- Statistics is the art and science of learning from data.

  ~Alan Agresti

- Statistics is the science whereby inferences are made about specific random phenomena on the basis of relatively limited sample materials.

  ~ Bernard Rosner

The **Population**

- The collection of **all** subjects of interest

The **Sample**

- A subset of the population

Parameters

- Numerical measures computed using population data

Statistics

- Numerical measures computed using sample data

Descriptive statistics

Collecting, summarizing, and presenting data

Inferential statistics

Drawing conclusions about a population based only on sample data

Some questions of interest:

- What is the average height of this class?
- What proportion of NUS students are female?
- What is the total sugar consumption (in a year) of all Singaporeans?

Descriptive Statistics

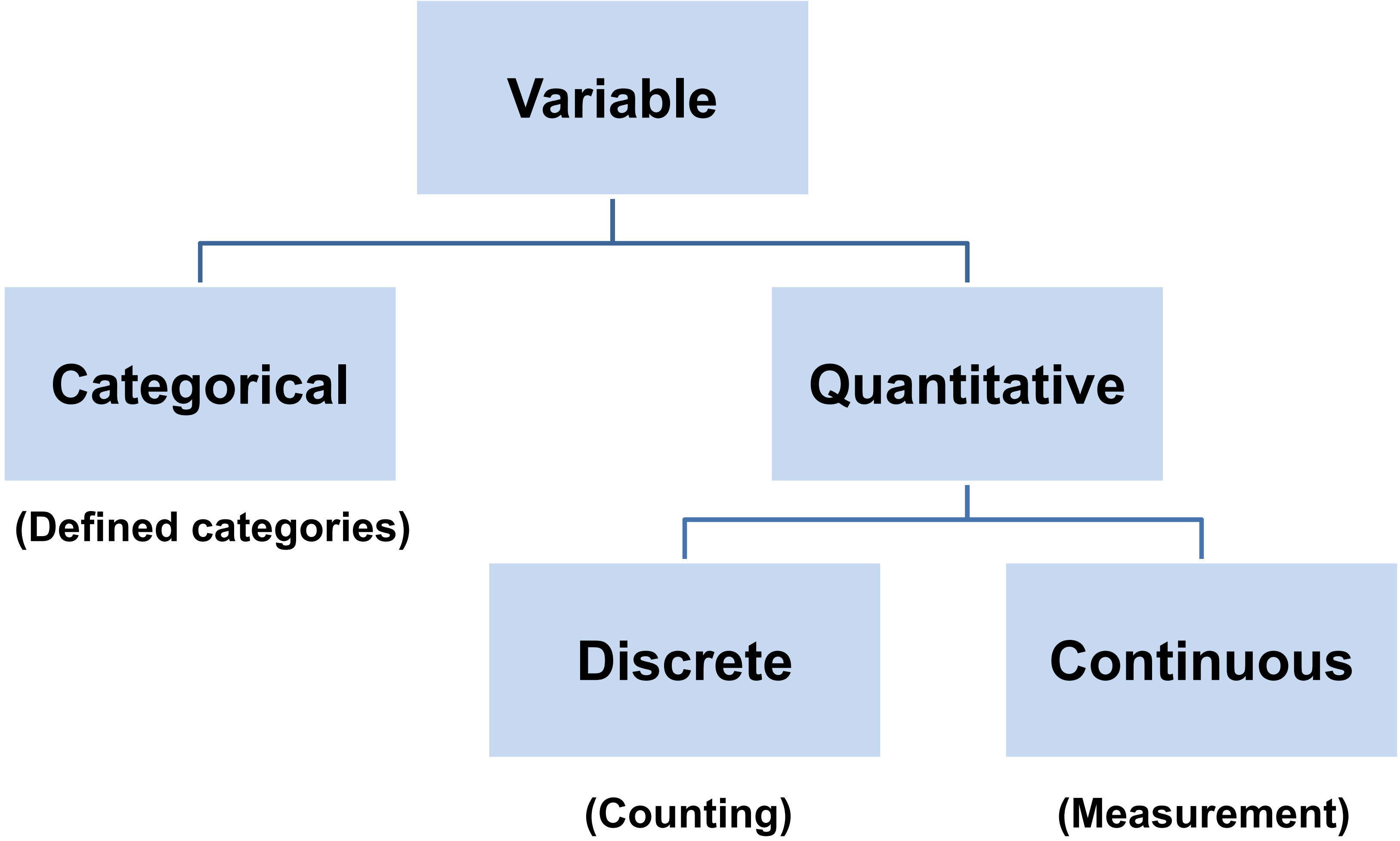
- Organize / graph
- Numerical summaries

Inferential Statistics

- Estimate / predict
- Decide /conclude

A variable is any characteristics that is recorded for subjects in the study.

The terminology variable highlights the fact that data vary.



Variable

Categorical

(Defined categories)

Quantitative

Discrete

(Counting)

Continuous

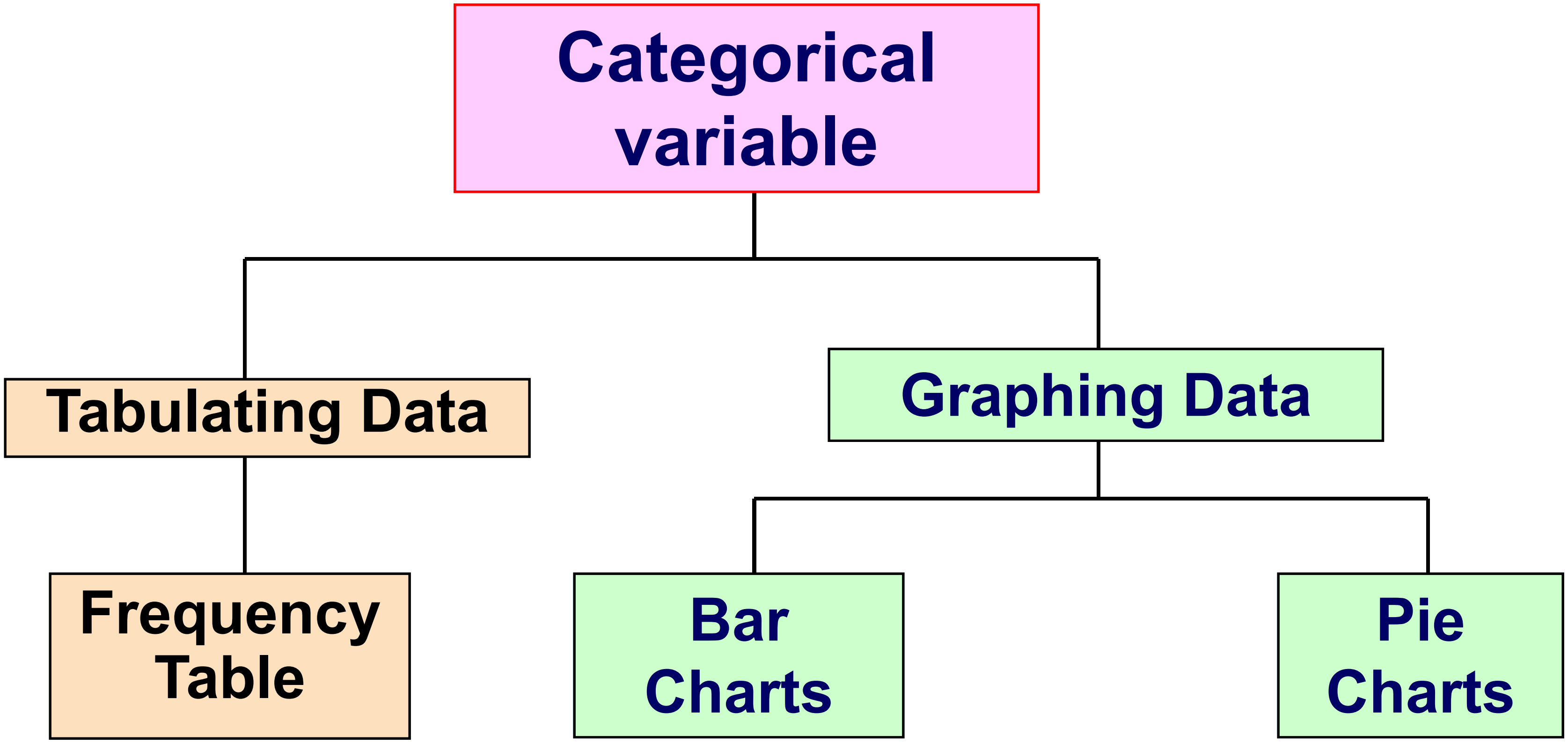
(Measurement)

Descriptive Statistics

- Graphical presentation
- Numerical Summary

Categorical

Quantitative



Categorical variable

Tabulating Data

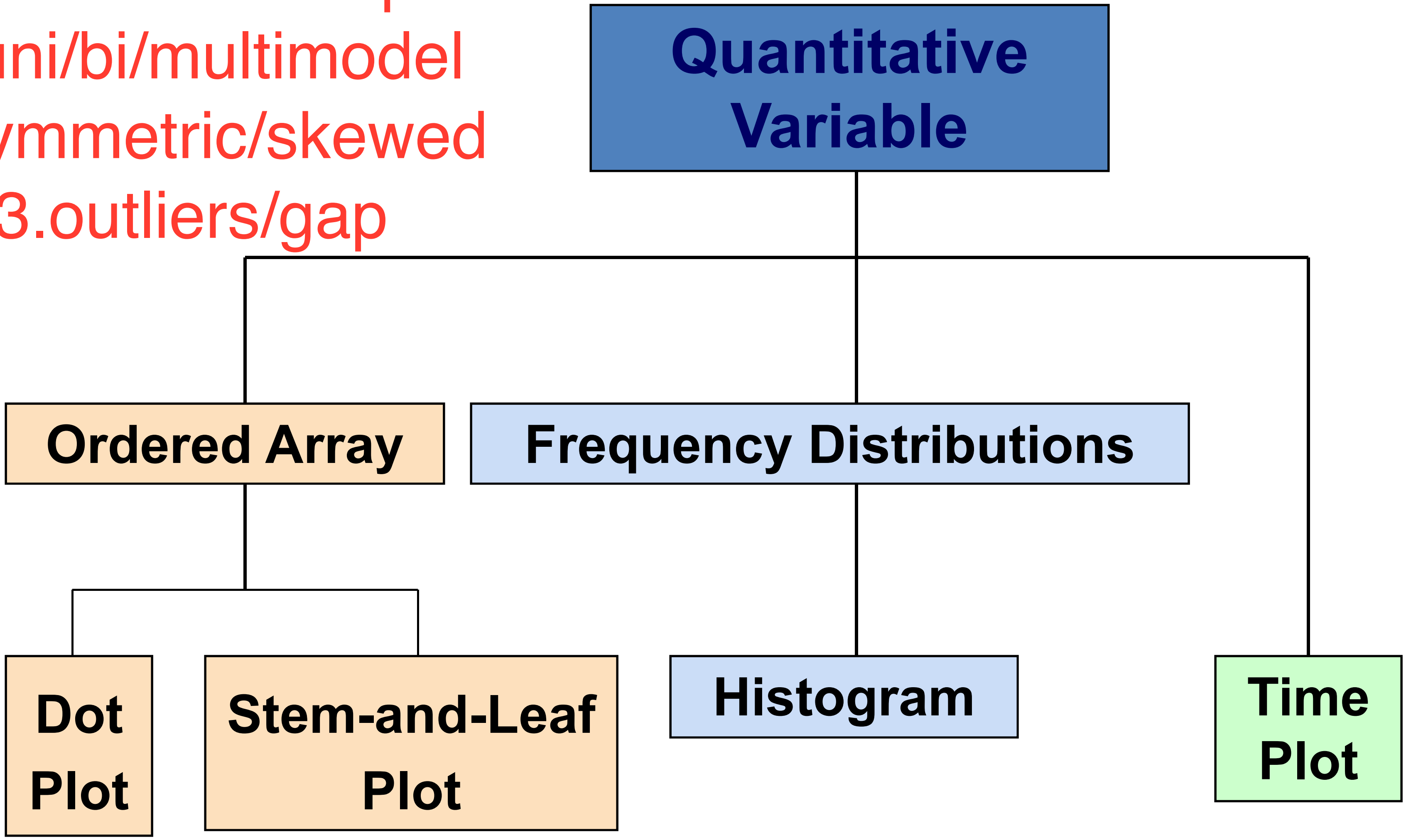
Graphing Data

Frequency Table

Bar Charts

Pie Charts

Distribution shape
1.uni/bi/multimodel
2.symmetric/skewed
3.outliers/gap



Quantitative Variable

Ordered Array

Frequency Distributions

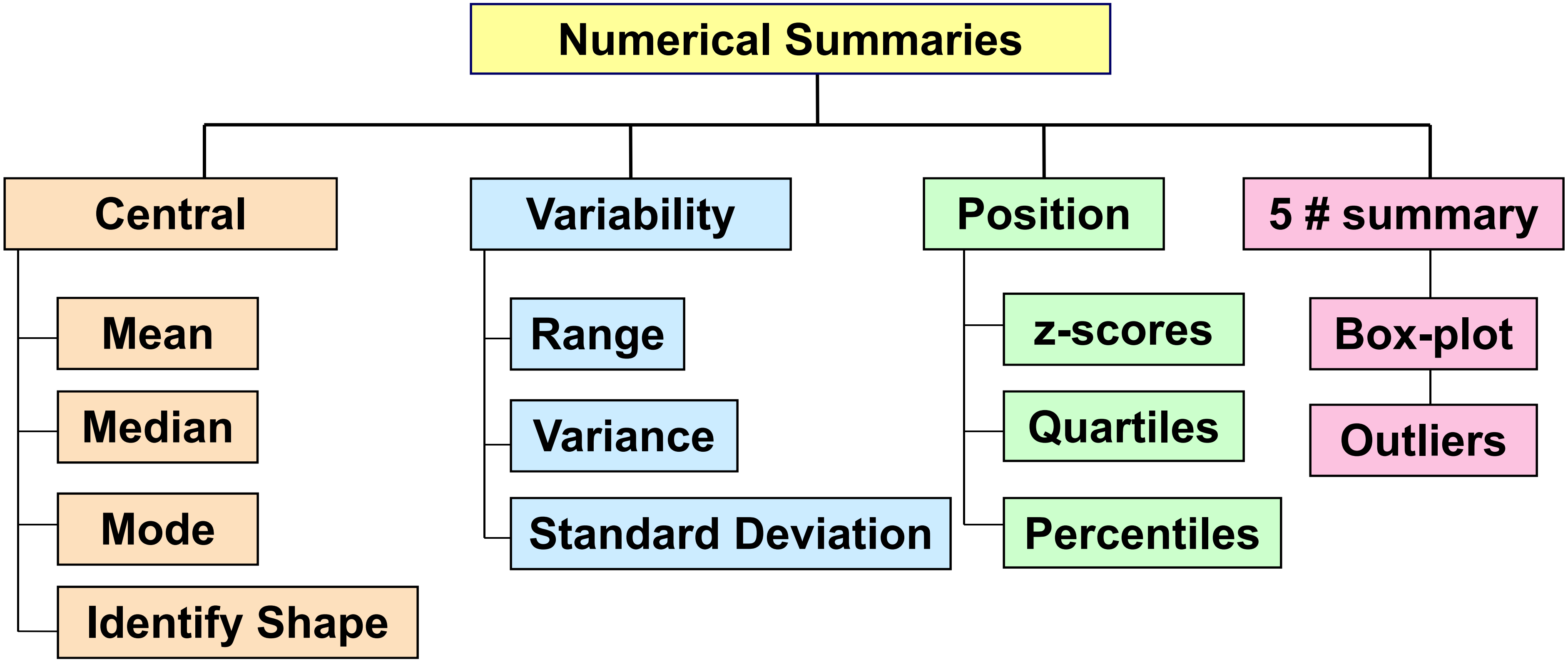
Dot Plot

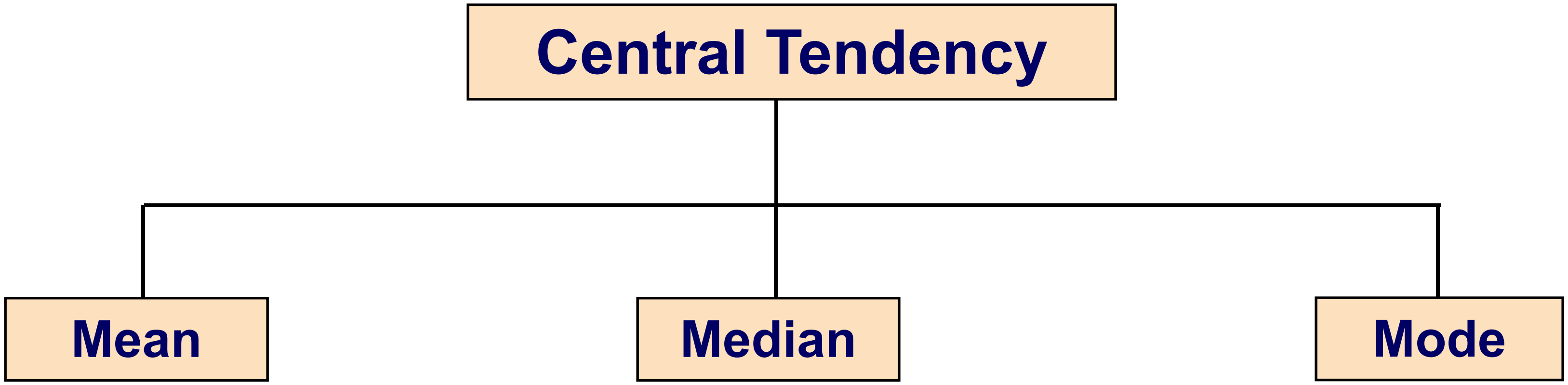
Stem-and-Leaf Plot

Histogram

Time Plot

Quantitative



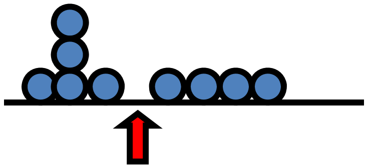


Central Tendency

Mean

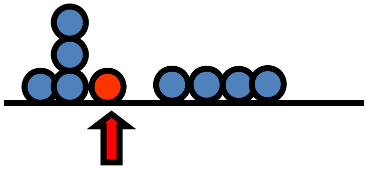
Median

Mode

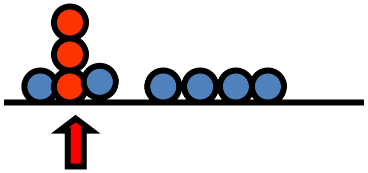


The center of gravity or the balance point

affected by outliars but utilize all info

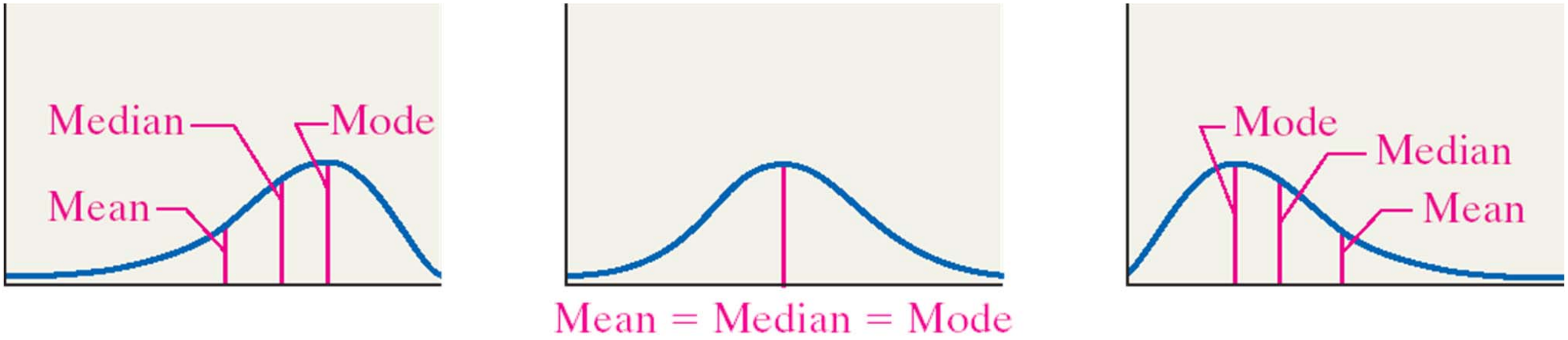


Midpoint of ranked values



Most frequently observed value

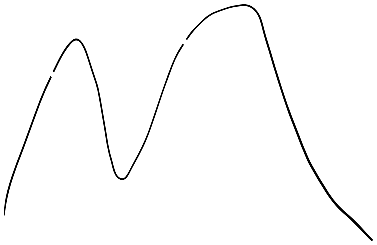
Distribution Shape



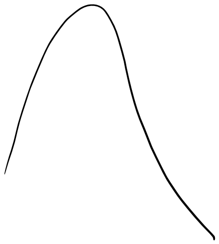
(a) Skewed Left
Mean < Median

(b) Symmetric
Mean = Median

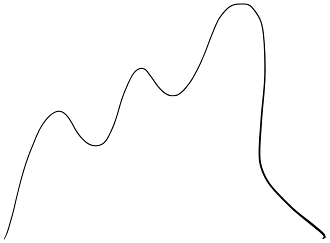
(c) Skewed Right
Mean > Median



bimodel

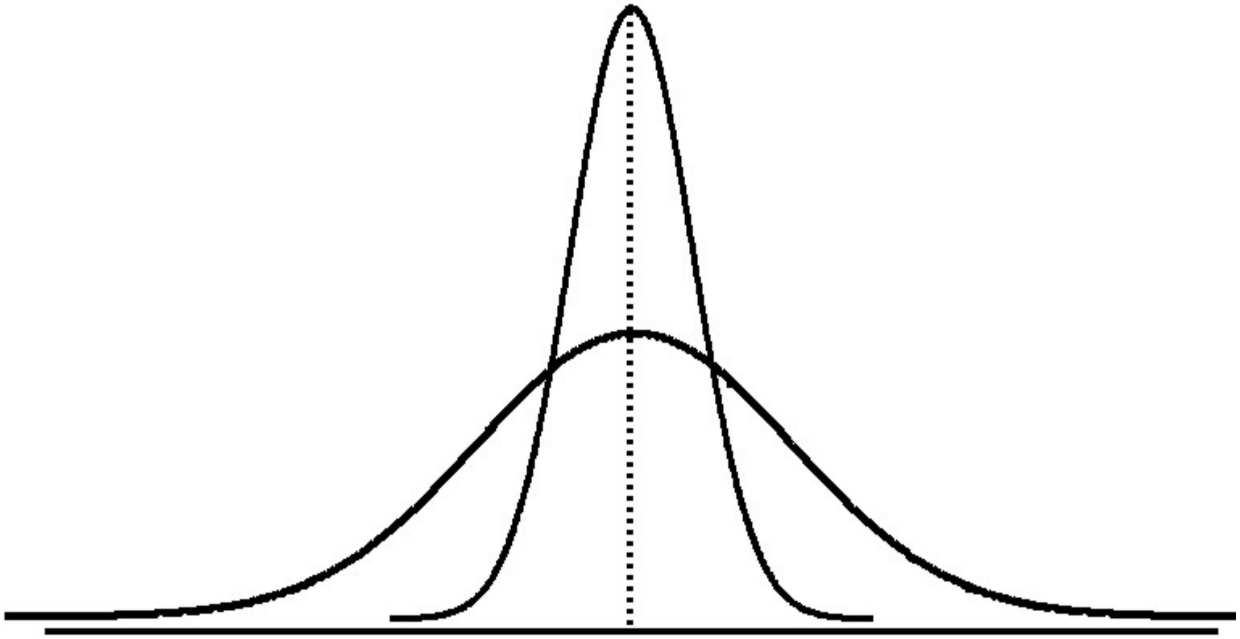


unimodel



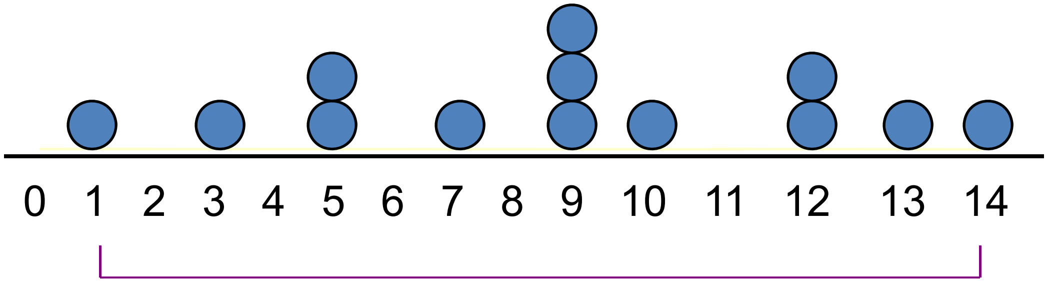
multimodel

Variability

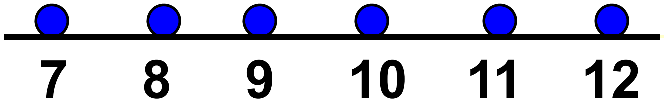


**Same center,
different variation**

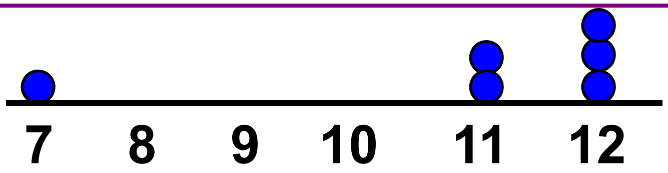
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$



Range =



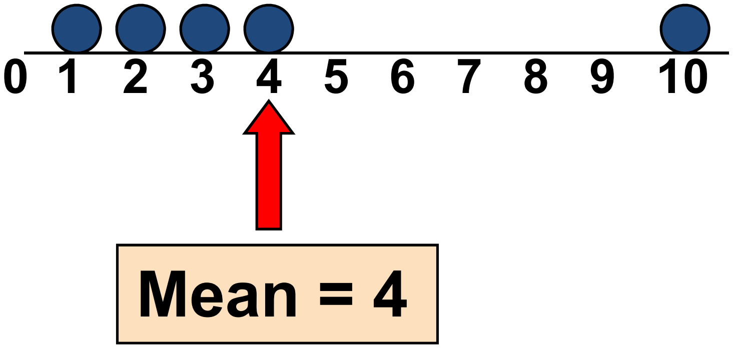
Range =



Range =

Deviation

Deviation = *X* - mean



av. dev. = 0

Variance

- The population variance, σ^2 of a variable is the sum of squared deviations divided by the number in the population

$$\boxed{\frac{\sum(x_i-\mu)^2}{N}=\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}}$$

mean=u(para),x bar (stat)

- The sample variance, s^2 of a variable is the sum of squared deviations divided by one less than the number in the sample

$$\boxed{\frac{\sum(x_i-\overline{x})^2}{n-1}=\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_n-\overline{x})^2}{n-1}}$$

proportion = p(para), p^ (stat) size= N (para), n (stat)

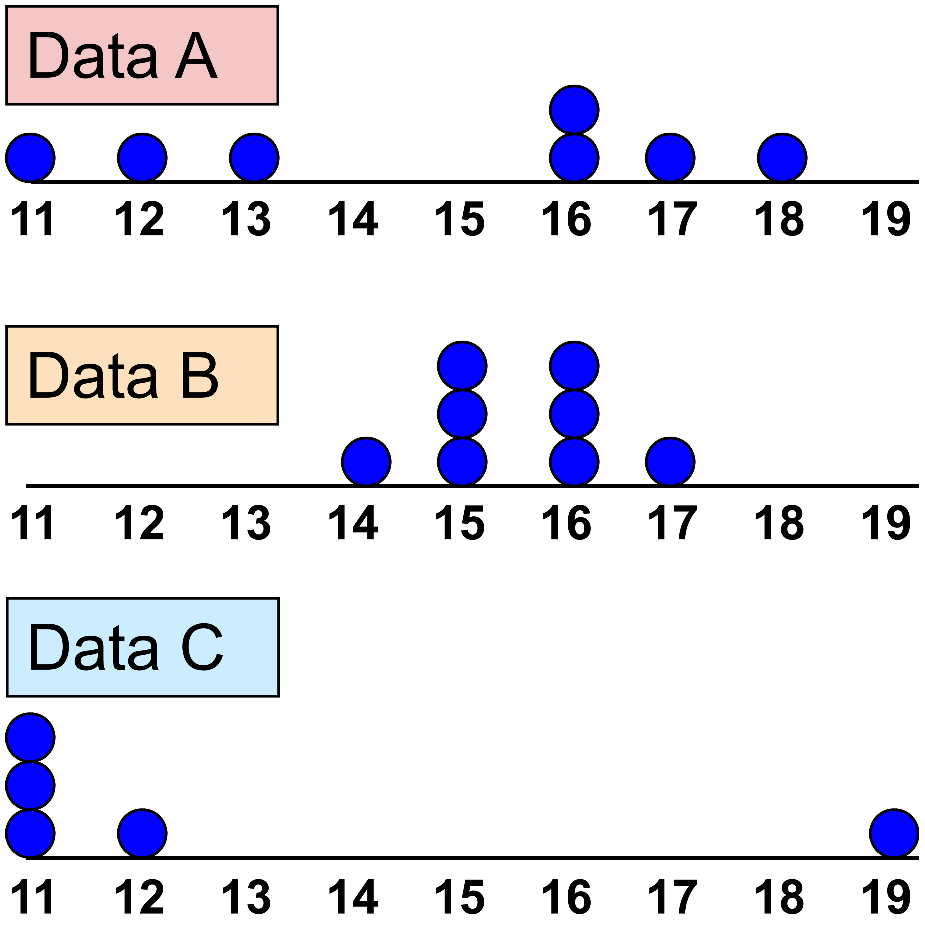
variance o^2(para), s^2 (stat)

Standard Deviation

- The standard deviation is the square root of the variance.
- σ is the population standard deviation.
- *s* is the sample standard deviation.

Standard Deviation

Comparing Standard Deviations



z-scores

- *z*-scores can be used to compare the relative positions of data values in different samples

$$z = \frac{x - \mu}{\sigma} \qquad\qquad z = \frac{x - \bar{x}}{s}$$

Pat received:

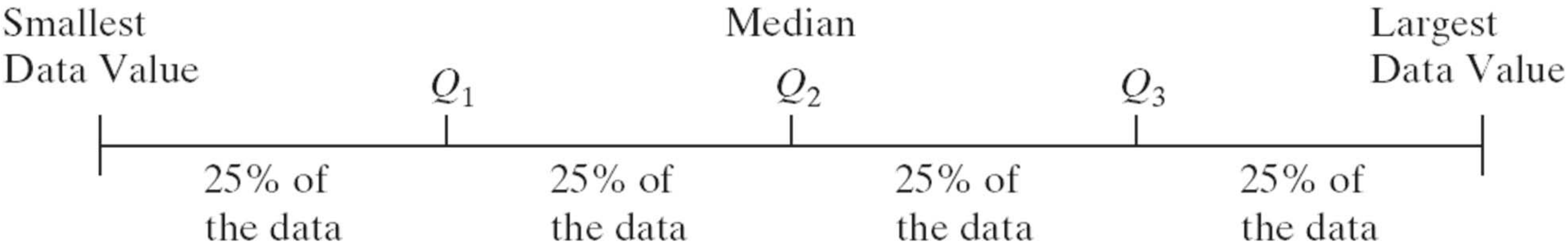
- A grade of 82 on her statistics exam where the mean grade was 74 and the standard deviation was 12
- A grade of 72 on her biology exam where the mean grade was 65 and the standard deviation was 10
- A grade of 91 on her kayaking exam where the mean grade was 88 and the standard deviation was 6

z-scores

- Statistics
  - Grade of 82
  - Z =
- Biology
  - Grade of 72
  - Z=
- Kayaking
  - Grade of 91
  - Z=
- ________________was the highest relative grade

Quartiles

- Quartiles divide the data set into four equal parts



Percentiles

- Percentiles divide the data set into 100 equal parts

Five-number Summary

- The five-number summary is the collection of
  - The smallest value
  - The first quartile (Q_1 or P_{25})
  - The median (M or Q_2 or P_{50})
  - The third quartile (Q_3 or P_{75})
  - The largest value
- These five numbers give a concise description of the distribution of a variable

median, IQR= OUTLIERS
mean, SD = NO OUTLIERS

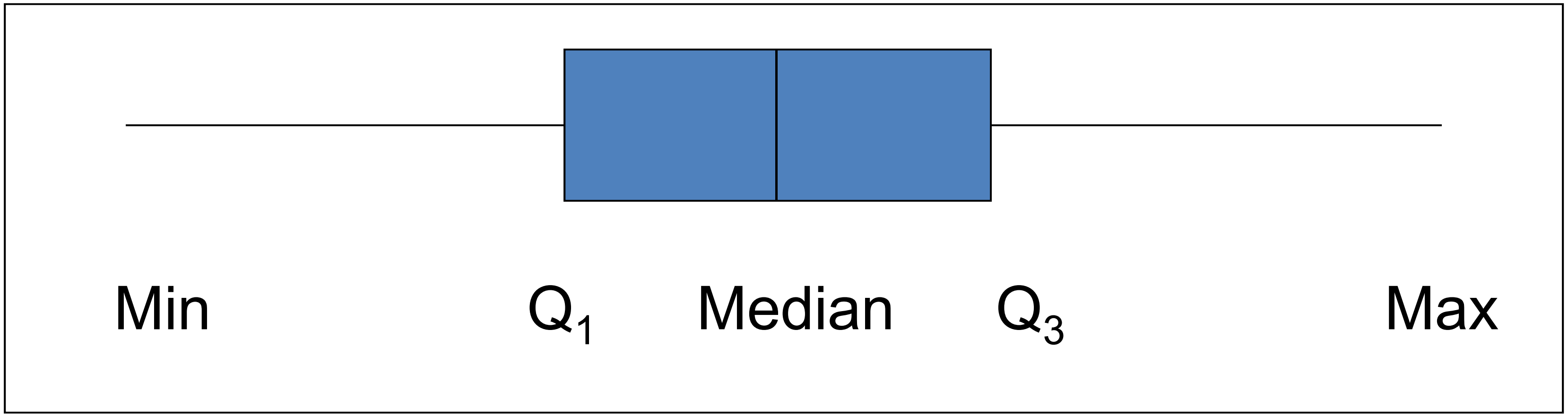
Five-number Summary

- Compute the five-number summary for

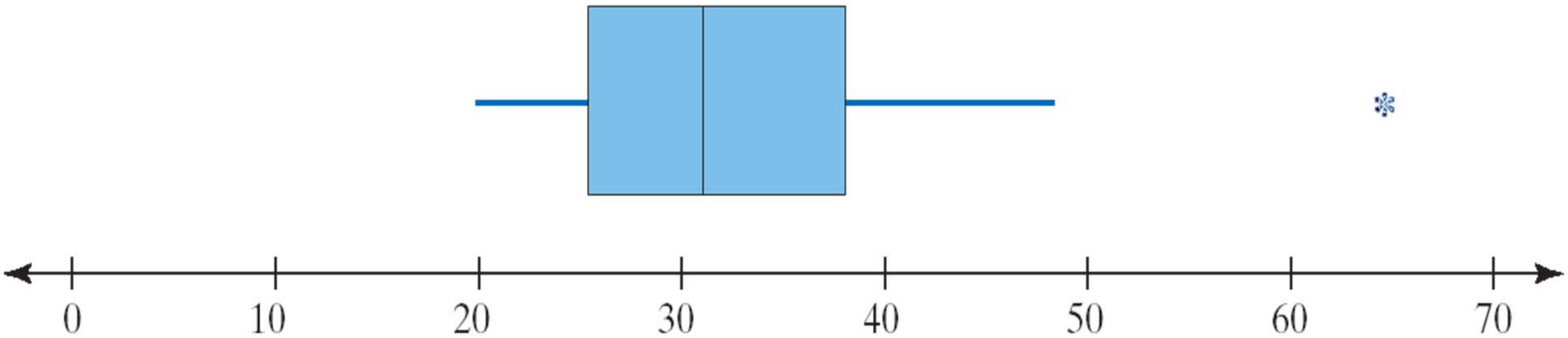
  1, 3, 4, 7, 8, 15, 16, 19, 23, 24, 27, 31, 33, 54

- Calculations
  - The minimum = 1
  - Q_1 = 7
  - M = 17.5
  - Q_3 = 27
  - The maximum = 54
- The five-number summary is

Boxplot



Min Q_1 Median Q_3 Max



Outliers

- Extreme observations in the data are referred to as outliers
- One way to check for outliers uses the inter-quartiles range, IQR = Q3- Q1
- The fences used to identify outliers are
  - **Lower fence = $LF = Q_1 - 1.5 \times IQR$**
  - **Upper fence = $UF = Q_3 + 1.5 \times IQR$**
- Values less than the lower fence or more than the upper fence could be considered outliers

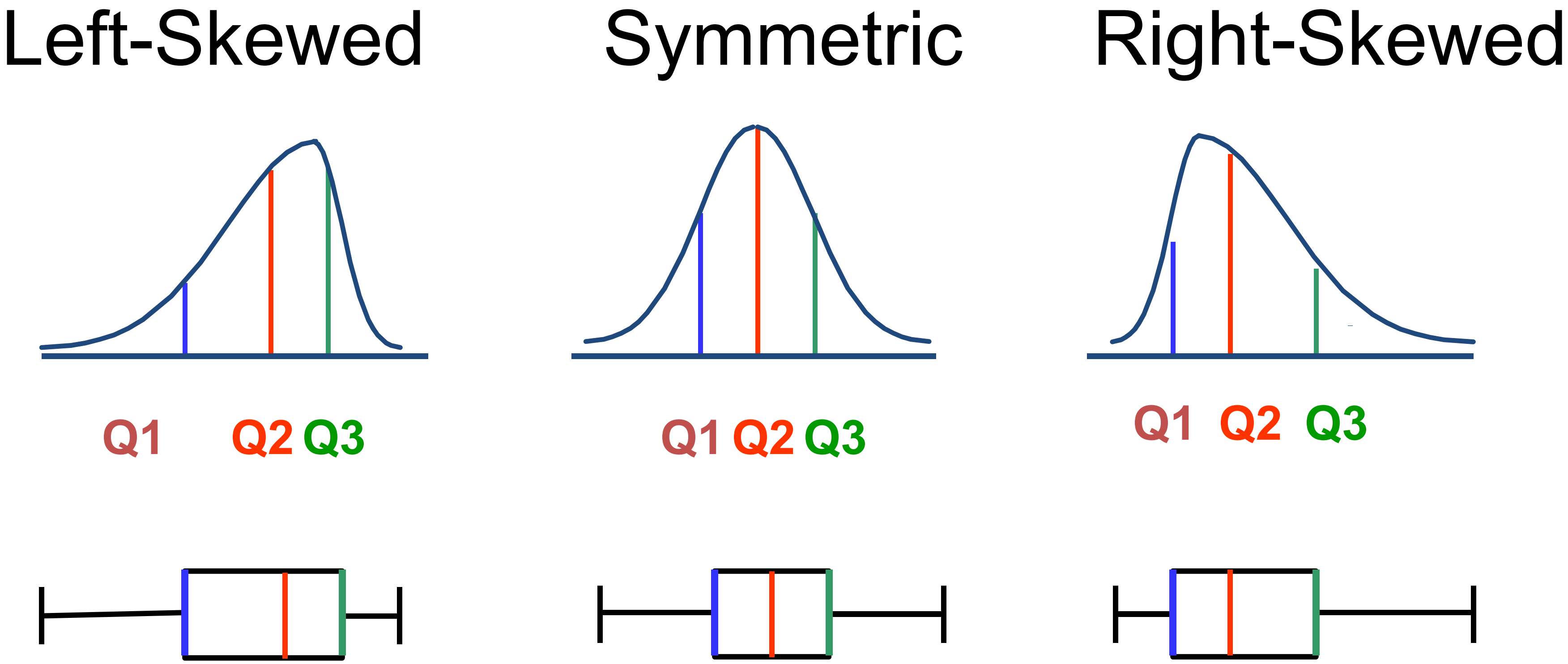
Outliers

- Is the value 54 an outlier?

  1, 3, 4, 7, 8, 15, 16, 19, 23, 24, 27, 31, 33, 54

- Calculations
  - Q_1 = 7
  - Q_3 = 27
  - *IQR* = 20
  - *UF* = 57
  - Using the fence rule, the value 54 is / is not an outlier

Boxplot



Left-Skewed Symmetric Right-Skewed

Q1 Q2 Q3 Q1 Q2 Q3 Q1 Q2 Q3