

Topic 1 Introduction to Statistics

Tutorial agenda:

1. go through key concepts
2. tutorial questions

Key concept

Week 2 L1

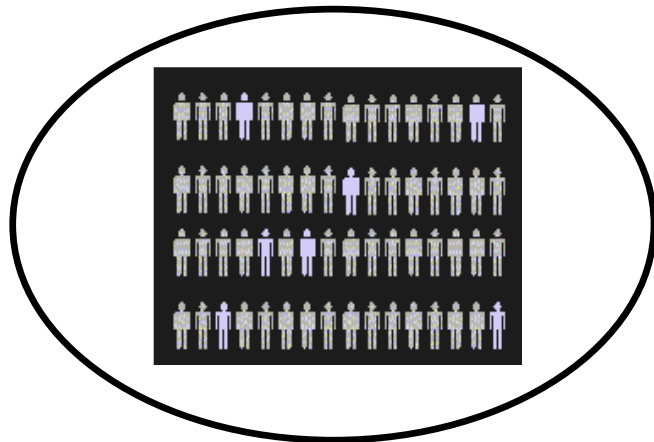
101001101001000010101
0011110111011011011010
101000011100101011001
010100111010100010101
0001011010110110110100
010101110001010100010
1000101110101100010011
010011010010000101010
0111101110110110110101
010000111001010110010
101001110101000101010
0010110101101101101001



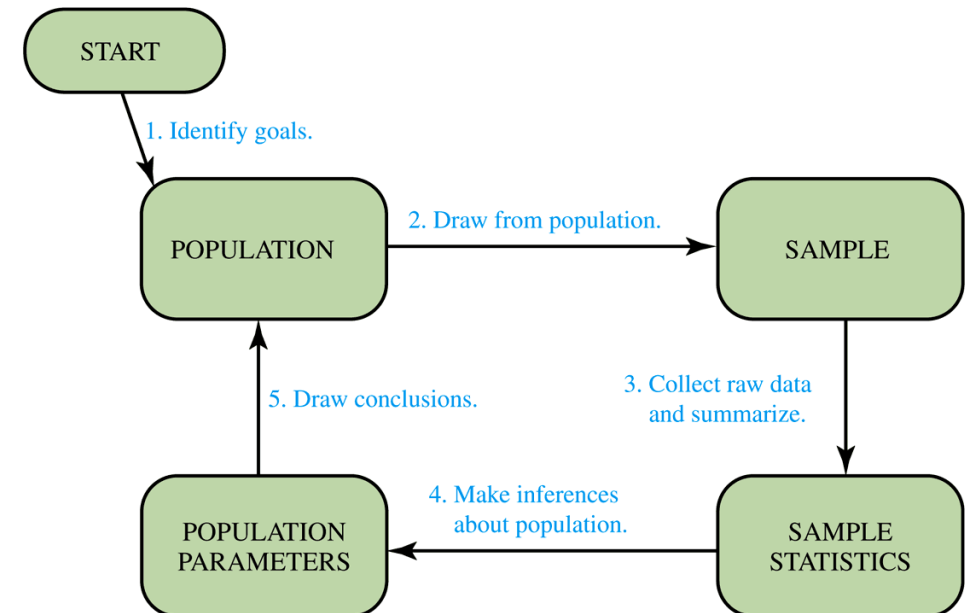
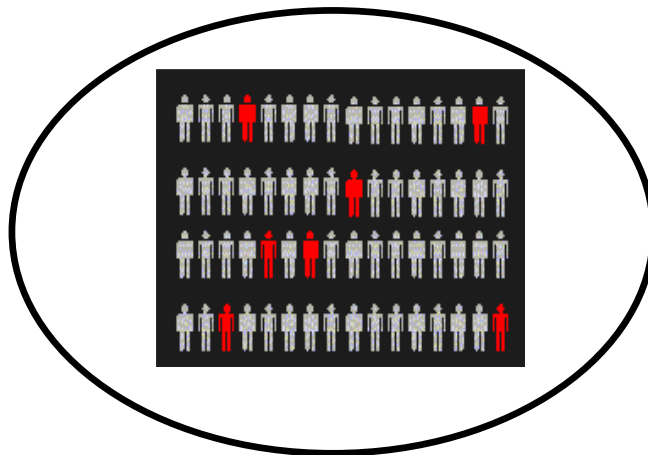
Statistical Study

- Step 1: **population**: determine the population and population parameters
- Step 2: **sampling**: use an appropriate sampling technique
- Step 3: **sample statistics**: collect raw data and summarize
- Step 4: **from sample to population**: infer population from sample statistics
- Step 5: **conclusions**

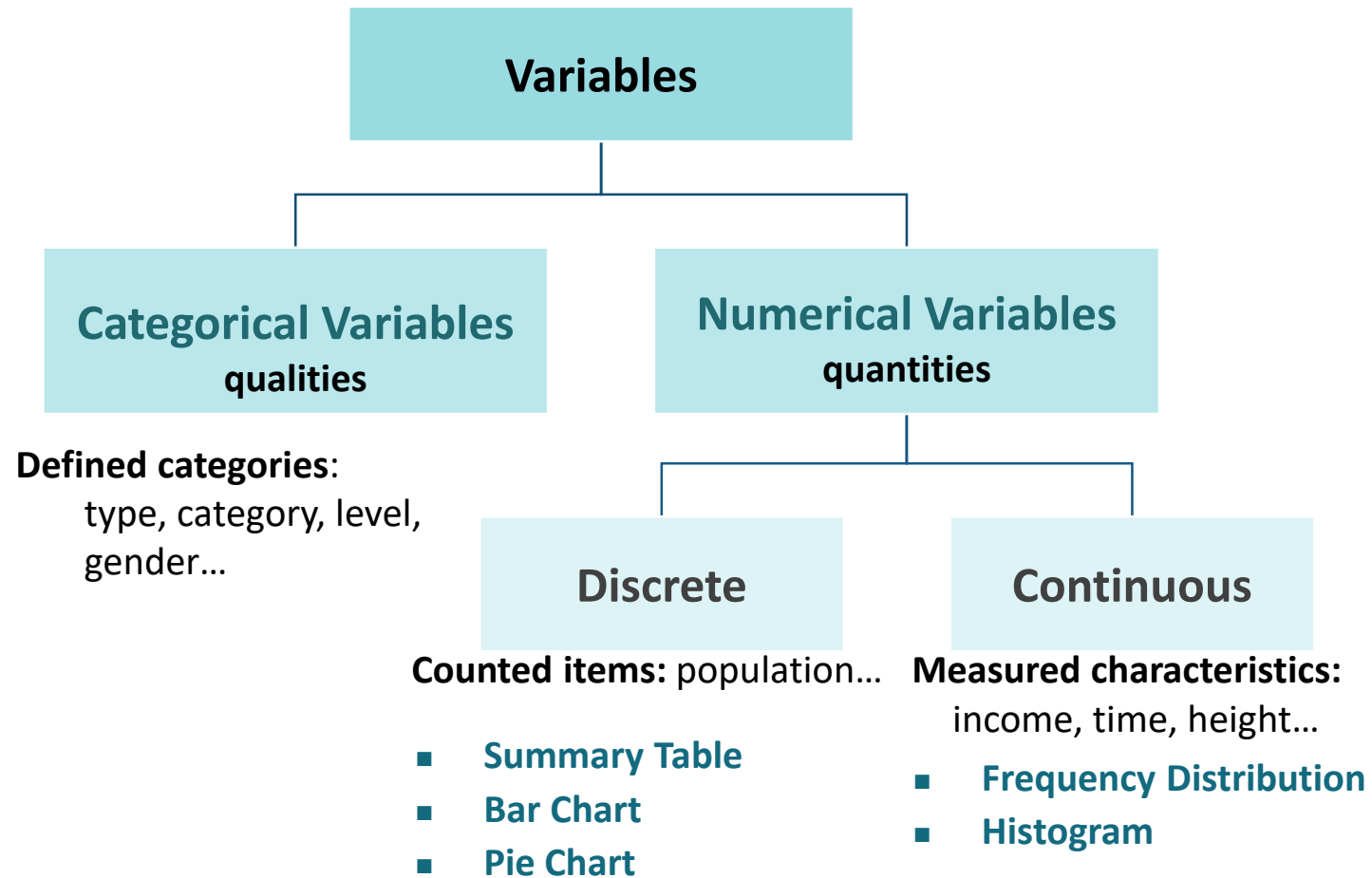
Population - Parameters



Sample - Statistics



Variables and Data



Data: values of a variable
Coding categorical data with numbers:
although the data values are numbers, the
variable is still categorical

Organizing and Visualizing Data

Type	Describe	Feature
Summary Table	count, percent	
Bar Chart	count, percent	Same gap, same width
Pie Chart	count, percent	mutually exclusive, total 100%
Frequency Distribution	range (lower & upper boundary, frequency)	Same interval (width) Steps: sort -> range -> no. of class -> class interval (width) -> class boundaries (limits) -> count
Histogram	frequency, percentage	no gap between bars, same width, made from frequency distribution

Count = frequency

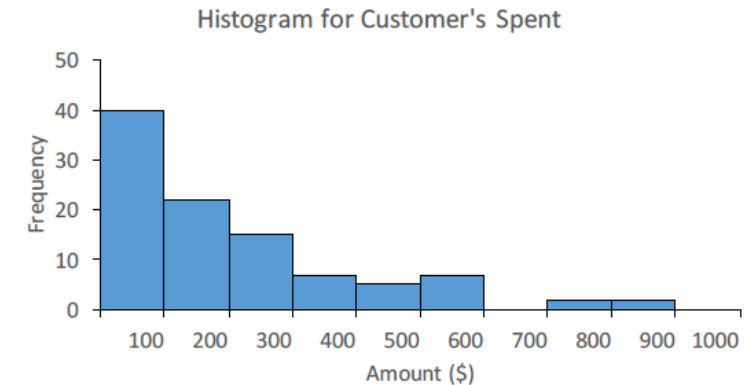
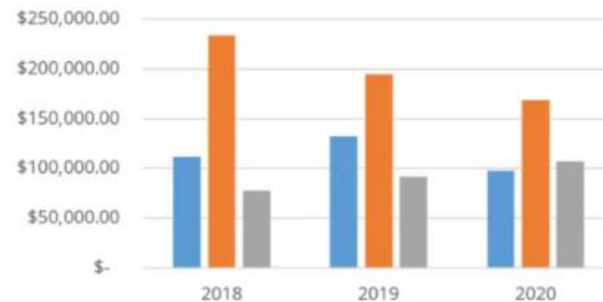
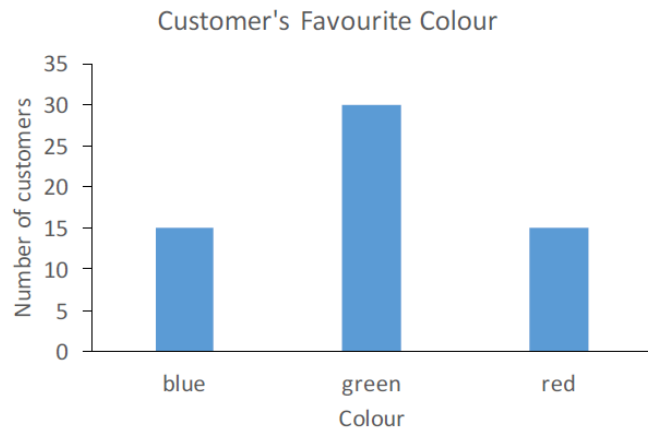
Percent = relative frequency

Bar Chart vs. Histogram?

Organizing and Visualizing Data

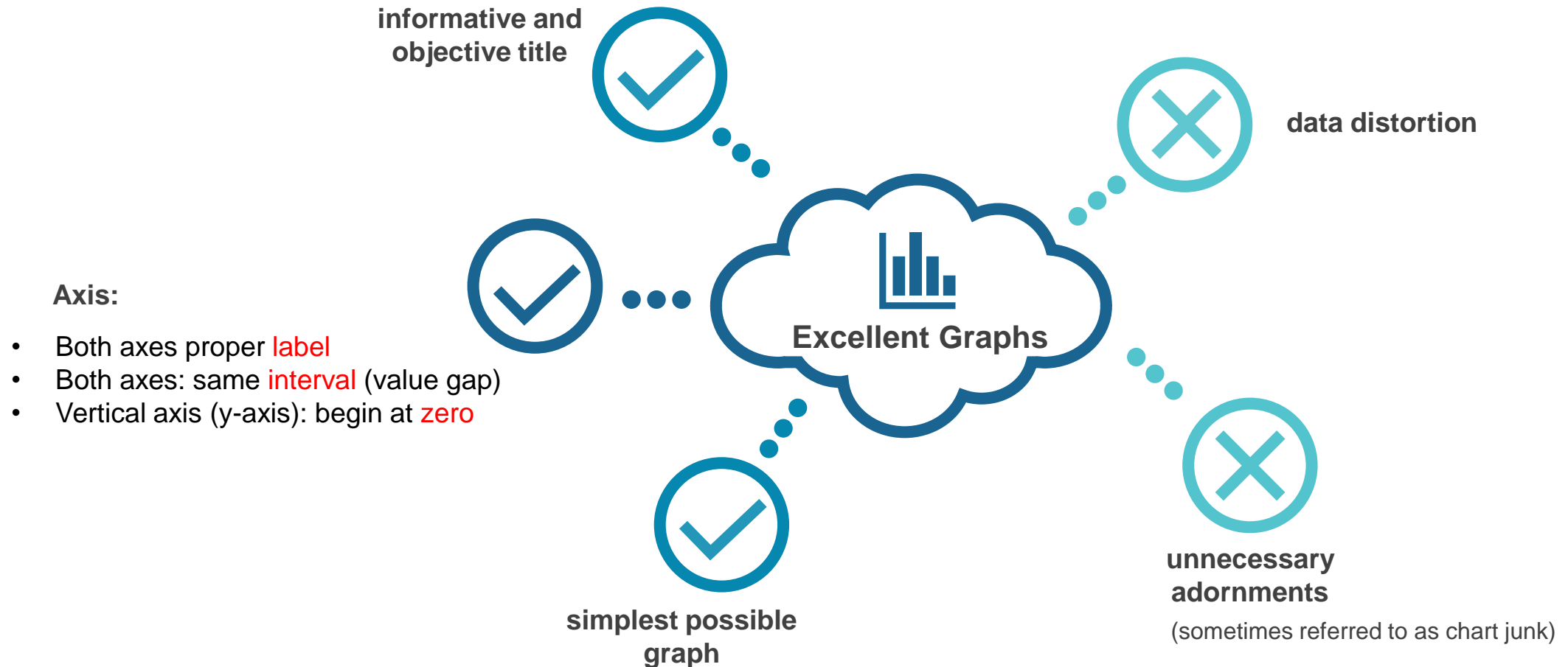
Type	Describe	Feature
Bar Chart	count, percent	Same gap, same width
Histogram	frequency, percentage	no gap between bars, same width, made from frequency distribution

Bar Chart vs. Histogram?



Histograms - show **distributions**, bin data nicely
Bar charts - used to **compare** between categories

Principles of Excellent Graphs



Summary Definitions I

- Central tendency: central value
 - **Mean**
 - population vs. sample
 - the only one affected by extreme values (outliers)
 - **Median**
 - order first
 - middle
 - **Mode**
 - most frequent
 - for both numeric & categorical data
 - may be no, or 1, or multiple

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$
$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

Summary Definitions II

- Variation: dispersion or scattering of values

- **Range**

- ignore distribution
 - sensitive to outliers

- **Quartiles**

- Q2=median
 - equal no. of values per segment
 - steps: sort -> find position -> find value (average/round)
 - not affected by outliers
 - interquartile range (i.e. *the midspread*) = Q3 - Q1

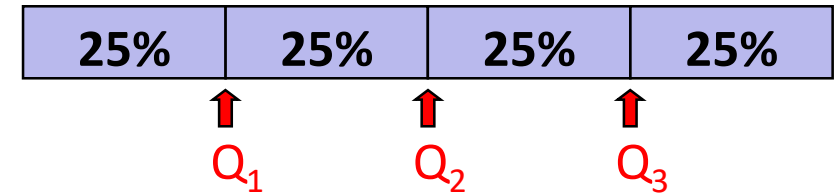
- **Variance**

- population vs. sample

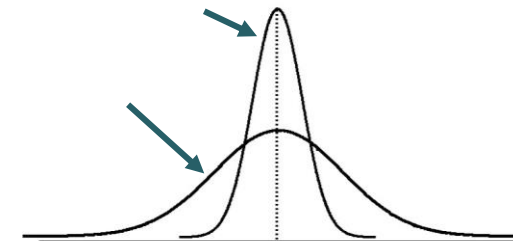
- **Standard deviation**

- square foot of variance

$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$



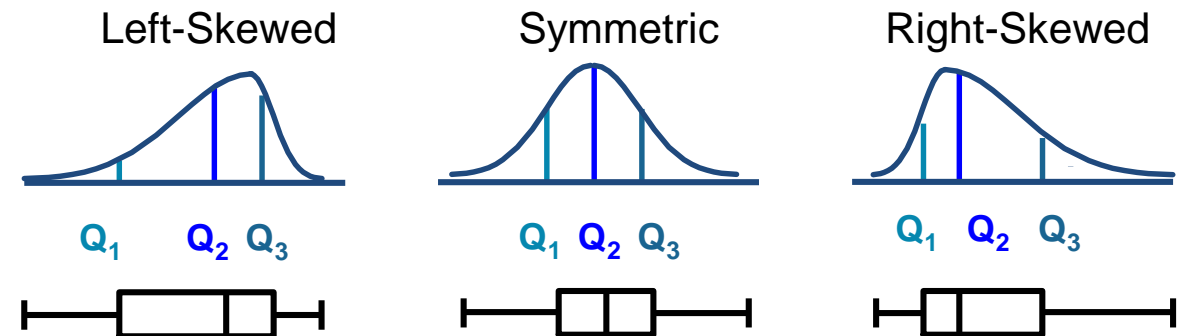
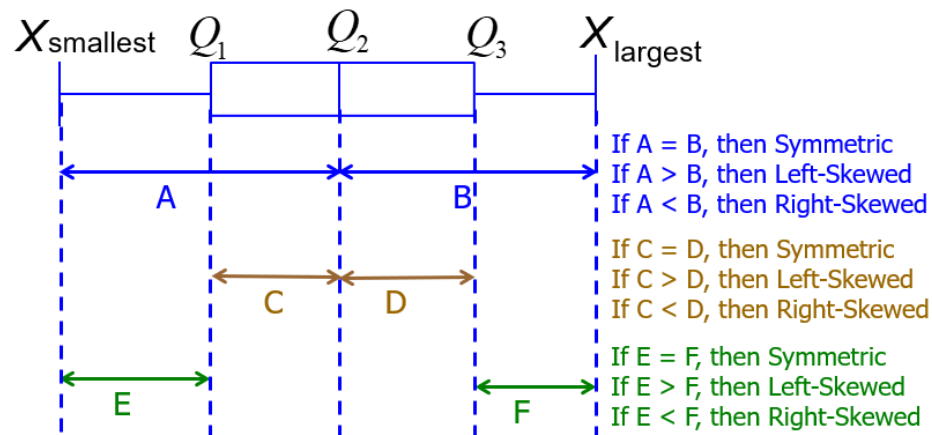
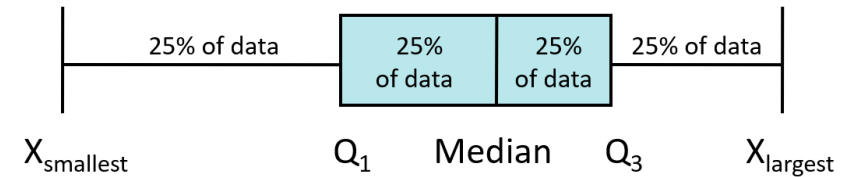
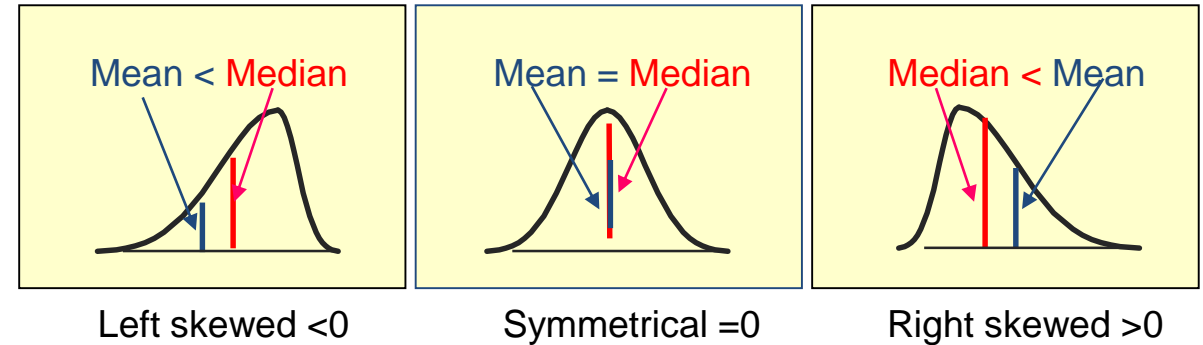
$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$



spread out => range ↑, variance ↑, standard deviation ↑
concentrated => range ↓, variance ↓, standard deviation ↓

Summary Definitions III

- Shape: distribution
 - **Skewness**
 - symmetrical level
 - median is preferred in skewed data
- **Five Number Summary and Boxplot**
 - what is the “5 numbers”?



Calculator

- Casio fx-50F
- Casio fx-82ES plus

(Your best friend: google)

Date Set:

163.6 156.2 166.3 179.3 157.8 165.4 159.5 161.7 160.4

1. Change to "Lin" mode

MODE **MODE** 5 1

2. Clear previous data

SHIFT **CLR** 1 **EXE**

3. Input data

163.6	M+	156.2	M+	166.3	M+	179.3	M+
157.8	M+	165.4	M+	159.5	M+	161.7	M+
160.4	M+						

4. Calculate descriptive statistics

Mean: **SHIFT** 2 1 1 **EXE** = 163.3555556

Population standard deviation: **SHIFT** 2 1 2 **EXE** = 6.459637417

Sample standard deviation: **SHIFT** 2 1 3 **EXE** = 6.851480132

No. of Data Input: **SHIFT** 1 3 **EXE** = 9

Tutorial questions

Week 2 T1 Q7-12

101001101001000010101
0011110111011011011010
101000011100101011001
010100111010100010101
0001011010110110110100
010101110001010100010
1000101110101100010011
010011010010000101010
0111101110110110110101
010000111001010110010
101001110101000101010
0010110101101101101001

