# INTRODUCTION TO STATISTICS

Session 3

April 24th 2017

Madoka Takeuchi

# Descriptive statistics

- When doing research, it is important to describe the data that you will be working with.
- The most fundamental information that should be reported are-
  - The size of the sample/data
  - The center of the sample/data
  - The spread of the the sample/data
  - The shape and distribution of the sample/data

# Review: Measures of Center

| | Always exist | Use all the data | Affected by extreme data |
|---|---|---|---|
| Mean | Yes | Yes | Yes |
| Median | Yes | Yes | No |
| Mode | No | No | No |

- The Mean is used in computing other statistics, such as the variance
  -It is often not appropriate for skewed distributions

- The Median is the center number and is good for skewed distributions because it is resistant to change.

# Review: Measures of Spread

- Range -simplest measure of variation
  - Range= Maximum-minimum
  - Since the range only uses the largest and smallest values, it is greatly affected by extreme values
- Variance- uses all the data
  - We want to see how much each point is deviated from the mean- could subtract each data point from the mean, sum all the differences and divide by N to find the average deviation in the data- *average deviation*

$$\Sigma(x-\mu) \ / \ N$$

  - The problem with the average deviation is that the summation is always zero

# Review: Measures of Spread cont`d

- To keep the deviation from the mean from summing to zero, square each individual deviations from the mean- called the *variance*

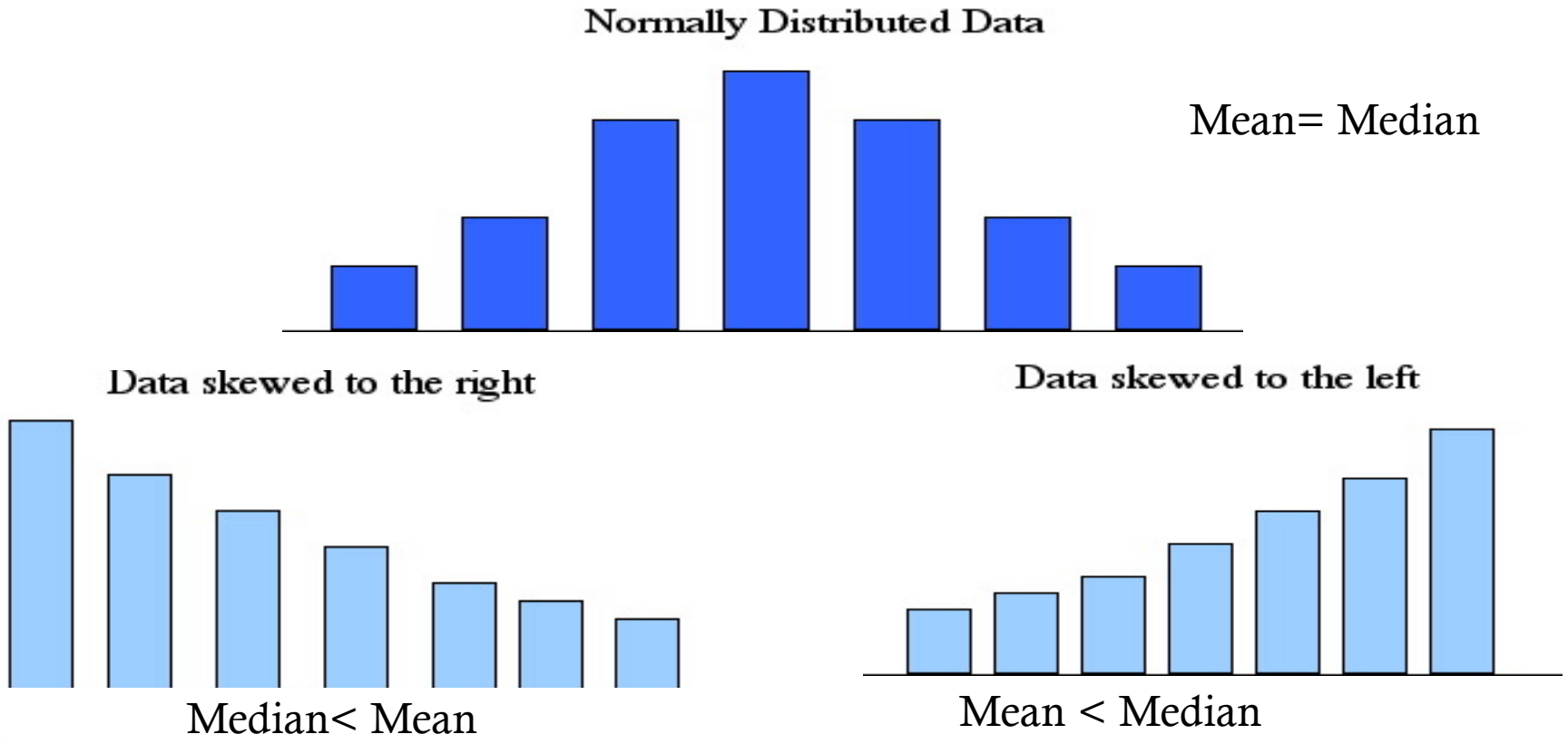
$$\sigma^2=\Sigma(x-\mu)^2/N$$

- There is a problem with variances since the deviations were squared resulting in the units also being squared- can be easily fixed by taking the square root- called the *Standard deviation*

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

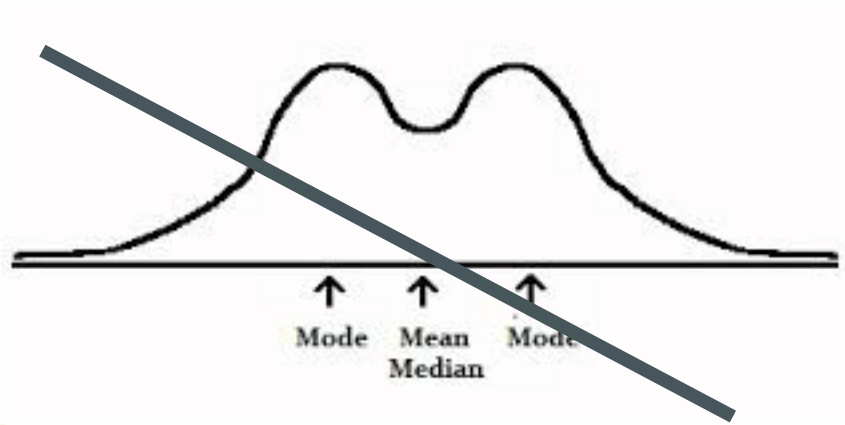
# Measures of Shape

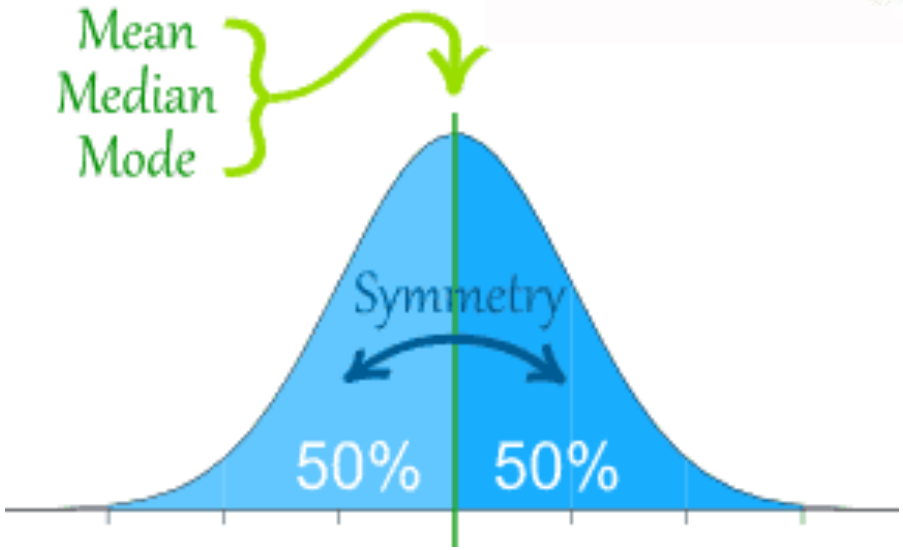
A histogram is a visual aid that can show the shape of the data



# Normal Distribution

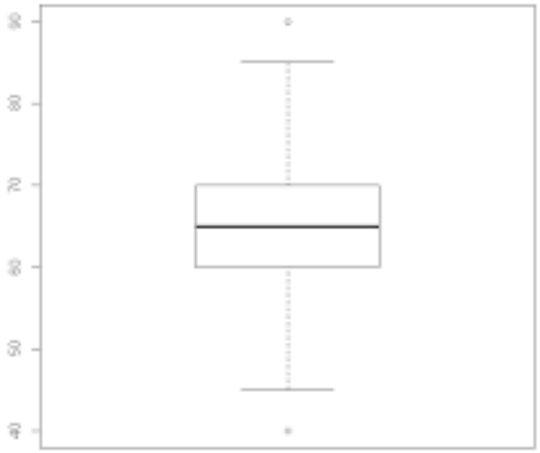
- Properties
  - Mean= median
  - Symmetrical
  - Unimodal- single peak

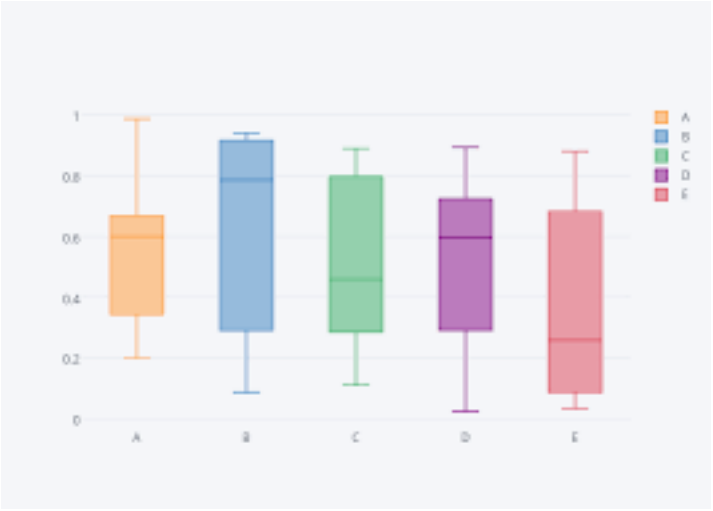


# Boxplot

- We saw that outliers affect the measures of center and spread.
  - Often, outliers are easiest to identify on a boxplot





# SPSS: Histogram

- ANALYZE> FREQUENCIES>CHART>HISTOGRAMS
- *Option to show normal curve on the histogram*
- Bar chart is very similar to histogram
  - Histogram is always in frequencies but bar chart can be in percentages
  - Build a bar chart to see the difference!

# SPSS: Boxplot

- GRAPHS>LEGACY DIALOGS>BOXPLOT
  - Choose simple
  - In the *data in chart are* box choose:
    - *summary of separate variables*- when want to get summary of the variable- i.e boxplot of weight
    - *Summary for groups of cases*- when want to get a boxplot for a variable separate by category- i.e. boxplot of weight by gender
      - Need to define category in the *category axis* box

# In class exercise- SPSS

- Data was collected on 12 students regarding average daily hours spent watching tv and hours of sleep
- Input the data into spss
- Describe the data-give the sample size, measures of center, measures of spread for each variable
- By looking at the relationship between the mean the median, what can be said about the distribution for the hours spent watching tv and hours of sleep?

| Initials | year | tv | sleep |
|---|---|---|---|
| AB | 1 | 3 | 6 |
| BC | 4 | 6 | 7 |
| CD | 1 | 2 | 5 |
| DE | 2 | 4 | 6 |
| EF | 2 | 5 | 8 |
| FG | 2 | 3 | 5 |
| GH | 2 | 4 | 6 |
| HI | 2 | 2 | 6.5 |
| IJ | 2 | 3 | 7 |
| JK | 1 | 1 | 3.5 |
| KL | 1 | 3 | 12 |
| LM | 2 | 3 | 6 |