Biostatistics

Chapter 2 Describing and Displaying Data

Jing Li

jing.li@sjtu.edu.cn

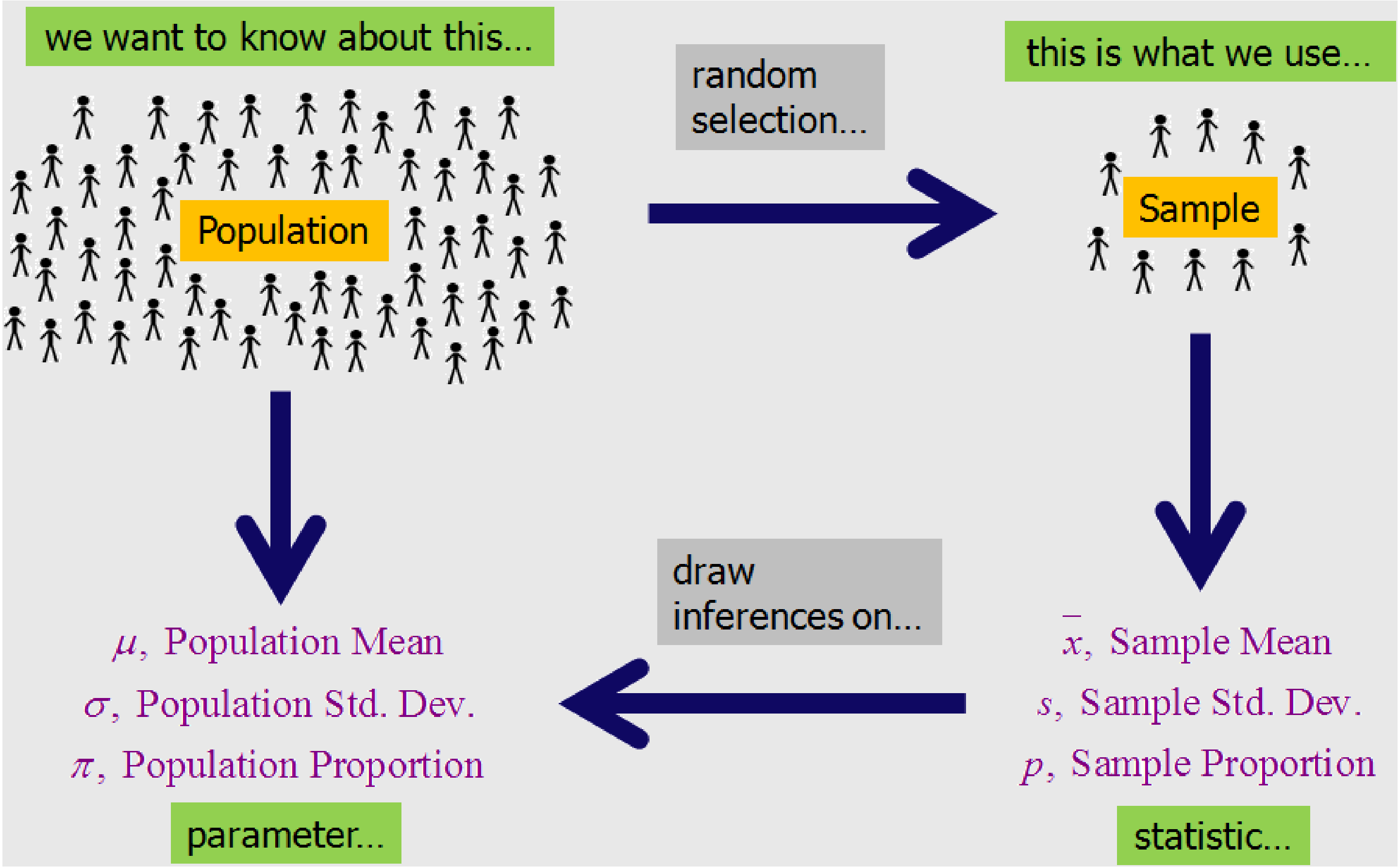
http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/

Dept of Bioinformatics & Biostatistics, SJTU

Review

- What's biostatistics ?
- Sample and population?

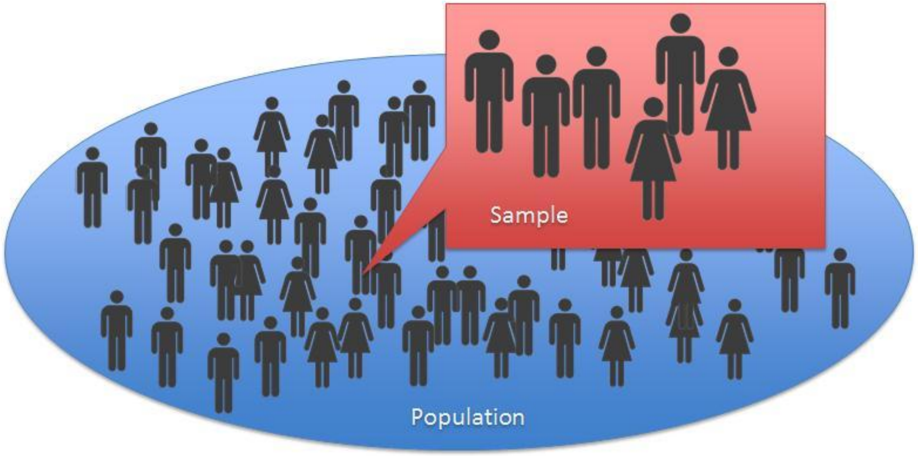
Sample and population



Homework1

- Please investigate the average height the undergraduates at SJTU (group work).

截至2016年12月，上海交通大学在校全日制本科生（国内）16195人

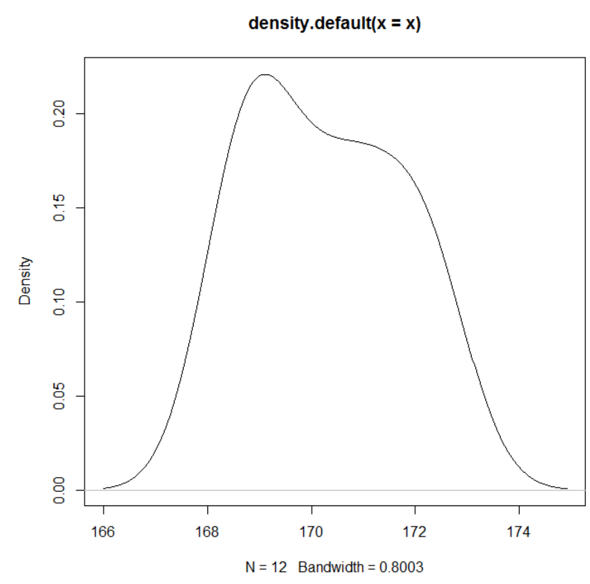


Homework1

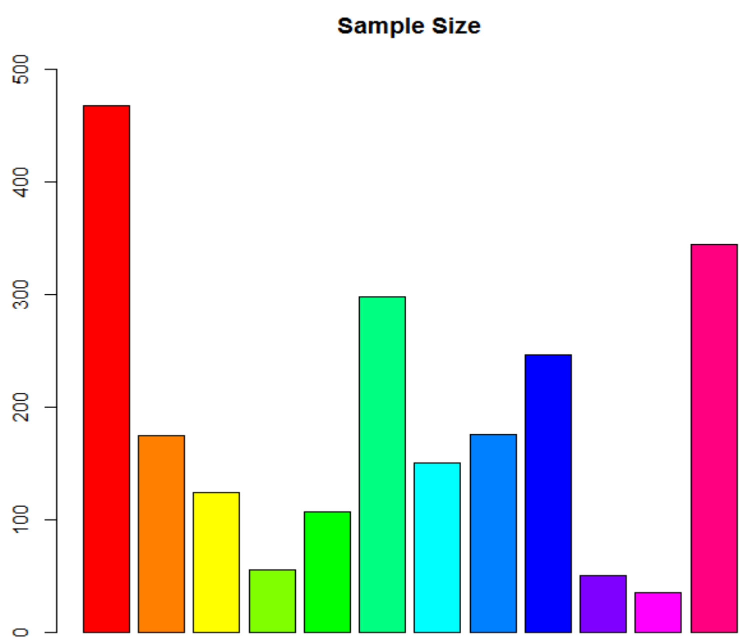
- Please investigate the average height the undergraduates at SJTU (group work).

截至2016年12月，上海交通大学在校全日制本科生（国内）16195人

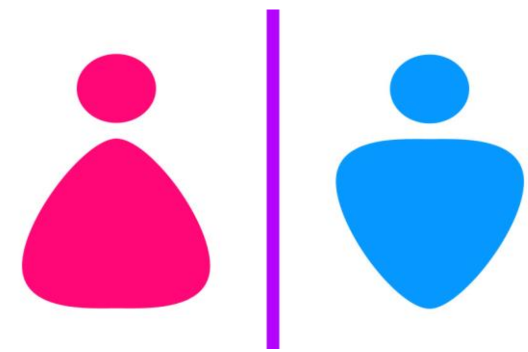
```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  168.4   168.9   170.2   170.3   171.3   172.5 
```



Homework1---- data collection




```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  168.4   168.9   170.2   170.3   171.3   172.5 
```



男女比
没有具体样本男女比例，默认学校给出数据1.9:1
没有具体样本男女比例，控制男女比例大致和全校男女比例1.9:1相等
98:26，用学校男女比2:1校正
没有具体样本男女比例
76:31
168：130
100:50
没有具体样本男女比例
2:1
没有具体样本男女比例
29:6
206:138；用学校男女比1.9:1校正

男生173.5cm ;女生163.5 cm
男生173.47cm;女生161.80cm

Chapter 2 Describing and Displaying Data

Topics:

- Displaying Data
- Describing Data

Looking at Data

Looking at Data

- How are the data distributed?
  - Where is the center?
  - What is the range?
  - What's the shape of the distribution (e.g., Gaussian, binomial, exponential, skewed)?
- Are there "outliers"?
- Are there data points that don't make sense?

Displaying Data

Frequency tables（频数表）

Used for displaying information about categorical variables or continuous variables chopped into categories.

Education	Count (millions)	Percent
Less than high school	4.6	12.1
High school graduate	11.6	30.5
Some college	7.4	19.5
Associate degree	3.3	8.7
Bachelor's degree	8.6	22.6
Advanced degree	2.5	6.6

Displaying Data

Frequency Plots

- Categorical variables
  - Bar Chart (条图)
- Continuous variables
  - Stem-and-Leaf Plot (茎叶图)
  - Histogram (直方图)
  - Box Plot (箱图)

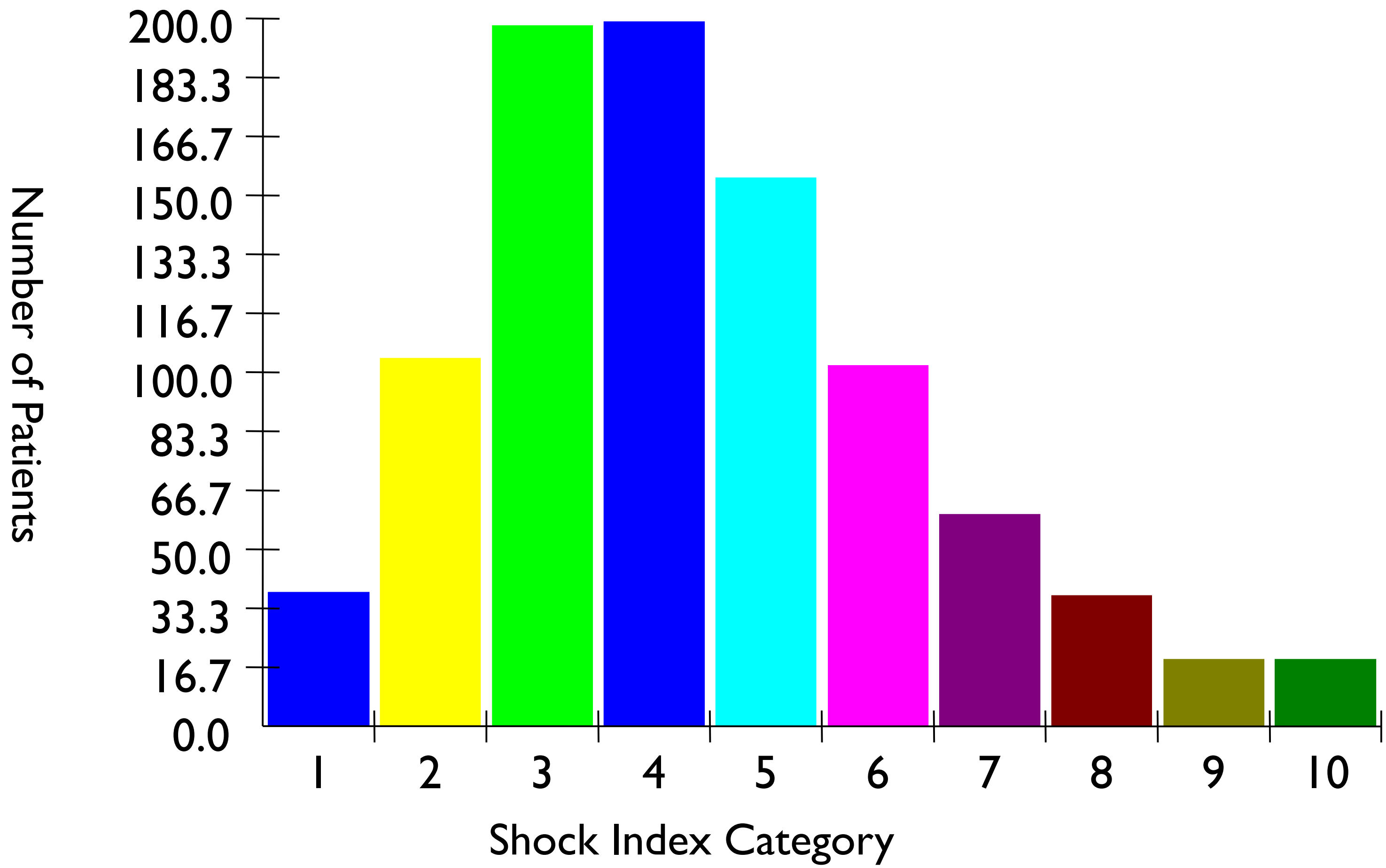
Display your data with graph

Displaying Data

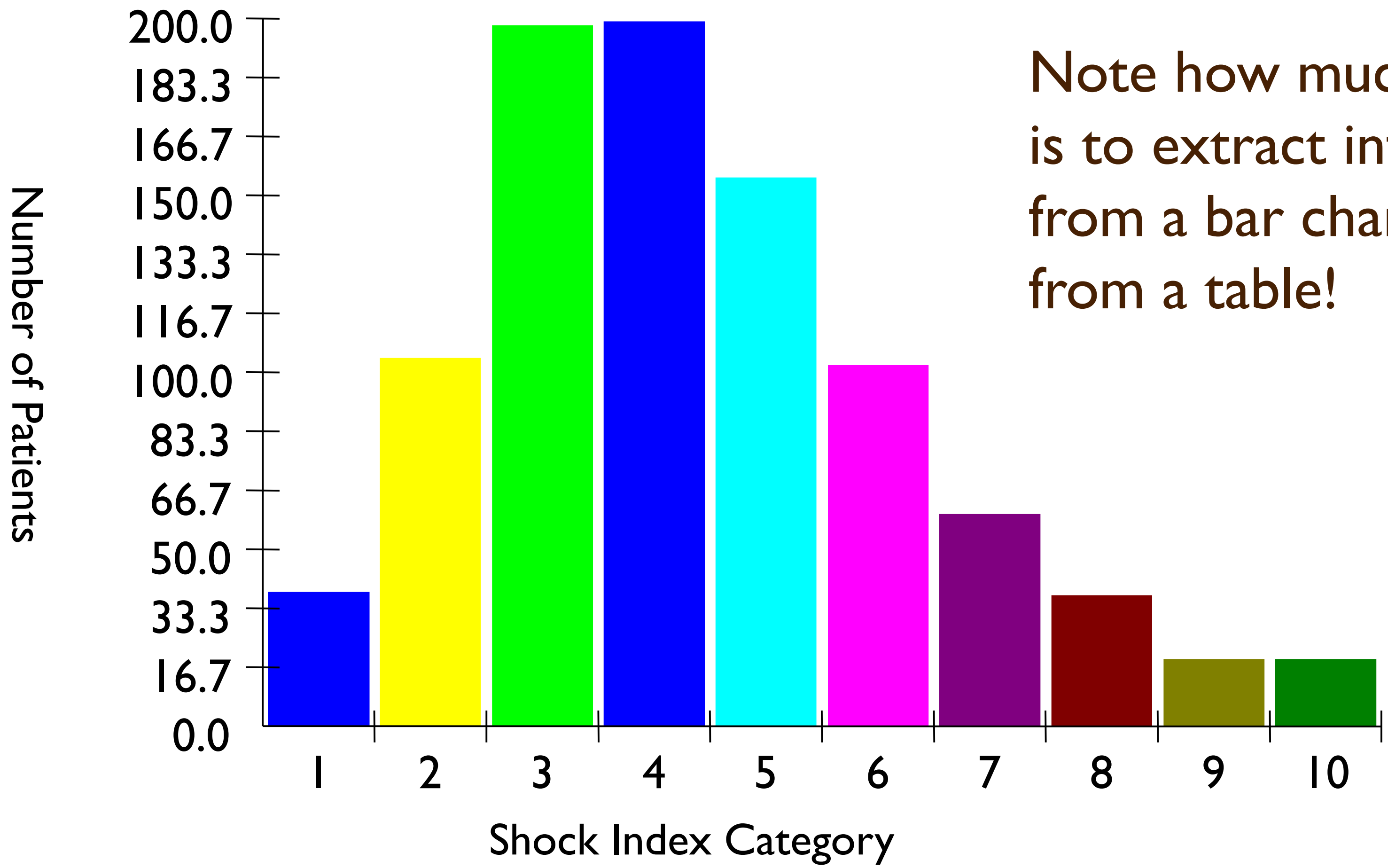
Bar Chart (条形图)

- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables into a pictorial representation…

Bar Chart for SI categories

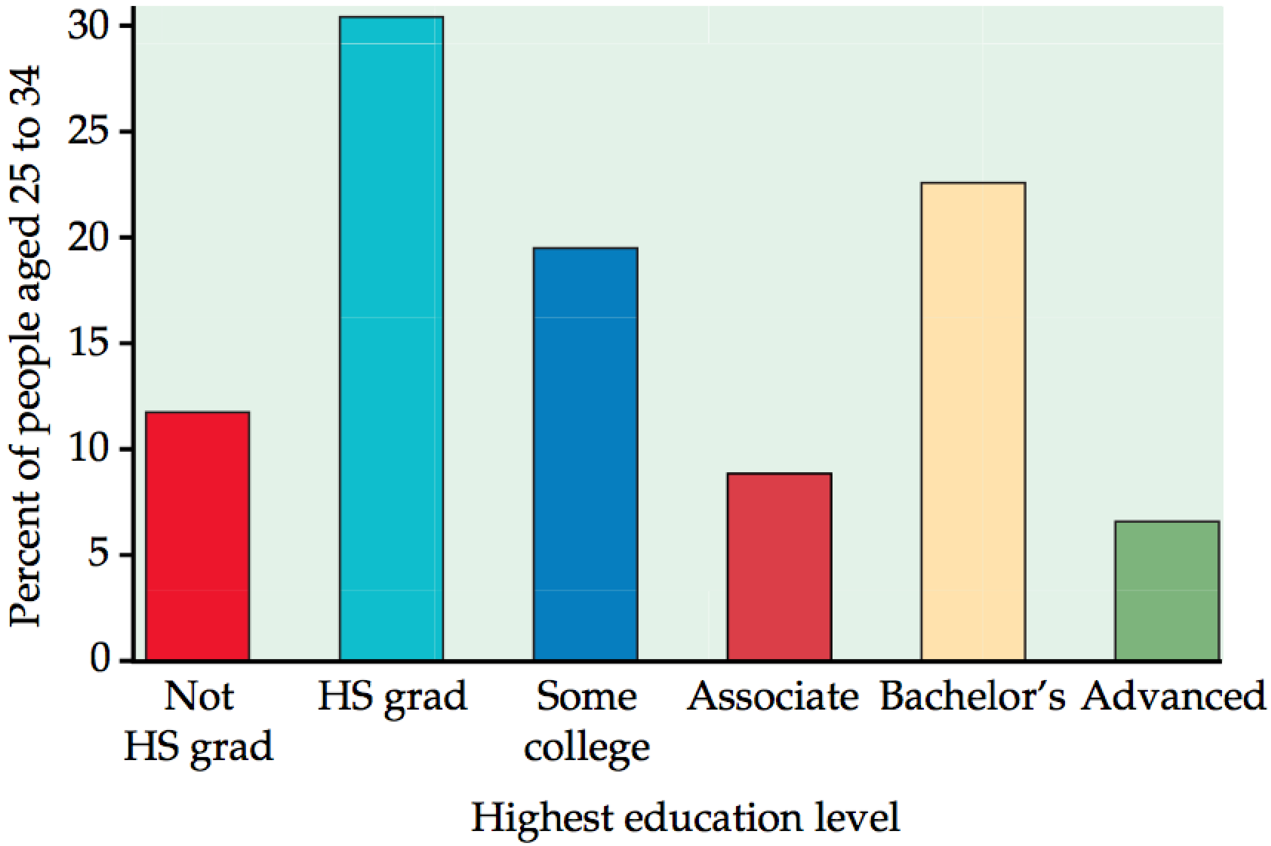


Bar Chart for SI categories



Note how much easier it is to extract information from a bar chart than from a table!

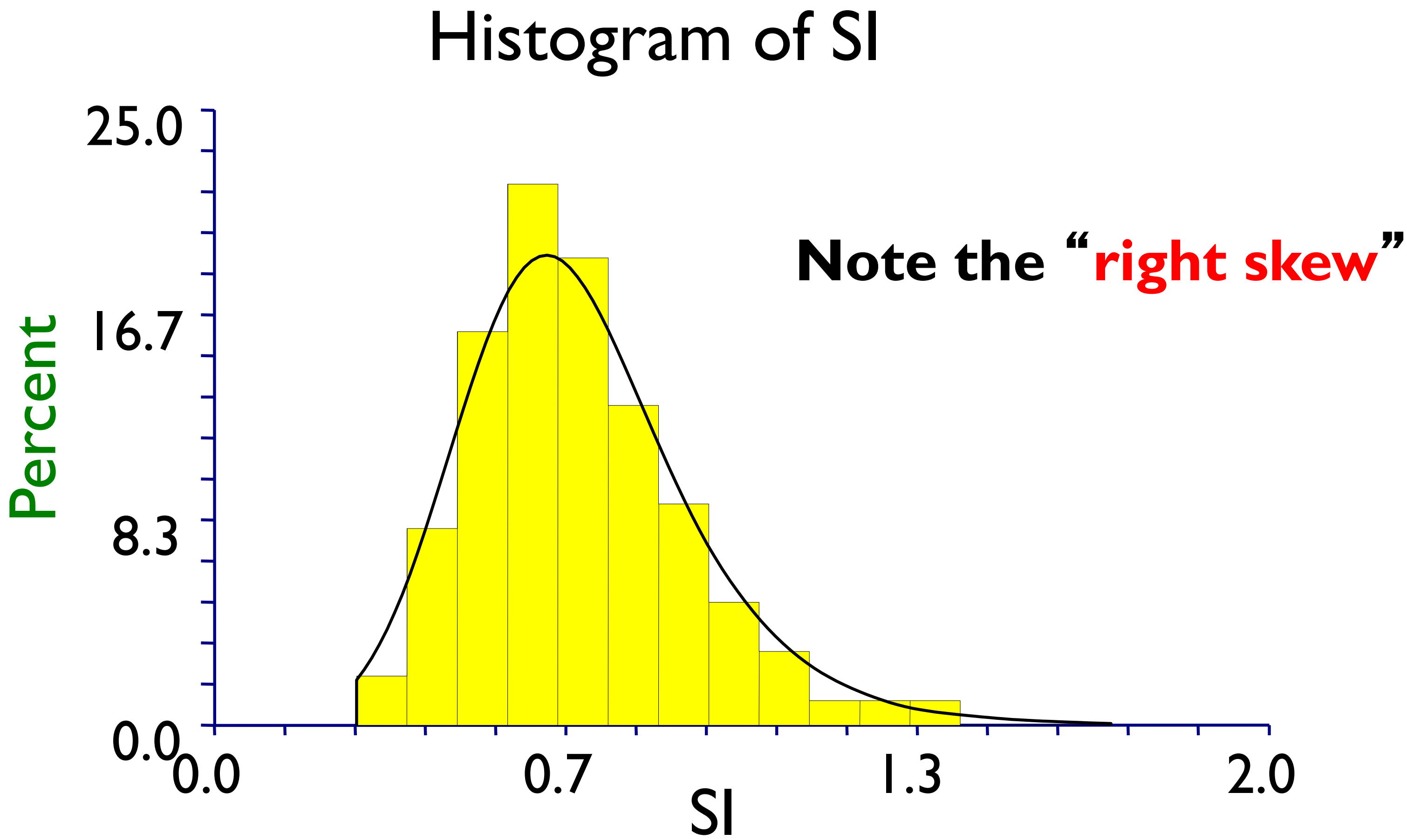
Another Example



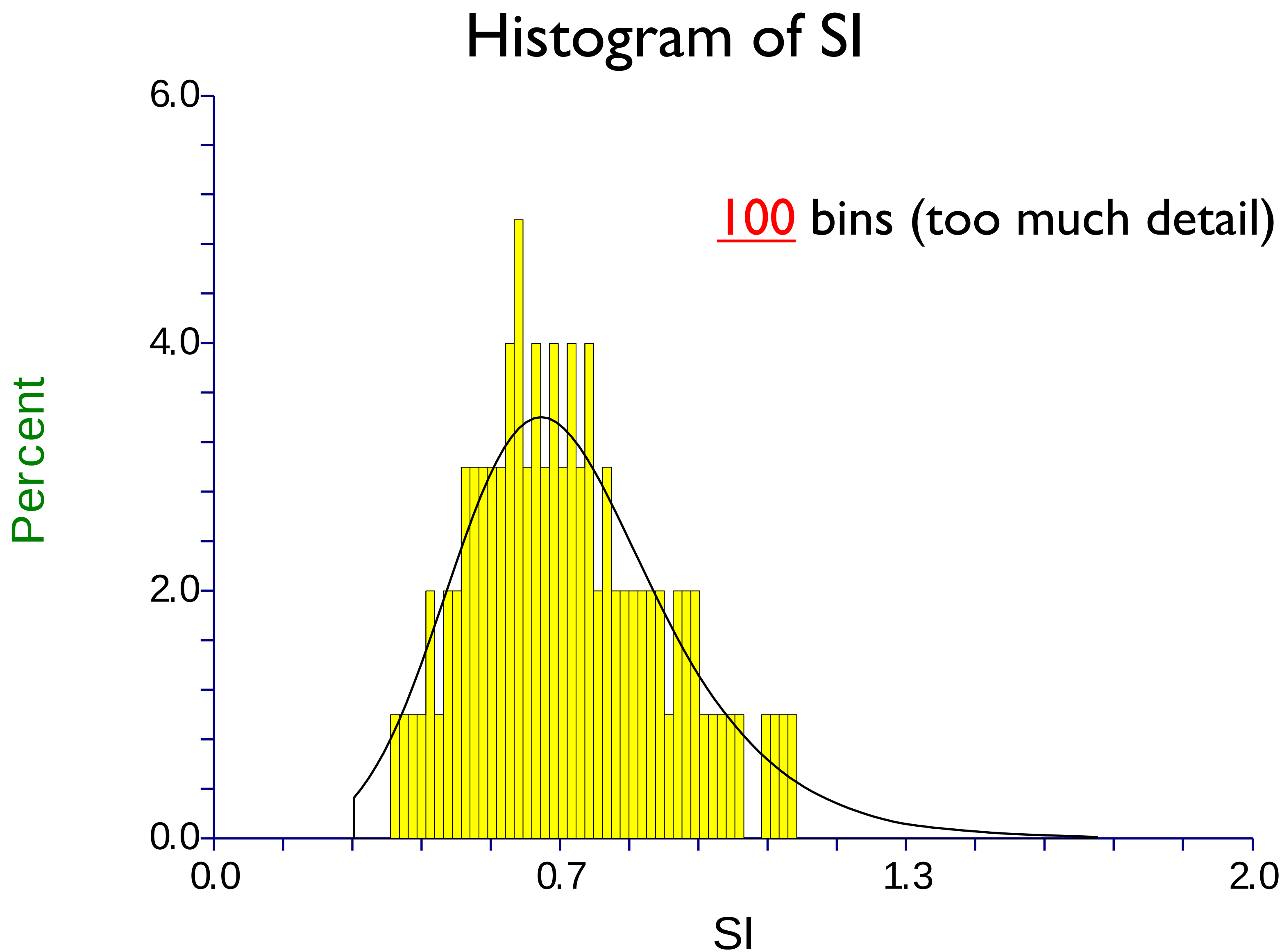
Displaying Data

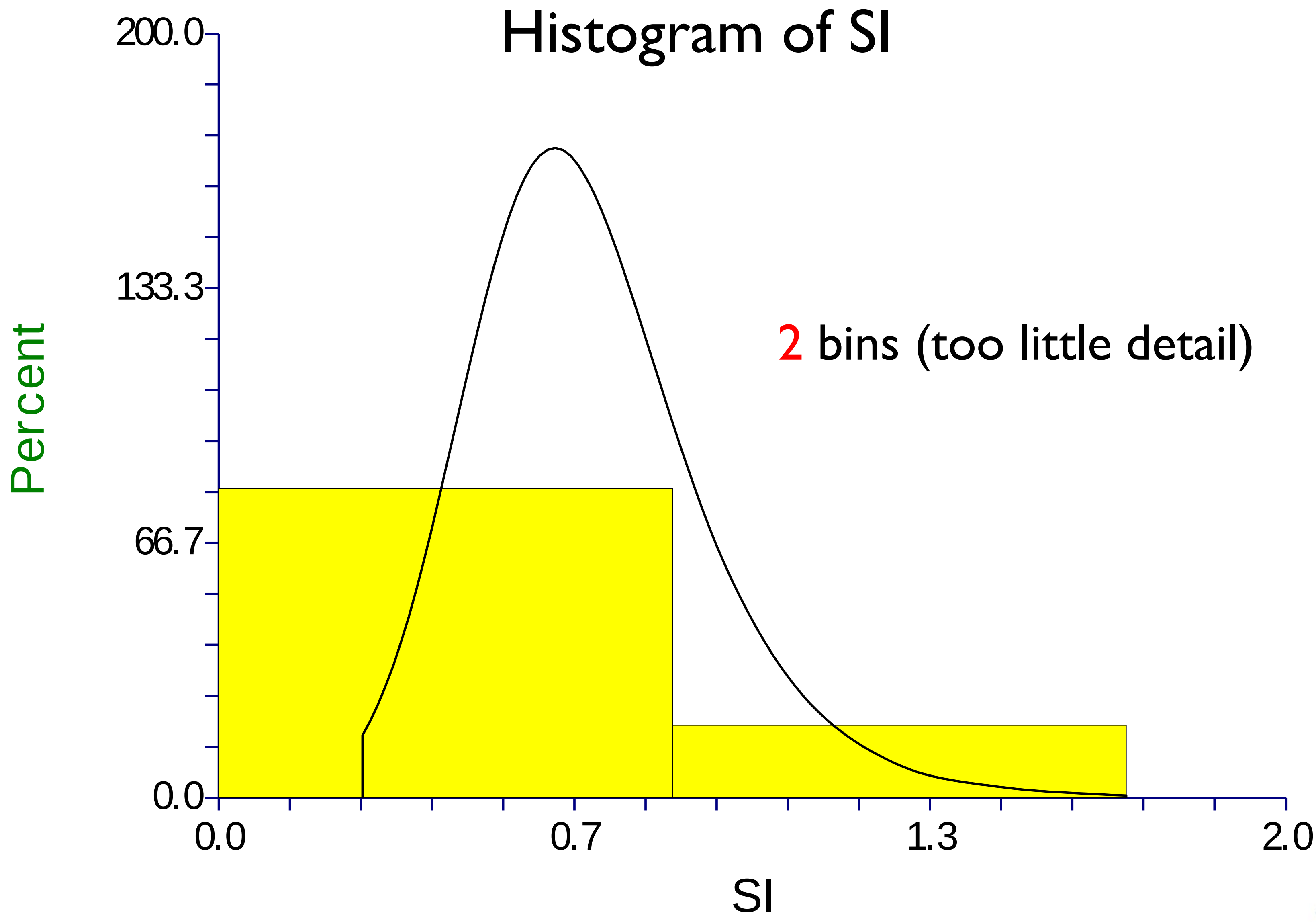
Histogram（直方图）

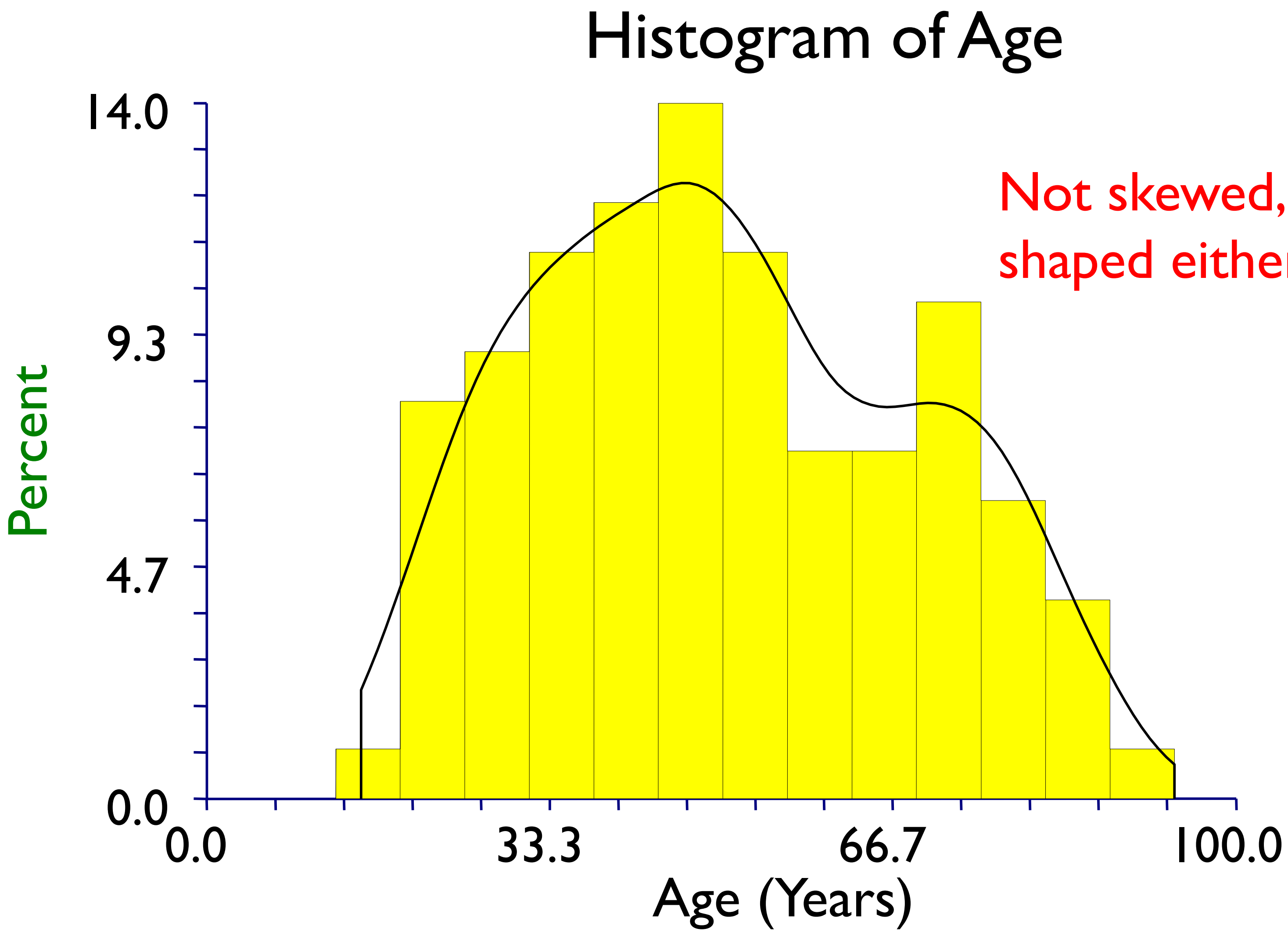
- To show the distribution (shape, center, range, variation) of continuous variables.



Bins of size 0.1 (automatically generated)

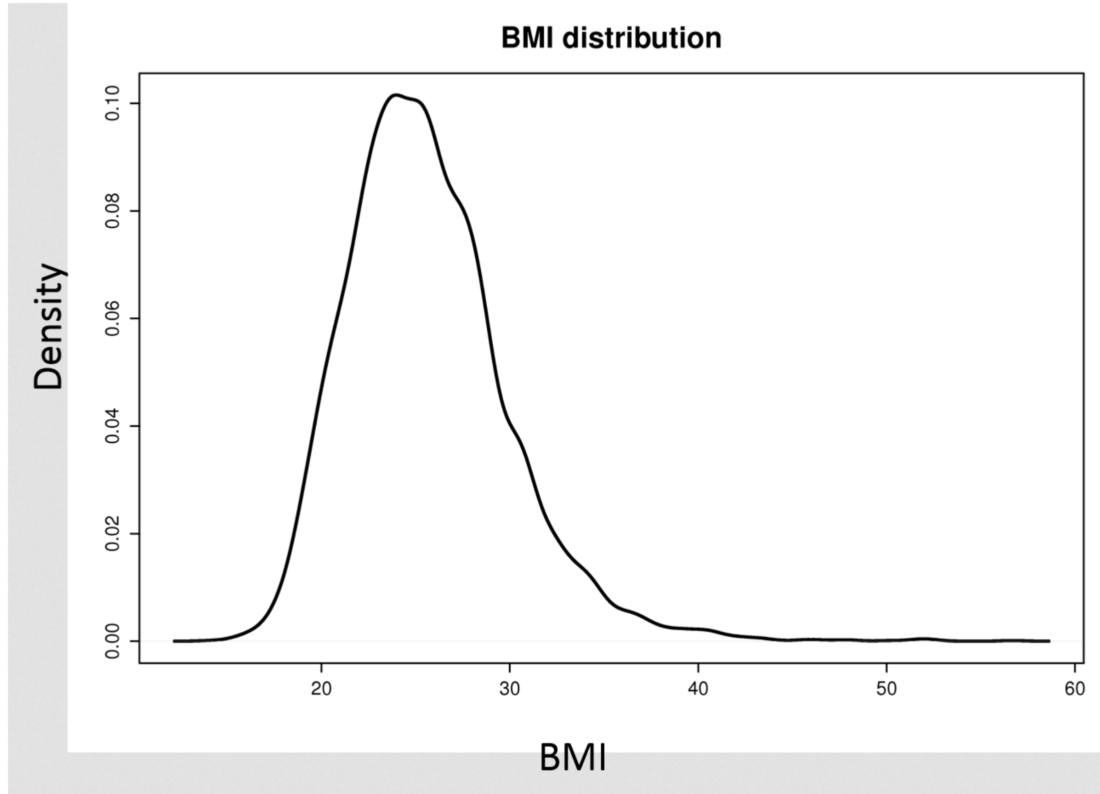

Histogram of SI
100 bins (too much detail)
Percent
6.0
4.0
2.0
0.0
0.0
0.7
1.3
2.0
SI


Histogram of SI
200.0
133.3
66.7
0.0
Percent
2 bins (too little detail)
0.0
0.7
1.3
2.0
SI



Not skewed, but not bell-shaped either…

Density plot



plot(density())

Review Question

What is the first thing you should do when you get new data?

a. Run a ttest
b. Calculate a p-value
c. Plot your data
d. Run multivariate regression

Review Question

What is the first thing you should do when you get new data?

a. Run a ttest
b. Calculate a p-value
c. **Plot your data!**
d. Run multivariate regression

Describing Data

Measures of central tendency

- Mean（均值）
- Median（中位数）
- Mode（众数）

Central Tendency

- Mean – the average; the balancing point

The sum of values divided by the sample size

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Central Tendency

Mean: example

Some data:

Age of participants: 17 19 21 22 23 23 23 38

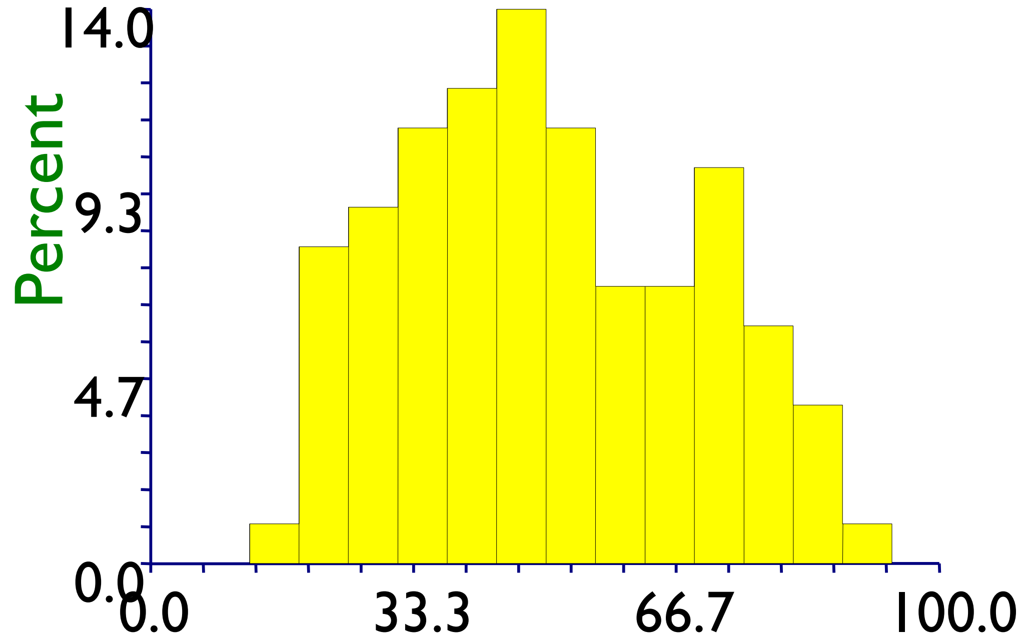
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{17+19+21+22+23+23+23+38}{8} = 23.25$$

Mean of age in Kline’s data

Descriptive Statistics Report

Means Section of AGE

Parameter	Mean	Median	Geometric Mean	Harmonic Mean	Sum	Mode
Value	50.19334	49	46.66865	43.00606	4673049	

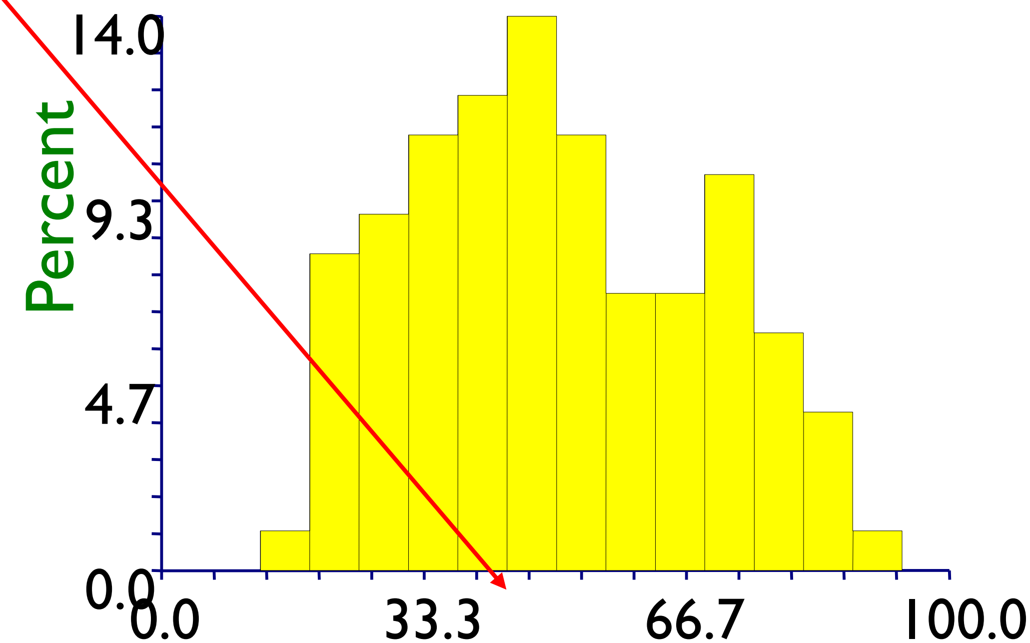


Mean of age in Kline’s data

Descriptive Statistics Report

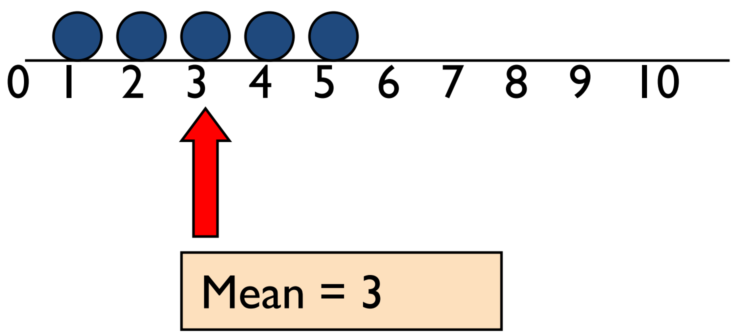
Means Section of AGE

Parameter	Mean	Median	Geometric Mean	Harmonic Mean	Sum	Mode
Value	50.19334	49	46.66865	43.00606	4673049	

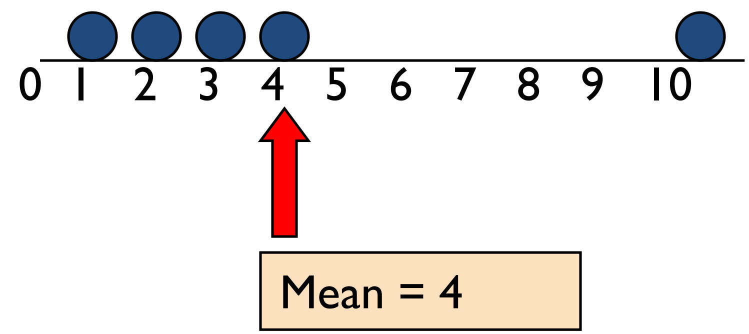


Central Tendency

- The mean is affected by extreme values (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Central Tendency

- Median – the exact middle value

Calculation:

- If there are an odd number of observations, find the middle value
- If there are an even number of observations, find the middle two values and average them

Central Tendency

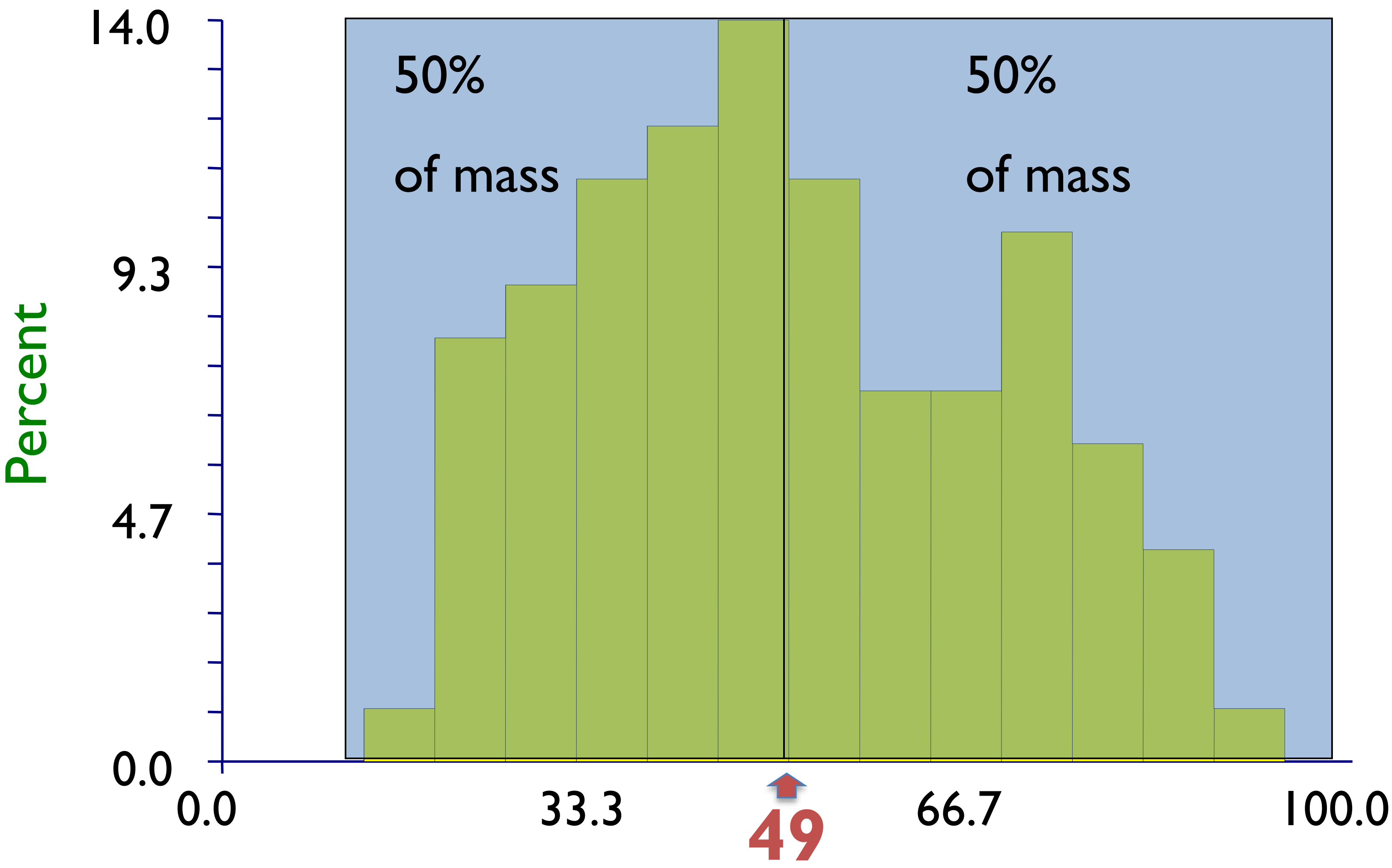
Median: example

Some data:

Age of participants: 17 19 21 22 23 23 23 38

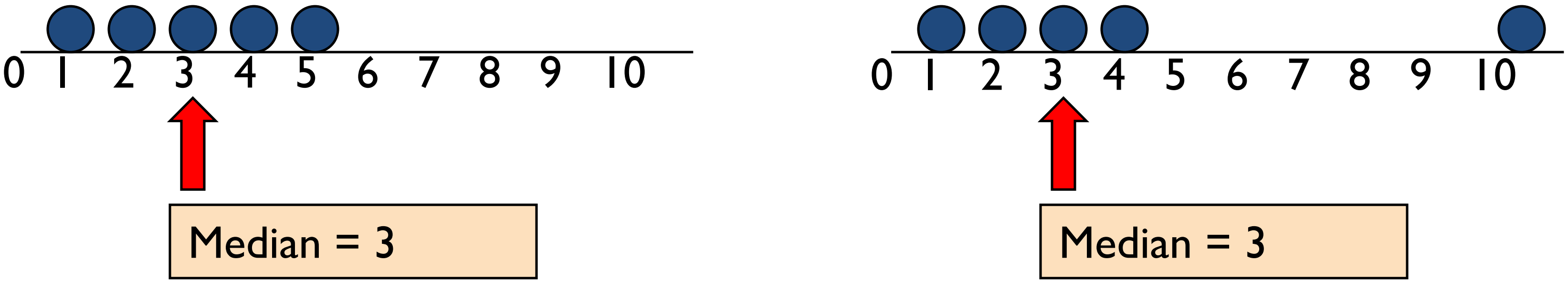
Median = (22+23)/2 = 22.5

Median of age in Kline's data



Central Tendency

- The median is not affected by extreme values (outliers).



Central Tendency

- Mode – the value that occurs most frequently

Some data:
Age of participants: 17 19 21 22 23 23 23 38

Central Tendency

- Mode – the value that occurs most frequently

Some data:
Age of participants: 17 19 21 22 23 23 23 38

Mode = 23 (occurs 3 times)

Review question

Some data:

Age of participants: 17 19 21 22 23 38

What's the mode ?

Review question

<u>Some data</u>:

Age of participants: 17 19 21 22 23 38

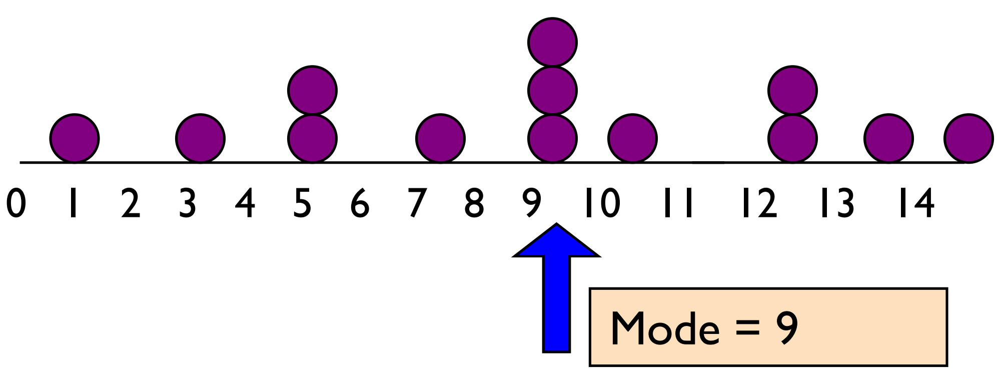
What's the mode ?

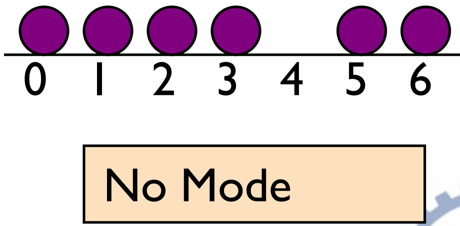
Answer: No mode

Central Tendency

Mode

- Not affected by extreme values
- Used for either numerical or categorical data
- There may may be no mode
- There may be several modes





Which measure of central tendency is best?

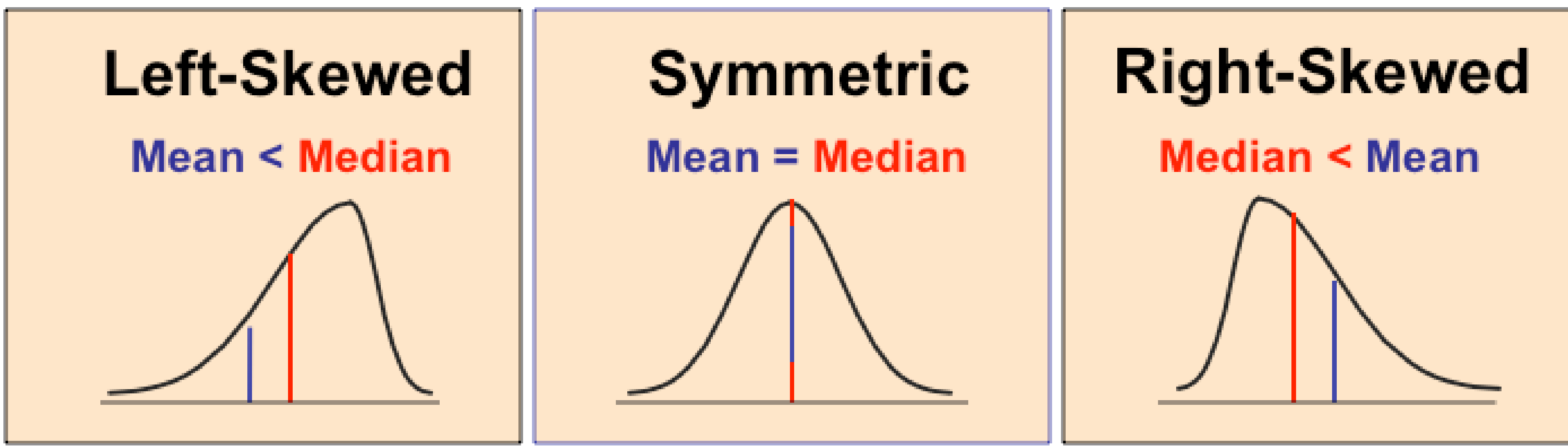
- Mean is generally used, unless extreme values (outliers) exist
- Then median is often used, since the median is not sensitive to extreme values.

Example:

median home prices may be reported for a region – less sensitive to outliers

Shape of a Distribution

- Describes how data are distributed
- Measures of shape
  - **Symmetric or skewed**



Shape of a Distribution

- Describes how data are distributed
- Measures of shape
- Skewness

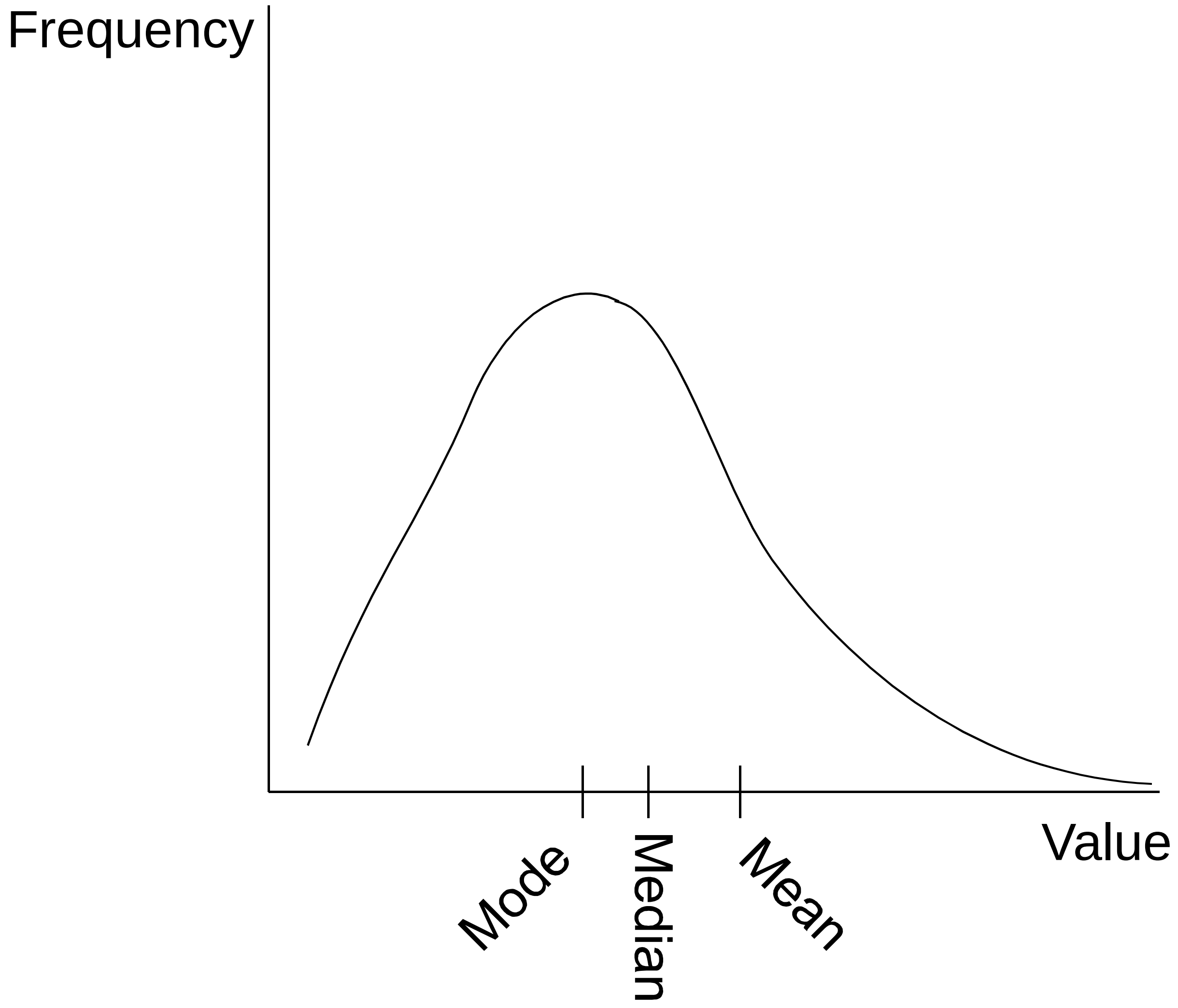
-样本偏度

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}},$$

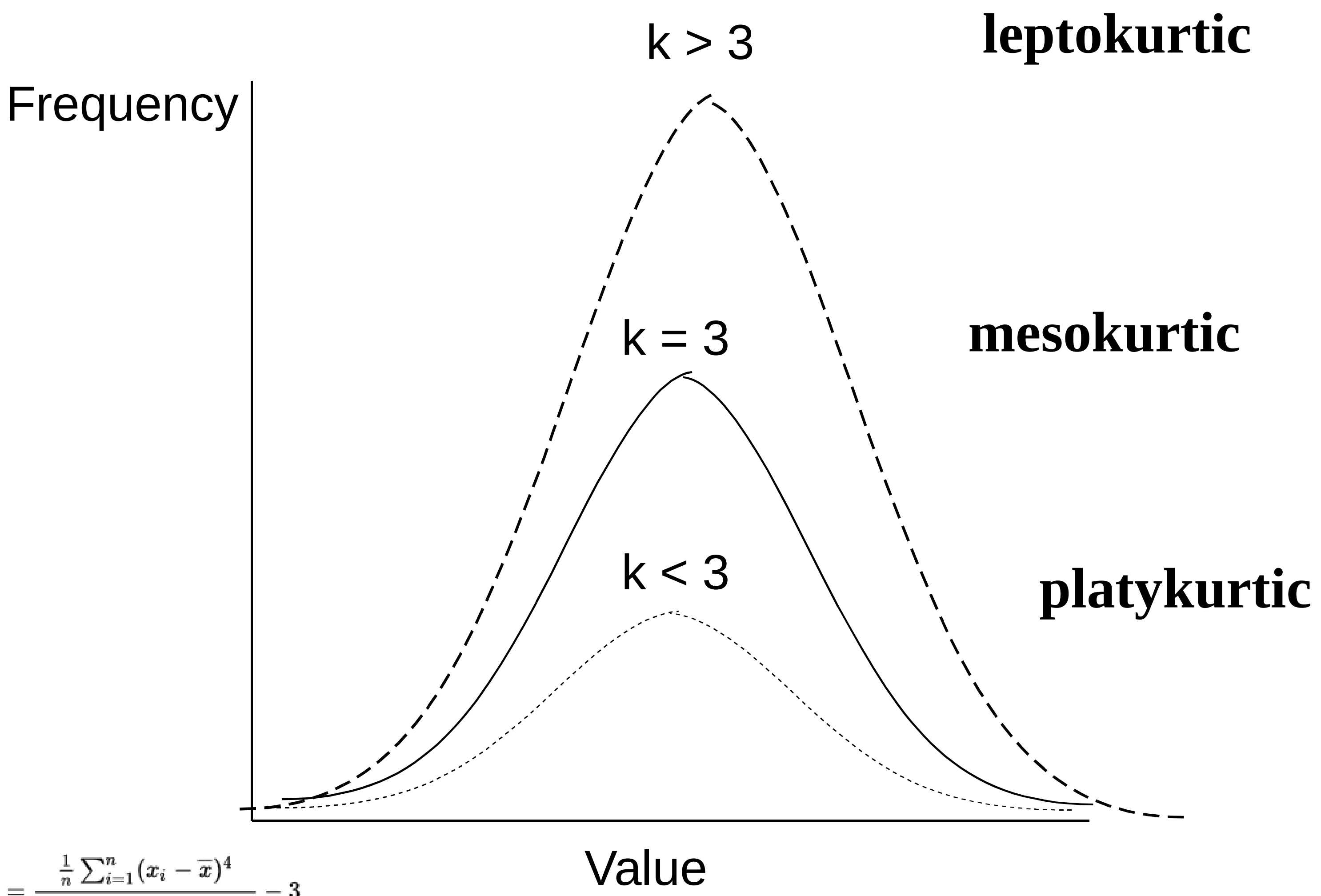
其中， m3是三阶样本中心矩，m2是二阶样本中心距，即样本方差。

正态分布：偏度为0；<0, 左偏，>0, 右偏

Skewness

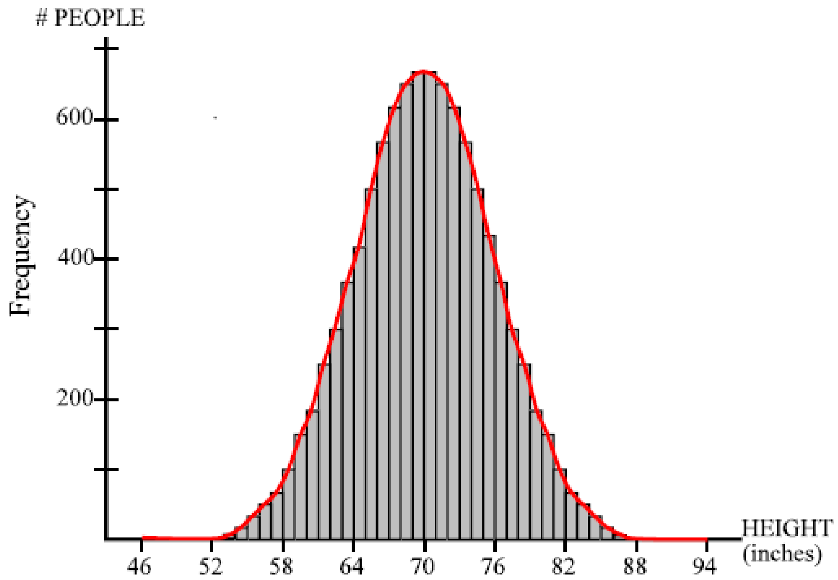


Kurtosis



$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$

Normal distribution



- Skewness = 0
- Kurtosis = 3

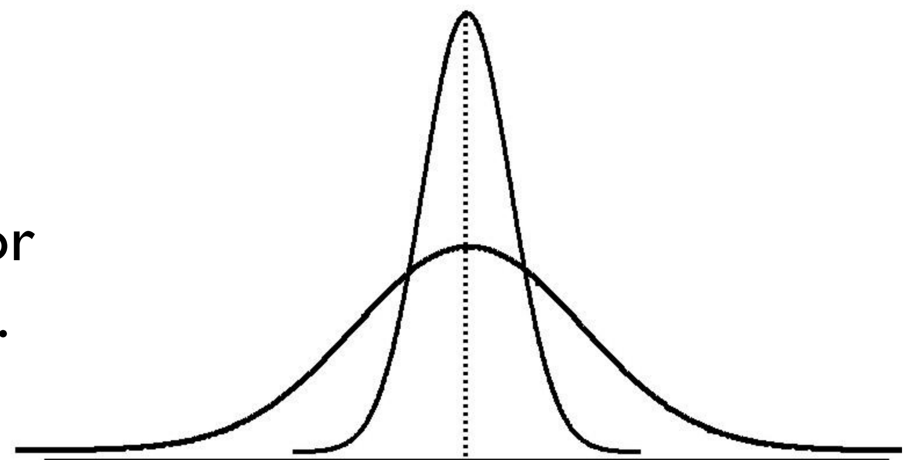
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

Measures of Dispersion

Dispersion

- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

- Measures of variation give information on the spread or variability of the data values.



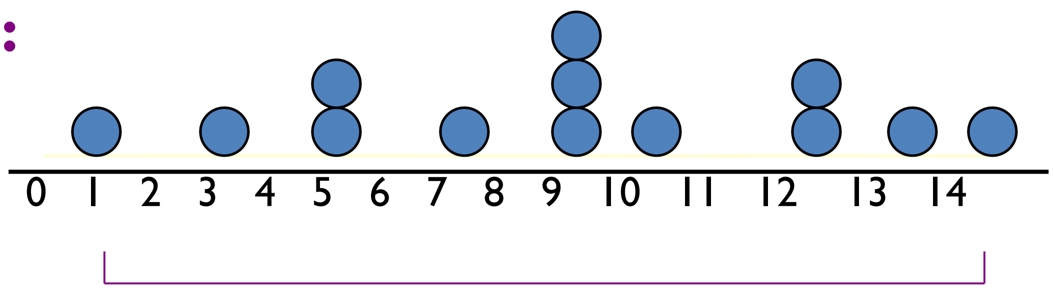
Same center, different variation

Measures of Dispersion

- Range
  - Simplest measure of dispersion
  - Difference between the largest and the smallest observations:

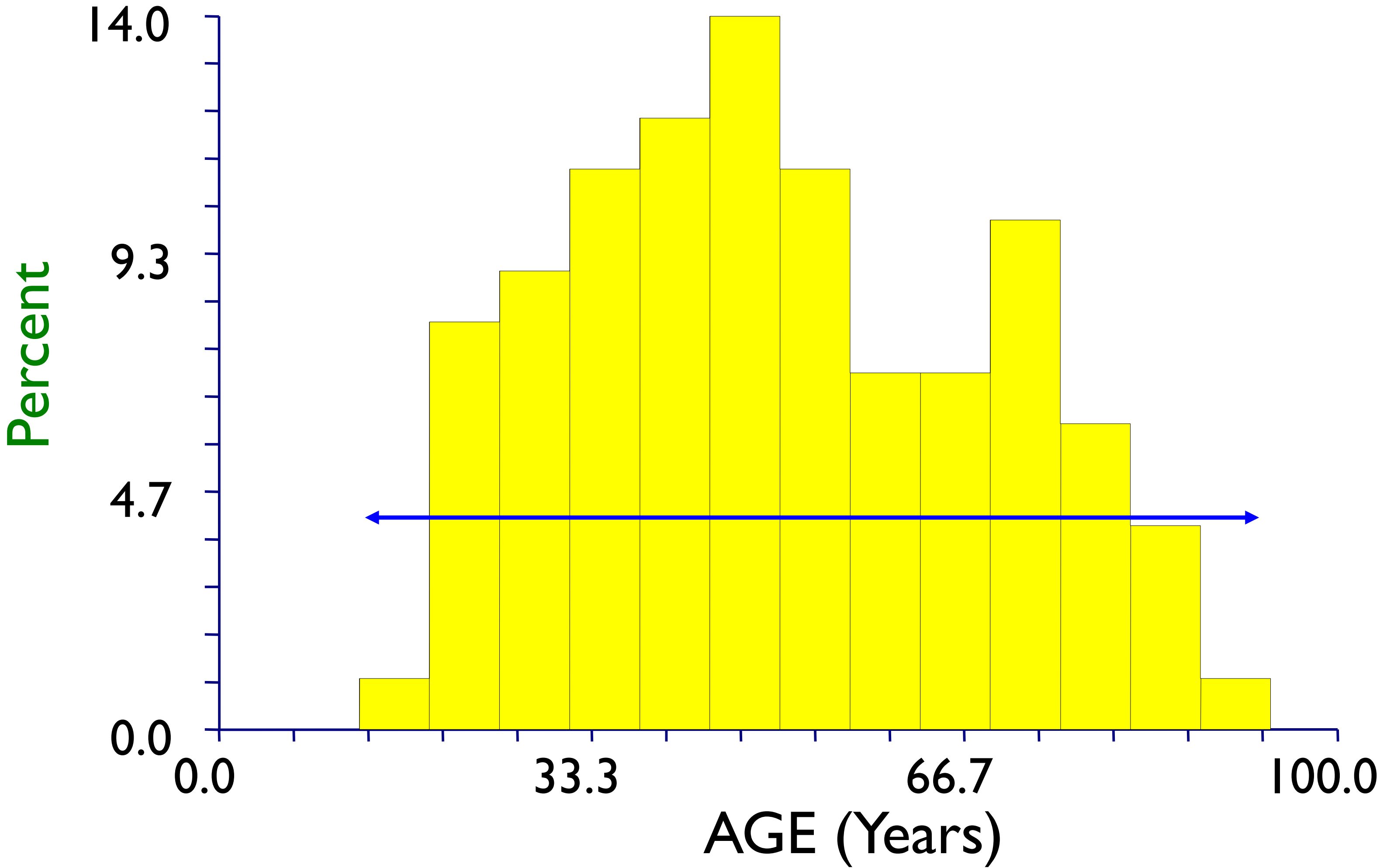
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:



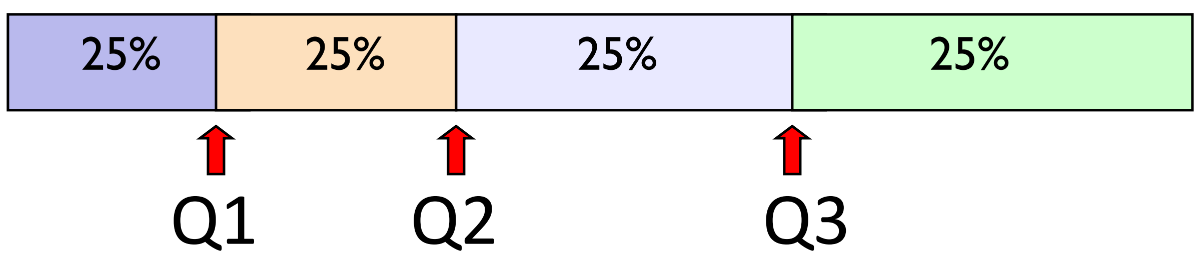
Range = 14 - 1 = 13

Range of age: 94 years-15 years = 79 years



Measures of Dispersion

- Quartiles



- The first quartile（下四分位数）, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile（上四分位数）

Measures of Dispersion

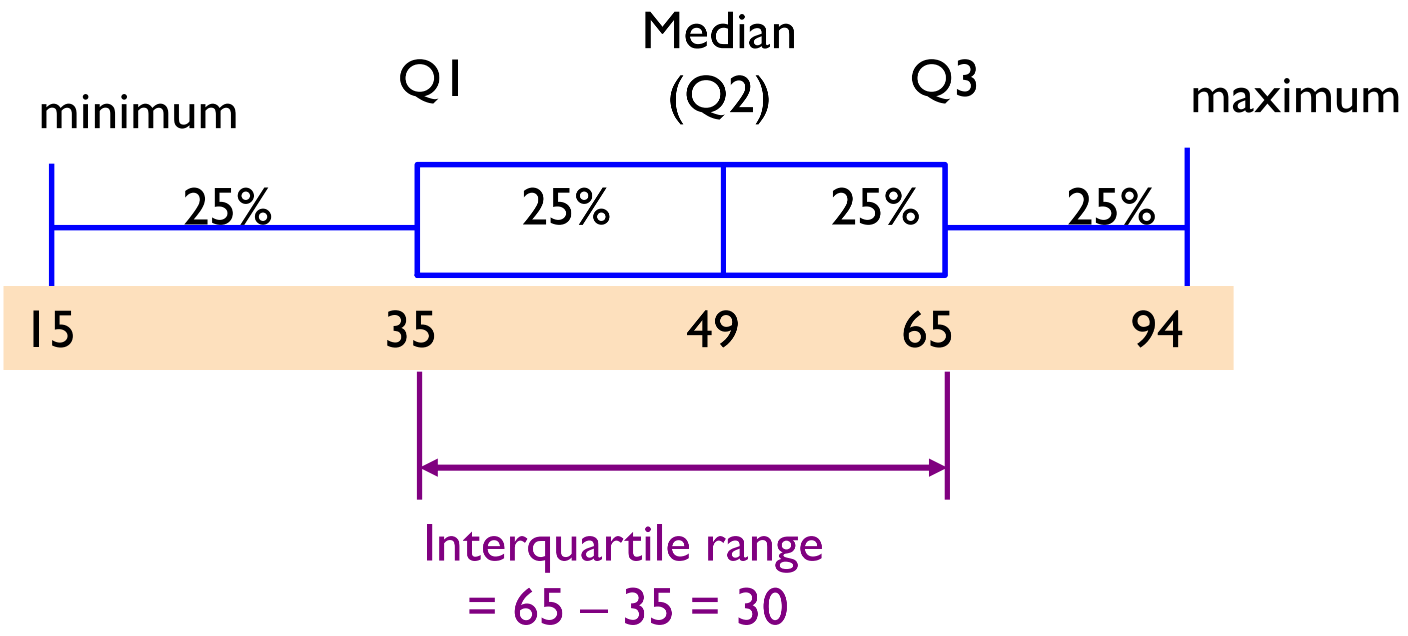
Interquartile Range (四分位数间距)

Interquartile range = 3rd quartile – 1st quartile

$$= Q_3 - Q_1$$

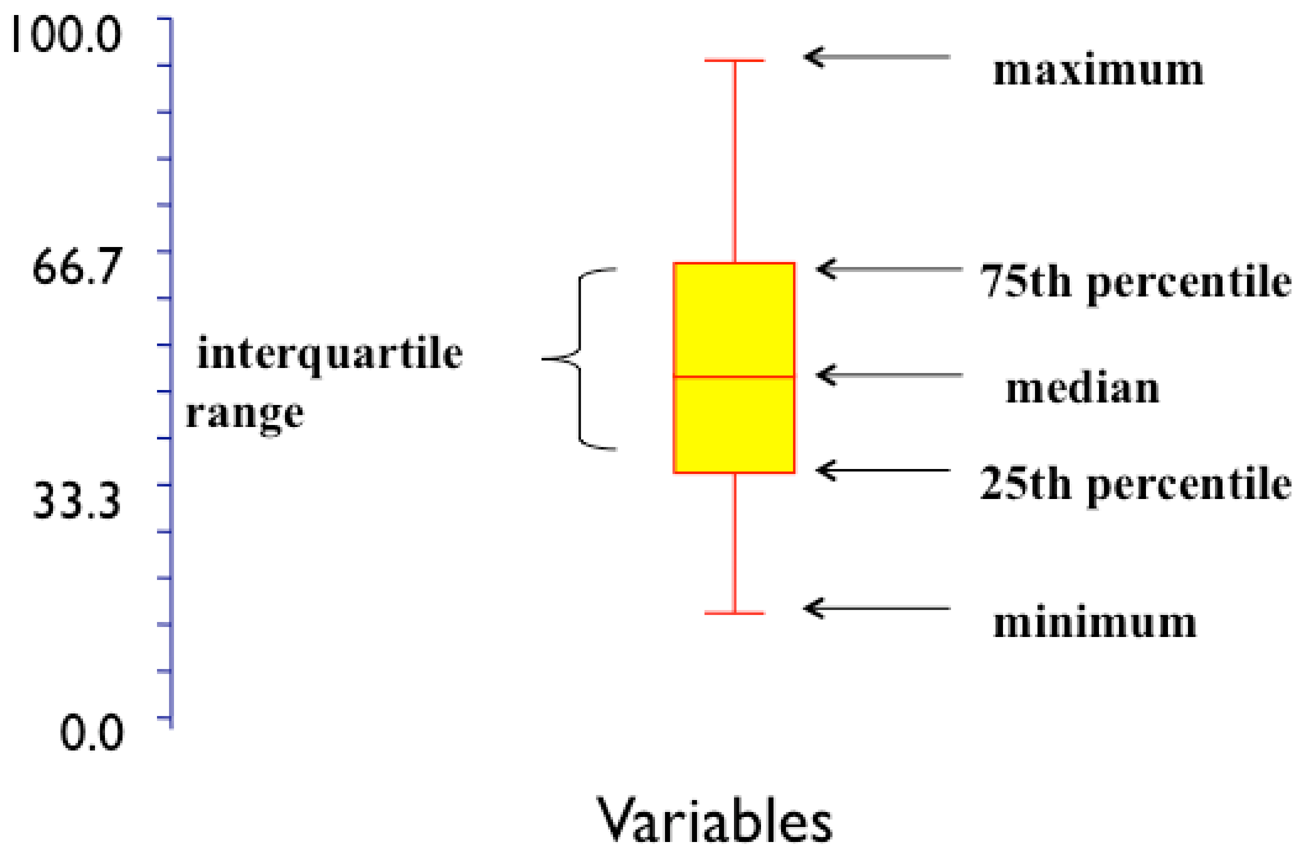
Example

Interquartile Range: age

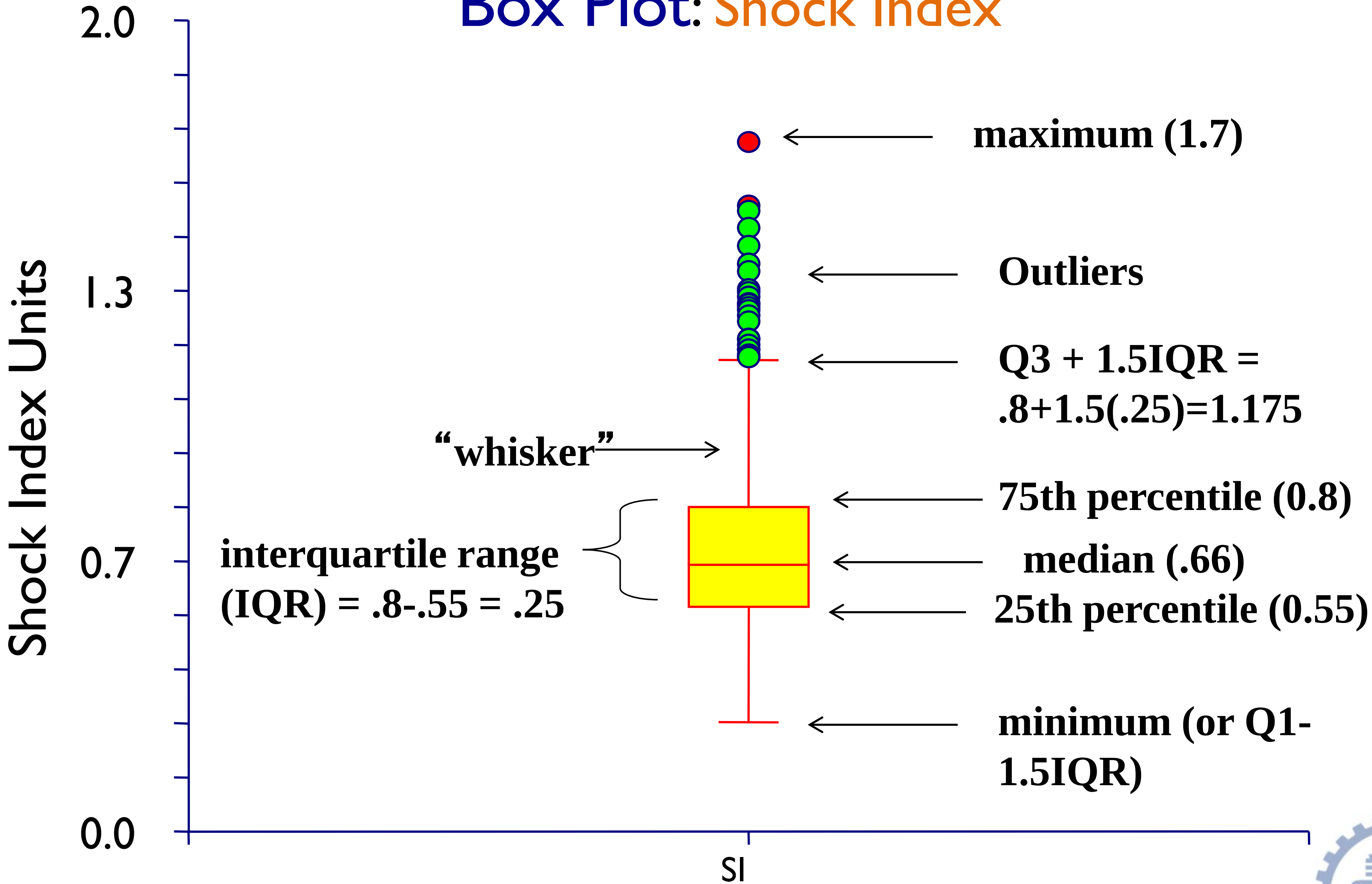


Displaying Data

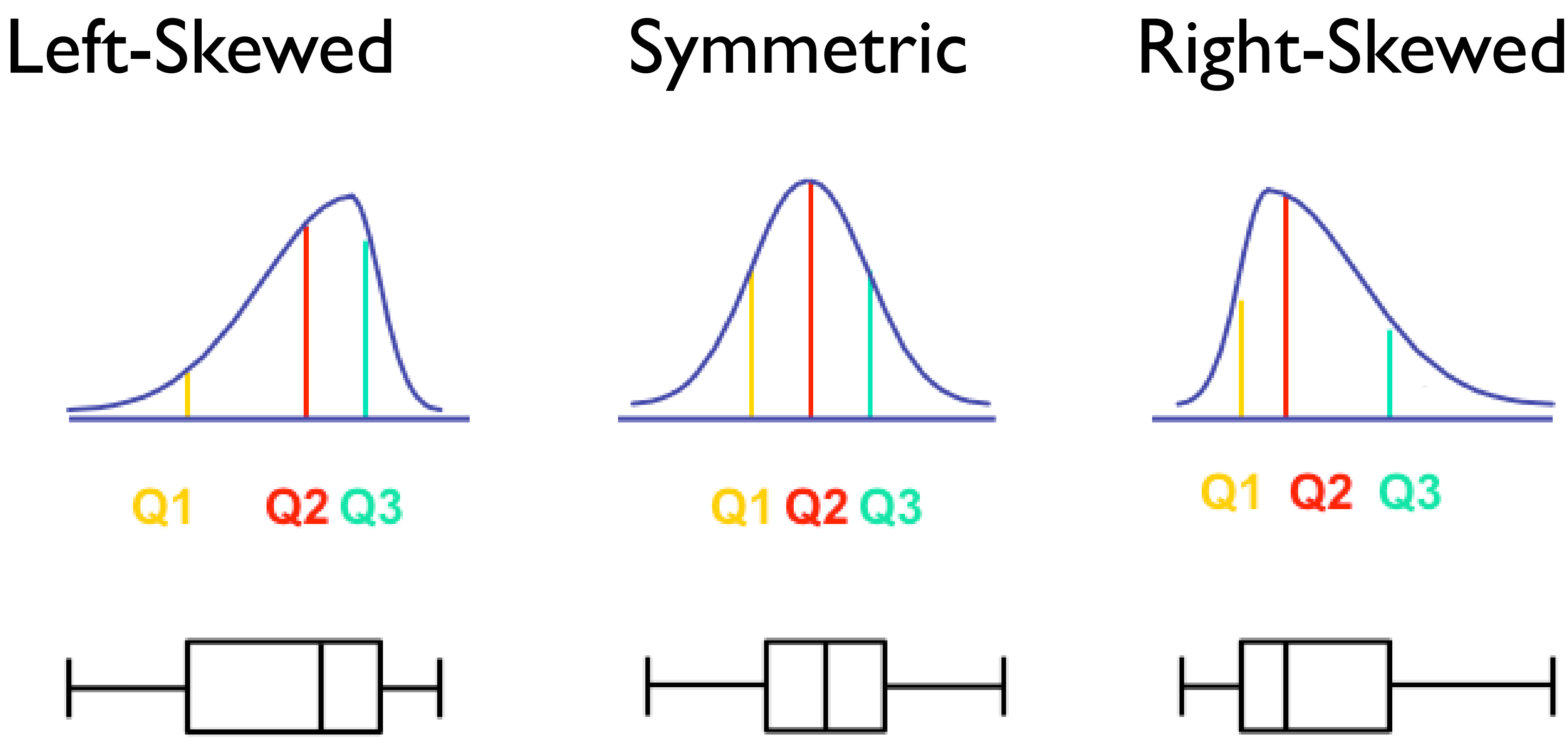
- Boxplot



Box Plot: Shock Index

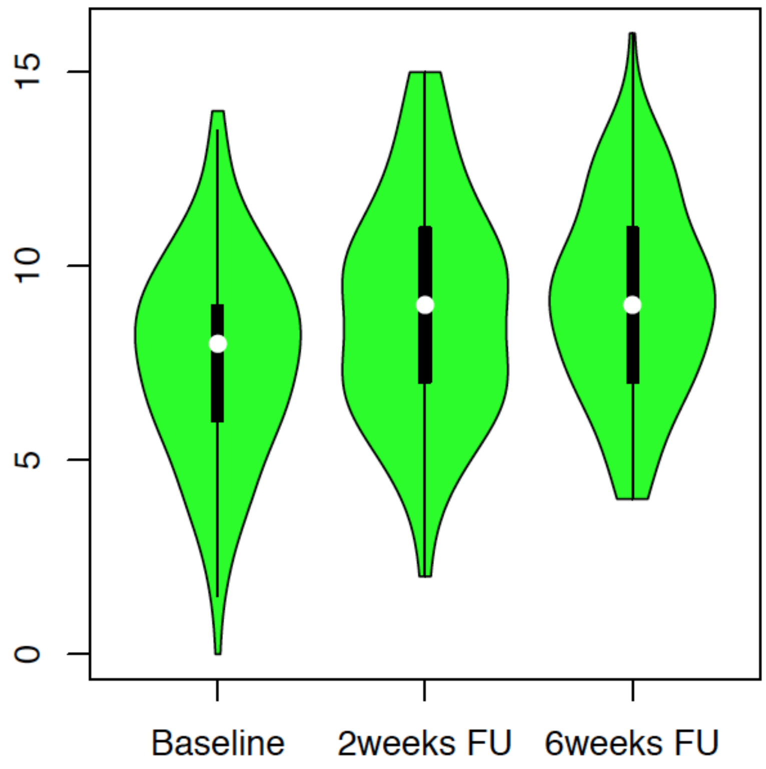


Distribution Shape and Box-and-Whisker Plot



Displaying data

- Violin PLot



Measures of Dispersion

- Variance/standard deviation **(方差/标准差)**

$\sigma^2 = Var(x) = E(x - \mu)^2$

"The expected (or average) squared distance (or deviation) from the mean"

$$\sigma^2 = Var(x) = E[(x-\mu)^2] = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i)$$

Measures of Dispersion

Sample Variance（样本方差）

Average (roughly) of squared deviations of values from the mean

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

Increasing contribution to the variance as you go farther from the mean

Measures of Dispersion

Degrees of freedom (自由度)

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

Degrees of freedom

df is (n-1) rather than n, since only (n-1) of the deviations are independent from each other. The last one can always be calculated from the others because all n of them must add up to zero

Measures of Dispersion

- Sample Standard Deviation (标准差)

✓ Most commonly used measure of variation
✓ Shows variation about the mean
✓ Has the same units as the original data

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

Example

Sample Standard Deviation

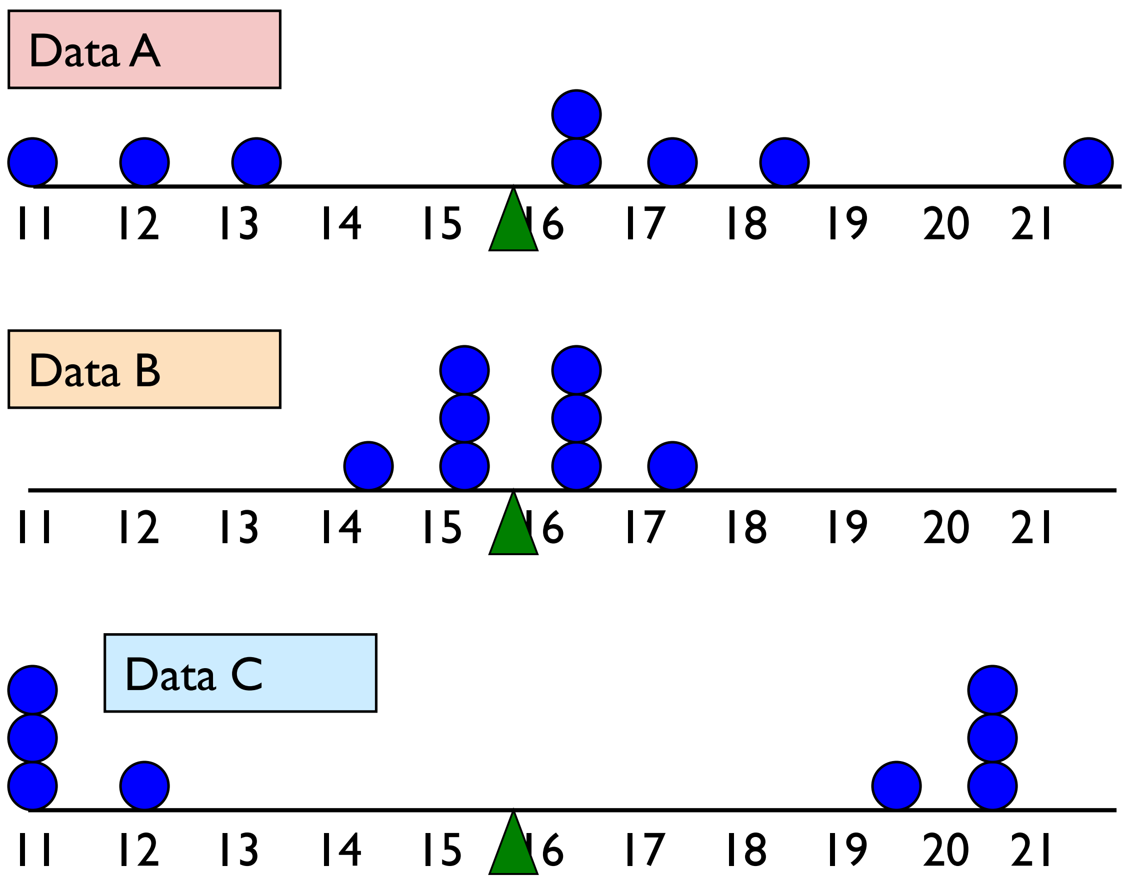
Age data (n=8) : 17 19 21 22 23 23 23 38

n = 8 Mean = $\overline{X}$ = 23.25

$$S = \sqrt{\frac{(17-23.25)^2 + (19-23.25)^2 + \cdots + (38-23.25)^2}{8-1}}$$

$$= \sqrt{\frac{280}{7}} = 6.3$$

Comparing Standard Deviations



Mean = 15.5
S = 3.338

Mean = 15.5
S = 0.926

Mean = 15.5
S = 4.570

Bienaymé-Chebyshev Rule （切比雪夫定律）

- Regardless of how the data are distributed, at least $(1 - 1/k^2)$ of the values will fall within k standard deviations of the mean (for $k > 1$)

Note use of μ (mu) to represent "mean".

Note use of σ (sigma) to represent "standard deviation."

At least	within
$(1 - 1/1^2) =$ 0%	$k=1$ $(\mu \pm 1\sigma)$
$(1 - 1/2^2) =$ 75%	$k=2$ $(\mu \pm 2\sigma)$
$(1 - 1/3^2) =$ 89%	$k=3$ $(\mu \pm 3\sigma)$

Measures of Dispersion

- Coefficient of variation (cv)

$$cv = \frac{s}{\bar{x}} \times 100\%$$

Describing Data

Summary Measures

- Describing Data Numerically
  - Central Tendency
    - Arithmetic Mean
    - Median
    - Mode
  - Dispersion
    - Range
    - Interquartile Range
    - Variance
    - Standard Deviation
    - Coefficient of Variation
  - Shape
    - Skewness

Summary of Symbols

- S^2= Sample variance
- S = Sample standard dev
- σ^2 = Population (true or theoretical) variance
- σ = Population standard dev.
- $\overline{X}$ = Sample mean
- μ = Population mean
- IQR = interquartile range (middle 50%)

Data collection from your classmates

- Each group choose one characteristic or variable.
- Collect the data from your classmates.
- Summary and display your data in homework.

Send your assignment to biostat_sjtu@163.com

Due to 4pm on Sunday