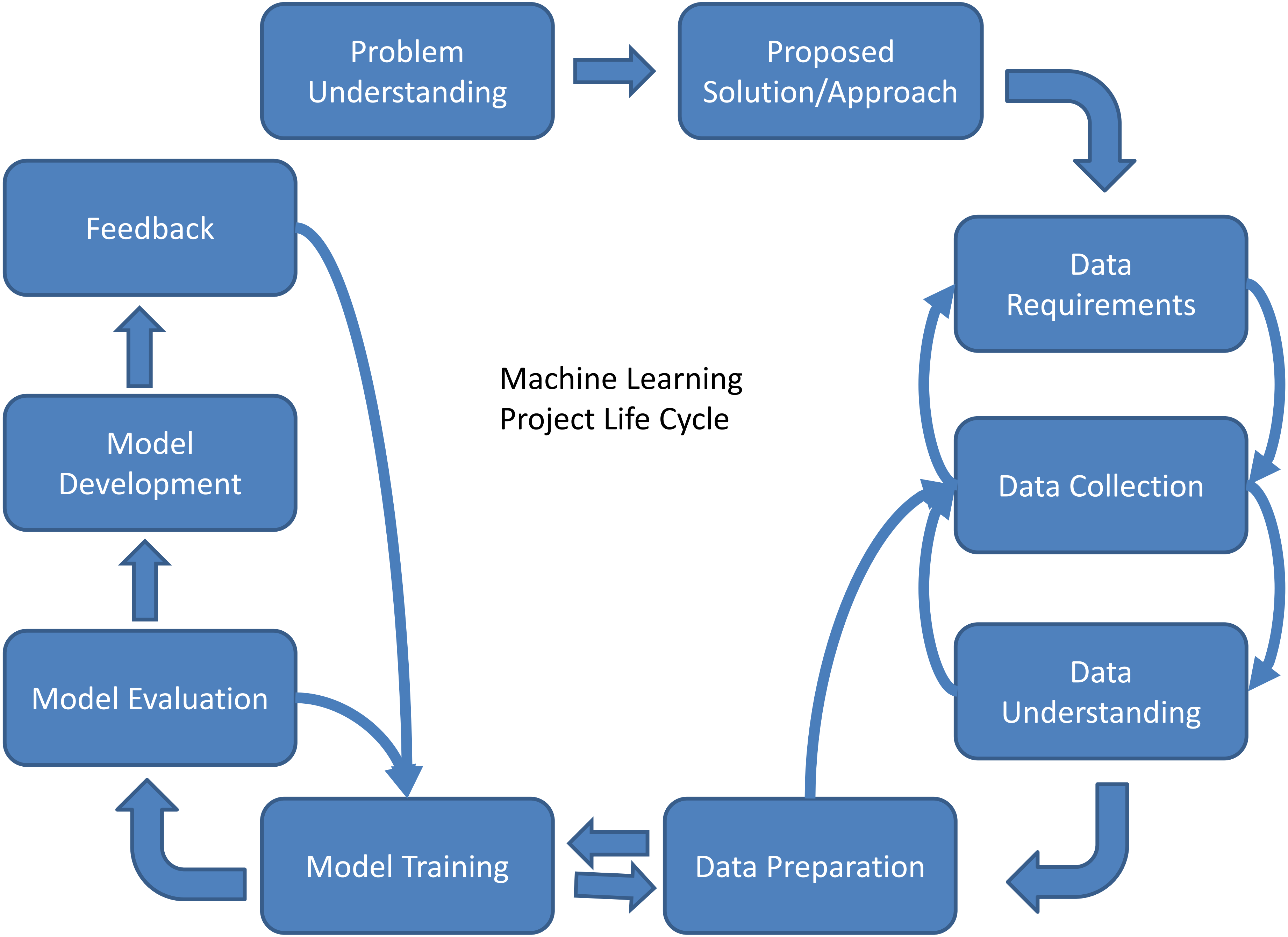
Data Understanding

Measurement & Characterization

Dr Uzair Ahmad

Agenda

- Measures of Central Tendency
- Data Distribution
- Measures of Spread
  - Range
  - Inter Quartile Range
  - Variance & Standard Deviation
- Covariance
- Correlation
- Histograms



Machine Learning
Project Life Cycle

Problem Understanding

Proposed Solution/Approach

Data Requirements

Data Collection

Data Understanding

Data Preparation

Model Training

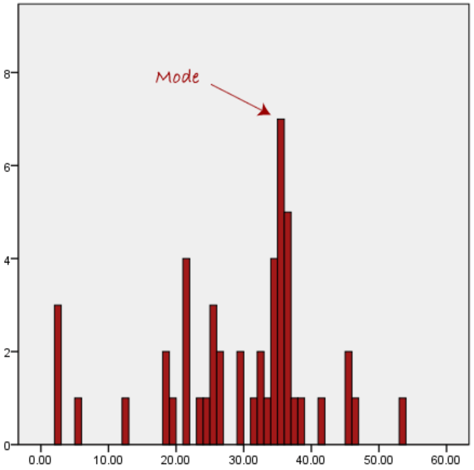
Model Evaluation

Model Development

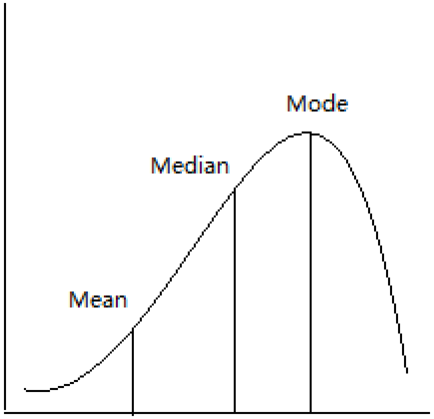
Feedback

Measures of Central Tendency

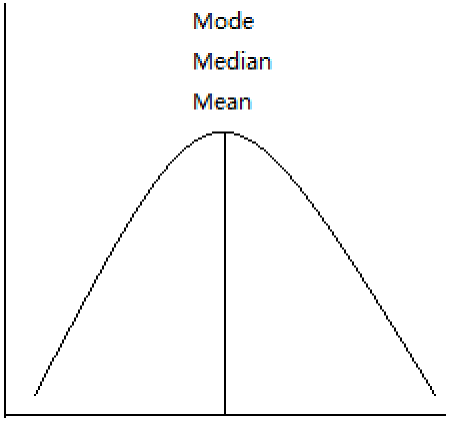
- Mean $$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$
- Median: Middle score for a arranged data
- Mode: The most frequent value in data



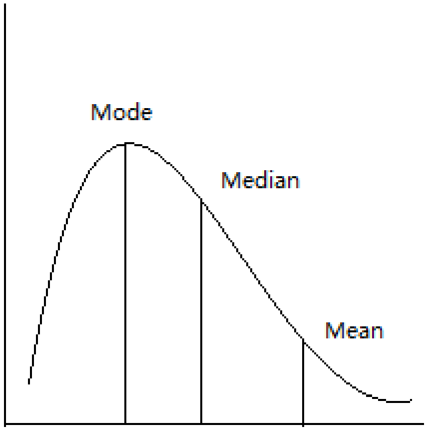
Data Distribution



Left skew



Normal Distribution



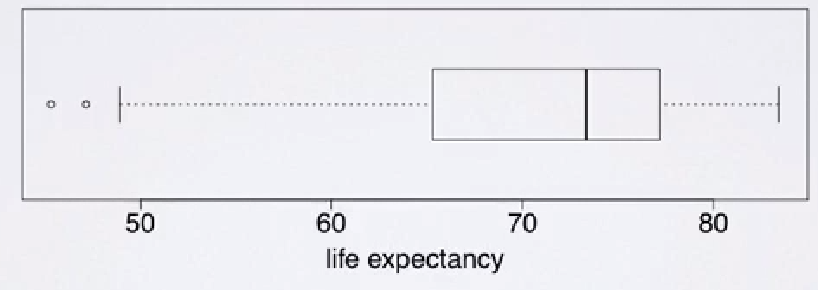
Right skew

Measures of Spread - Range

- Range = Max - Min

Inter Quartile Range

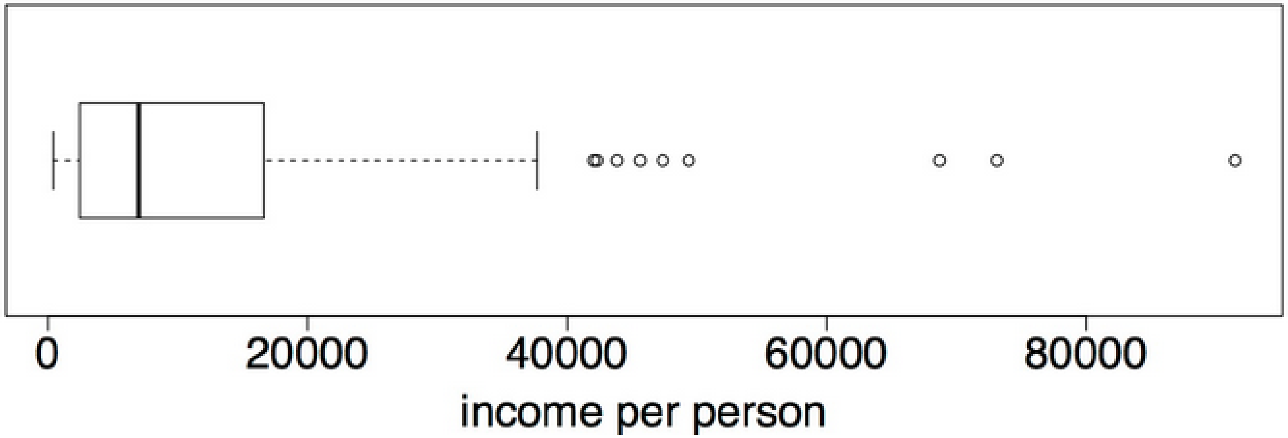
- Box Plot
- Middle Line in the Box: 50% of the Data
- Box: IQT = Q3 – Q1
    - Q3 → 75% of the data
    - Q1 → 25%



- Life Expectancy Data
    - Q1 = 65, Q3 = 77, IQT = 12
- IQT doesn’t rely on end points

Inter Quartile Range

Which of the following is false about the distribution of income per person in countries?



Min. = $403, Q1 = $2438

Median = $6975, Q3 = $16650

Max. = $91490

○ The mean is expected to be greater than the median since the distribution is right skewed.

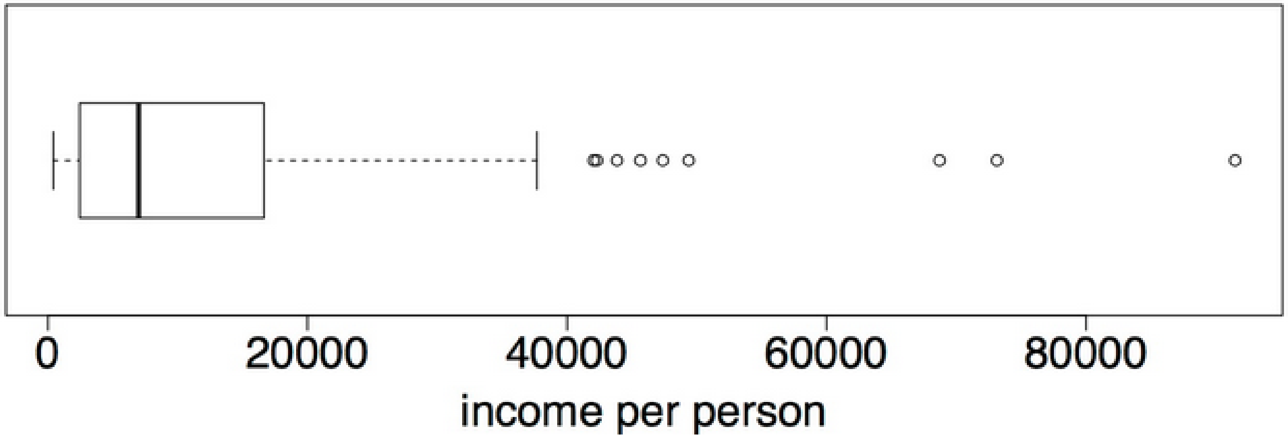
○ IQR is 14212.

○ 25% of the countries have incomes per person below $2438.

○ 75% of the countries have incomes per person above $16650.

Inter Quartile Range

Which of the following is false about the distribution of income per person in countries?



Min. = $403, Q1 = $2438

Median = $6975, Q3 = $16650

Max. = $91490

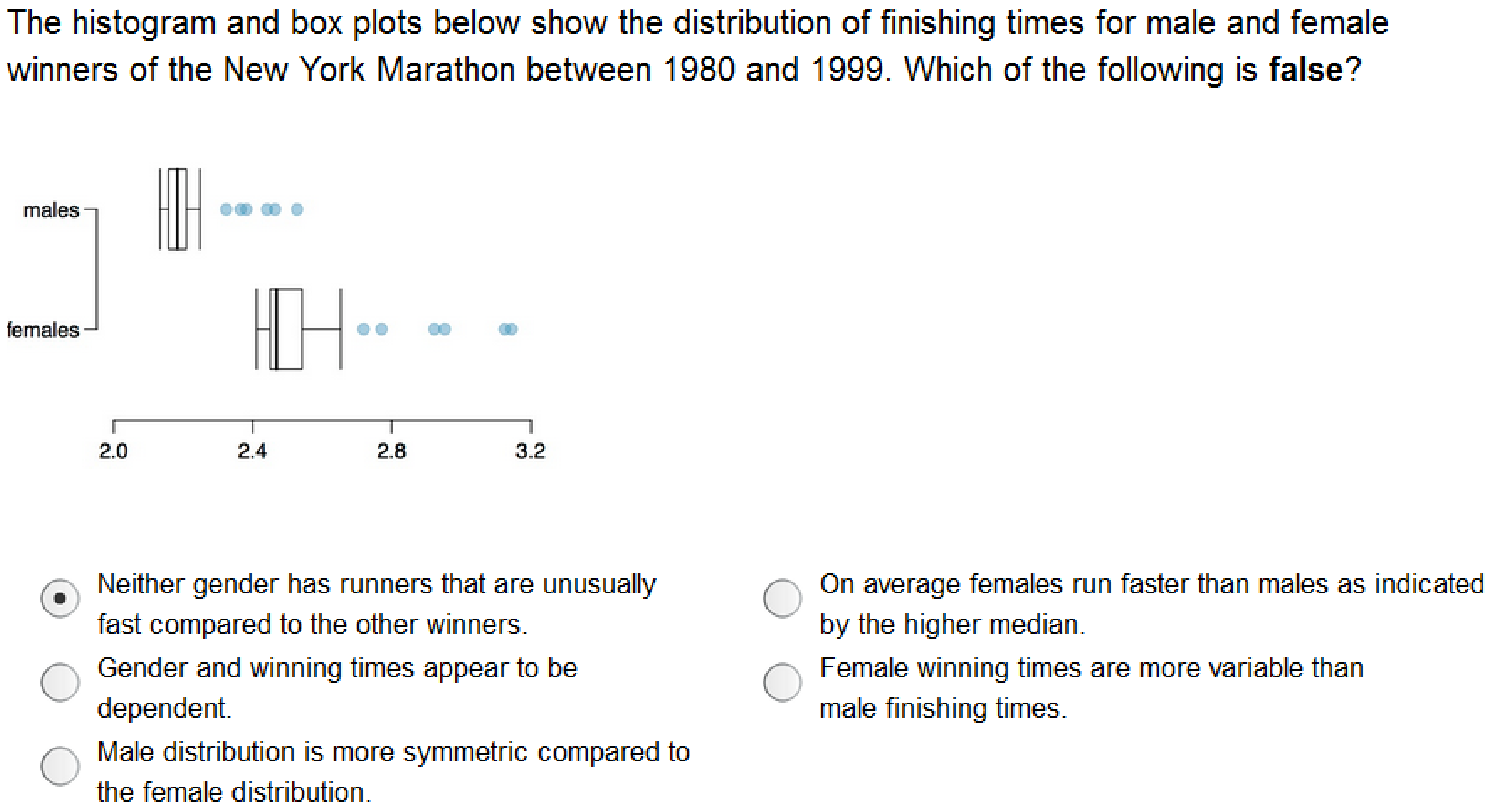
○ The mean is expected to be greater than the median since the distribution is right skewed.

○ IQR is 14212.

○ 25% of the countries have incomes per person below $2438.

🚫 75% of the countries have incomes per person above $16650.

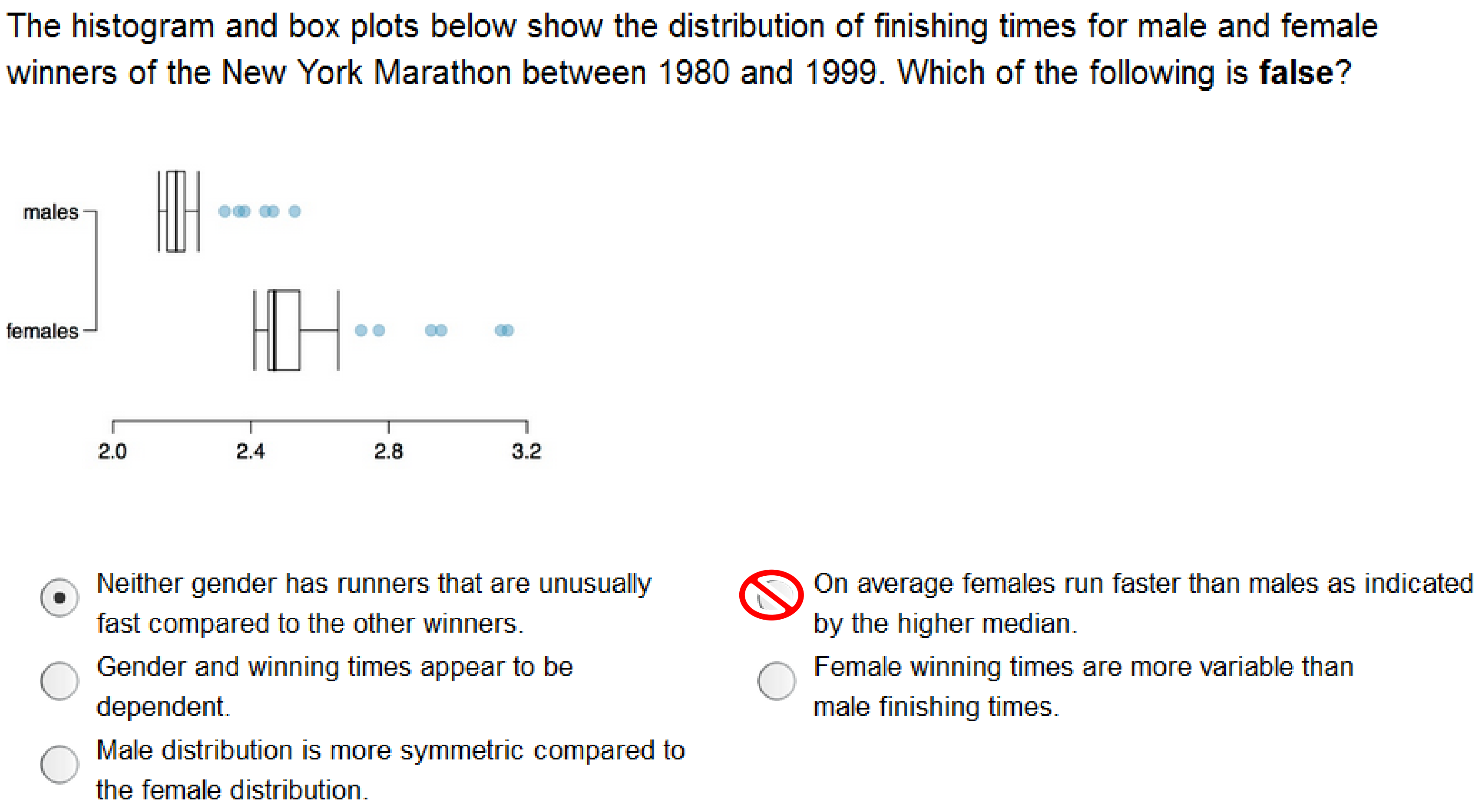
Inter Quartile Range



The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1980 and 1999. Which of the following is **false**?

- ● Neither gender has runners that are unusually fast compared to the other winners.
- ○ Gender and winning times appear to be dependent.
- ○ Male distribution is more symmetric compared to the female distribution.
- ○ On average females run faster than males as indicated by the higher median.
- ○ Female winning times are more variable than male finishing times.

Inter Quartile Range



The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1980 and 1999. Which of the following is **false**?

● Neither gender has runners that are unusually fast compared to the other winners.

○ Gender and winning times appear to be dependent.

○ Male distribution is more symmetric compared to the female distribution.

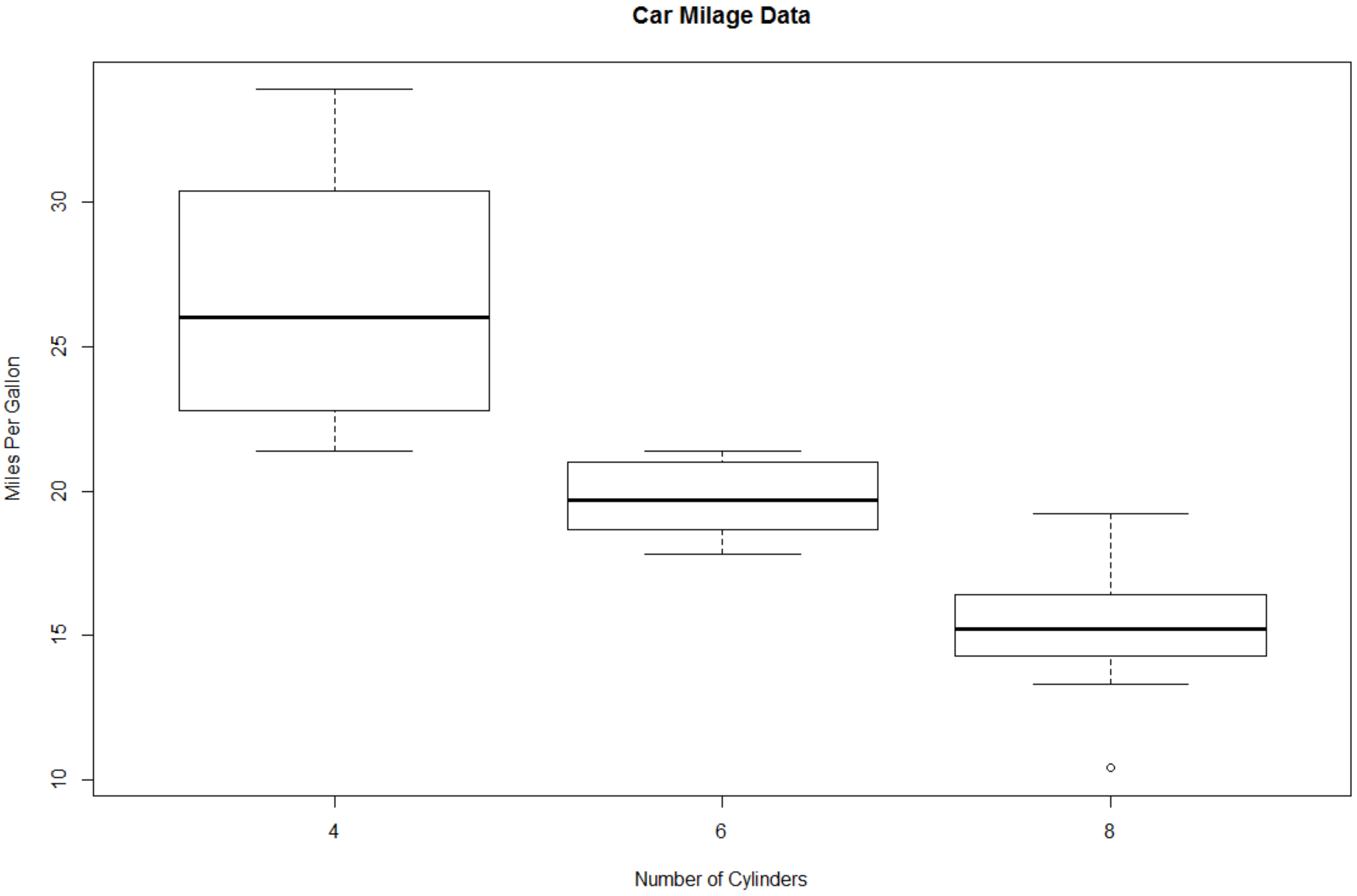
⊘ On average females run faster than males as indicated by the higher median.

○ Female winning times are more variable than male finishing times.

Creating Box Plots

- boxplot(mpg~cyl,data=mtcars, main="Car Milage Data", xlab="Number of Cylinders", ylab="Miles Per Gallon")

Creating Box Plots



Measures of Spread - Variance

- Mean: A numerical representation of data

Assignment	Score X	Score Y
1	3	7
2	5	7
3	7	7
4	10	7
5	10	7
Mean	7	7

- Same mean for different data sets?

Variance

- How to represent data spread ?

Assignment	Score X	Score Y
1	3	7
2	5	7
3	7	7
4	10	7
5	10	7
Mean	7	7

Variance

- The average deviation, or difference, of the values from the mean. ?

$$\frac{\sum(x-\bar{X})}{N}$$

Assignment	Score	$x-\bar{X}$
1	3	3-7=-4
2	5	5-7=-2
3	7	7-7=0
4	10	10-7=3
5	10	10-7=3
$\bar{X}$	7	0

- But… Its always going to be Zero …

Variance

- The average of absolute differences of the values from the mean?

$$\frac{\sum \left| x - \bar{X} \right|}{N}$$

Assignment	Score	$\left\| x - \bar{X} \right\|$
1	3	3-7 = 4
2	5	5-7 = 2
3	7	7-7 = 0
4	10	10-7 = 3
5	10	10-7 = 3
$\bar{X}$	7	12

- But… It does not support further inferential formulas …

Variance

- The average of squared differences of the values from the mean.

$$\sigma^2 = \frac{\sum \left(x - \overline{X}\right)^2}{N}$$

Assignment	Score	$x - \overline{X}$	$\left(x - \overline{X}\right)^2$
1	3	3-7 = -4	16
2	5	5-7 = -2	4
3	7	7-7 = 0	0
4	10	10-7 = 3	9
5	10	10-7 = 3	9
$\overline{X}$	7		
σ^2			38/5 = 7.6

Variance

Dive	X	Y
1	28	27
2	22	27
3	21	28
4	26	6
5	18	27
Mean	23	23
Variance	12.8	72.4

Variance

- How to represent data spread ?

Dive	X	Y
1	28	27
2	22	27
3	21	28
4	26	6
5	18	27
Mean	23	23
Variance	12.8	72.4

Outlier

- What does lower values of variance mean ?

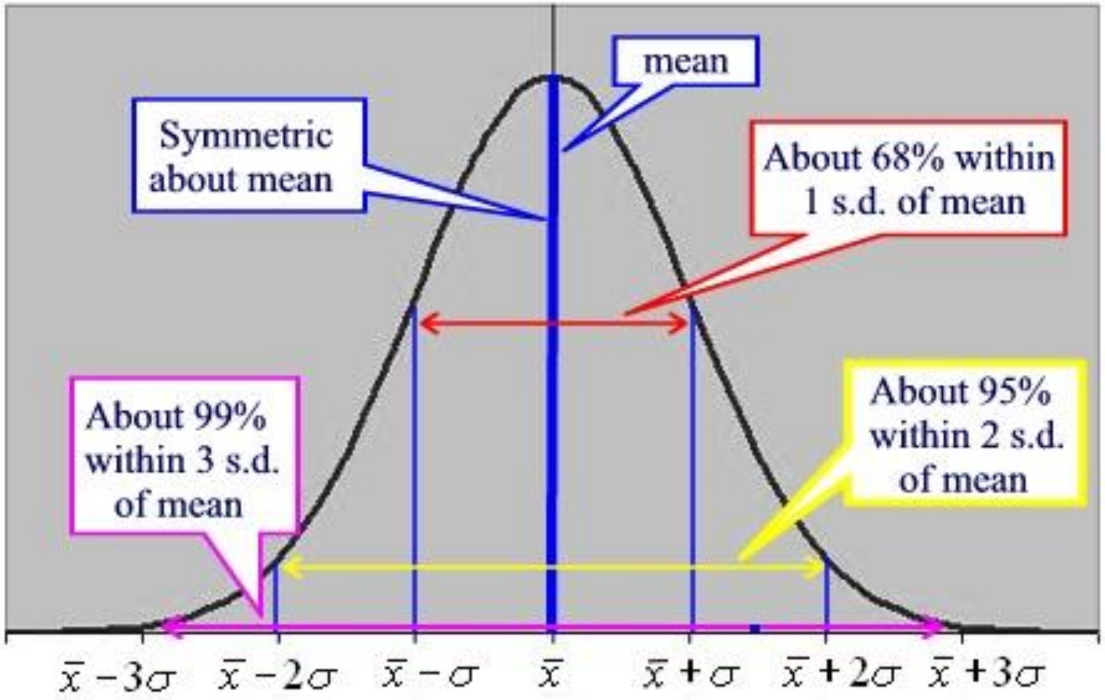
Standard Deviation

- Can think of standard deviation as the **average distance to the mean**,
  - although that's not numerically accurate, it's conceptually helpful.
- All ways of saying the same thing:
  - higher standard deviation indicates higher spread, less consistency, and less clustering.

Population $$\sigma = \sqrt{\frac{\sum\left(x - \overline{X}\right)^2}{N}}$$

Sample $$s = \sqrt{\frac{\sum\left(x - \overline{X}\right)^2}{n-1}}$$

Standard Deviation



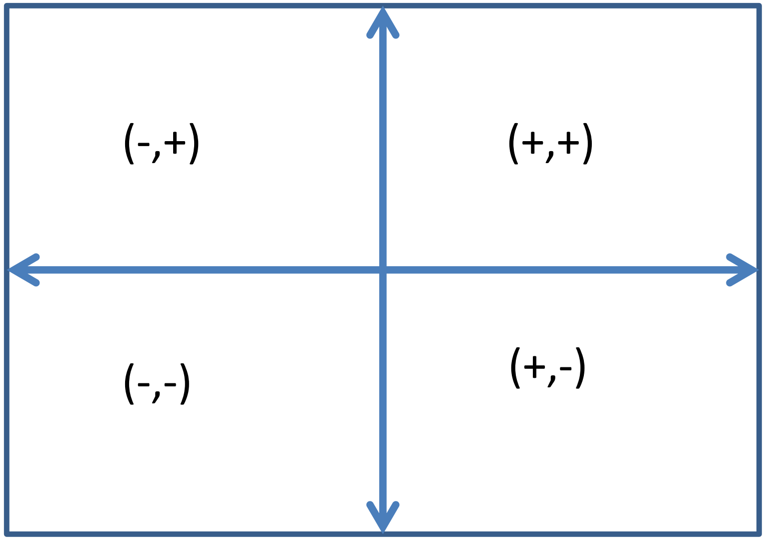
<u>empirical rule for data (68-95-99)</u> -
only applies to a set of data having a distribution that is approximately bell-shaped

CoVariance

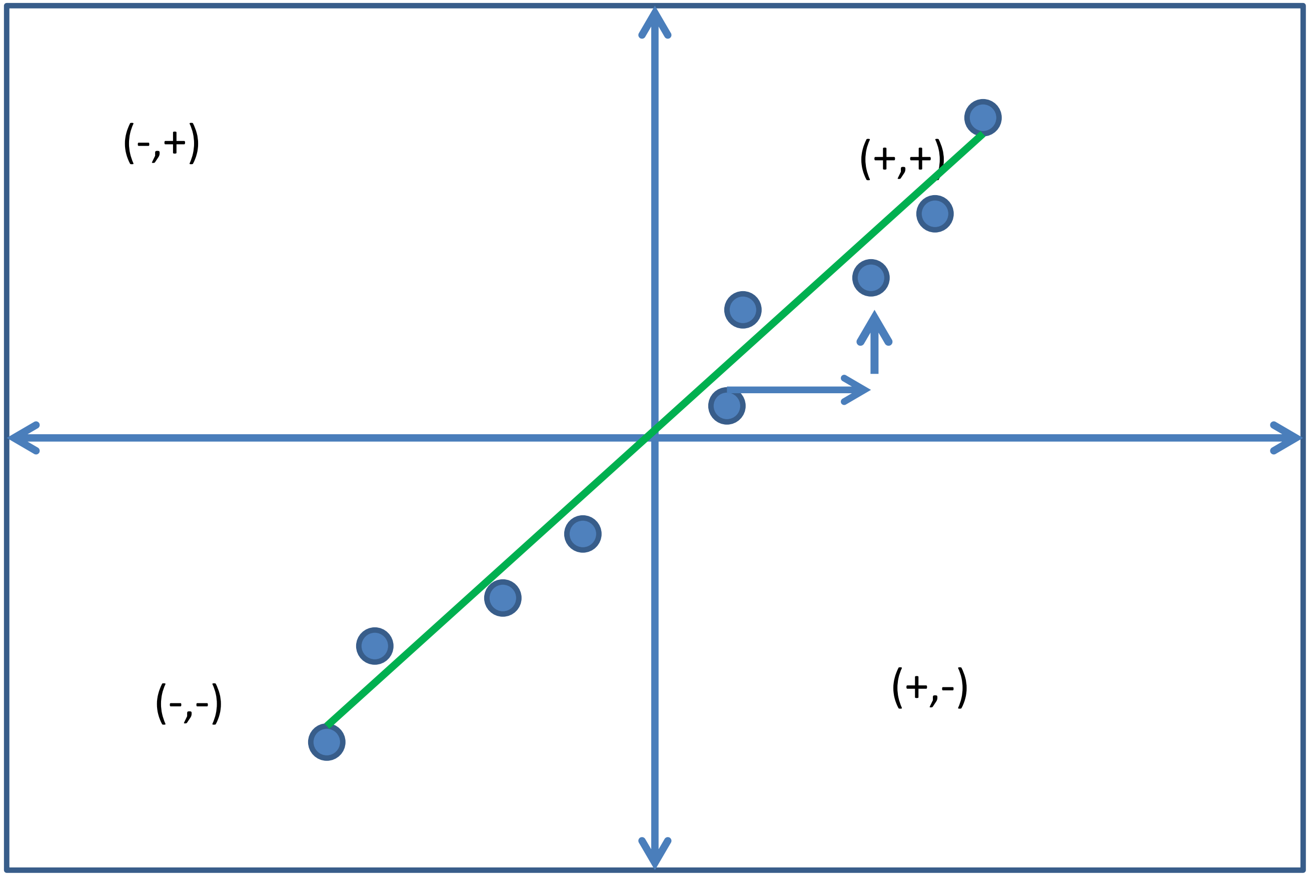
- Variables may change in relation to each other
  - How variables relate in pairs?
- *Covariance* measures how much the movement in one variable predicts the movement in a corresponding variable
- Bivariate Distribution
  - Relationship between two variables

CoVariance

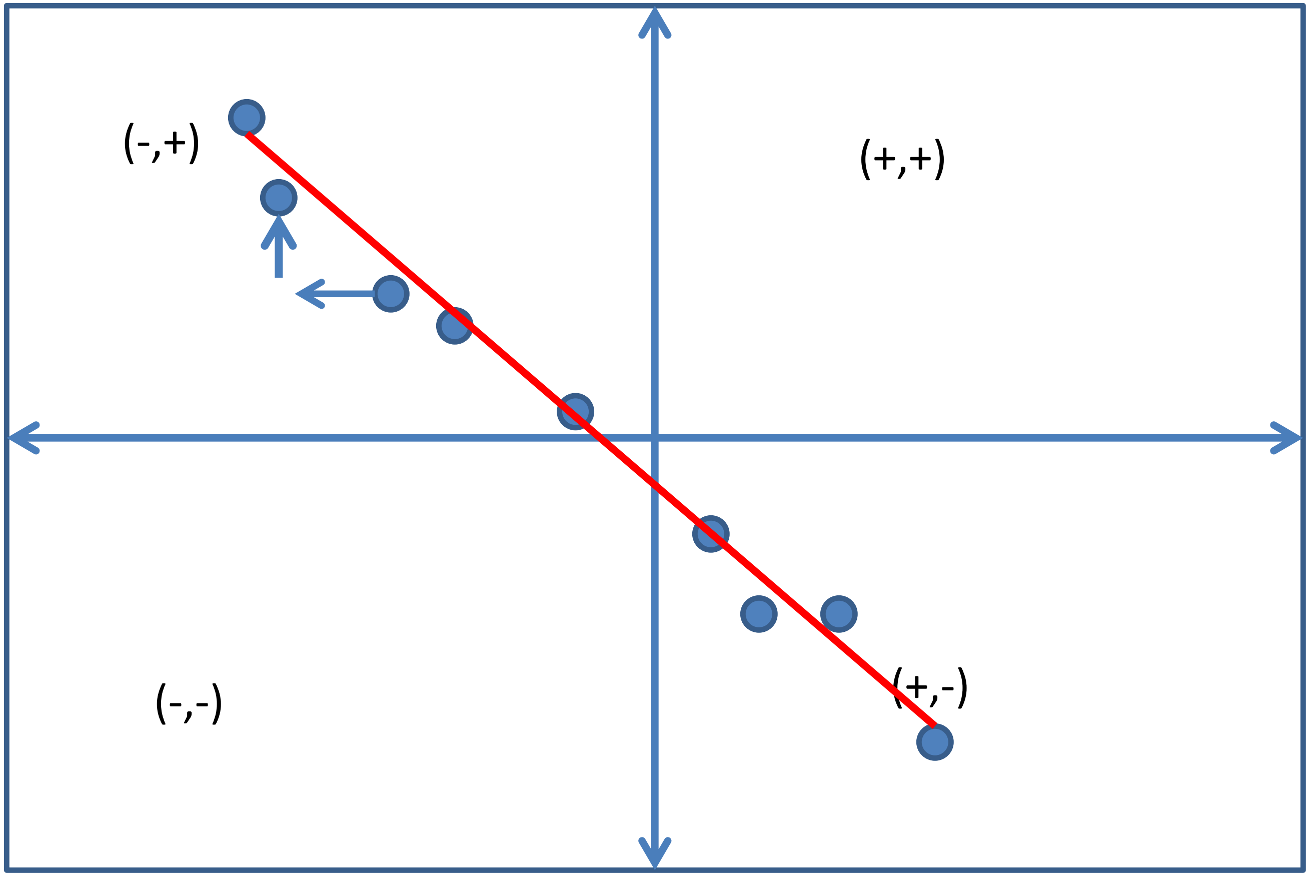
- A descriptive measure of linear association
- Direction of relationship
  - Positive: One Moves Up/Down, the other moves Down/Up
  - Negative: One Moves Up/Down, the other Moves Up/Down



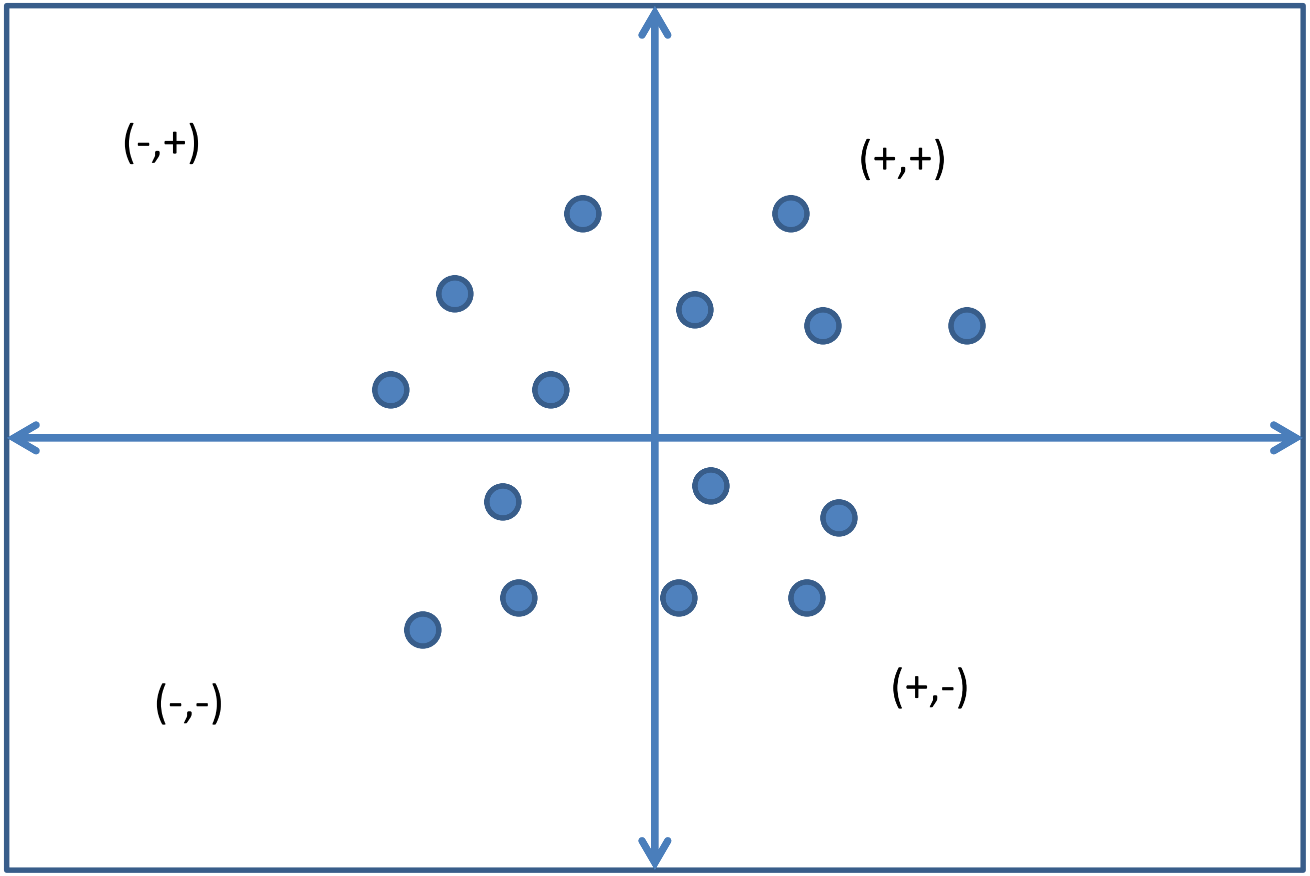
CoVariance



CoVariance



CoVariance



No Linear Relationship:

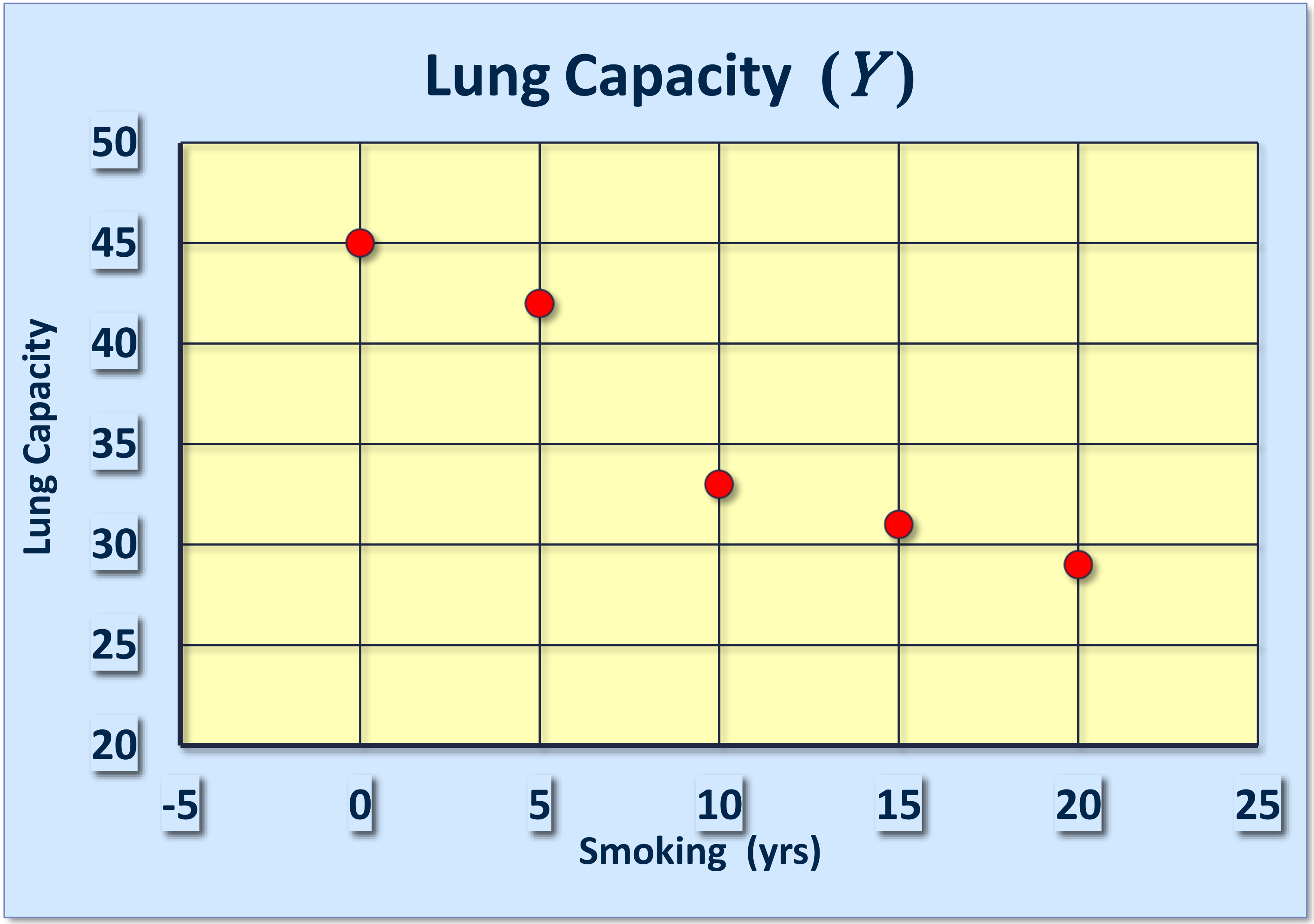
Example Bivariate Distribution

- Example: investigate relationship between *cigarette smoking* and *lung capacity*
- Data: sample group response data on smoking habits, *and* measured lung capacities, respectively

Example Bivariate Distribution

N	**Cigarettes (X)**	**Lung Capacity (Y)**
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

Example Bivariate Distribution



Example Bivariate Distribution

- Observe that as smoking exposure goes up, corresponding lung capacity goes down
  - Variables *covary* inversely
  - A negative linear relationship
- *Covariance* quantifies relationship of two variables

Covariance

- Variables that *covary* inversely, like smoking and lung capacity, tend to appear on opposite sides of the group means
  - When smoking is above its group mean, lung capacity tends to be below its group mean.
- Average *product of deviation* measures extent to which variables covary, the degree of linkage between them

The Sample Covariance

- Similar to variance, for theoretical reasons, average is typically computed using (N-1), not N. Thus,

$$S_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right)$$

Calculating Covariance

Cigs (X)	**Lung Cap (Y)**
0	45
5	42
10	33
15	31
20	29

$\bar{X} = 10$	$\bar{Y} = 36$

Calculating Covariance

Cigs (X)	$(X-\bar{X})$	$(X-\bar{X})(Y-\bar{Y})$	$(Y-\bar{Y})$	Cap (Y)
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29

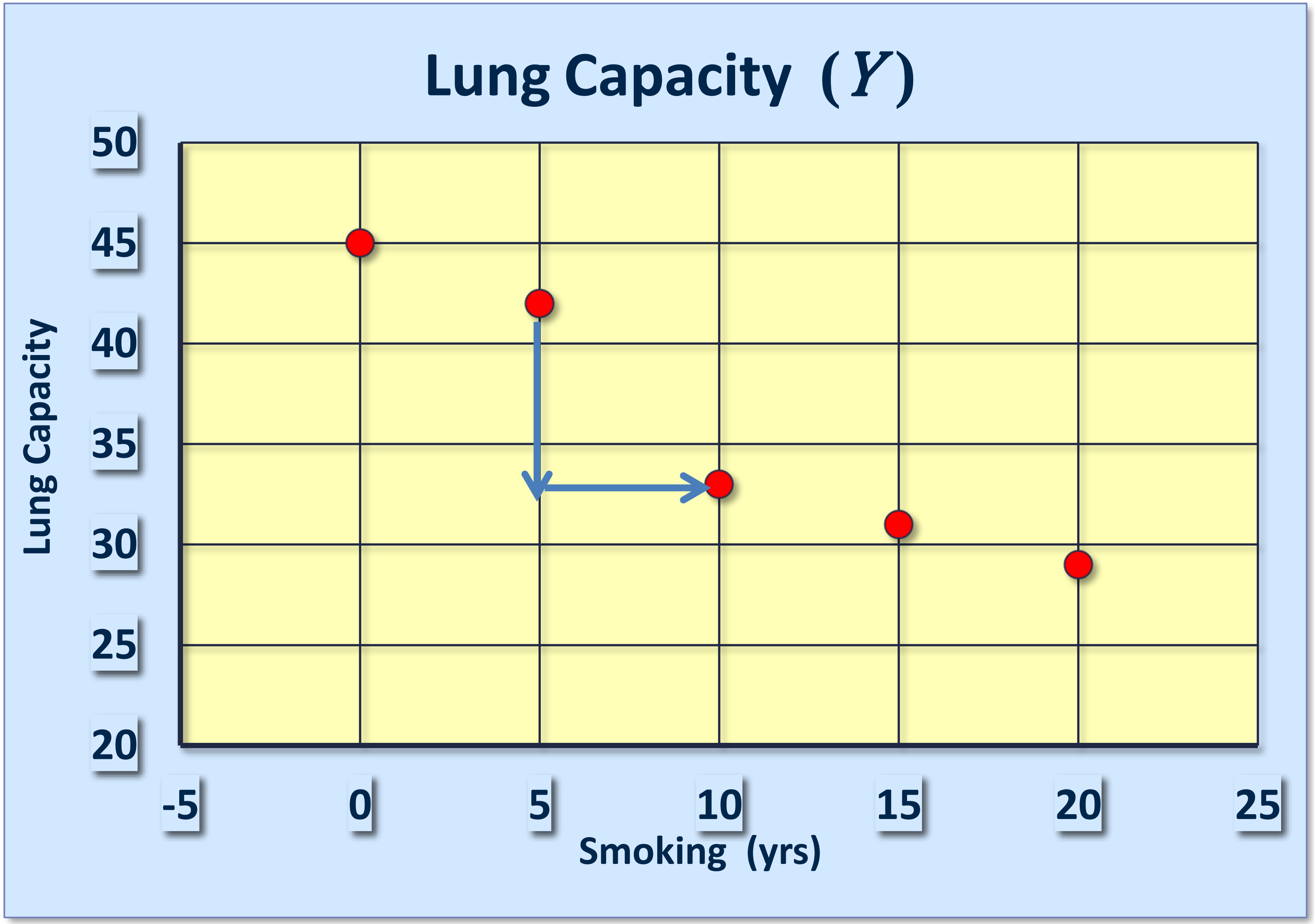
$\sum$= -215

Calculating Covariance

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

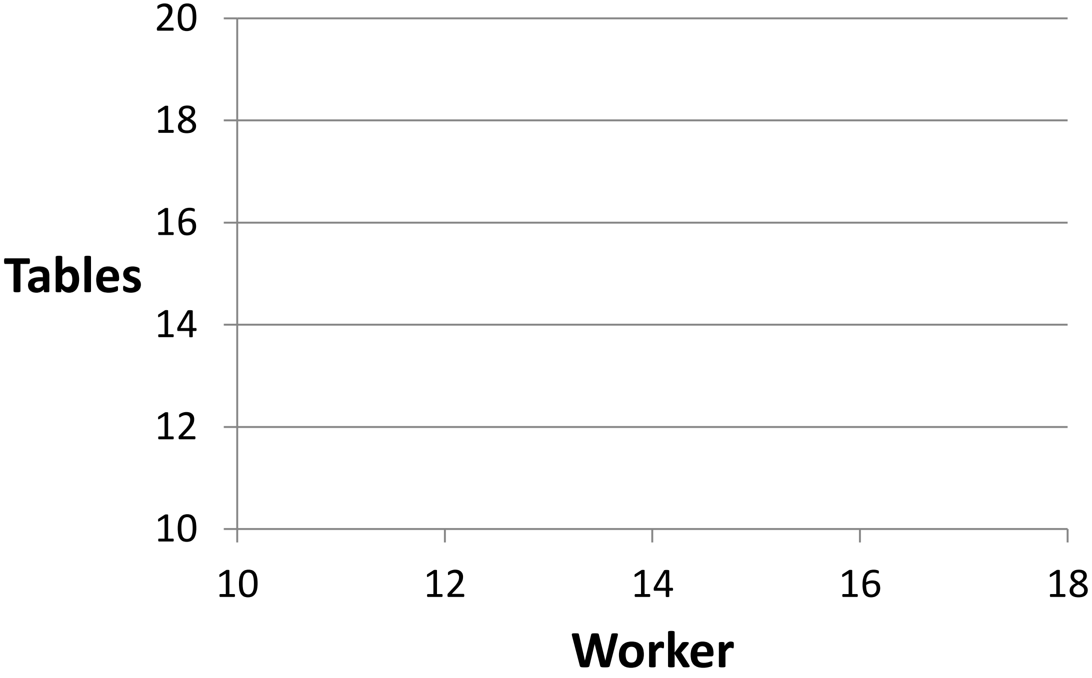
$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

Example Bivariate Distribution



Another Example

Workers	Tables
12	20
30	60
15	27
24	50
14	21
18	30
28	61
26	54
19	32
27	57



What relationship do you see ?

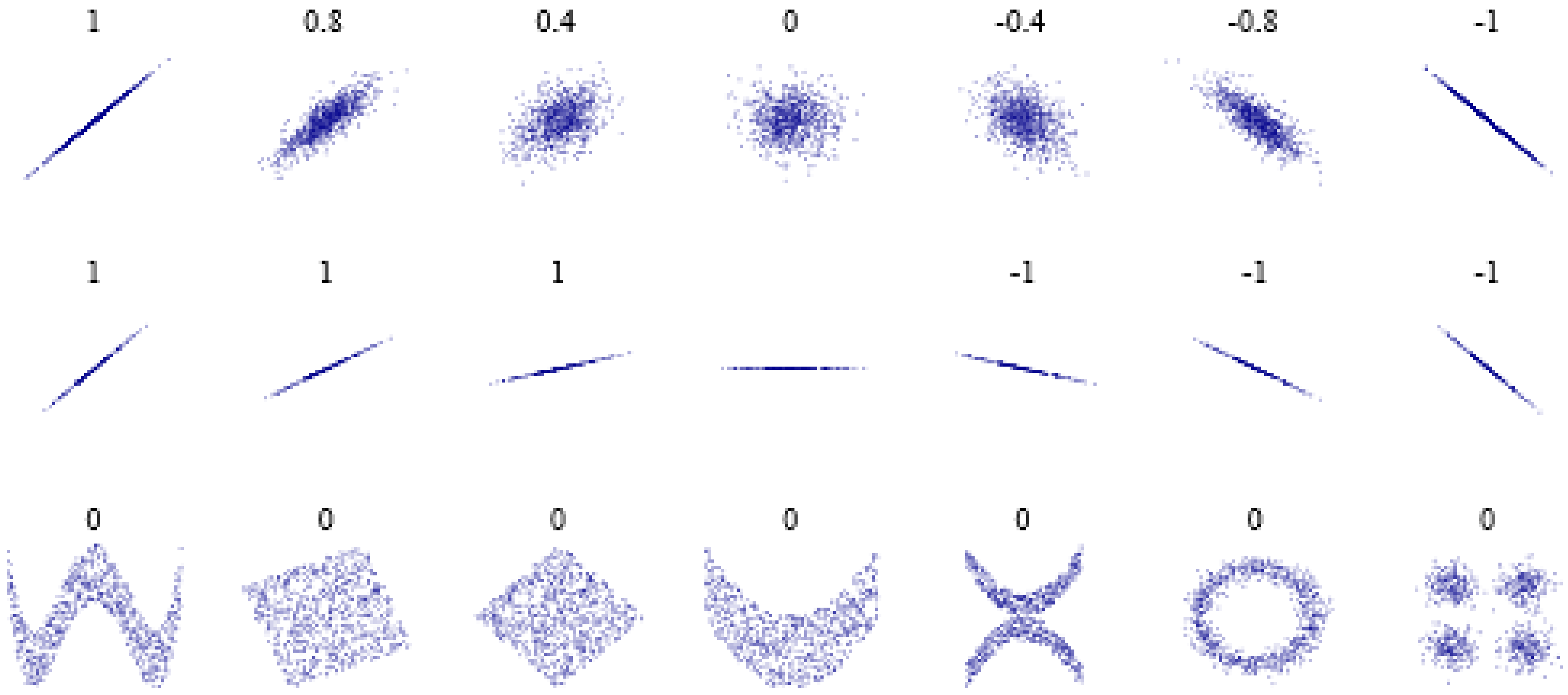
Correlation

- A cousin of Covariance
- Covariance provides direction of relationship
- Correlation provides (direction + strength) of relationship
- Covariance has no lower or upper bounds
  - It depends of the scales of the variables
- Correlation is always between -1 and +1
  - Independent of the scale of the variables
- Covariance not standardized
  - Correlation standardized measure

Correlation

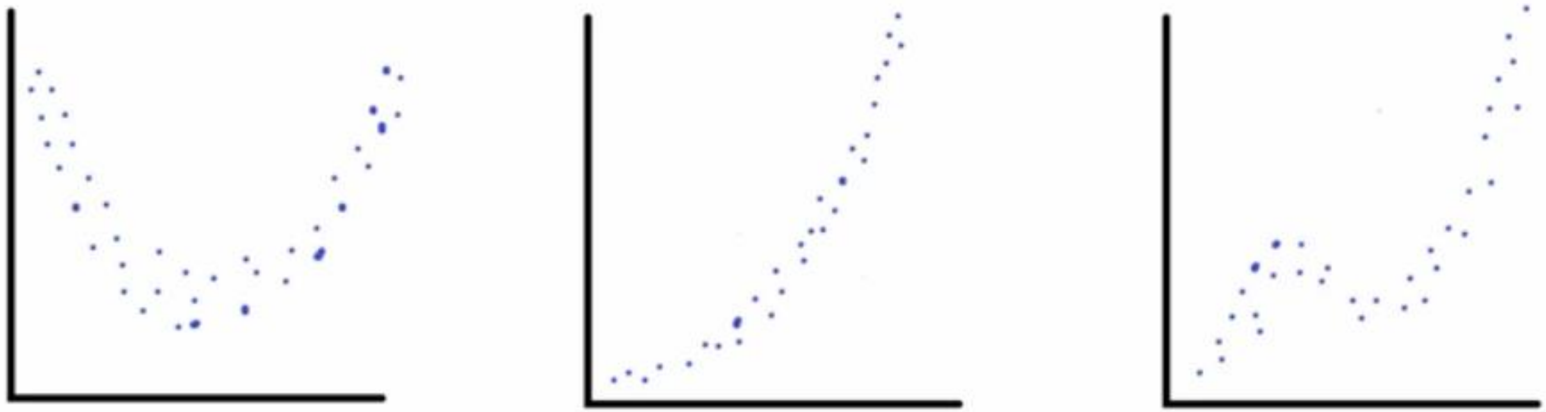
- Visualize your data before calculating Correlation
  - Scatter plot
- Correlation is only applicable of linear relationships
- Correlation does not mean causation
- Correlation strength does not imply its statistically significance

Correlation



Correlation

Non-linear relationships



Quadratic

Exponential

Polynomial

Correlation Formula

- Pearson Correlation Coefficient

$$r = \frac{Cov(x, y)}{S_x S_y}$$

Calculating Correlation

	Workers X	Tables Y	$(x-\bar{X})$	$(y-\bar{Y})$	$(x-\bar{X})(y-\bar{Y})$
	12	20	-9.3	-21.2	197.16
	30	60	8.7	18.8	163.56
	15	27	-6.3	-14.2	89.46
	24	50	2.7	8.8	23.76
	14	21	-7.3	-20.2	147.46
	18	30	-3.3	-11.2	36.96
	28	61	6.7	19.8	132.66
	26	54	4.7	12.8	60.16
	19	32	-2.3	-9.2	21.16
	27	57	5.7	15.8	90.06
Mean	21.3	41.2			962.4
STD	6.4816	16.685		Cov(X,Y)	106.9333333
				Correlation	0.98877256

Does a relationship Exists ?

$$if \Rightarrow |r| \geq \frac{2}{\sqrt{n}}$$

Number of samples n

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

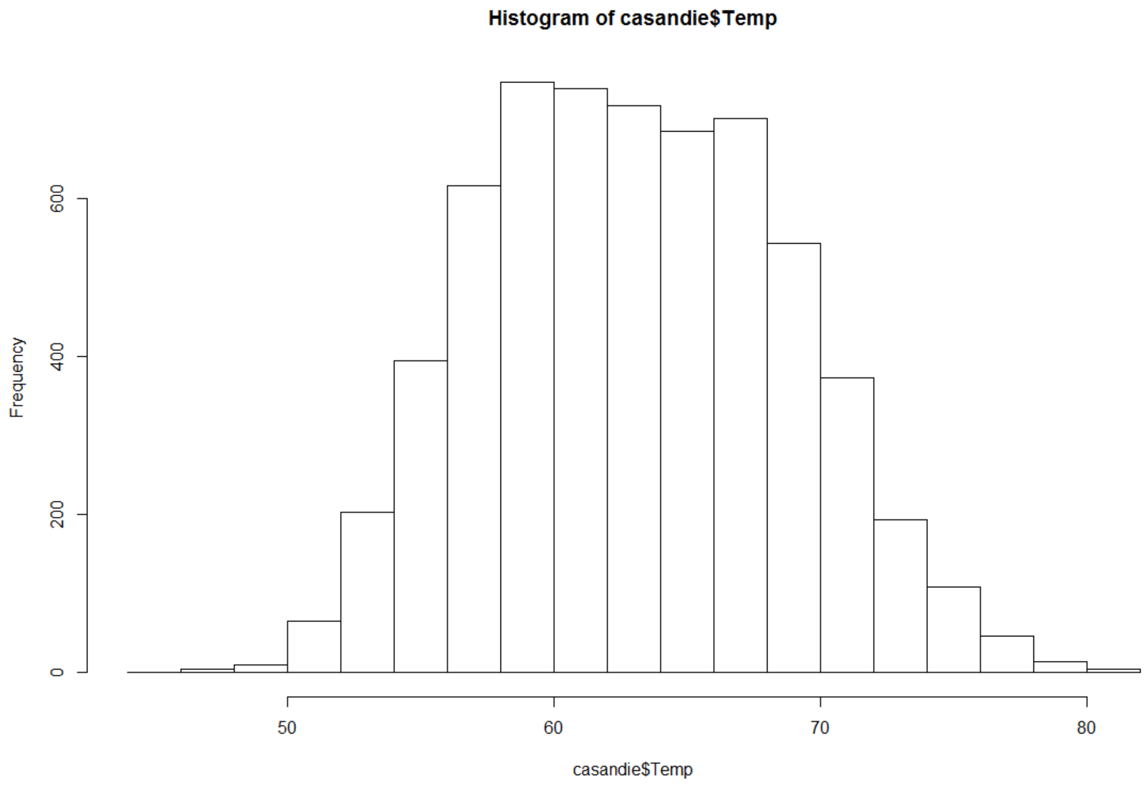
$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Histograms – Univariate

- R > hist(casandie$Temp)

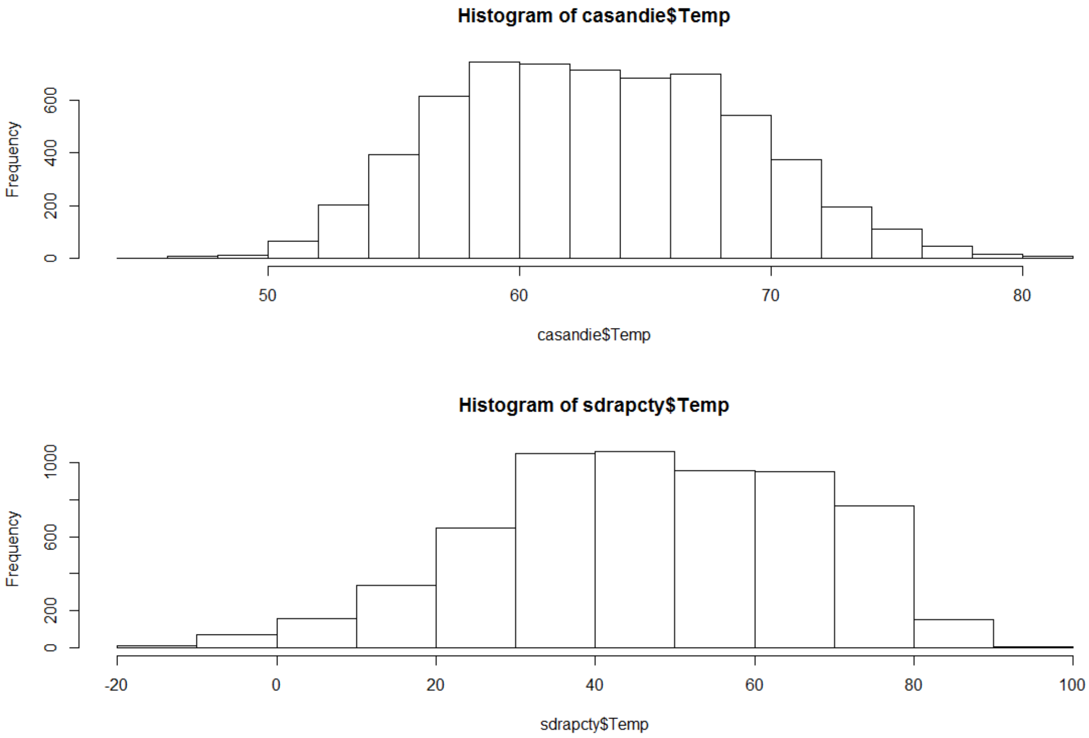


http://jgscott.github.io/STA371H_Spring2016/data.html

Histograms - Bivariate

- R Script

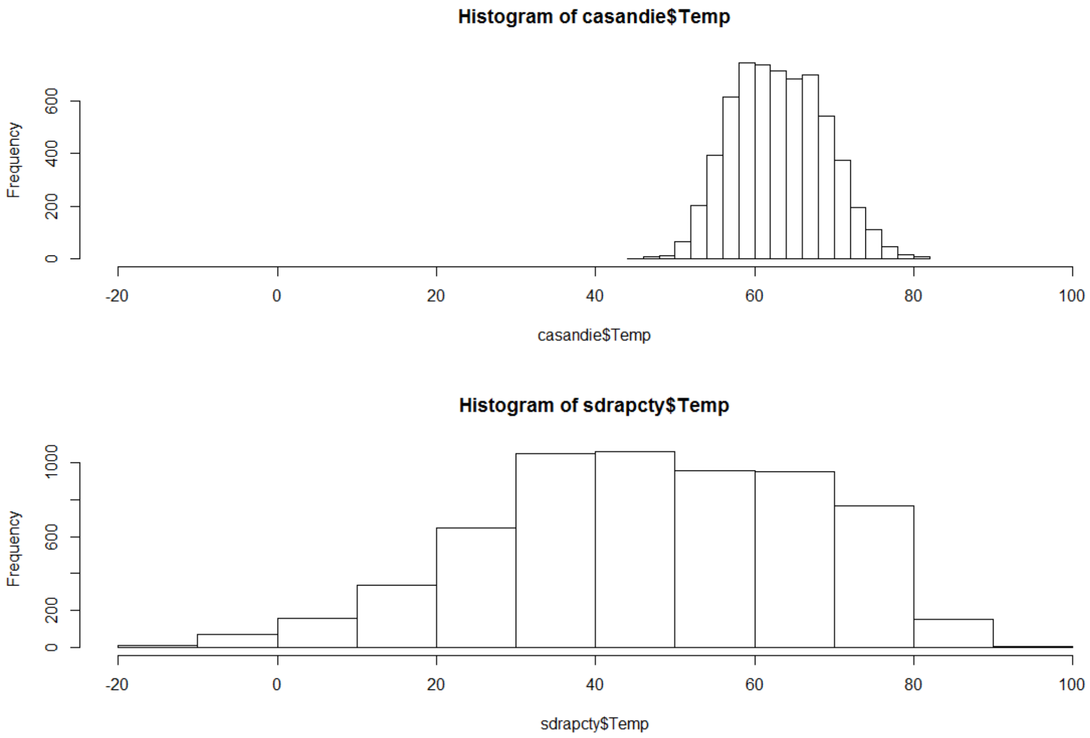
```
> par(mfrow=c(2,1))
> hist(casandie$Temp)
> hist(sdrapcty$Temp)
```



Comparability ?

Histograms - Bivariate

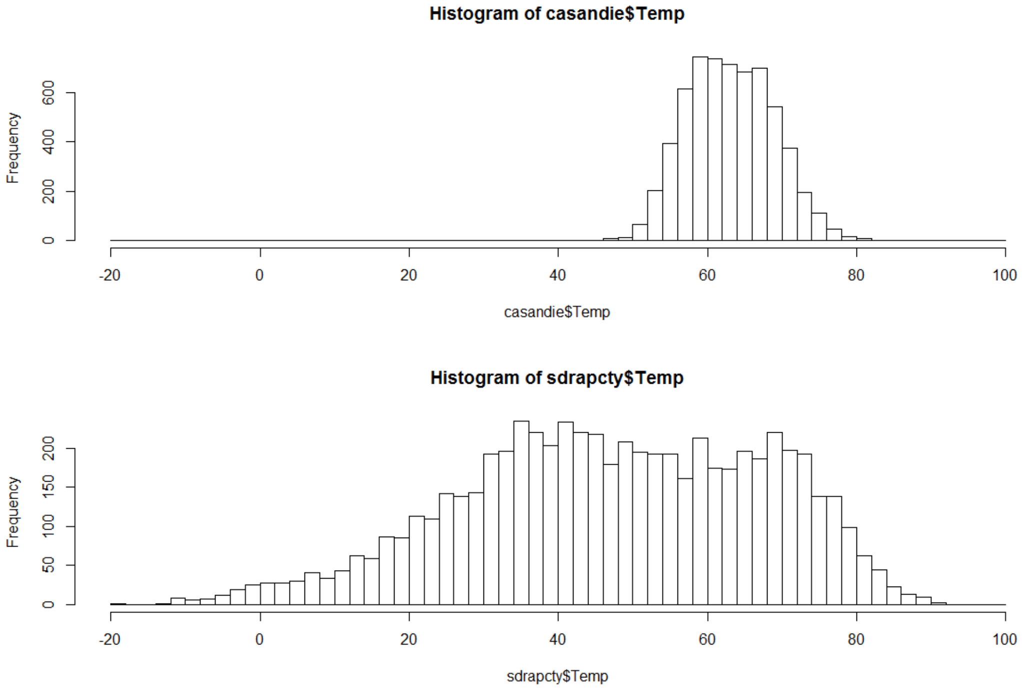
```
par(mfrow=c(2,1))
hist(casandie$Temp, xlim = c(-20,100))
hist(sdrapcty$Temp, xlim = c(-20,100))
```



Comparability ?

Histograms - Bivariate

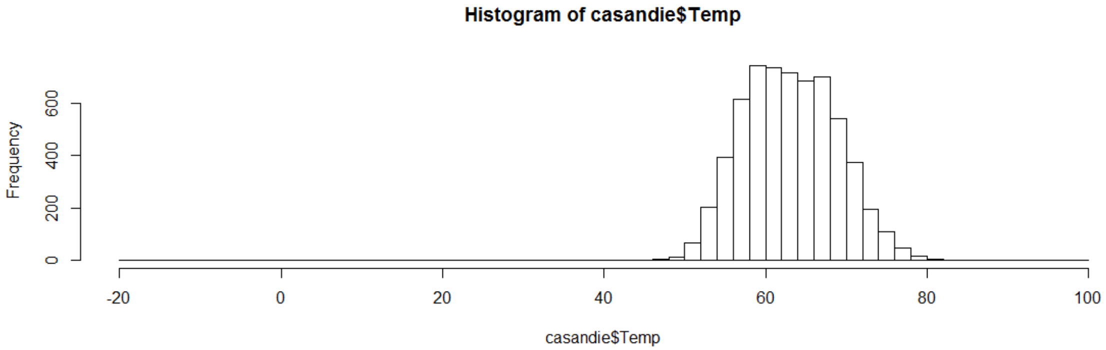
```
par(mfrow=c(2,1))
bins <- seq(-20, 100, 2)
hist(casandie$Temp, xlim = c(-20,100), breaks=bins)
hist(sdrapcty$Temp, xlim = c(-20,100), breaks=bins)
```

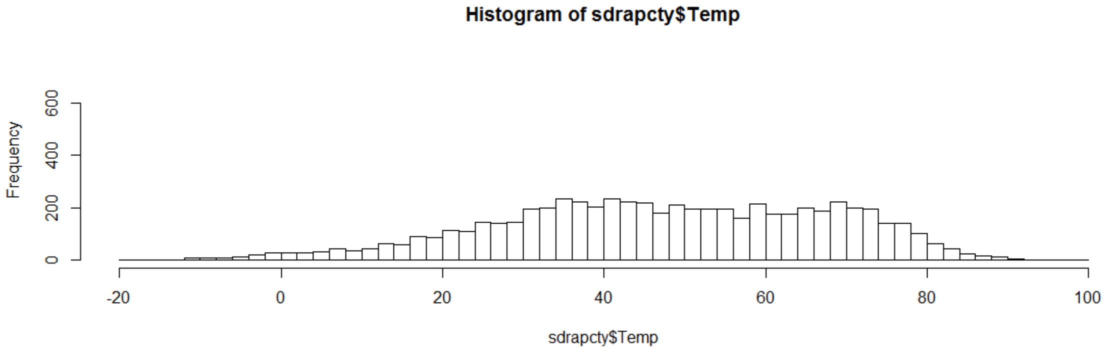


Comparability ?

Histograms - Bivariate

```
par(mfrow=c(2,1))
bins <- seq(-20, 100, 2)
bins <- seq(-20, 100, 2)
hist(casandie$Temp, xlim = c(-20,100), ylim=c(0,760), breaks=bins)
hist(sdrapcty$Temp, xlim = c(-20,100), ylim=c(0,760), breaks=bins)
```





Comparability ?

Average Temperature
&
Variability

Tabulation

```
library(effects)
library(mosaic)

data("TitanicSurvival", package = "effects")
names(TitanicSurvival)
head(TitanicSurvival)

# stratification of data (group by )
xtabs(~sex + survived, TitanicSurvival)
# tally from mosaic
tally(~sex + survived, data = TitanicSurvival)

#How about numeric variables like age
AgeFactor <- cut(TitanicSurvival$age, c(0,13,19,Inf), labels= c("Child", "Teen", "Adult"))
AgeSexFactor <- factor(AgeFactor:TitanicSurvival$sex)

tally(~ survived + AgeSexFactor, data = TitanicSurvival)
tally(~ survived + AgeSexFactor:passengerClass, data = TitanicSurvival)
```

Analytic Graphs: Beautiful Evidence

- Show Comparison
- Show Causality or Causal Framework
- Multivariate Analysis
- Integration of Evidence
- Credibility: Appropriate Labels, Scales, Sources
- Content Quality, Relevance and Integrity

https://www.edwardtufte.com/tufte/index