# Statistical Modelling and Hypothesis Testing

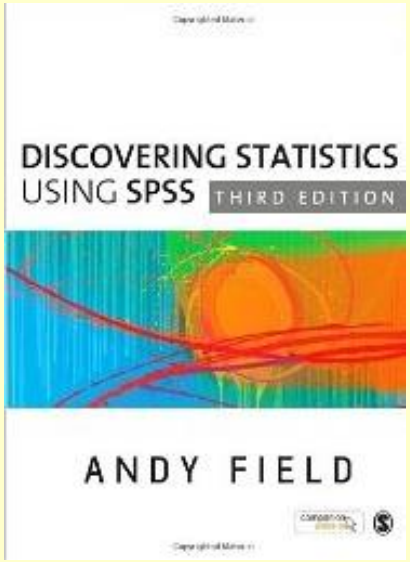
$$\text{Outcome}_i = (\text{Model}) + \text{error}_i$$

http://www.pelagicos.net/classes_biometry_fa16.htm

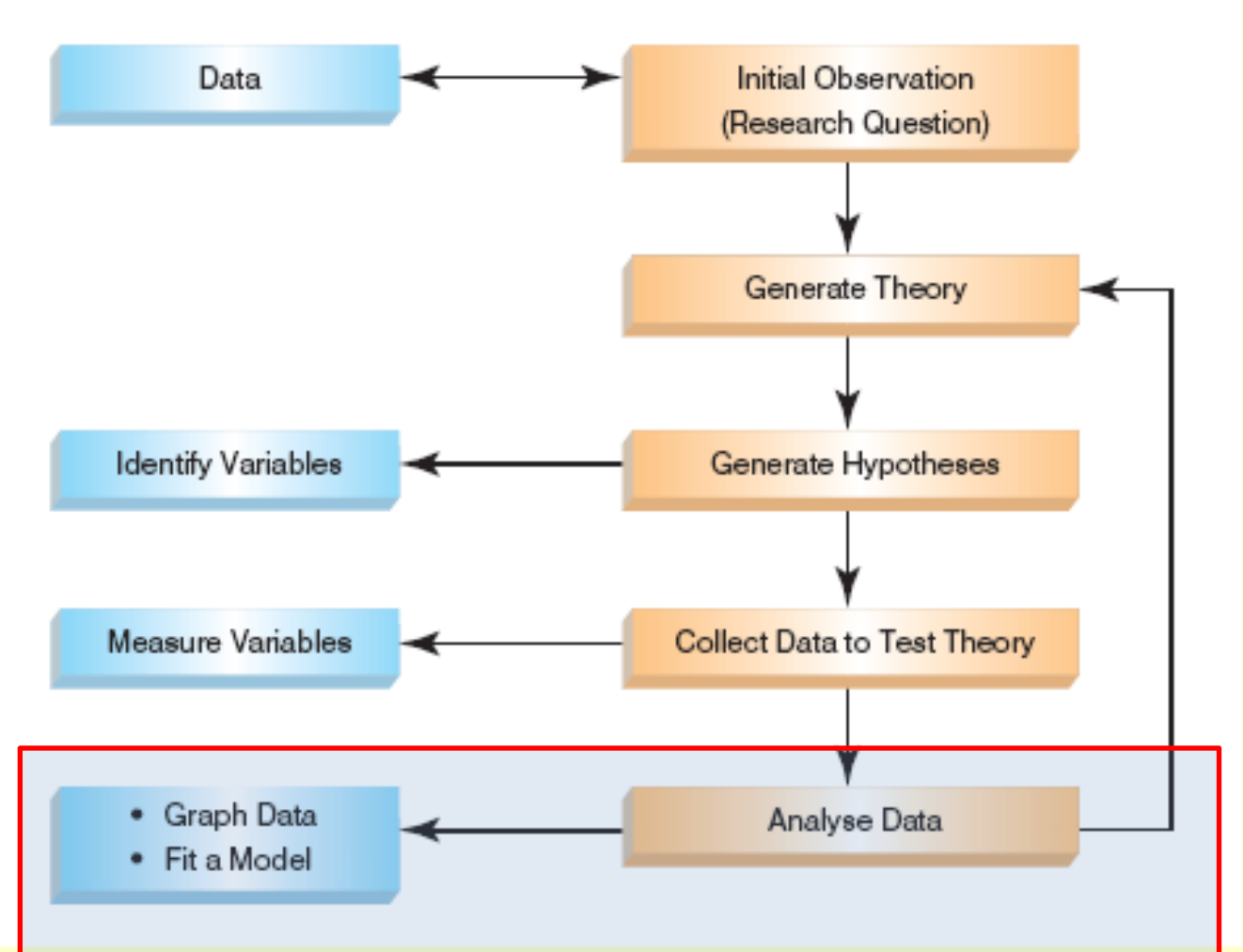
# Reading - Field: Chapters 1 & 2

## AIMS

- Understand what are statistical models and why we use them.
- Know what the 'fit' of a model is and why it is important.
- Understand how to quantify the 'fit' and statistical significance of models.

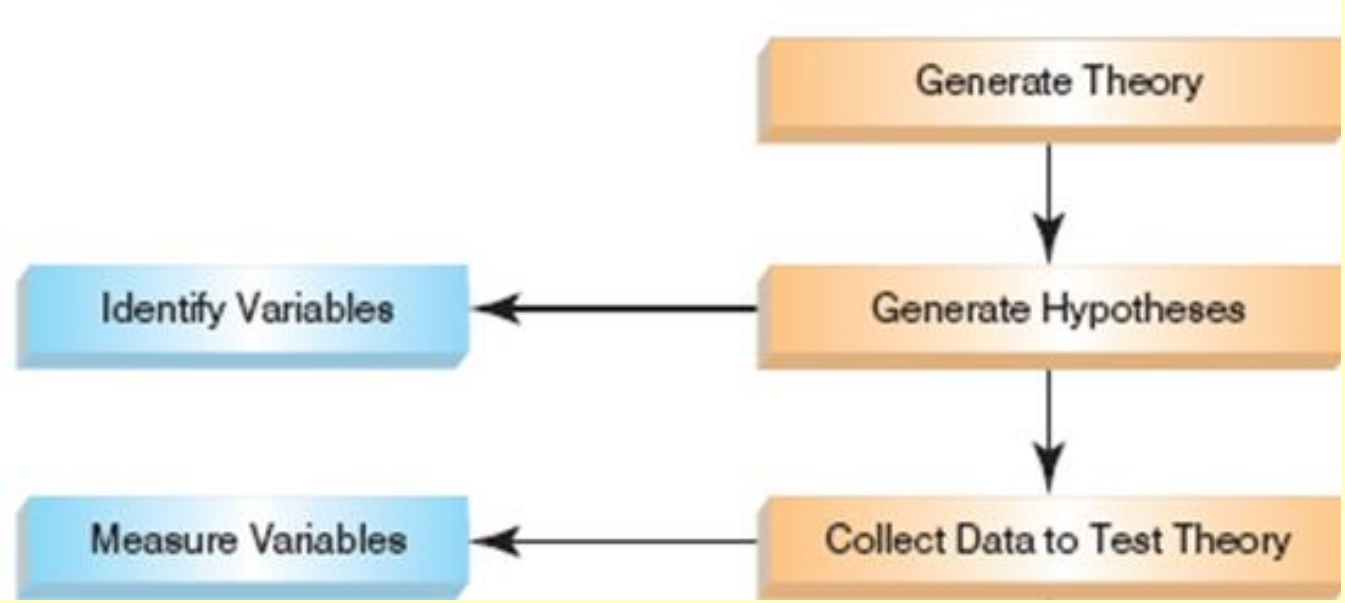


# Modeling in the Research Process



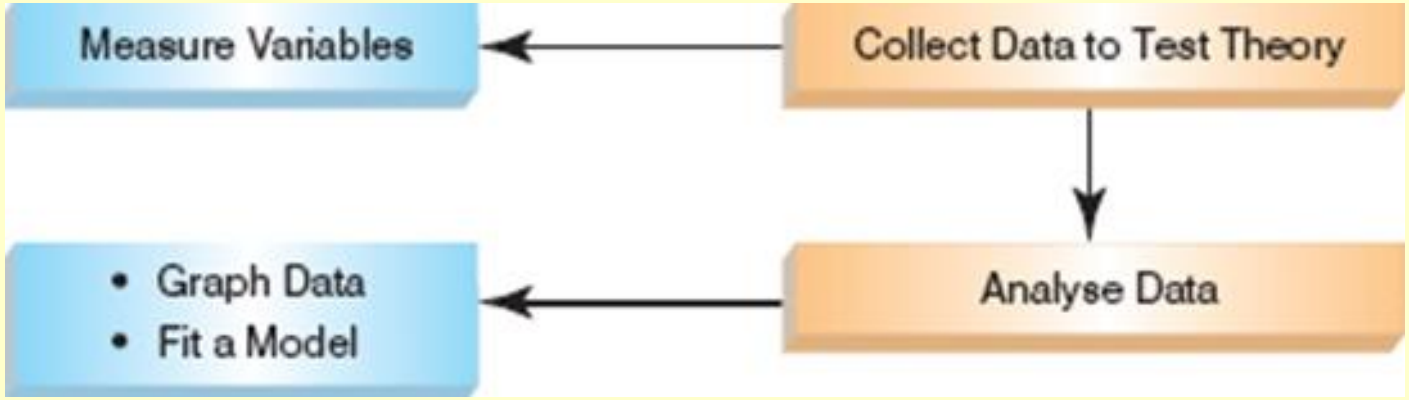

**Statistics**

# Step 1 – Generate Ideas




- Observations (reading the literature, going in the field, studying previous ecological theories, ...) help to generate hypotheses
- Predictions needed to test these hypotheses

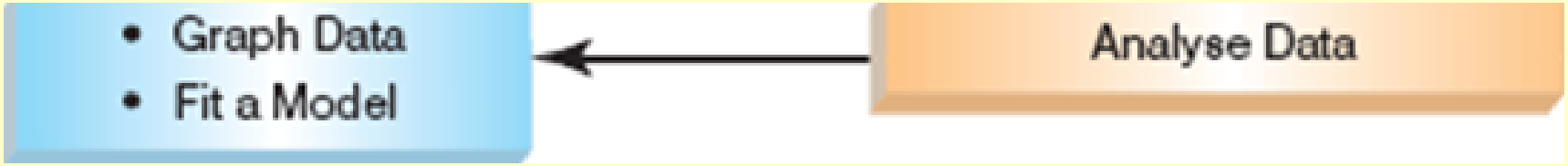
# Step 2 – Sample the Population




Sampling is needed to collect the data for comparing predictions with observations

- Population: The collection of units (people, plants, ...) to which we want to generalize findings or statistical model
- Sample: A smaller (but representative) collection of units used to estimate parameters from that population

# Step 3 – Model Fitting




- Models are a simplified description or representation of a certain aspect of a real process, phenomenon or object
- Model fitting balances the accuracy of the model (how well predictions match observations) and the interpretation of the results (simplicity of the model)
- This entails penalizing models with more parameters (more later on … in the class)

# Measure the "Fit" of a Model

Fitting models to the data teaches us about patterns and mechanisms (i.e. we use models to represent what is happening in the real world).

Model "Fit" quantifies the agreement between the model predictions and the observations.

Different models make specific assumptions about the relationships between the different variables (e.g. linear vs non-linear regression) and the frequency distribution of the errors.

# Model "Fit" – A Simple Example

- The mean is a hypothetical central tendency value (i.e. it does not have to be a value that actually exists in the data set).
- As such, the mean is simple statistical model.
- It is not a perfect representation of the data.
- How well does the mean represent reality ?

# The Mean - A Simple Example

$$\text{Mean}(\bar{X}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean is the value from which the sum of squares scores deviates the least. It has the least error.

## For Example:

- Collect data (5 random samples): 1, 3, 4, 3, 2
- Add them up: $\sum_{i=1}^{n} x_i = 1+3+4+3+2 = 13$
- Divide by number of scores, *n*: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{13}{5} = 2.6$

# The Mean - A Simple Model

$$\text{Outcome}_i = (\text{Model}) + \text{error}_i$$

**For Example:** For our data: 1, 3, 4, 3, 2

$$\text{Outcome}_{lecturer1} = (\bar{X}) + \text{error}_{lecturer1}$$

$$1 = 2.6 + \text{error}_{lecturer1}$$

$$\text{error}_{lecturer1} = -1.6$$

# Each Observation Has Error

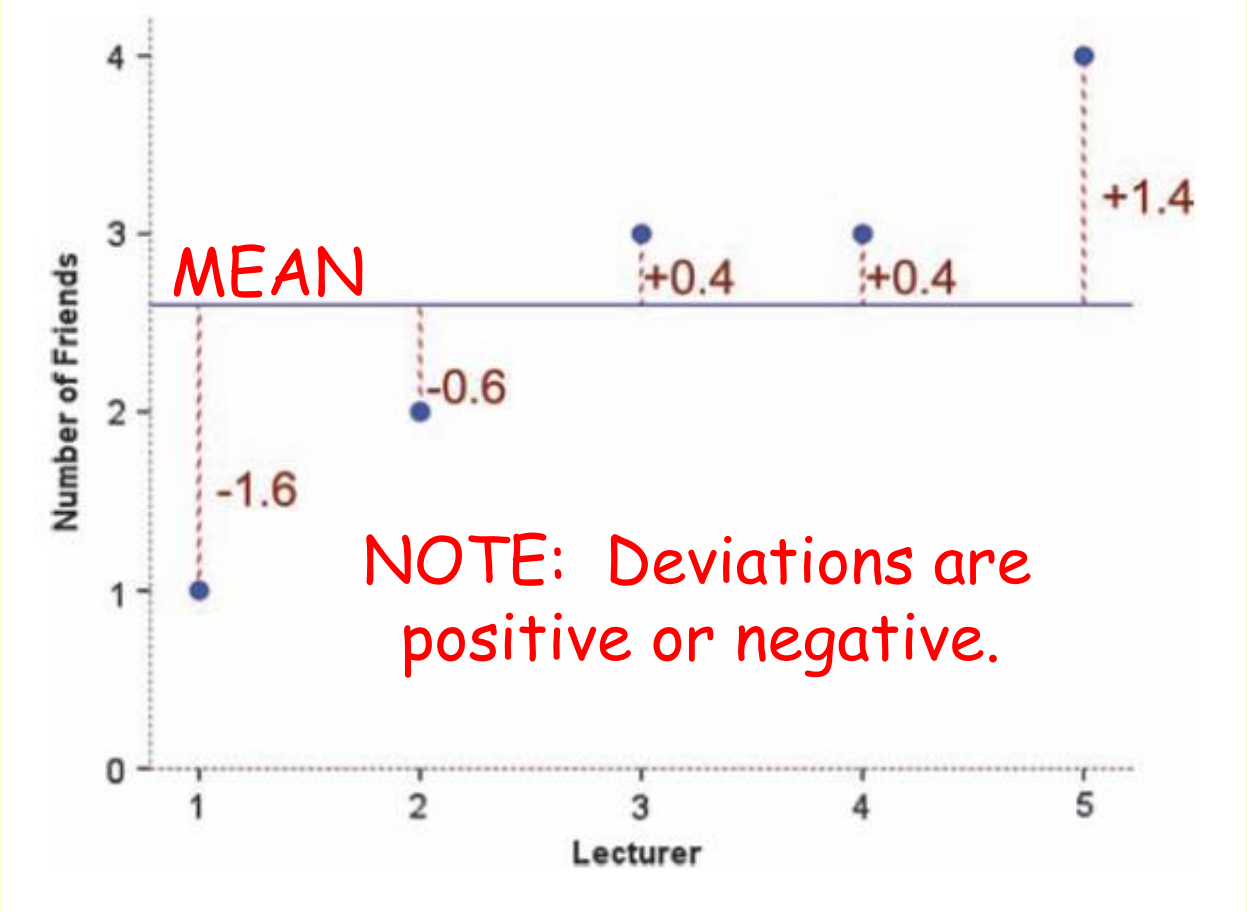
- A deviation is the difference between the mean and an actual data point.
- Deviations are calculated by subtracting the mean from each observation:

$$\text{Deviation} = x_i - \bar{x}$$

**NOTE:** More generically, deviations are calculated between any model predictions and the observations.

# Calculating the Deviations

Plot shows the difference between the observations (number of friends five lecturers have) and the model estimate (mean number of friends)



# What is the Total Error?

- If we calculate the error between the mean and each observation and add them.

| Score | Mean | Deviation |
|---|---|---|
| 1 | 2.6 | -1.6 |
| 2 | 2.6 | -0.6 |
| 3 | 2.6 | 0.4 |
| 3 | 2.6 | 0.4 |
| 4 | 2.6 | 1.4 |
| | Total = | 0 |

$$\sum (X - \bar{X}) = 0$$

NOTE: The sum of all the deviations always adds to "zero"

# Statistical Modeling: Inference & Reliability

- We square each deviation because it does not matter if the observations are larger / smaller than the mean.
- The larger the difference from the mean, the larger the total deviation.
- If we add these squared deviations we get the Sum of Squared Errors (*SS*).

# Sum of Squares

| Score | Mean | Deviation | Squared Deviation |
|---|---|---|---|
| 1 | 2.6 | -1.6 | 2.56 |
| 2 | 2.6 | -0.6 | 0.36 |
| 3 | 2.6 | 0.4 | 0.16 |
| 3 | 2.6 | 0.4 | 0.16 |
| 4 | 2.6 | 1.4 | 1.96 |
| | | Total | 5.20 |

$$SS = \sum (X - \overline{X})^2 = 5.20$$

NOTE: Note the similarity between the variance and the Sum of Squares

# Sum of Squares

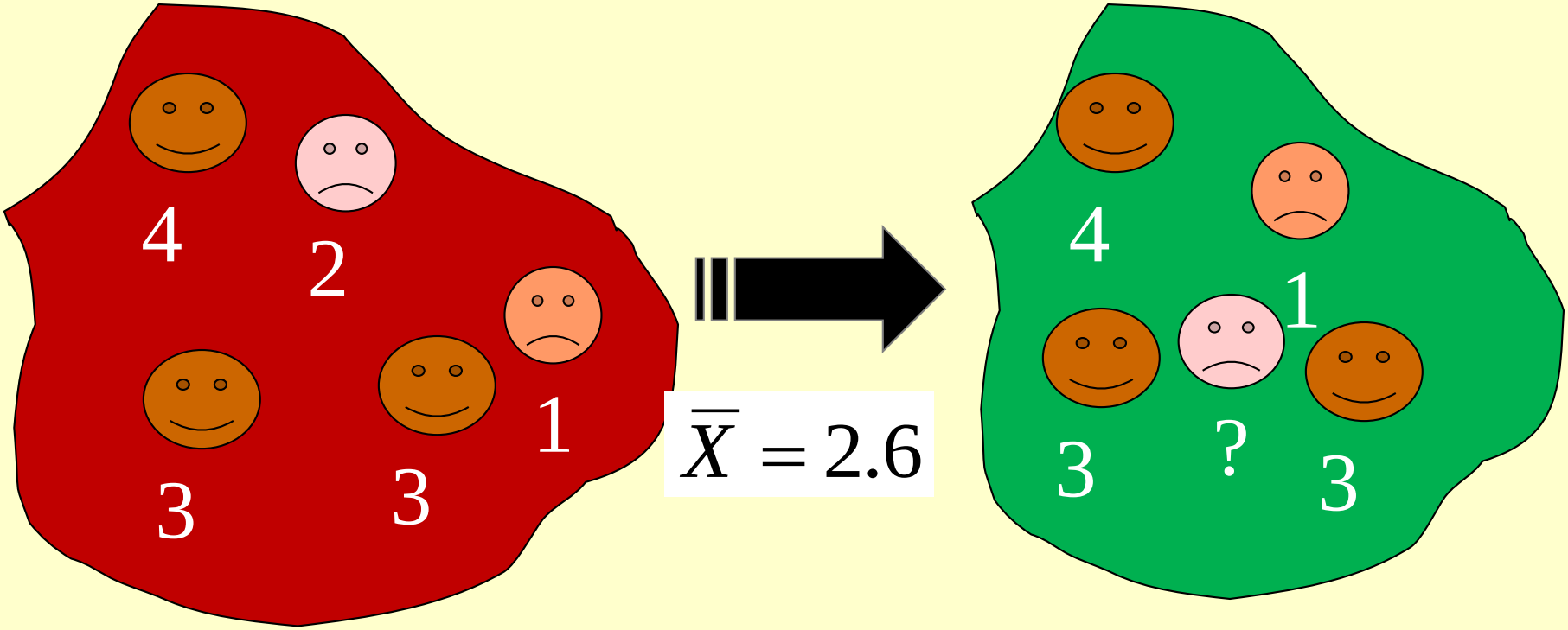
- The sum of squares is a good measure of overall variability, but is dependent on the overall number of scores.
- We calculate the average variability by dividing by the degrees of freedom (*n -1*).
- Remember: This is called the variance ($s^2$)

$$\text{variance } (s^2) = \frac{SS}{N-1} = \frac{\sum (x_i - \overline{x})^2}{N-1} = \frac{5.20}{4} = 1.3$$

# Side Note: Degrees of Freedom

MEAN

VARIANCE




n = 5

df = 4

Note: Because calculating the variance requires a mean … once three values are free, the fifth observation is fixed.

# So What Does this All Mean ?

Sum of Squares provides a robust measure to assess the deviation between observations (data) and predictions (model).

Dividing the Sum of Squares by the Degrees of Freedom yields the Mean Square Error.

The simplest example of the MS Error is the use of the variance to assess the fit of the observations to their mean.

# Summary

We fit models to the data to learn about patterns and mechanisms. Because there often are competing explanations for the data, we want to assess how well different models describe the observations.

Model "Fit" quantifies the agreement between the model predictions and the observations.

We use Sum of Squares to assess deviation.

Mean Squares terms are Sum of Squares standardized by the degrees of Freedom.

# Summary

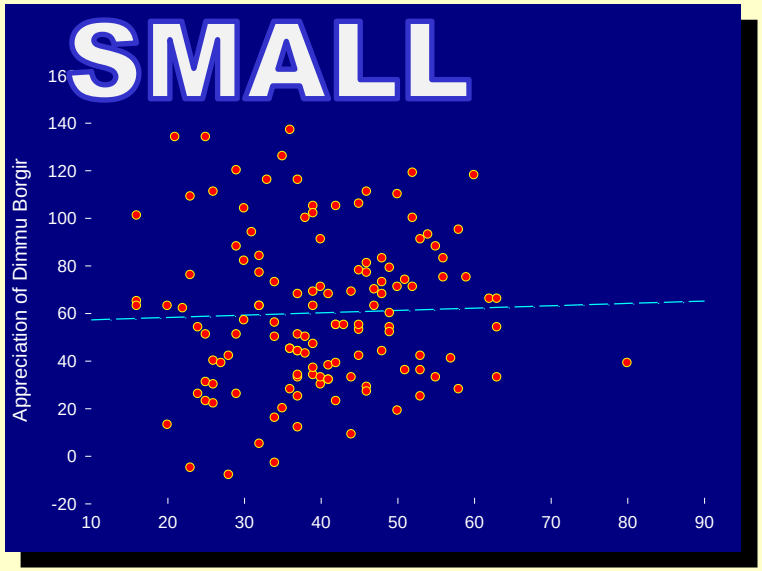
Assessing model performance involves comparing proportion the total variance they explain ($R^2$)

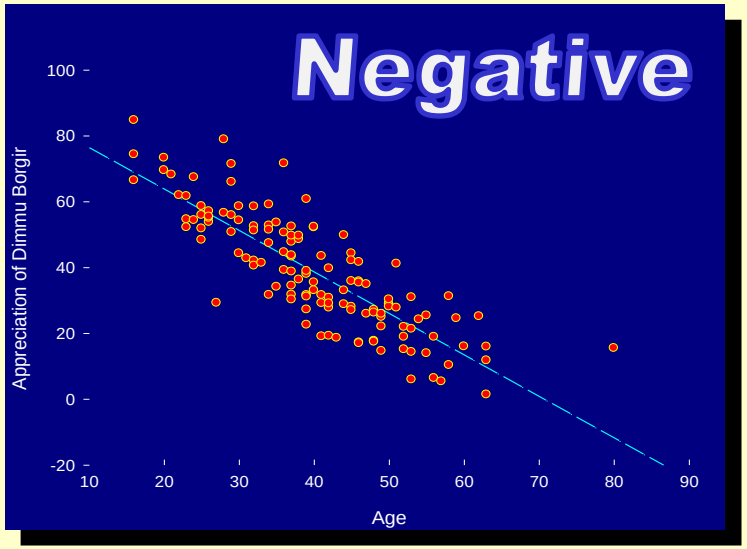
Statistical testing of models involves comparing the Mean Square ratio, F = MS Model / MS Error

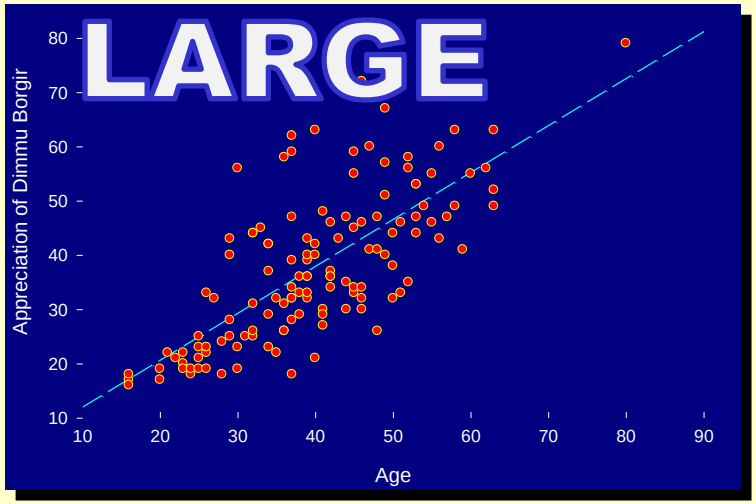
This ratio balances the magnitude of the signal (variation described by the model) and the noise (unresolved variation inherent in the system)

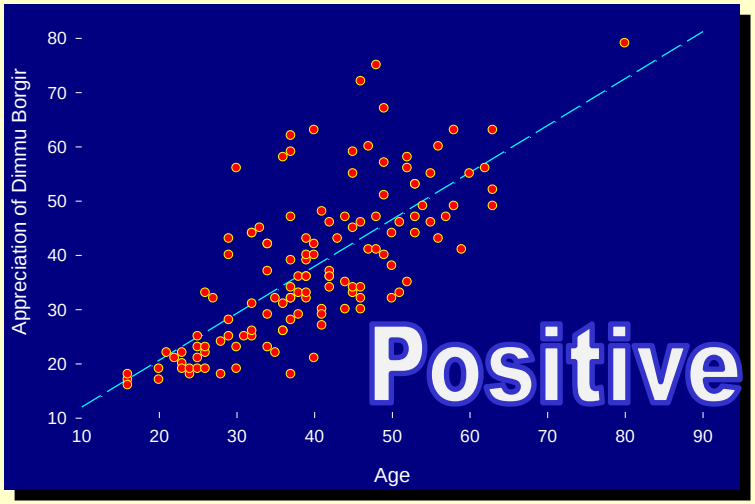
$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

# Correlation



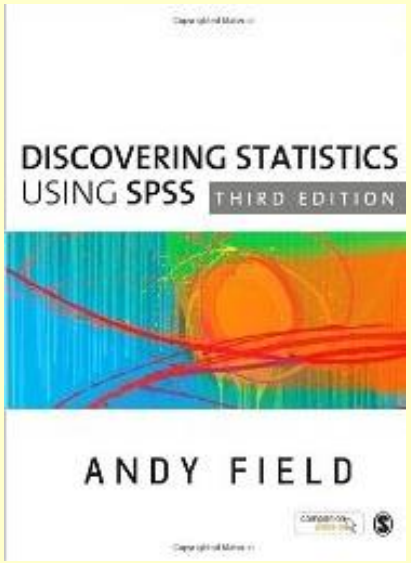




http://www.pelagicos.net/classes_biometry_fa16.htm

# Reading - Field: Chapter 6

## AIMS

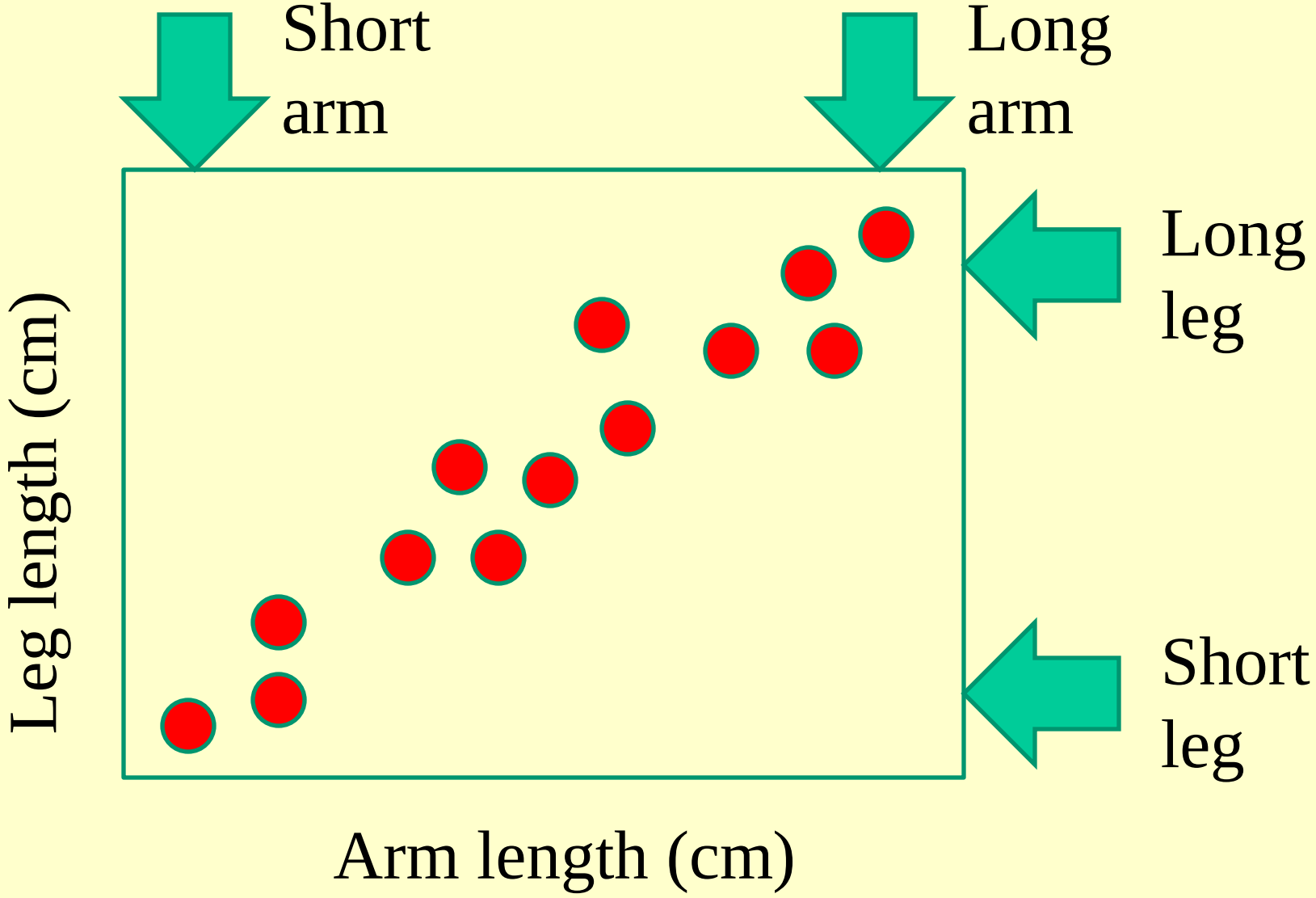
- Measuring Relationships
  - Scatterplots
  - Covariance
  - Pearson's Correlation Coefficient
- Nonparametric measures
  - Spearman's Rho
  - Kendall's Tau
- Interpreting Correlations
- Partial Correlations



# What is a Correlation

- *Correlation* refers to departure of two variables from independence
- Quantifies extent to which two variables are related (co-dependent)
- Measures concurrent changes in the variables: same / no response / opposite
- Specifically it refers to several types of relationship between variable values

# What is a Correlation



# Measuring Correlation

- Assessed using multiple statistics:
  some parametric, some non-parametric

- Most common is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables

- Other correlation coefficients are more sensitive to nonlinear relationships

# Pearson Correlation - Assumptions

Pearson correlation makes five assumptions:

- (In addition to reliance on "random sampling")

- Variables either interval or ratio measurements.

- Variables normally distributed – No Outliers.

- The two distributions have equal variances
  (also known as homoscedasticity).

- Linear relationship between the two variables.

# Measuring the Pearson Correlation

- Determine whether as one variable increases, the other one increases, decreases or stays the same.

- This is done by calculating the Covariance
  - Determine how much each observation (x,y pair) deviates from the mean
  - If both variables deviate from their means by similar amounts, they are correlated

# The Mean & The Variance

Data Series X: 1,2,3,4,5

Data Series Y: 2,4,6,8,10

Mean X:
3

Mean Y:
6

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Variance X:
2.5

Variance Y:
10

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

**Variance = sum of squared deviations from mean**

**degrees of freedom**

# Calculating the Co-variance

- The variance quantifies how single variable scores deviate from mean
- Based on the sum of squares

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

- Covariance is similar:
  quantifies how scores of two variables differ from their respective means

# Covariance – How to

- Calculate deviation between each score for first variable ($x$) and their mean.
- Calculate deviation between each score for second variable ($y$) and their mean.
- Multiply these two deviation values to calculate the cross product deviations.
- The covariance is the average cross-product of the deviations:

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

# Measuring Correlations

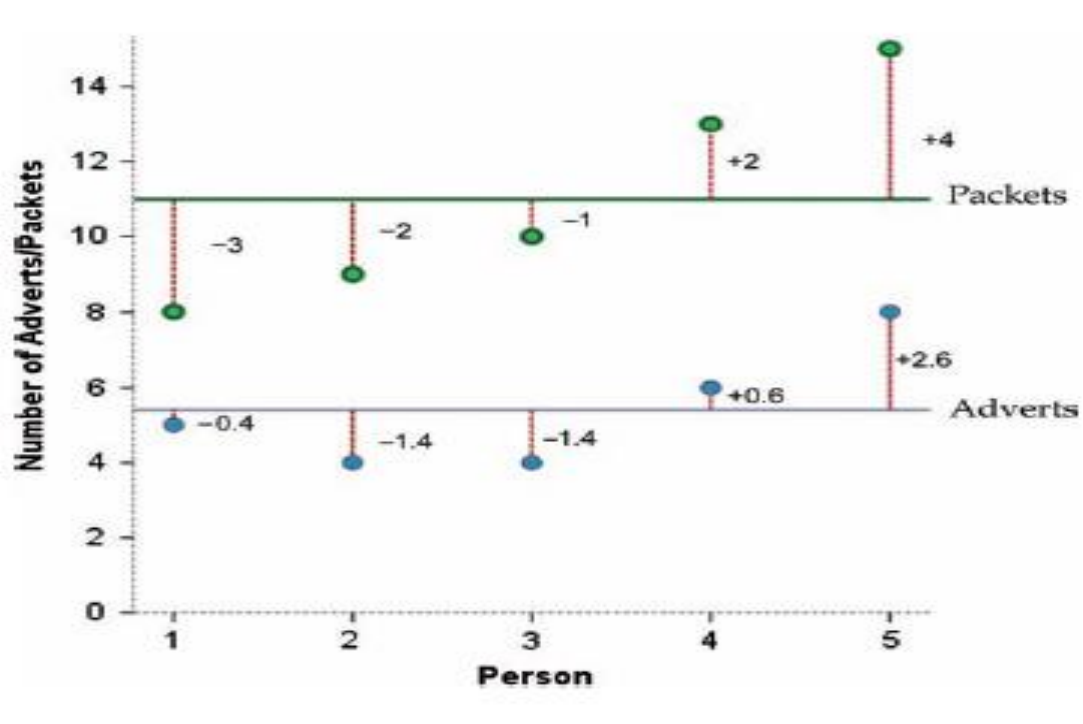
| Subject | 1 | 2 | 3 | 4 | 5 | Mean | S |
|---|---|---|---|---|---|---|---|
| Adverts Watched | 5 | 4 | 4 | 6 | 8 | 5.4 | 1.67 |
| Packets Bought | 8 | 9 | 10 | 13 | 15 | 11.0 | 2.92 |

Deviations:

Observed Value

-

Group Mean




Positive / Negative

Sum(deviations) = 0

# Covariance – How to

$$\mathrm{cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4}$$

$$= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4}$$

$$= \frac{17}{4}$$

$$= 4.25$$

What if units were milli-ads and milli-packets?
(Multiply observations and means by 1000)
(Covariance would be multiplied by 1000)

# Covariance – How to

- Depends upon the units of measurement.
  - E.g. The Covariance of two variables measured in Miles is 4.25, but if same scores are converted to Km, the Covariance is 11.

- One solution: standardise it!
  - Divide by S.D.s of both variables.

- The standardised version of Covariance is known as the Correlation coefficient.

# Covariance – How to

$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$= \frac{4.25}{1.67 \times 2.92}$$

$$= .87$$

# Correlation: Scaled Covariance

- Covariance of X and Y divided by SD in X and SD in Y
- Quantifies intensity of association between two variables

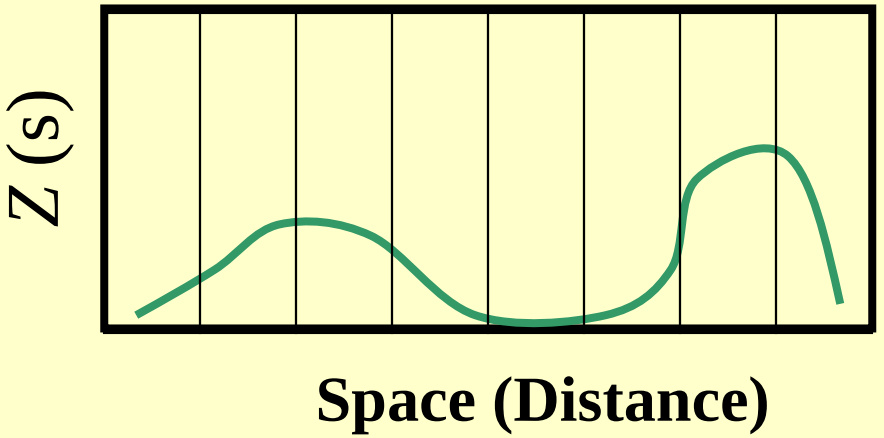
$$r = \frac{Covariance}{\sqrt{(Variance\ X)(Variance\ Y)}}$$

$$r = \frac{\sum (Xi - X)(Yi - Y)}{\sqrt{\sum (Xi - X)^2 \sum (Yi - Y)^2}}$$

# Covariance = Variance of Two Variables

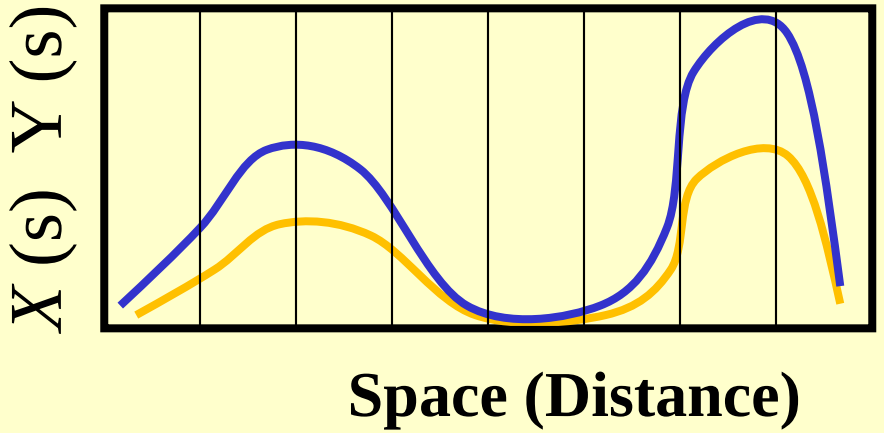
**The variance used to assess variability in one variable**

$$Variance = \frac{\sum (Zi - \bar{Z})(Zi - \bar{Z})}{(n-1)}$$

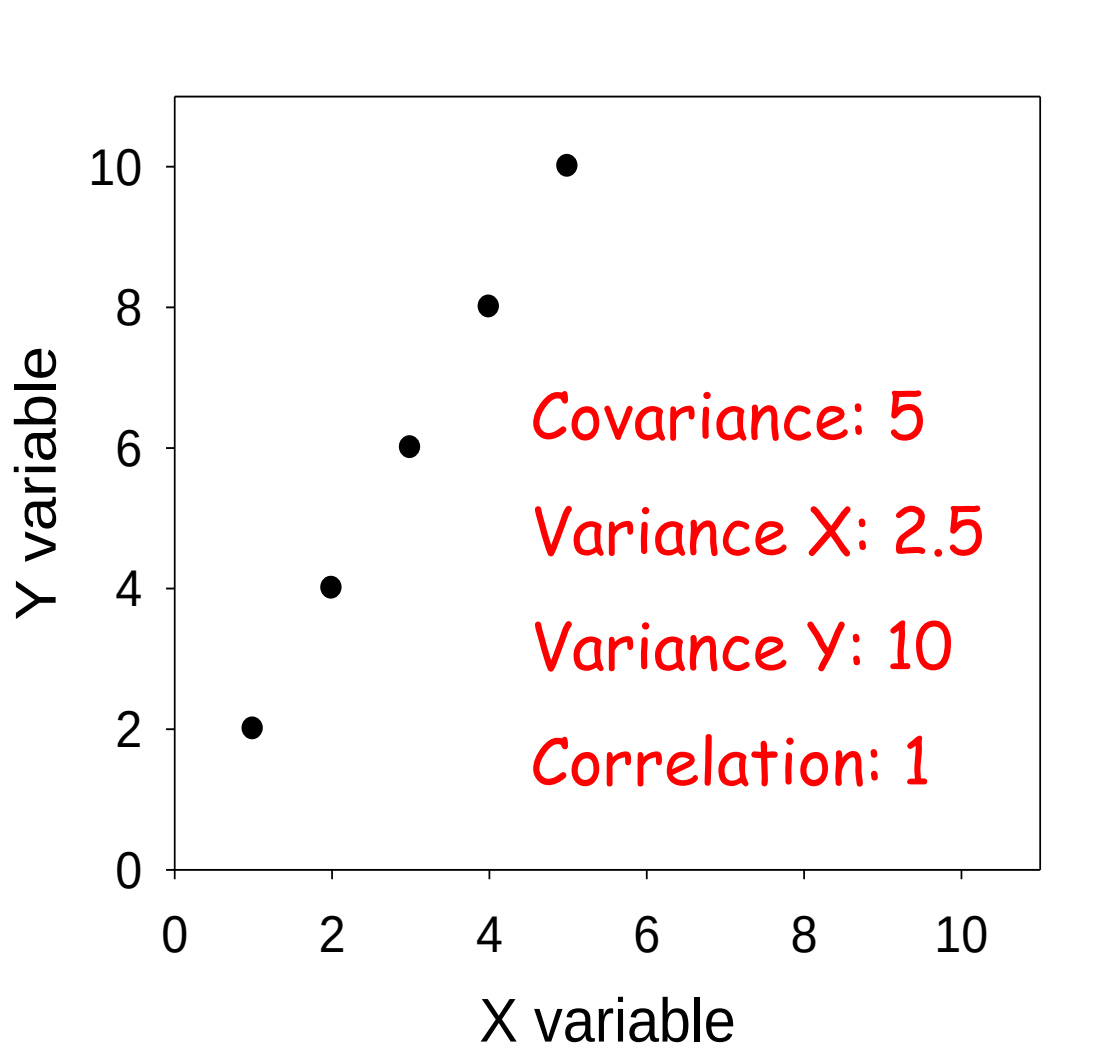



**How to quantify variability shared by two variables?**

$$Covariance = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{(n-1)}$$



# Work out this Example Yourselves




| X | Y |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 6 |
| 4 | 8 |
| 5 | 10 |

Correlation coeff:

$$\frac{5}{\text{sqrt}(2.5)^* \text{ sqrt}(10)}$$

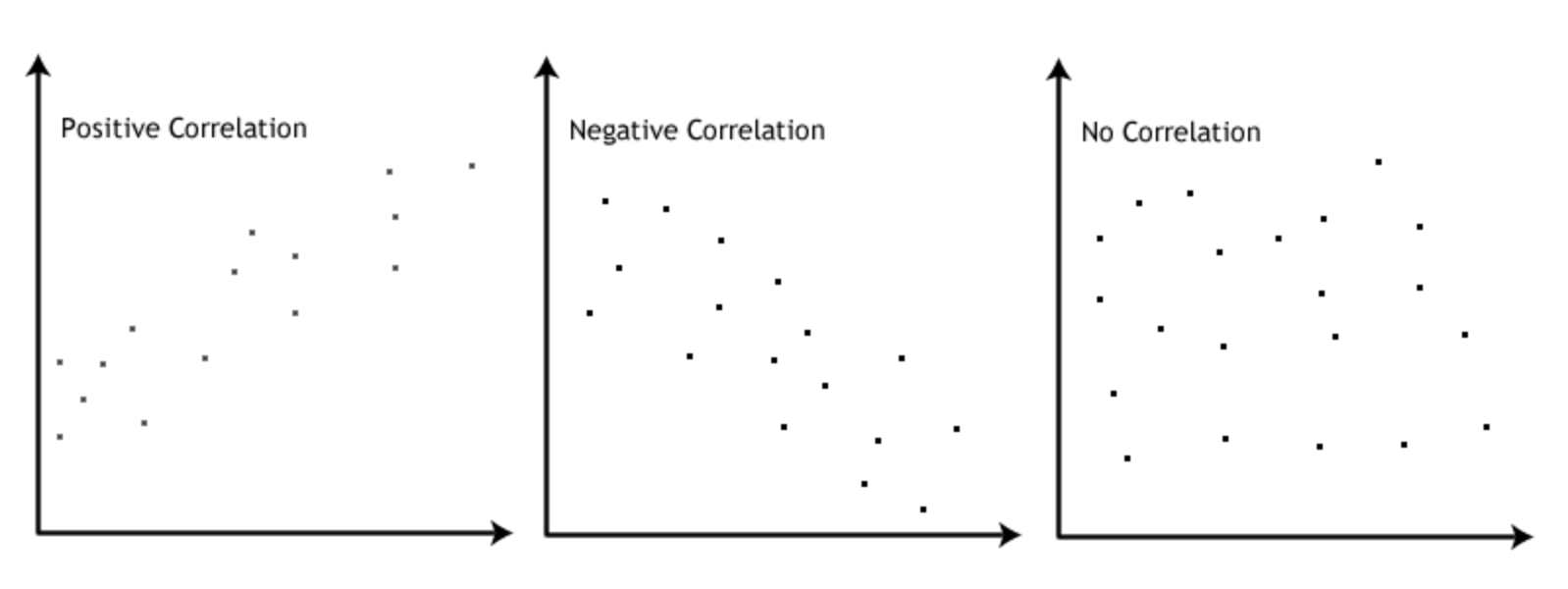
$r = +1$

$r^2 = 1$

# Pearson Correlation Coefficient (r)

A measure of the sign and strength of a linear association between two variables.

The Pearson correlation coefficient, *r*, indicates how far away the data points fall of a best fit line describing relationship between variables .



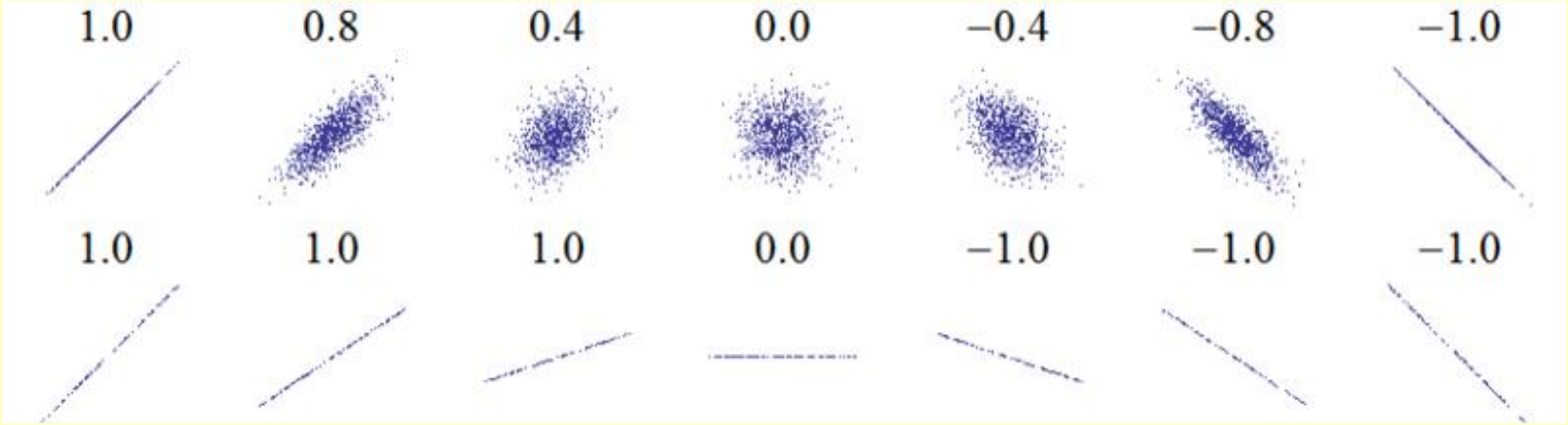
# Interpretation of Pearson Correlation

The correlation coefficient (r) indicates the strength and the direction of a linear relationship between two random variables

The coefficient of determination ($r^2$) indicates the % of the variance in one variable that is explained by the other (bidirectional)

$-1 \leq r \leq 1$

$0 \leq r^2 \leq 1$



# Pearson Correlation Coefficient (r)

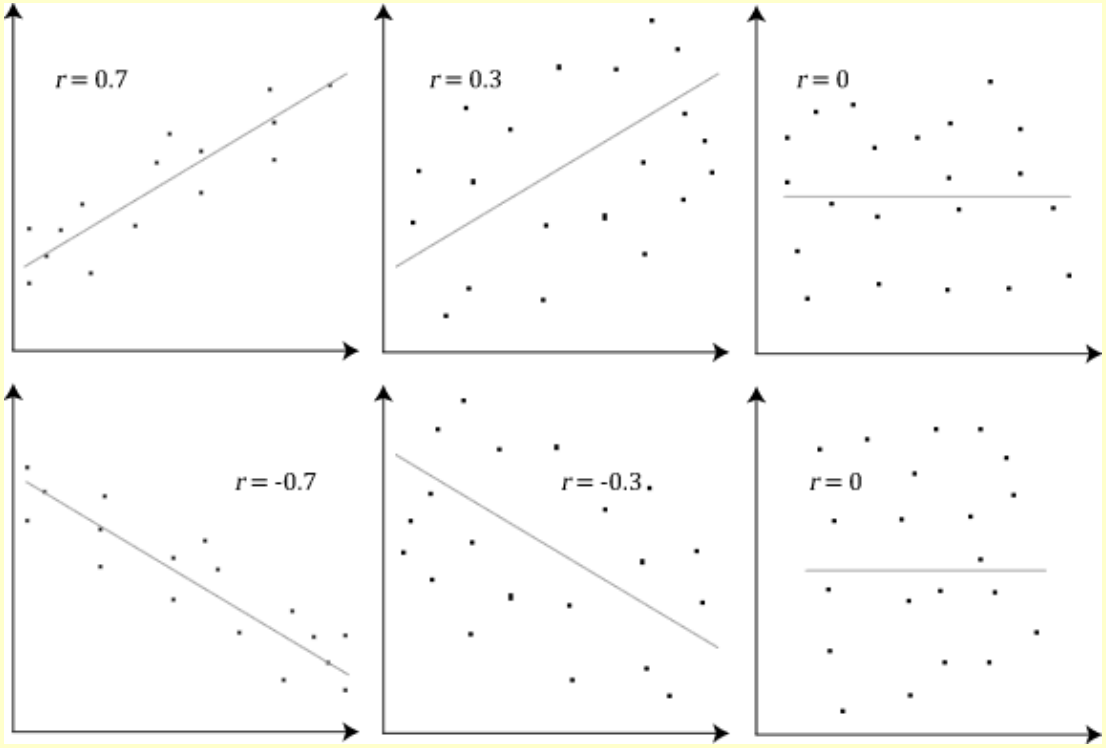
The stronger the association of the two variables, the closer the correlation coefficient, will be to +1 or -1 depending on whether the relationship is positive or negative, respectively.

A value of +1 or -1 means that all your data points fall along the line of best fit - no data points show any variation away from this line.

The closer the value of *r* to 0 the greater the variation around the line of best fit and the weaker the fit of the linear relationship.

# Pearson Correlation Coefficient (r)

| | Coefficient, *r* | |
|---|---|---|
| Strength of Association | Positive | Negative |
| Small | .1 to .3 | -0.1 to -0.3 |
| Medium | .3 to .5 | -0.3 to -0.5 |
| Large | .5 to 1.0 | -0.5 to -1.0 |




There are guidelines to determine the "strength" of the correlation coefficient.

**NOTE:** But significance of r depends on the sample size (d.f.).