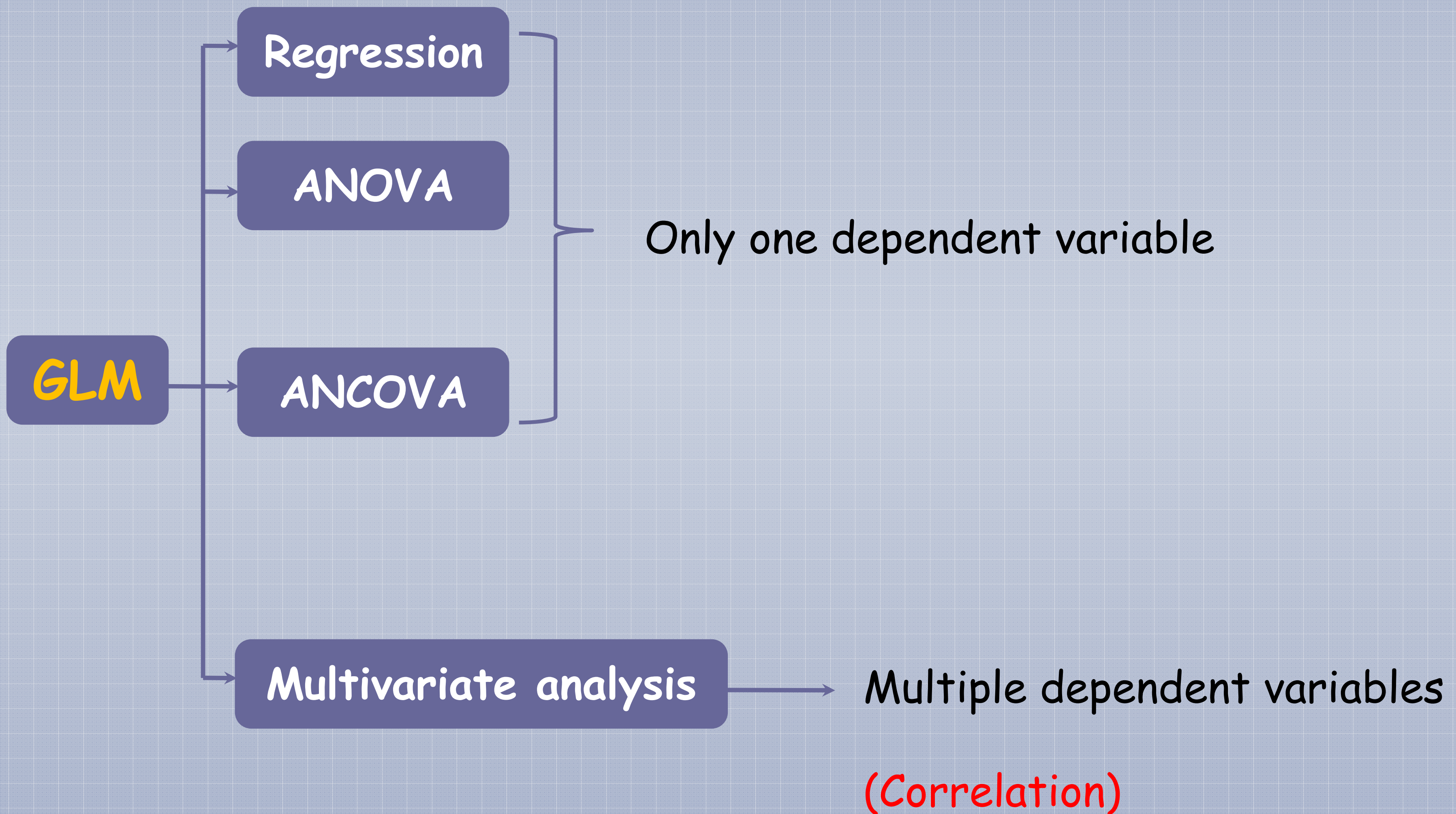
BIOL 4605/7220
CH 20.1 Correlation

GPT Lectures
Cailin Xu

November 9, 2011

GLM: correlation



Correlation

- Two variables associated with each other?
- No casual ordering (i.e., NEITHER is a function of the other)

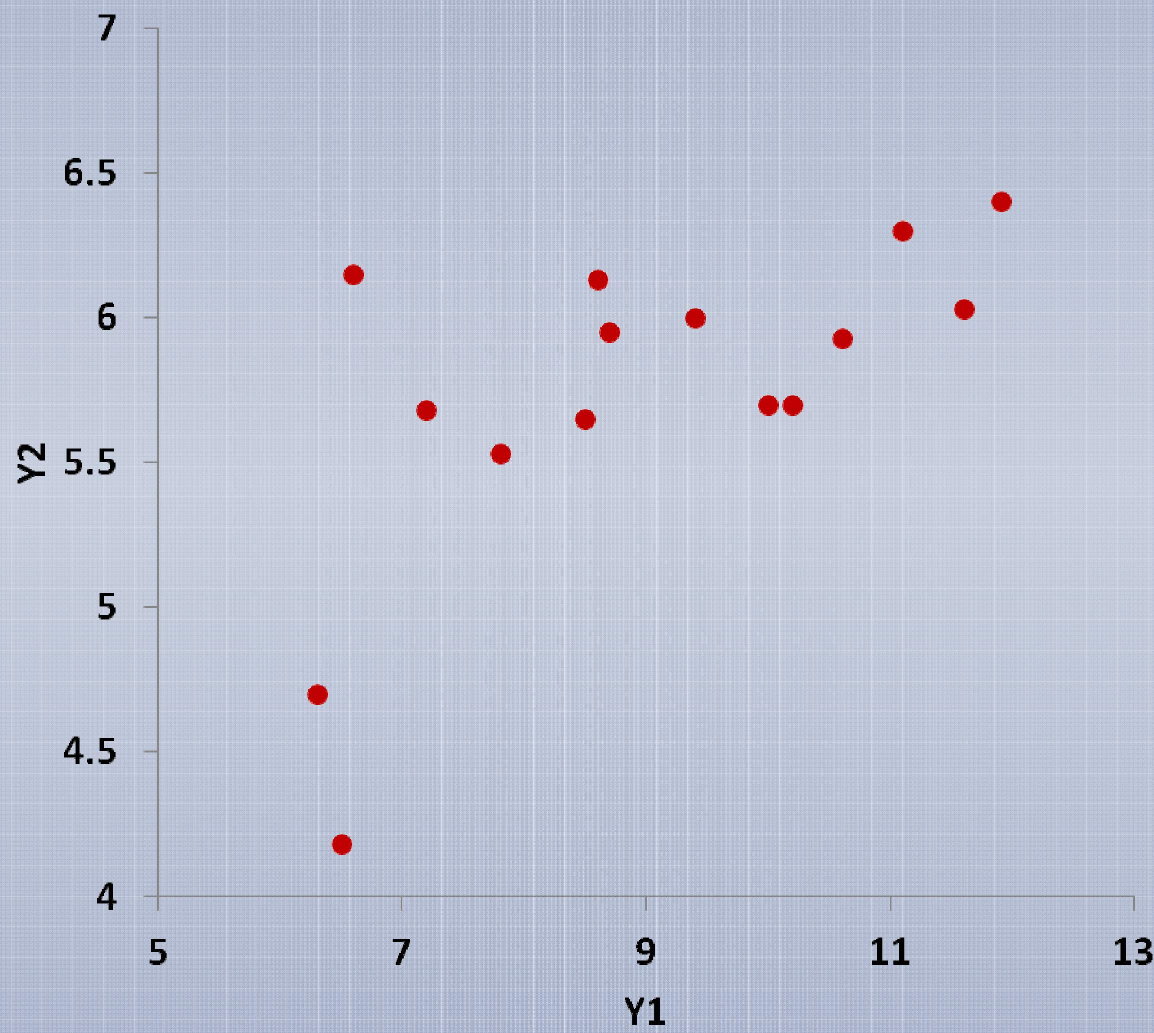
Y_1 – Total length of aphid stem mothers

Y_2 – Mean thorax length of their parthenogenetic offspring

Data from Box 15.4 Sokal and Rohlf 2012

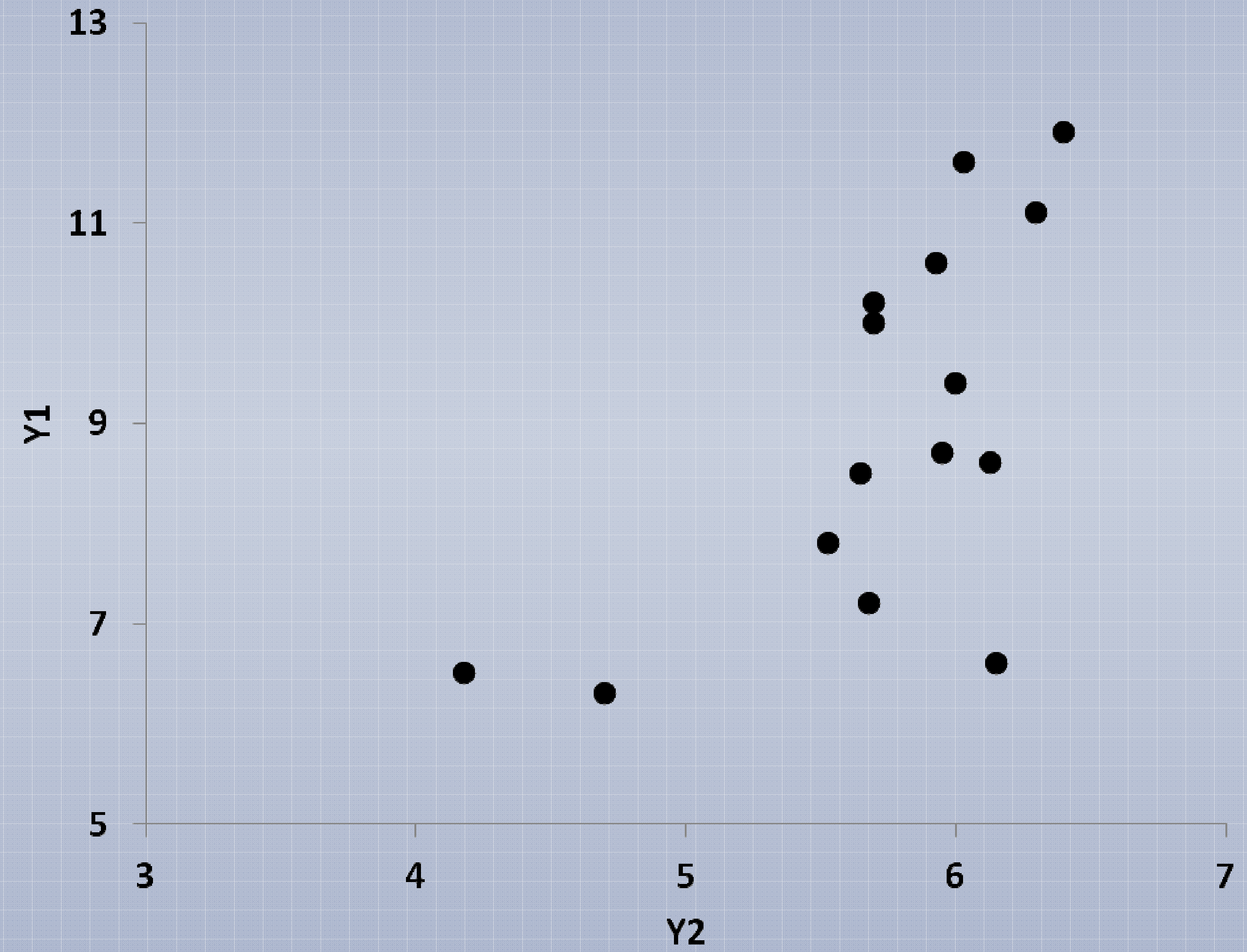
Correlation

$Y_2 \; vs. \; Y_1$



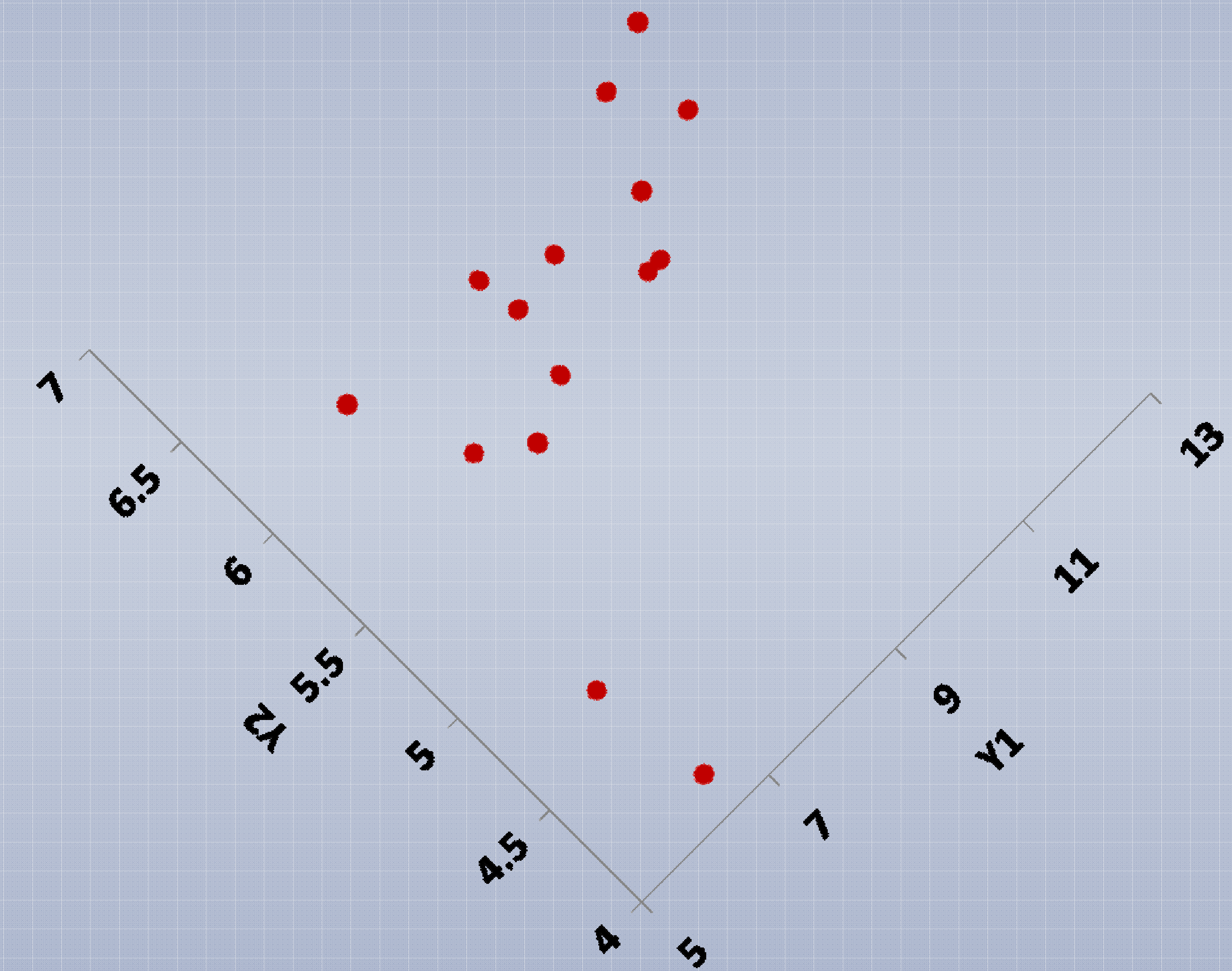
Correlation

$Y_1 \ vs. Y_2$



Correlation

Rotate



Regression vs. Correlation

Regression

- Does Y depend on X?
  (describe func. relationship/predict)
- Usually, X is manipulated & Y is a random variable
- Casual ordering Y=f(X)

Correlation

- Are Y1 and Y2 related?
- Both Y1 & Y2 are random variables
- No casual ordering

Correlation: parametric vs. non-parametric

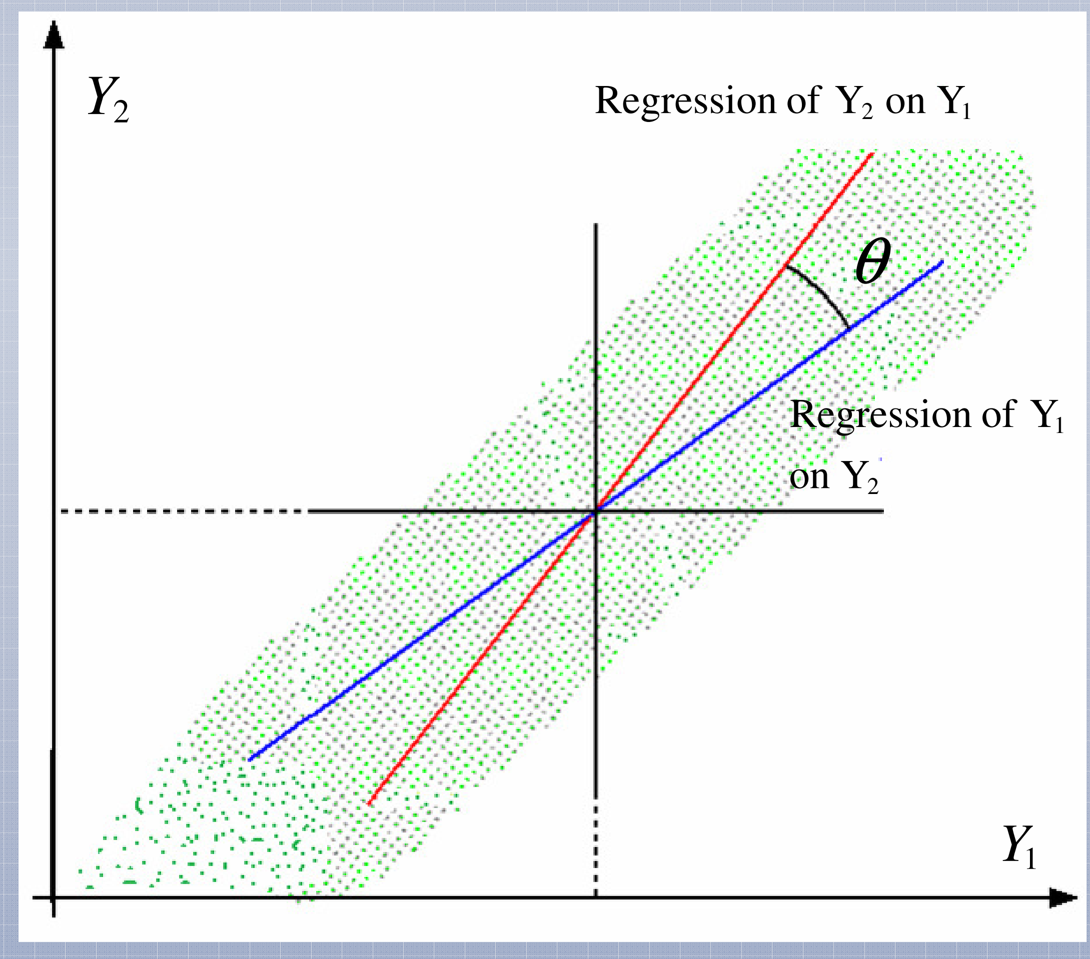
Parametric measures: Pearson's correlation

Nonparametric measures: Spearman's Rho, Kendall's Tau

Type of data	Measures of correlation
Measurements (from Normal/Gaussian Population)	Parametric: Pearson's correlation
Ranks, Scores, or Data that do not meet assumptions for sampling distribution (t, F, χ^2)	Nonparametric: Spearman's Rho, Kendall's Tau

Pearson's Correlation Coefficient (ρ)

- Strength of relation between two variables Y_1 & Y_2
- Geometric interpretation



$$\rho = \cos(\theta)$$

- **Perfect positive association:**

  Θ =0° ρ=1
- **No association:**

  Θ =90° ρ=0
- **Perfect negative association:**

  Θ =180° ρ=-1

$-1 \leq \rho \leq 1$, true relation

Pearson's Correlation Coefficient (ρ)

- Strength of relation between two variables Y_1 & Y_2
- Geometric interpretation
- Definition

$$\rho_{Y_1, Y_2} = \frac{\operatorname{cov}(Y_1,\ Y_2)}{\sigma_{Y_1}\ \sigma_{Y_2}} = \frac{E\left[\left(Y_1 - \mu_{Y_1}\right)\left(Y_2 - \mu_{Y_2}\right)\right]}{\sigma_{Y_1}\ \sigma_{Y_2}}$$

Covariance of the two variables divided by the product of their standard deviations

Pearson's Correlation Coefficient (ρ)

- Strength of relation between two variables Y_1 & Y_2
- Geometric interpretation
- Definition
- Estimate $(\hat{\rho} = r)$ from a sample

Parameter		Estimate
Name	Symbol	
Mean of Y_1	μ_{Y_1}	$\bar{Y}_1$
Mean of Y_2	μ_{Y_2}	$\bar{Y}_2$
Variance of Y_1	$\sigma^2_{Y_1}$	$s^2_{Y_1}$
Variance of Y_2	$\sigma^2_{Y_2}$	$s^2_{Y_2}$

Pearson's Correlation Coefficient (ρ)

- Strength of relation between two variables Y_1 & Y_2
- Geometric interpretation
- Definition
- Estimate $(\hat{\rho} = r)$ from a sample

Parameter	Estimate
μ_{Y_1}	$\bar{Y}_1$
μ_{Y_2}	$\bar{Y}_2$
$\sigma^2_{Y_1}$	$s^2_{Y_1}$
$\sigma^2_{Y_2}$	$s^2_{Y_2}$

$$\rho_{Y_1, Y_2} = \frac{\operatorname{cov}(Y_1,\ Y_2)}{\sigma_{Y_1}\ \sigma_{Y_2}} = \frac{E\left[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})\right]}{\sigma_{Y_1}\ \sigma_{Y_2}}$$

$$\downarrow$$

$$r = \hat{\rho} = \frac{1}{n-1} \cdot \frac{\sum_i (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}{s_{Y_1}\ s_{Y_2}} = \frac{\sum_i (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}{\sqrt{\sum_i (Y_{1i} - \bar{Y}_1)^2 \sum_i (Y_{2i} - \bar{Y}_2)^2}}$$

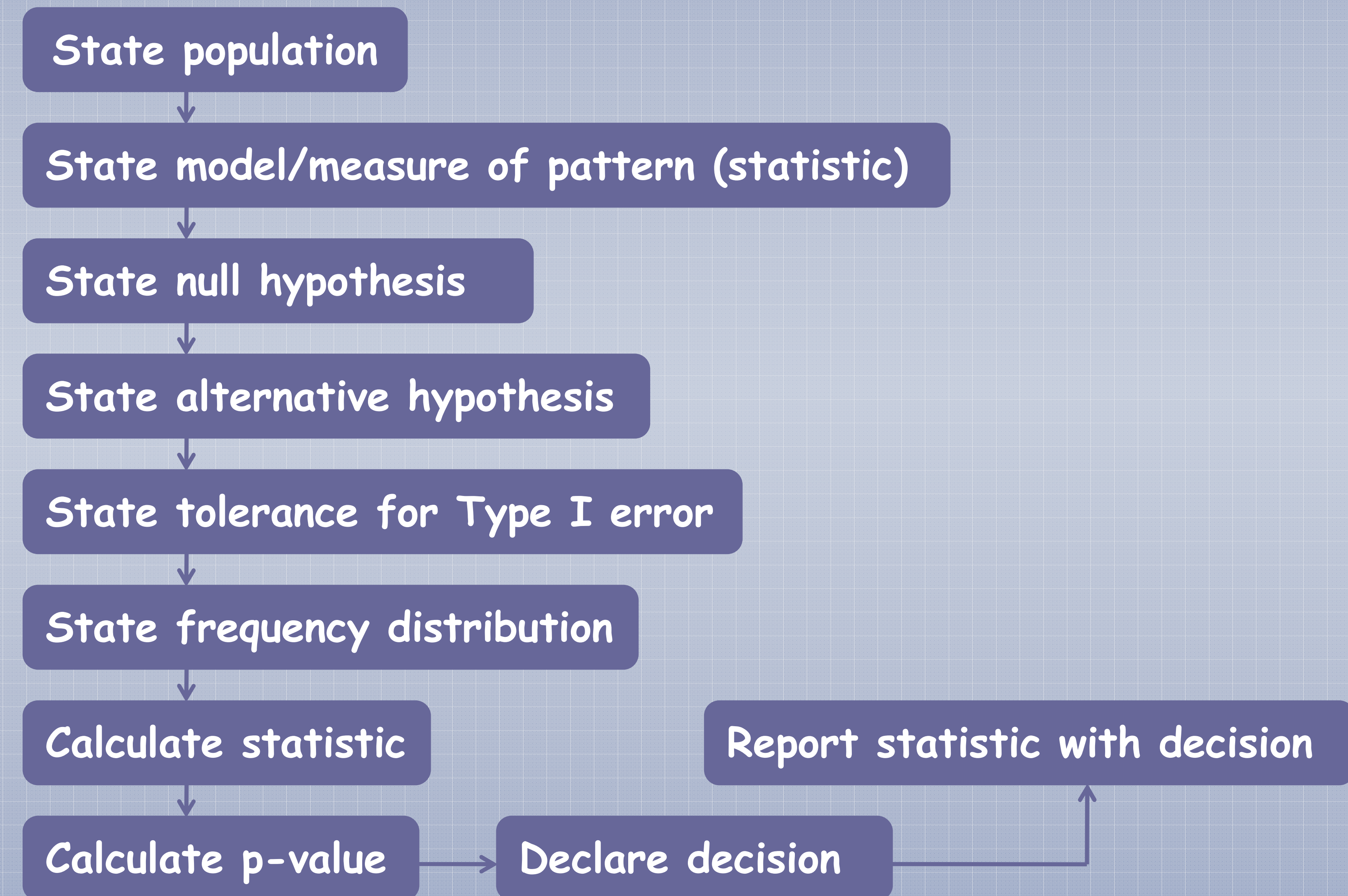
Pearson's Correlation: Significance Test

- Determine whether a sample correlation coefficient could have come from a population with a parametric correlation coefficient of ZERO

- Determine whether a sample correlation coefficient could have come from a population with a parametric correlation coefficient of CERTAIN VALUE $\neq 0$

- Generic recipe for Hypothesis Testing

Hypothesis Testing --- Generic Recipe



Hypothesis Testing --- Generic Recipe

State population

All measurements on total length of aphid stem mothers & mean thorax length of their parthenogenetic offspring made by the same experimental protocol

1). Randomly sampled
2). Same environmental conditions

Hypothesis Testing --- Generic Recipe

State population

State model/measure of pattern (statistic)

- Correlation of the two variables, ρ
- In the case $H_0 : \rho = 0$

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$\left(r = \hat{\rho}, \ \sim \begin{cases} 1)\ N\left(0, \frac{1 - r^2}{n - 2}\right), \text{if } n \text{ LARGE} \\ 2)\ t - \text{distribution}, df = n - 2, \text{otherwise} \end{cases} \right)$$

- In the case $H_0 : \rho = \rho_1 \quad (\rho_1 \neq 0)$

$$\eta = \frac{1}{2} \ln\left(\frac{1 + \rho_1}{1 - \rho_1}\right)$$

$$t = \frac{z - \eta}{1 / \sqrt{n - 3}}$$

$$\left(\text{where } z = \frac{1}{2} \ln\left(\frac{1 + r}{1 - r}\right), E(z) = \eta, \operatorname{var}(z) = \frac{1}{n - 3} \right)$$

z: Normal/tends to normal rapidly as n increases for $\rho \neq 0$

t-statistic: N(0, 1) or t (df = ∞)

Hypothesis Testing --- Generic Recipe

State population

↓

State model/measure of pattern (statistic)

- Correlation of the two variables, ρ
- In the case $H_0 : \rho = 0$

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$\left(r = \hat{\rho}, \sim \begin{cases} 1)\ N\left(0, \frac{1-r^2}{n-2}\right), if\ \ n\ LARGE \\ 2)\ t - distribution, df\ = n-2 \end{cases} \right)$$

Hypothesis Testing --- Generic Recipe

State population

State model/measure of pattern (statistic)

State null hypothesis

$$H_0 : \rho = 0$$

Hypothesis Testing --- Generic Recipe

State population

State model/measure of pattern (statistic)

State null hypothesis

State alternative hypothesis

$$H_A : \rho \neq 0$$

Hypothesis Testing --- Generic Recipe

- State population
- State model/measure of pattern (statistic)
- State null hypothesis
- State alternative hypothesis
- State tolerance for Type I error
  - $\alpha = 5\%\ (conventional\ level)$

Hypothesis Testing --- Generic Recipe

- State population
- State model/measure of pattern (statistic)
- State null hypothesis
- State alternative hypothesis
- State tolerance for Type I error
- State frequency distribution
  - t-distribution

Hypothesis Testing --- Generic Recipe

State population
↓
State model/measure of pattern (statistic)
↓
State null hypothesis
↓
State alternative hypothesis
↓
State tolerance for Type I error
↓
State frequency distribution
↓
Calculate statistic

- t-statistic
- correlation coefficient estimate, $r = 0.65$
- $t = (0.65 - 0)/0.21076 = 3.084$

Hypothesis Testing --- Generic Recipe

- State population
- ↓ State model/measure of pattern (statistic)
- ↓ State null hypothesis
- ↓ State alternative hypothesis
- ↓ State tolerance for Type I error
- ↓ State frequency distribution
- ↓ Calculate statistic
- ↓ Calculate p-value
  - $t = 3.084$, $df = 13$
  - $p = 0.0044$ (one-tail) & 0.0088 (two-tail)

Hypothesis Testing --- Generic Recipe

- State population
- State model/measure of pattern (statistic)
- State null hypothesis
- State alternative hypothesis
- State tolerance for Type I error
- State frequency distribution
- Calculate statistic
- Calculate p-value → Declare decision
  - $p = 0.0088 < \alpha = 0.05$
  - reject H_0
  - accept $H_A : \rho \neq 0$

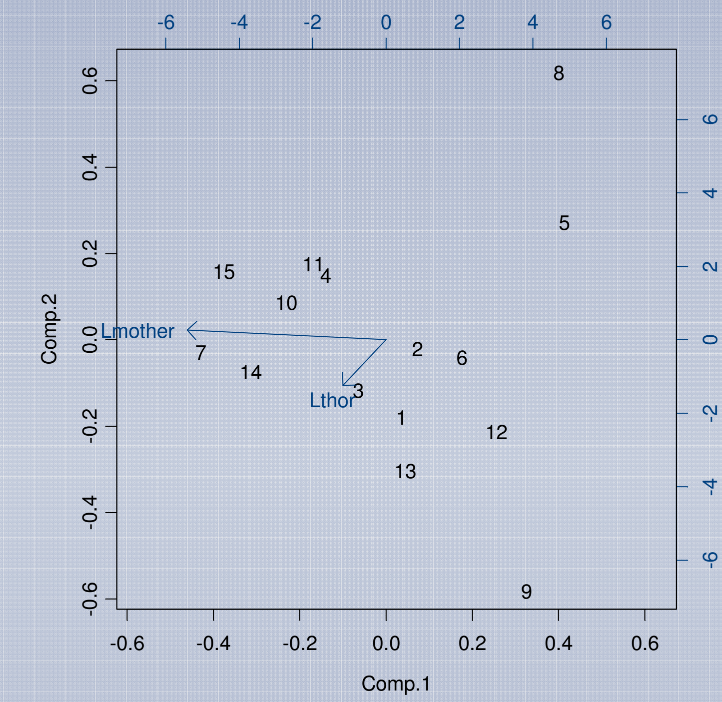
Hypothesis Testing --- Generic Recipe

1. State population
2. State model/measure of pattern (statistic)
3. State null hypothesis
4. State alternative hypothesis
5. State tolerance for Type I error
6. State frequency distribution
7. Calculate statistic
8. Calculate p-value
9. Declare decision
10. Report statistic with decision

- r = 0.65, n = 15, p = 0.0088
- Total length & offspring thorax length are related

Pearson's Correlation – Assumptions

- Assumptions
- Normal & independent errors
- Homogeneous around straight line

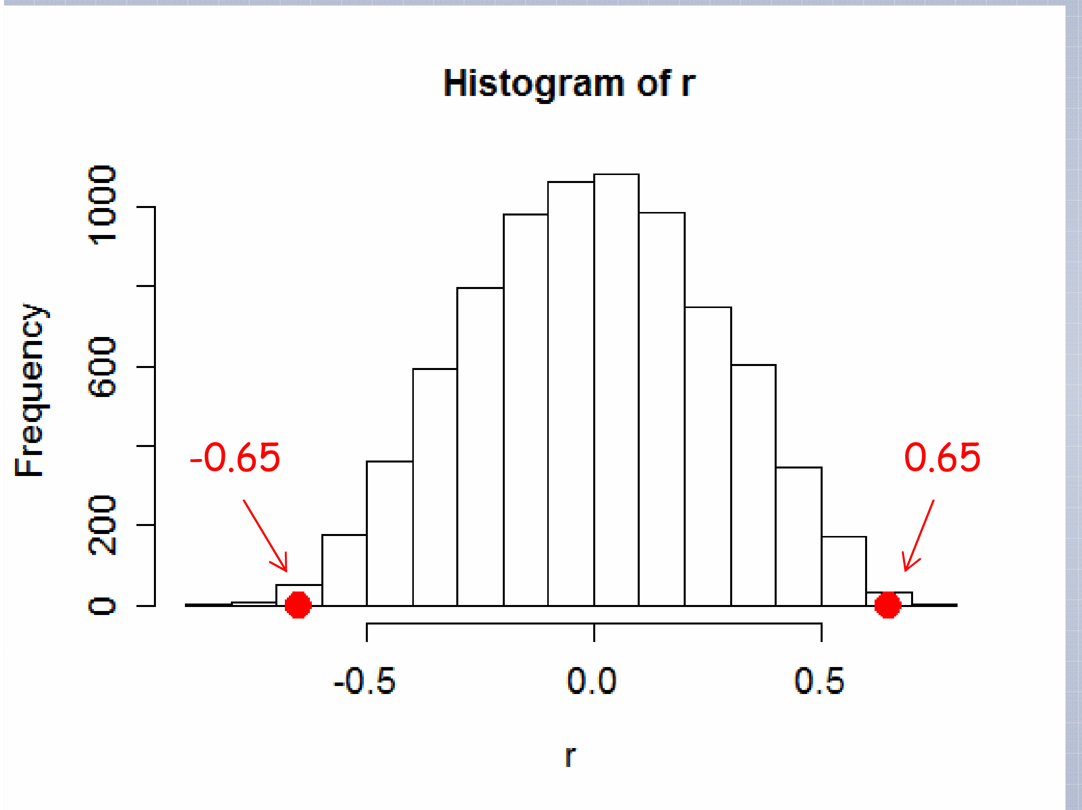


- What if assumptions for Pearson test not met?
- Here are the observations relative to the correlation line (comp 1)
- Not homogeneous, due to outliers (observations 8 & 9)

Pearson's Correlation – Randomization test

- Significance test with no distributional assumptions
- Hold one variable, permute the other one many times
- A new r from each new permutation
- Construct empirical frequency distribution
- Compare the empirical distribution with the observed r

Pearson's Correlation – Randomization test



- 8000 times
- p1 = p(r > 0.65) = 0.001875
- p2 = p (r < -0.65) = 0.003875
- p = p1 + p2 = 0.00575 < α = 0.05
- Reject Null, accept alternative
- Consistent with testing result from theoretical t-distribution, for this data

Pearson's Correlation coefficient - Confidence Limit

- 95% confidence limit (tolerance of Type I error @ 5%)
- t-distribution (df = n - 2)   (NO)

a). H0: ρ = 0 was rejected

b). Distribution of r is negatively skewed

c). Fisher's transformation

- $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right); \quad \frac{z-\eta}{1/\sqrt{n-3}} \sim N(0,1)\ or\ t_{[\infty]}$

$$\eta = \frac{1}{2}\ln\left(\frac{1+\rho_1}{1-\rho_1}\right)$$

Pearson's Correlation coefficient – Confidence Limit

C. I. for *η*:

- $$\begin{cases} z_l = z - z_{(1-\alpha/2)} \cdot \sqrt{1/(n-3)} \\ z_u = z + z_{(1-\alpha/2)} \cdot \sqrt{1/(n-3)} \end{cases}$$, $z_{(1-\alpha/2)}$, critical value from N(0, 1) at p = 1-α/2

C. I. for *ρ*:

- $$\begin{cases} r_l = \tanh(z_l) = \dfrac{\exp(2z_l)-1}{\exp(2z_l)+1} \\ r_u = \tanh(z_u) = \dfrac{\exp(2z_u)-1}{\exp(2z_u)+1} \end{cases}$$

For our example:

***95 percent confidence interval*:**

$r_l = 0.207$

$r_u = 0.872$

Nonparametric: Spearman's Rho

- Measure of monotone association used when the distribution of the data make Pearson's correlation coefficient undesirable or misleading
- Spearman's correlation coefficient (Rho) is defined as the Pearson's correlation coefficient between the ranked variables
- $$Rho = \frac{\sum_i (y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)}{\sqrt{\sum_i (y_{1i} - \bar{y}_1)^2 \sum_i (y_{2i} - \bar{y}_2)^2}}, \quad where\ y_{1i},\ y_{2i}\ are\ ranks\ of\ Y_{1i}, Y_{2i}$$
- $$If\ no\ ties,\ Rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)},\ where\ d_i = y_{1i} - y_{2i}$$
- Randomization test for significance (option)

Nonparametric: Kendall's Tau

- Concordant pairs (Y_{1i}, Y_{2i}) *and* (Y_{1j}, Y_{2j}):

  $$If\ Y_{1i} > Y_{1j}\ and\ Y_{2i} > Y_{2j}\ or\ if\ Y_{1i} < Y_{1j}\ and\ Y_{2i} < Y_{2j}$$

  (if the ranks for both elements agree)

- Discordant pairs (Y_{1i}, Y_{2i}) *and* (Y_{1j}, Y_{2j}):

  $$If\ Y_{1i} > Y_{1j}\ and\ Y_{2i} < Y_{2j}\ or\ if\ Y_{1i} < Y_{1j}\ and\ Y_{2i} > Y_{2j}$$

  (if the ranks for both elements disagree)

- Neither concordant or discordant

  $$If\ Y_{1i} = Y_{1j}\ or\ Y_{2i} = Y_{2j}$$

Nonparametric: Kendall's Tau

- Kendall's Tau =

$$\begin{cases} \dfrac{n_c - n_d}{\frac{1}{2}n(n-1)} & \text{(no ties)} \\ \\ \dfrac{n_c - n_d}{n_c + n_d} & \text{(in the case of ties)} \end{cases}, \; where\; n_c = number\ of\ concordant\ pairs$$

$$n_d = number\ of\ discordant\ pairs$$

Gamma coefficient or Goodman correlation coefficient

Properties:

- The denominator is the total number of pairs, $-1 \leq tau \leq 1$
- tau = 1, for perfect ranking agreement
- tau = -1, for perfect ranking disagreement
- tau ≈ 0, if two variables are independent
- For large samples, the sampling distribution of tau is approximately normal

Nonparametric

For more information on nonparametric test of correlation

e.g., significance test, etc.

References:

- Conover, W.J. (1999) "Practical nonparametric statistics", 3rd ed. Wiley & Sons
- Kendall, M. (1948) "Rank Correlation Methods", Charles Griffin & Company Limited
- Caruso, J. C. & N. Cliff. (1997) "Empirical Size, Coverage, and Power of Confidence Intervals for Spearman's Rho", Ed. and Psy. Meas., 57 pp. 637–654
- Corder, G.W. & D.I. Foreman. (2009) "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach", Wiley

Data Total length of aphid stem mothers (Y1) Vs. Mean thorax length of their parthenogenetic offspring (Y2)

#	Y_1	Y_2
1	8.7	5.95
2	8.5	5.65
3	9.4	6.00
4	10.0	5.70
5	6.3	4.70
6	7.8	5.53
7	11.9	6.40
8	6.5	4.18
9	6.6	6.15
10	10.6	5.93
11	10.2	5.70
12	7.2	5.68
13	8.6	6.13
14	11.1	6.30
15	11.6	6.03

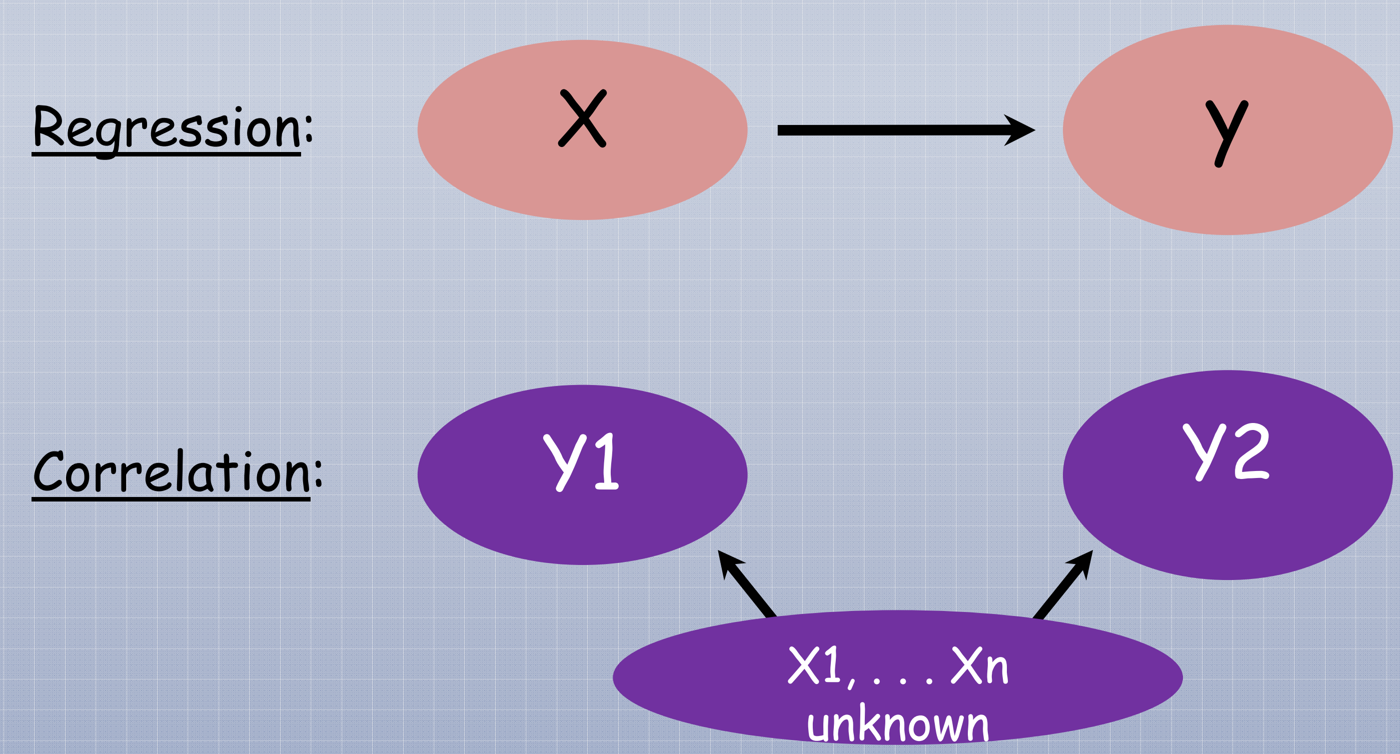
Total length of mothers Vs. Mean thorax length of offspring

	RAW		RANK	
#	Y_1	Y_2	y_1	y_2
1	8.7	5.95	8	9
2	8.5	5.65	6	4
3	9.4	6.00	9	10
4	10.0	5.70	10	6.5
5	6.3	4.70	1	2
6	7.8	5.53	5	3
7	11.9	6.40	15	15
8	6.5	4.18	2	1
9	6.6	6.15	3	13
10	10.6	5.93	12	8
11	10.2	5.70	11	6.5
12	7.2	5.68	4	5
13	8.6	6.13	7	12
14	11.1	6.30	13	14
15	11.6	6.03	14	11

Group Activity

Activity Instructions

- Question: **REGRESSION or CORRELATION?**
- Justification guideline:



Activity Instructions

- Form small groups or 2-3 people.
- Each group is assigned a number
- Group members work together on each example for 5 minutes, come up with an answer & your justifications
- A number will be randomly generated from the group #'s
- The corresponding group will have to present their answer & justifications
- Go for the next example . . .

Activity Instructions

There is NO RIGHT/WRONG ANSWER (for these examples), as long as your justifications are LOGICAL

Example 1

Height and ratings of physical attractiveness vary across individuals. Would you analyze this as regression or correlation?

Subject	Height	Phy
1	69	7
2	61	8
3	68	6
4	66	5
5	66	8
.	..	.
48	71	10

Example 2

Airborne particles such as dust and smoke are an important part of air pollution. Measurements of airborne particles made every six days in the center of a small city and at a rural location 10 miles southwest of the city

(Moore & McCabe, 1999. Introduction to the Practice of Statistics).

Would you analyze this relation as regression or correlation?

Example 3

A study conducted in the Egyptian village of Kalama examined the relation between birth weights of 40 infants and family monthly income

(El-Kholy et al. 1986, Journal of the Egyptian Public Health Association, 61: 349).

Would you analyze this relation as regression or correlation?