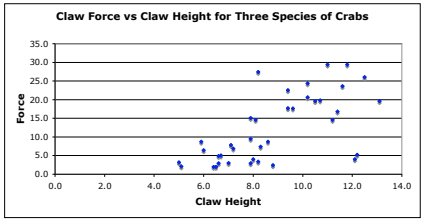
Correlation and Covariance

Dr Tom Ilvento
Department of Food and Resource Economics

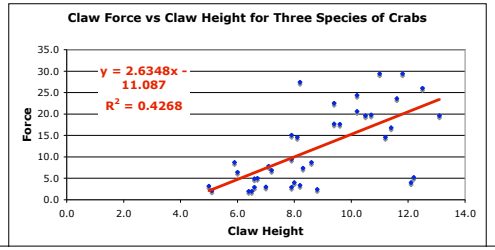
What is Next?
Correlation and Regression

- **Correlation**
  - A measure of association between two variables
  - Expressed as a linear relationship
  - Based on the Co-variance - how two variables vary about their means together
  - Can be shown in a visual way via a scatterplot



What is Next?
Correlation and Regression

- **Regression**
  - We specify a dependent variable as a function of one or more independent variables, as a linear function and based on co-variance
  - Regression provides estimates of the relationship between the dependent variable and the independent variable(s)
  - The estimates, called coefficients, can be based on a sample and can be tested via a hypothesis test or confidence interval



Correlation and Regression

- **A focus on the variance**

$$\sum (X - \bar{X})^2 = \text{TSS Total Sum of Squares Deviations}$$

$$\sum \frac{(X - \bar{X})^2}{n-1} = \text{MS Mean Squared Deviation}$$

- **A focus on the co-variance**

$$Cov_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

- **A focus on the equation of a line**
  - Y = a + b*X where a is the intercept and b is the slope

A note about Measures of Association

- It is useful to have a summary measure that represents the relationship of one variable to another
- We call these **Measures of Association**
- There are many used across many fields
  - Conditional probability
  - Odds Ratio
  - Correlation coefficient
  - Regression coefficient
  - R-square

Measures of Association - Properties

- A measure of Association should express
  - Direction of the association - positive or negative
  - Strength of the relationship
- There are many properties of Measures of Association that help define how useful they are, and how we can interpret them. We should ask:
  - Is it bounded with an upper and lower limit?
  - If so, what is the range?
  - Is it symmetrical?
  - How to interpret it within the range

Let's revisit the Variance

- We have been interested in how a variable varies about its mean
- We represented this as the Variance - the **Mean Squared Deviation**

$$\sum (X - \overline{X})^2 = \text{TSS Total Sum of Squares Deviations}$$

$$\sum \frac{(X - \overline{X})^2}{n-1} = \text{MS Mean Squared Deviation}$$

The Co-Variance

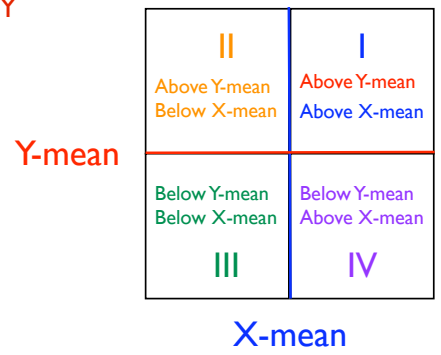
- The **Covariance** looks at how two variables, X and Y, vary about their means together
- We express it as an average, divided by n (not n-1)

$$Cov_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n} \qquad Cov_{XY} = \frac{SS_{XY}}{n}$$

The covariance is a basic building block of correlation, regression, and the General Linear Model

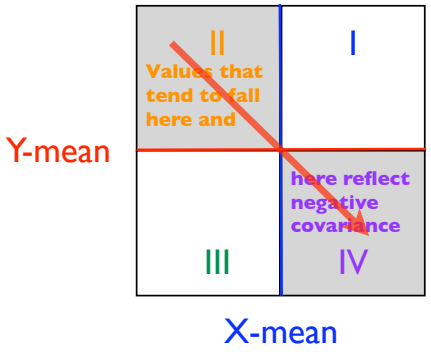
Basics of Co-Variance

- Let's start with a basic graph of a Y-variable vs an X variable.
- I will dissect the graph with the mean of X and the Mean of Y



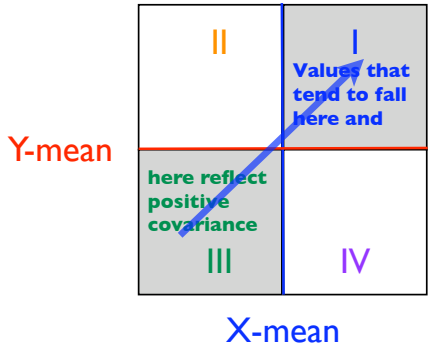
Basics of Co-Variance

Negative Covariance



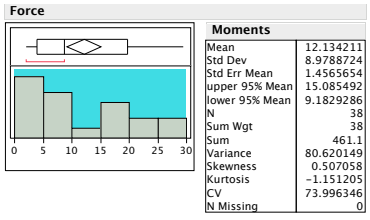
Basics of Co-Variance

Positive Covariance

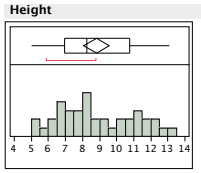


Crab Data Example

- This is some data on three species of crabs
- The key variables we will focus on are FORCE of the crab claw and the HEIGHT of the claw
- Here are JMP's summary statistics on both of these variables



Moments	
Mean	12.134211
Std Dev	8.9788724
Std Err Mean	1.4565654
upper 95% Mean	15.085492
lower 95% Mean	9.1829286
N	38
Sum Wgt	38
Sum	461.1
Variance	80.620149
Skewness	0.507058
Kurtosis	-1.151205
CV	73.996346
N Missing	0

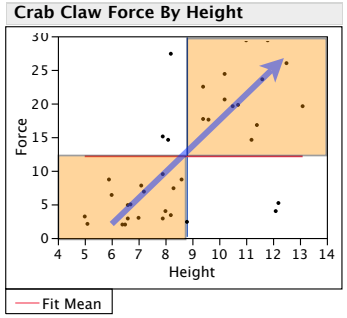


Moments	
Mean	8.8131579
Std Dev	2.2262769
Std Err Mean	0.3611498
upper 95% Mean	9.5449169
lower 95% Mean	8.0813989
N	38
Sum Wgt	38
Sum	334.9
Variance	4.9563087
Skewness	0.208142
Kurtosis	-1.013624
CV	25.26083
N Missing	0

	Mean	Variance
FORCE	12.134	80.620
HEIGHT	8.313	4.956

Crab Data

- Most of the data points fall into quadrants I and III
  - **Positive co-variance**
  - As the crab HEIGHT increase, so does crab FORCE



Covariance Matrix

	Force	Height
Force	80.62015	13.05900
Height	13.05900	4.95631

	Mean	Variance
FORCE	12.134	80.620
HEIGHT	8.313	4.956

Shortcomings of Co-Variance

- The covariance between two variables is a useful concept – it is the building block for regression and other multivariate techniques
- But as a measure of association it has limits
  - It is symmetrical - not a bad thing
  - It is unbounded – unknown high or low
  - It is difficult to determine what the represents - a lot? a little? just how much????
  - Expressed in awkward cross-product units

Covariance Matrix

	Force	Height
Force	80.62015	13.05900
Height	13.05900	4.95631

Pearson Correlation Coefficient - r

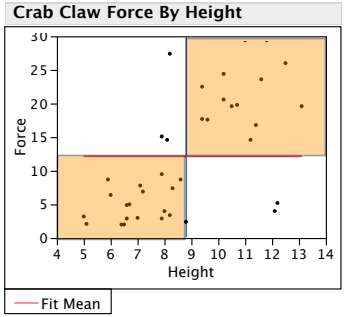
- The correlation coefficient (r) is the co-variance adjusted for the standard deviations of both variables
- The adjustment is simple, and it makes it so much easier to interpret

$$r = \frac{Cov_{XY}}{s_X s_Y}$$

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}} \qquad r = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}}$$

Crab Data

- Most of the data points fall into quadrants I and III
  - **Positive co-variance**
  - As the crab HEIGHT increase, so does crab FORCE



13.059/(80.62015*4.95631)$^{.5}$ = .6533

13.059/(8.9789*2.2263) = .6533

Covariance Matrix

	Force	Height
Force	80.62015	13.05900
Height	13.05900	4.95631

Correlation Coefficient r

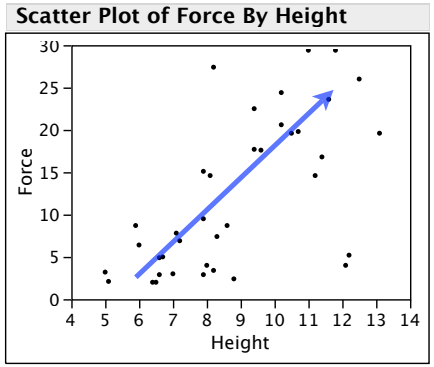
- **Properties of r**
  - Based on a linear measure of association
  - Bounded between -1 and 1
  - Symmetrical relationship: $r_{XY} = r_{YX}$
  - Easier to interpret
  - Invariant to linear scaling
    - add/subtract or multiply/divide by a constant does not change the value of r between two variables
    - Example: The correlation between the respondent's education and income does not change if you express income in total dollars or per $1000

Interpretation of r

- **The closer the correlation is to 1:**
  - the more perfect positive linear relationship
  - If r = 1 then all values would fall on a straight line, upward slope
- **The closer the correlation is to –1:**
  - The more perfect negative linear relationship
  - If r = -1 then all values would fall on a straight line, downward slope
- **The scatterplot is a visual depiction of the correlation coefficient**

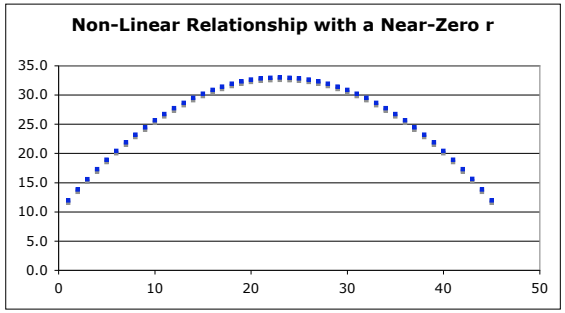
Scatter Plot Of Crab Force by Height

- The correlation is **.6533**, a moderately strong positive correlation



Interpretation of r

- 0 means **no linear relationship**



Interpretation of r

- **One other interesting interpretation of r**
- The square of r → **r^2**
- can be interpreted as the percent of variability in one variable that is "explained" by the other variable
- In the two-variable case, r^2 is equal to R-square, a measure of association in Regression
  - Only in the case of a bivariate regression - one independent variable
  - And it moves us toward defining one variable as explaining the other

Interpreting a correlation coefficient: Rules of Thumb for Narratives

- The following is a table giving guideline for narratives involving correlations. For simplicity sake, the table is based on the absolute value of the correlation (|r|)
- And the exact description depends upon the subject and discipline

Correlation Range	Percent Variability Explained (r^2)	Description
.00 to .33	0 to 10%	Weak
.34 to .49	11% to 24%	Moderate
.50 to .75	25% to 56%	Moderately Strong
.76 to 1.00	57% to 100%	Strong

Summary

- Covariance is the basic building block for more advanced statistical techniques
- It is an extension of the variance, now including how two variables vary together about their means
- Correlation is a re-expression of the covariance so that it is bounded and more easy to interpret
- Correlation and covariance are both Measures of Association, which show how two variables are related to each other
- Correlation can be visually represented via a scattergram.