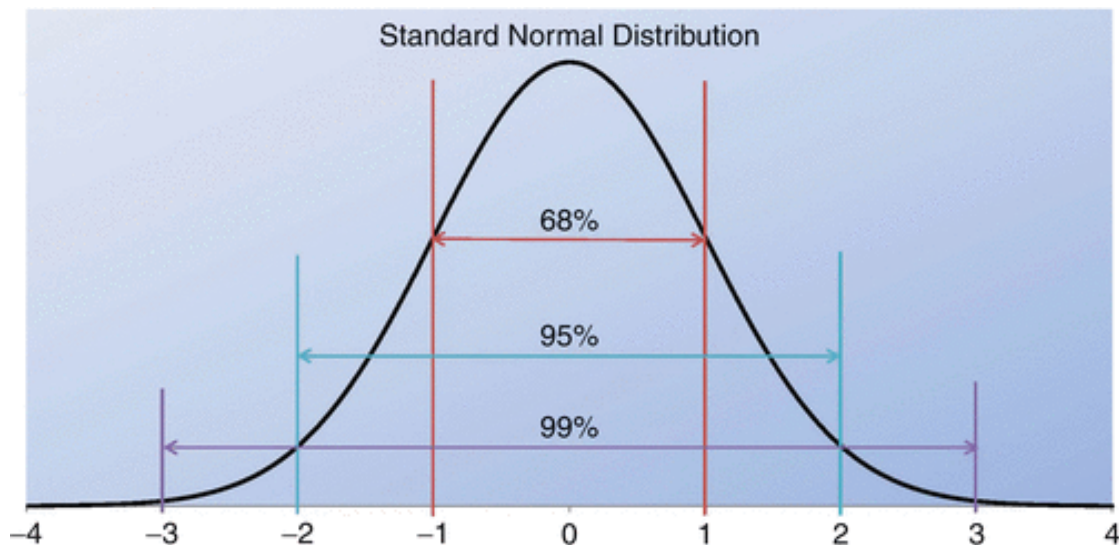


Module 3a – Scatterplots and Correlation

Reviewing Module 2

- **Normal Distributions**

- Example 3.4 on page 85
- Applying the 68/95/99.7 Rule (label it with values)
- Use z-score/standard score formula to find z (what does this tell us about a woman who is 60" tall?)
- Use Table A to calculate a percentile (What does this tell us?)



Stat Procedure Diagram – Where we are

		Descriptive Statistics (Describing Pops or Samples)		Inferential Statistics (from Samples)
	Variable Types	Display	Describe	Estimation
Univariate	categorical (nominal or ordinal)*	Bar Graph/Pie Chart	Counts/Percentages	When binary/dichotomous: Confidence interval for proportions
	quantitative/continuous	Histogram/Stem & Leaf Box Plot	Mean/St Dev (normal) Median/Min, Q1, Q3, Max (skewed)	Confidence interval for means
		Display	Describe	Significance Tests/Hypothesis Tests
Bivariate	2 categorical	Tables or Bar Graphs	Two-way tables/Crosstabulation	Chi-square test (for goodness of fit)
	1 categorical, 1 quant.	Bar Graphs	Comparison of means/averages	T-test (one sample/group, two samples/groups) ANOVA (two or more samples/groups)
	2 quant.	Scatterplot	Correlation Coef. (Coef. of determ)/ Regression Line	T-test for correlation
		Display	Describe	Significance Tests/Hypothesis Tests
Multivariate	Response Variable is Quant.	-	Ordinary Least Squares Regression (OLS)	F-test for overall model T-tests for each explanatory variable
	Response Variable is categorical (dichotomous)	-	Logistic Regression	Chi-square tests of significance

NOTE: Items highlighted in yellow are covered in this course.

*When a categorical variable has two categories, it is called dichotomous.

Scatterplot – Displays Relationship between two quant/continuous variables

- Watch Against all Odds, Unit 10, Unit 11, and Unit 12 (recommended)
- Scatterplots
 - Explanatory/Independent variable goes on x-axis
 - Response/Dependent goes on y-axis
- Scatterplots tell us three things about the relationship:
 - 1. The **Form**: (Linear/Curvilinear/Diffuse)
 - 2. The **Direction**: (Positive or Negative)
 - 3. The **Strength**: (Weak/Strong – how close are the points to forming a straight line)

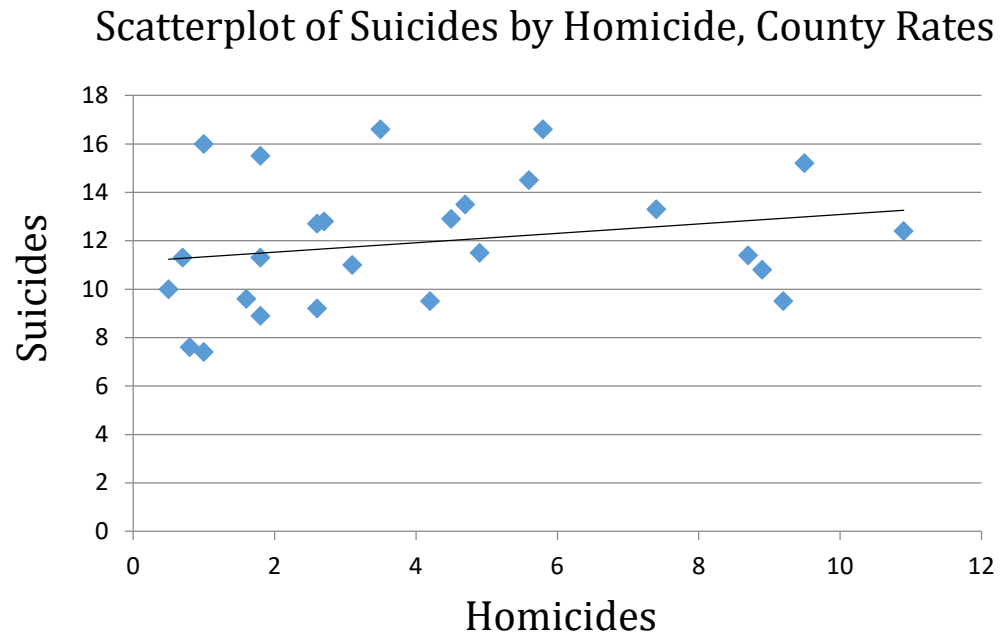
Example 4.4 on Page 104

- Two Quantitative variables: Rate of Homicide and Rate of Suicide
- What/Who are the “individuals” in this problem?

	Homicide Rate	Suicide Rate
County 1	4.2	9.5
County 2	1.8	15.5
County 3	2.6	12.7
County 4	1	16
County 5	5.6	14.5
County 6	3.5	16.6
County 7	9.2	9.5
County 8	0.8	7.6
County 9	8.7	11.4
County 10	2.7	12.8
County 11	8.9	10.8
County 12	1.8	11.3
County 13	4.5	12.9
County 14	3.1	11
County 15	7.4	13.3
County 16	10.9	12.4
County 17	0.5	10
County 18	2.6	9.2
County 19	9.5	15.2
County 20	1.6	9.6
County 21	4.7	13.5
County 22	4.9	11.5
County 23	5.8	16.6
County 24	0.7	11.3
County 25	1.8	8.9
County 26	1	7.4

Correlation Coefficient

0.23

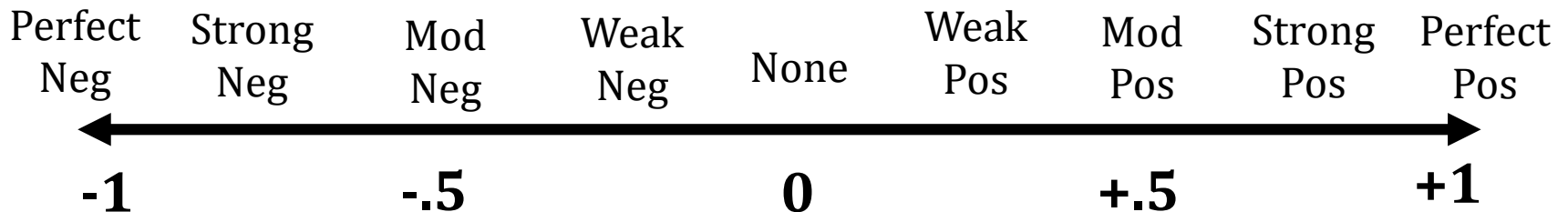


- Form?
- Direction?
- Strength?

Correlation Coefficient – Describes Relationship between two quant/continuous variables (standardized)

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

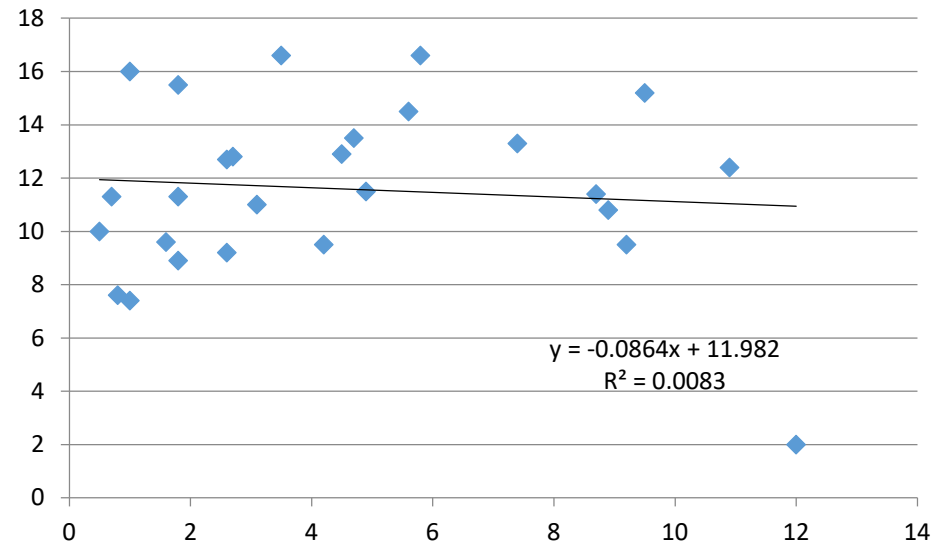
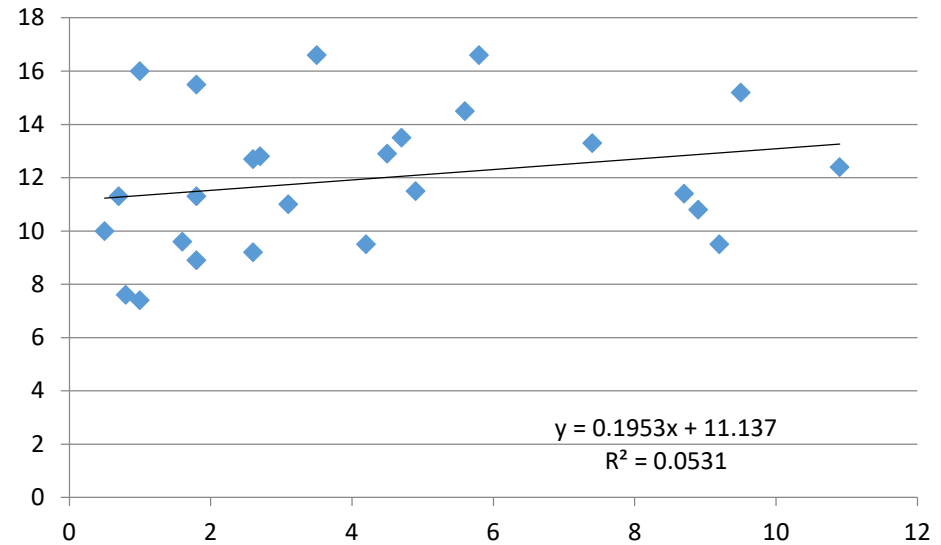
- Correlation coefficient or Pearson's R
 - Tells us **strength** and **direction** of association
 - Range: -1 to +1 (perfect neg to perfect pos association/correlation)
 - Movie preference exercise for out of class activity
 - (I will place you in pairs)



Beware the effect of outliers

	Homicide Rate	Suicide Rate
County 1	4.2	9.5
County 2	1.8	15.5
County 3	2.6	12.7
County 4	1	16
County 5	5.6	14.5
County 6	3.5	16.6
County 7	9.2	9.5
County 8	0.8	7.6
County 9	8.7	11.4
County 10	2.7	12.8
County 11	8.9	10.8
County 12	1.8	11.3
County 13	4.5	12.9
County 14	3.1	11
County 15	7.4	13.3
County 16	10.9	12.4
County 17	0.5	10
County 18	2.6	9.2
County 19	9.5	15.2
County 20	1.6	9.6
County 21	4.7	13.5
County 22	4.9	11.5
County 23	5.8	16.6
County 24	0.7	11.3
County 25	1.8	8.9
County 26	1	7.4
County 27 (outlier)	12	2

Correlation Coefficient without outlier 0.23
 Correlation Coefficient with outlier -0.09



Coefficient of Determination (R-Squared)

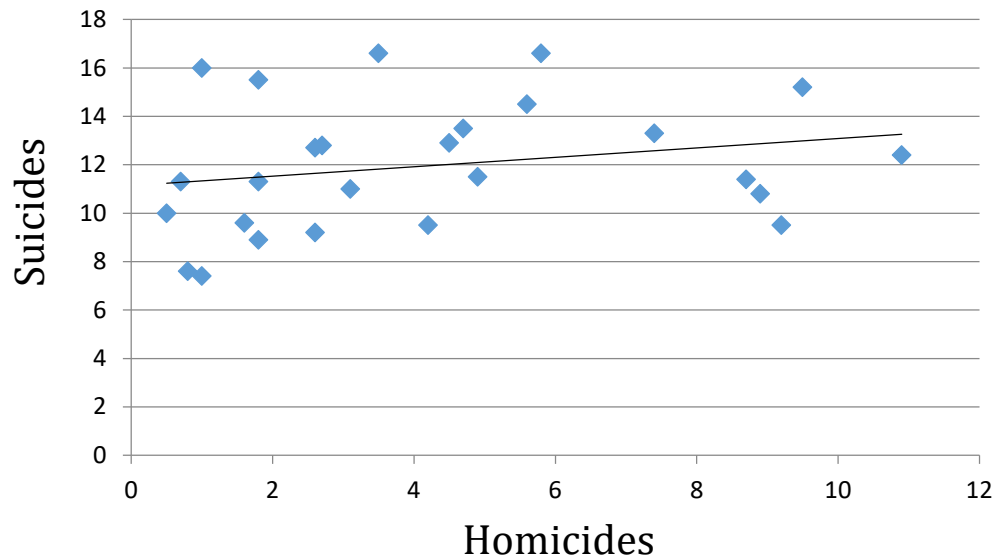
- Literally, the correlation coefficient value squared, it tells us the “percent of the response/dependent variable that can be explained by the explanatory/independent variable.”

	Homicide Rate	Suicide Rate
County 1	4.2	9.5
County 2	1.8	15.5
County 3	2.6	12.7
County 4	1	16
County 5	5.6	14.5
County 6	3.5	16.6
County 7	9.2	9.5
County 8	0.8	7.6
County 9	8.7	11.4
County 10	2.7	12.8
County 11	8.9	10.8
County 12	1.8	11.3
County 13	4.5	12.9
County 14	3.1	11
County 15	7.4	13.3
County 16	10.9	12.4
County 17	0.5	10
County 18	2.6	9.2
County 19	9.5	15.2
County 20	1.6	9.6
County 21	4.7	13.5
County 22	4.9	11.5
County 23	5.8	16.6
County 24	0.7	11.3
County 25	1.8	8.9
County 26	1	7.4

Correlation Coefficient

0.23

Scatterplot of Suicides by Homicide, County Rates



- ***Correlation coefficient***=.23
- ***Coefficient of determination***=.053 (5.3%)
- 5.3% of var in suicide rate can be explained by variation in homicide rate