Data Organization (Chapter 3) Date: DAY 1

Standards/Objectives:
1) understand **bivariate data** and their **linear trends**

Agenda/Activities:
1A) HANDOUT - Chapter 3 Key/Typical Questions
1B) VIDEO - World Population
1C) ACTIVITY - GPA vs. Study Time
1D) The Correlation Coefficient
1E) PWRPT - Vocabulary
1F) practice answering Key Questions
1G) Ministers vs. Rum

1A) Chapter 3 Key/Typical Questions for Bivariates → 2 variables (examine the relationship between 2 variables)

1. What does "r" mean (in context of the problem)? What is its name?

There is a *strong/mild/weak* (*strength*) *positive/negative* (*direction*) **linear** relationship between . . .

This is the **correlation coefficient**.

2. What does r^2 mean (in context of the problem)? What is its name?

???% (*r^2 value*) **of the variability of** *response variable definition* (*Y*) is due to *explanatory variable definition* (*X*).

This is the **coefficient of determination**.

3. What is the equation of the least squares regression line? Define all variables.

$\hat{y} = a + bx$, where $\hat{y}$ is the **predicted** *response variable definition* (*Y*) and x is *explanatory variable definition* (*X*).

4. Place a value in X to produce a Y.

5. Place a value in Y to produce an X.

6. What does the slope mean (in context of the problem)? What is its value?

For every *explanatory variable def.* (*X*), the *response variable def.* (*Y*) is **predicted** to *increase/decrease* (*direction*) by *???* (*slope value*).

7. What is the formula that involves slope, correlation, and standard deviation?

$b_1 = r\frac{s_y}{s_x}$, where b_1 = slope, r = correlation coefficient, s_y and s_x are the standard deviations of the y's and x's

8. What does resistant and non-resistant mean? Name things that are resistant / non-resistant.

Resistant → outliers **WILL NOT** greatly affect the value.

Non-resistant → outliers **WILL** greatly affect the value.

9. Is the line a good fit for the data? Justify your answer.

1) Justify using the **residual plot**, should be **randomly scattered**. 2) Justify using r^2.

10. What is the meaning of least squares?

Minimizes the **sum of residuals** (vertical distances) **squared** of all observations from the **prediction line**.

11. What does an outlier influence in the problem?

Artificially reduces the value of r.

12. What does an influential point do to the problem?

Artificially inflates the value of r.

13. What does "s" mean on the Minitab printout?

The root square mean error, which is the **standard deviation of the residuals**.

In other words, **your predicted** *response variable def.* (*Y*) **is typically off by** *value* (*s*).

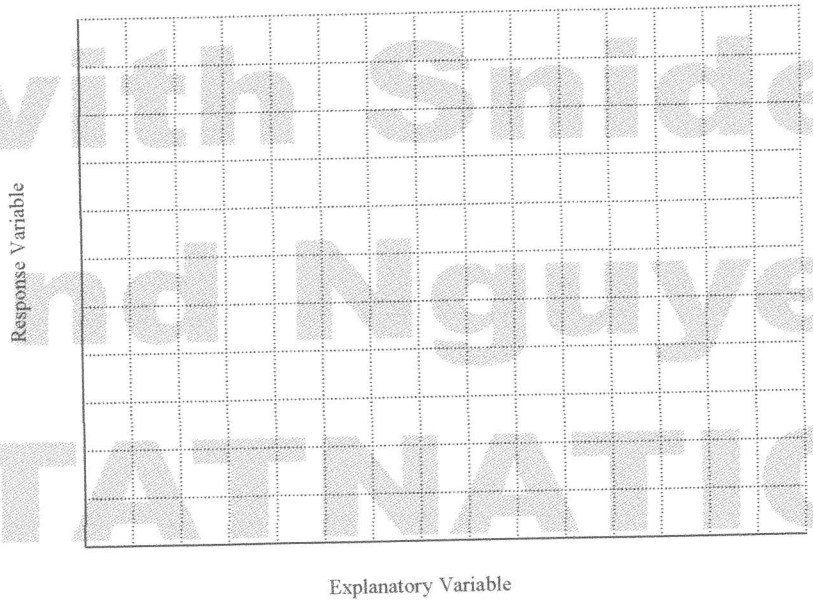
RECALL: Explanatory variable = Response variable =

1B) VIDEO - World Population (explanatory variable = , response =)

1C) ACTIVITY - GPA vs. Study Time
Make scatterplot (label variables):

Grade Point Average and Study Time

Student	Hours per Night	GPA
Johnny	3.8	3.8
Sue	4.8	3.9
Phat	0.8	2.5
Richard	0.5	1.5
Jose	3.2	2.5
Jeannie	5.2	3.7
Phyllis	3.8	2.4
Jane	3.5	3.6
Juan	3.9	4
Timmy	3.2	3.8
Annie	6.2	4.5



Examine the relationship between these two variables.

1) **direction** of association: _____

2) **strength** of association: _____

3) **form** of association: _____

What is:

- **regression analysis**?
  Regression Line – mathematical model that _____

- the meaning/purpose of **LSRL** (Least Squares Regression Line)?
  line that _____

  $\hat{y} = a + bx$ (predicted response $\hat{y}$ for any x)

  $a = \bar{y} - b\bar{x}$

  $b = r\frac{s_y}{s_x}$ a change in *1* std dev in x corresponds to a change of *r* std dev in y

  error = _____

- a **residual**?
  the difference between _____

  residual = _____

- a **correlation coefficient**?
  <<see below>>

- a **coefficient of determination**?
  the proportion of the _____

  $r^2 = \frac{SSM - SSE}{SSM}$, where SSM = Sum of Squares about the Mean $\bar{y} = \sum(y - \bar{y})^2$

  SSE = Sum of Squares for Error = $\sum(y - \hat{y})^2$

1D) The **Correlation Coefficient**

Key Properties:

1) strength: $|r| \approx 1$ indicates strong association; $|r| \approx 0$ indicates weak/no assocation

2) direction: $r > 0$ indicates positive association; $r < 0$ indicates negative association

r: −1 ——— 0 ——— 1

Strong Negative association | Weak/ No | Strong Positive association

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

*r ≠ slope, just strength of linearity
*r does not indicate any "curved" relationship
*r doesn't change if there's a change in units of measurements (ex: feet into meters)
*r is strongly affected by *outliers*

1E) PWRPT - Vocabulary

1F) practice answering Typical Questions

1G) Ministers vs. Rum

Data Organization (Chapter 3) Date: DAY 2

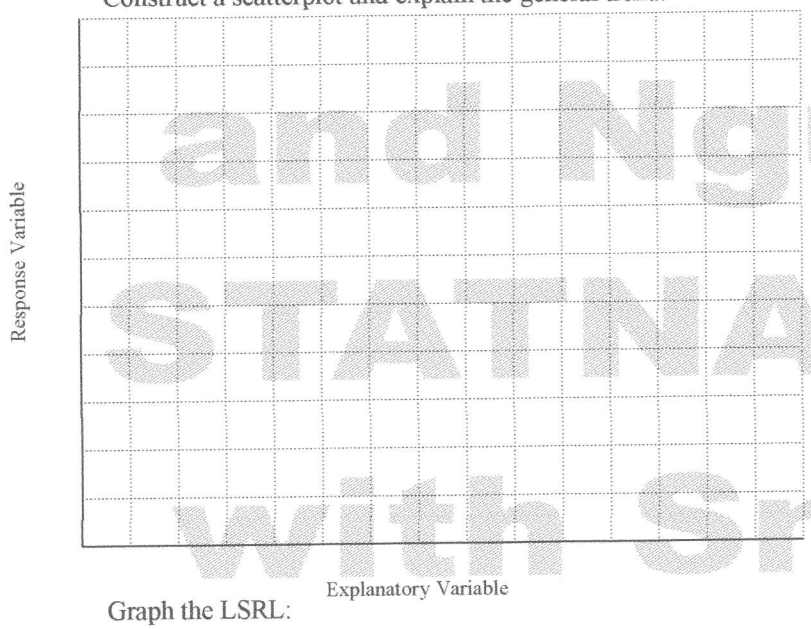
Standards/Objectives:
1) practice **typical questions** regarding bivariate data
2) discussion on **residual plots**

Agenda/Activities:
1A) JMP: TV vs. Life Expectancy
2A) Scatterplots and Linear Regressions

1A) JMP: TV vs. Life Expectancy

Identify the explanatory variable: _____ Identify the response variable: _____

Construct a scatterplot and explain the general trend:



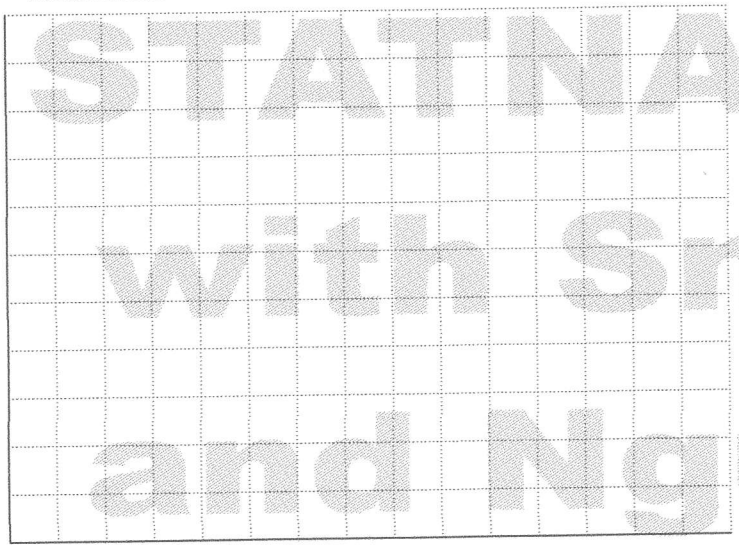
Graph the LSRL:

Typical Questions:

1. *What does "r" mean (in context of the problem)? What is its name?*
2. *What does r^2 mean (in context of the problem)? What is its name?*
3. *What is the equation of the least squares regression line? Define all variables.*
4. *Place a value in X to produce a Y.*
5. *Place a value in Y to produce an X.*
6. *What does the slope mean (in context of the problem)? What is its value?*
7. *What is the formula that involves slope, correlation, and standard deviation?*
8. *What does resistant and non-resistant mean? Name things that are resistant / non-resistant.*
9. *Is the line a good fit for the data? Justify your answer.*
10. *What is the meaning of least squares?*
11. *What does an outlier influence in the problem?*
12. *What does an influential point do to the problem?*
13. *What does "s" mean on the Minitab printout?*

2A) Scatterplots and Linear Regressions - Graph the scatterplot, then calculate the linear regression and graph the LSRL.

Diet (weeks)	X	1	3	5	9	10
Pounds Lost	Y	6	8	11	9	13



FUN FACT: "r", the correlation coefficient, has no units of measure/dimension and thus is unchanged if x and y are interchanged

Data Organization (Chapter 3) Date: DAY 3

Standards/Objectives:
1) practice typical questions

Agenda/Activities:
1A) AP Flashcard FR
1B) JMP → alcohol and heart disease

1A) AP Flashcard

1B) JMP - Alcohol and Heart Disease

Typical Questions:

1. *What does "r" mean (in context of the problem)? What is its name?*
2. *What does r^2 mean (in context of the problem)? What is its name?*
3. *What is the equation of the least squares regression line? Define all variables.*
4. *Place a value in X to produce a Y.*
5. *Place a value in Y to produce an X.*
6. *What does the slope mean (in context of the problem)? What is its value?*
7. *What is the formula that involves slope, correlation, and standard deviation?*
8. *What does resistant and non-resistant mean? Name things that are resistant / non-resistant.*
9. *Is the line a good fit for the data? Justify your answer.*
10. *What is the meaning of least squares?*
11. *What does an outlier influence in the problem?*
12. *What does an influential point do to the problem?*
13. *What does "s" mean on the Minitab printout?*

BONUS: Would you recommend more alcohol to help with heart disease? Explain.

Data Organization (Chapter 3) Date: DAY 4

Standards/Objectives:
1) practice typical questions

Agenda/Activities:
1A) JMP

1A) select random data from JMP

Typical Questions:

1. *What does "r" mean (in context of the problem)? What is its name?*
2. *What does r^2 mean (in context of the problem)? What is its name?*
3. *What is the equation of the least squares regression line? Define all variables.*
4. *Place a value in X to produce a Y.*
5. *Place a value in Y to produce an X.*
6. *What does the slope mean (in context of the problem)? What is its value?*
7. *What is the formula that involves slope, correlation, and standard deviation?*
8. *What does resistant and non-resistant mean? Name things that are resistant / non-resistant.*
9. *Is the line a good fit for the data? Justify your answer.*
10. *What is the meaning of least squares?*
11. *What does an outlier influence in the problem?*
12. *What does an influential point do to the problem?*
13. *What does "s" mean on the Minitab printout?*

Data Organization (Chapter 3) Date: DAY 5

Standards/Objectives:
1) understand **correlation vs. causation**
2) understand **interpolation vs. extrapolation**

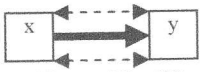
Agenda/Activities:
1A) PWRPT 4.2 - Interpreting Correlation
1B) State Proficiency vs. NAEP
2A) Interpolation vs. Extrapolation

1A) PWRPT 4.2 - Interpreting Correlation

Correlation → there is an association between variables, but not necessarily a cause-and-effect relationship

3 reasons for the occurrence of an association (**correlation**):

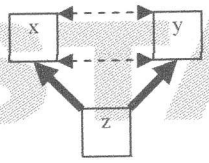
1) **cause-effect**, which can't be confirmed without _____



2*) **common response** - which indicates a BIG extraneous variable

_____, both of which responded in sync to a (third) lurking variable

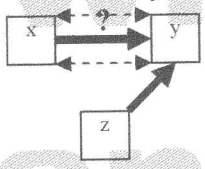
ex: SAT and GAP



3*) **confounding response** - which indicates a BIG extraneous variable

_____ affect our response, but we didn't recognize all of them

ex: ethnicity and GPA



1B) correlation between State Proficiency vs. NAEP test scores

2A) Interpolation vs. Extrapolation

Interpolation – use of regression line/curve for prediction inside the domain of values (of the explanatory variable x)

Extrapolation – use of regression line/curve for prediction outside the domain of values (of the explanatory variable x)

*extrapolated predictions cannot be trusted

Data Organization (Chapter 3) Date: DAY 6

Standards/Objectives:
1) Quiz 4.3
2) understand **marginal vs. conditional counts**/distributions in tables

Agenda/Activities:
1A) Quiz 4.3C
2A) Worksheet 4.3A
2B) Worksheet 4.3B

1A) Quiz 4.3C

2A) Worksheet 4.3A
marginal count = sum of each row, sum of each column
conditional count = individual "relevant" cell(s)

marginal distribution = marginal count ÷ table total
conditional distribution = cell ÷ marginal count

2B) Worksheet 4.3B

Data Organization (Chapter 3) Date: DAY 7

Standards/Objectives:
1) know **Simpson's paradox** and why/how it occurs
2) practice Chapter 3 FRQ

Agenda/Activities:
1A) PWRPT - Simpson's Paradox
2A) practice FRQ - Chapter 3 2006 #2

1A) PWRPT - Simpson's Paradox

Simpson's Paradox – reversal of the direction of a comparison or an association when data from several groups are combined to form a single group (i.e., *conditional* marginal percents may indicate one outcome, but *total* marginal percents may indicate another)

Question: How can we "win every battle" yet "lose the war"?

Answer: This is due to _____

Ex:

	A	B
	9/10 =	72/90 =
	18/90 =	1/10 =
Total	27/100 =	73/100 =

*this occurs because A's _____ and B's _____ are heavily weighted

1B) practice FRQ – Chap 3 2006 #2

Data Organization (Chapter 3) Date: DAY 8

Standards/Objectives:
1) practice Chapter 3 FRQ
2) review 4.3 tables

Agenda/Activities:
1A) WS - Tables
2A) JMP

1A) WS - Tables

2A) use JMP and practice 13 typical questions

Typical Questions:

1. *What does "r" mean (in context of the problem)? What is its name?*
2. *What does r^2 mean (in context of the problem)? What is its name?*
3. *What is the equation of the least squares regression line? Define all variables.*
4. *Place a value in X to produce a Y.*
5. *Place a value in Y to produce an X.*
6. *What does the slope mean (in context of the problem)? What is its value?*
7. *What is the formula that involves slope, correlation, and standard deviation?*
8. *What does resistant and non-resistant mean? Name things that are resistant / non-resistant.*
9. *Is the line a good fit for the data? Justify your answer.*
10. *What is the meaning of least squares?*
11. *What does an outlier influence in the problem?*
12. *What does an influential point do to the problem?*
13. *What does "s" mean on the Minitab printout?*

RECALL: segmented bar graphs, which acts as a "vertically stacked" pie chart

RECALL: given r^2, find r and assign + or - sign

FUN FACT: the center of gravity, $(\bar{x}, \bar{y})$, is included in every LSRL

Data Organization (Chapter 3) Date: DAY 9

Standards/Objectives:
1) Test Chapter 3

Agenda/Activities:
1A) Test Chapter 3

Data Organization (Chapter 3) Date: DAY 10

Standards/Objectives:
1) review test

Agenda/Activities:
1A) go over Multiple Choice and FRQs