Statistical Association Assignments for Pre-Service Teachers

Stephanie Casey and Andrew Ross, Mathematics Department, Eastern Michigan University

Table of Contents

Introduction

The following homework assignments are intended for college students who are studying to be K-12 mathematics teachers (mostly at the secondary level) and are enrolled in a statistics class. Often, such a class consists of such pre-service teachers along with other non-teaching majors (sciences, engineering, etc.). We have adopted the practice of giving a first assignment on each topic that both teaching and non-teaching majors do, usually consisting of typical procedural problems from a textbook. Then the second assignment on each topic is differentiated, with pre-service teachers working on the homeworks included in this document, and non-teaching majors working on science- and engineering-focused statistics problems. The assignments for non-teaching majors are not included in this file.

Many of the assignments are designed to be submitted electronically, with students typing their answers in between prompts, rather than leaving space for students to write responses by hand.

Association of Categorical Variables

1) The chi-square test statistic is defined as the sum of $\frac{(O-E)^2}{E}$ for all cells.

A. Define what O and E represent in this formula & how to find each.

B. Explain why the chi-square test statistic is defined this way. Include reasoning for every aspect of this formula, including the difference of *O-E*, the squaring of this difference, the division by *E,* and why these values are summed.

C. Look at a chi-square distribution when the degrees of freedom is 6. What is its shape? Why?

D. Describe the shape of the chi-square distribution. Include graphs of the distribution for various degrees of freedom. How does its shape change as the degrees of freedom increases? Why?

E. What is the smallest possible value chi-square can be in any situation? Why?

2) Adapted from Burrill, G., Franklin, C., Godbold, L., & Young, L. (2003). *Navigating through data analysis in grades 9-12.* Reston, VA: NCTM.

Scenario: A bank had 48 candidates for promotion, with equal numbers of males and females. A total of 35 people were recommended for promotion.

A. Suppose that the recommendations of male & female candidates showed no evidence of gender discrimination. Fill in the table below with a possible scenario that shows this.

	Recommended for Promotion	Not Recommended for Promotion	Total
Male			24
Female			24
Total	35	13	48

B. Now suppose that the recommendations for promotion showed strong evidence of discrimination against the female candidates for promotion. Complete the table to show a possible example of this case.

	Recommended for Promotion	Not Recommended for Promotion	Total
Male			24
Female			24
Total	35	13	48

C. Suppose the evidence of discrimination against the women fell into a “gray” area, making any discrimination against the women not clearly obvious without further study. Complete the table below to show such a case.

	Recommended for Promotion	Not Recommended for Promotion	Total
Male			24
Female			24
Total	35	13	48

D. It turned out that at the actual bank, of the 24 females, 14 candidates were recommended for promotion, and 21 of the 24 male candidates were recommended for promotion. Complete the table below with this actual data.

	Recommended for Promotion	Not Recommended for Promotion	Total
Male			24
Female			24
Total	35	13	48

E. Some of the female employees felt like there was gender discrimination in this decision. Set up and carry out a simulation to investigate this claim. Note: Do not investigate the claim through a hypothesis test. Describe your simulation here.

F. Based on your simulation results, is there an association between gender and promotion recommendation at this bank? Back this up with numerical justification based on your simulation.

G. Suppose a similar study at another company showed a clearly strong, statistically significant association between gender and promotion recommendation. Pretend you are the company's lawyer. Argue that this is not sufficient evidence that gender discrimination took place.

3) Adapted from College Board (2009). *AP Statistics Free-Response Questions.* New York, NY: Author.

A simple random sample of 100 high school seniors was selected from a large school district. The gender of each student was recorded, and each student was asked the following questions:

1. Have you ever had a part-time job?
2. If you answered yes to the previous question, was your part-time job in the summer only?

The responses are summarized in the table below.

Job Experience	Male	Female	Total
Never had a part-time job	21	31	52

Had a part-time job during summer only	15	13	28
Had a part-time job but not only during summer	12	8	20
Total	48	52	100

(a) Construct a graphical display that represents the association between gender and job experience for the students in the sample. You may use the grid below or paste a graphic in from a program of your choice.

(b) Write a few sentences summarizing what the display in part (a) reveals about the association between gender and job experience for the students in the sample.

(c) Give 3 different numerical comparisons you can make using the data in the table to determine if there is an association between gender & job experience for the students in the sample.

(d) Which test of significance should be used to test if there is an association between gender and job experience for the population of high school seniors in the district?

State the null and alternative hypotheses for the test, but do not perform the test.

4) A student named Martha comes to you for help. She says she knows you use chi-square tests with categorical data, but she doesn't know when to use which one (goodness of fit, independence, & homogeneity). Write below your response to Martha as her teacher, perhaps including examples to illustrate your points. Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.

5) The Common Core State Standards for Mathematics includes the following standard: "Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies).Recognize possible associations and trends in the data."

Research has found that students commonly have three incorrect conceptions about association of categorical variables.

- Determinist: students believed that an association meant all cases must show an association with no exceptions. These students believed that the cells in the two-way table that did not agree with the association should have zero frequency.
- Unidirectional: students believed dependence occurred only when it was direct. This could be explained by the tendency of students to give more relevance to positive cases than negative cases that confirm a given hypothesis.
- Localist: students looked at part of the data to determine if an association existed, often only looking at the cell with the highest frequency or at only one conditional distribution.

Design a classroom lesson focused on the given standard which addresses these misconceptions. You do not need to create homework or exam questions; only the classroom lesson.

6) You are leading your students in an activity in which they complete all 4 steps of the GAISE framework (formulate a research question, collect their own data, analyze the data, and interpret the results) in a situation relevant to chi-square tests. Students have brainstormed the following research questions. For each question, determine if it is relevant to chi-square tests and provide your response to the student. Consider data collection issues as well. For those that are not relevant and/or have data collection issues, suggest some modifications.

LAURA: I want to study if girls are smarter than boys, so I am going to compare the GPAs of boys and girls at school.

BILL: I think people with higher GPAs are less likely to have had a car crash while driving. I am going to ask a senior with drivers licenses their GPAs and if they've crashed while driving.

ANNE MARIE: I want to see whether boys are more likely to own smartphones than girls. So I'm going to count during passing periods the number of boys I see with smartphones and the number of girls I see with no phones.

NICK: I think girls are more likely to bring their lunch to school than boys. So I'm going to stand at the entrance to the lunchroom and count the number of boys and girls that bring their lunches.

JEFF: I want to study the relationships between gender, race, grade level, political party and whether students make honor roll or not at our school.

KRISTIN: I read that 77% of high school students go on to college. I want to see if that's true of the students here.

Association of Quantitative Variables, Part I

1. The regression line $\hat{y} = 1.6 + 2.2x$ for the five points given in the table below was computed using the method of least squares.

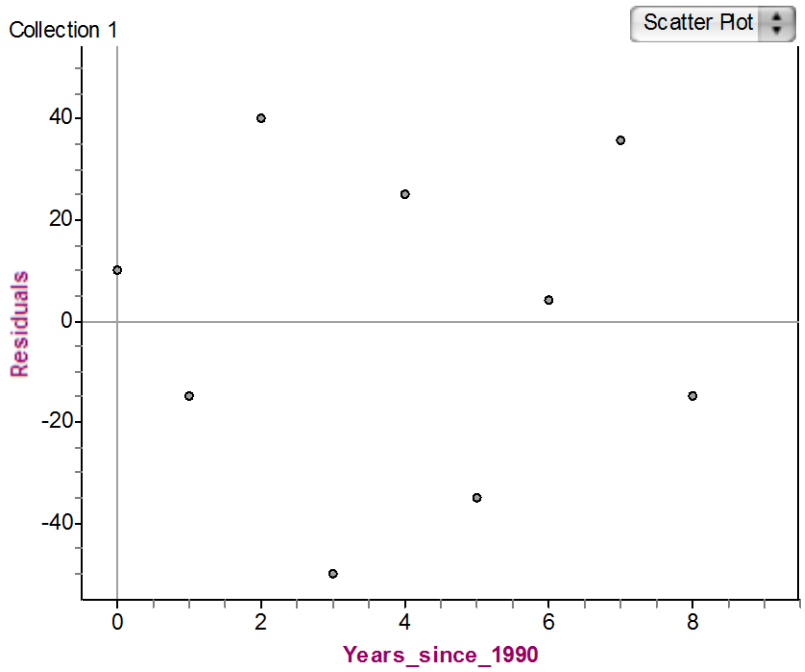
x	1	2	3	4	5
y	6	3	10	7	15

   A. On graph paper, plot these five points and the regression line.
   B. Using your work from part A, use pictures and words to demonstrate the meaning of the term "least squares", i.e., why is it called the *least squares* regression line?
   C. Why is the least squares criterion appropriate for a best fit line?

2. You are leading the class in analyzing the relationship between GPA and ACT scores for a data set. You have asked your students to use technology to find the correlation coefficient, coefficient of determination, and regression line for the data set. One pair of students asks for your help. They have done their work independently and are now comparing their answers. They have the same correlation coefficient & coefficient of determination, but their regression lines are different. They are asking you for help to understand how this could be possible. What is your response? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.

3. Adapted from College Board (1999). *AP Statistics Free-Response Questions.* New York, NY: Author.

   Martha and Barry were searching the Internet to find information on air travel in the United States. They found data on the number of commercial aircraft flying in the United States during the years 1990-1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



Predictor	*Coef*	*Stdev*	*t-statistic*	*p*
Constant	2939.93	20.55	143.09	0.000
Years	233.517	4.316	54.11	0.000
s=33.43				

A. Is a line an appropriate model to use for this data set? What information tells you this?
B. What is the value of the slope of the least squares regression line?
   Interpret the slope in the context of this situation.
C. What is the value of the y-intercept of the least squares regression line?
   Interpret the y-intercept in the context of this situation.
D. What was the actual number of commercial aircraft flying in 1992?
E. Explain the meaning of the t-statistic value of 54.11 in the computer output, i.e. what does this numerical value of 54.11 tell you?

4. To start a lesson on lines of best fit, you are following the curriculum guide and have presented your students the following data set about the pounds of beans eaten by families of different sizes when traveling on the Overland Trail:

Number of people	5	8	6	7	11	10	5	7	10	5	8	7	9	12	10
Pounds of beans	61	95	56	75	125	135	80	100	103	75	100	105	125	150	125

Note: This is a data set used in a mathematics curriculum, but is it unknown whether it is real data.

The first question in the activity asks the students to discuss ways in which you could use the information in the table to decide how many pounds of beans a family with 20 people would actually eat.

A. Mary responds first, stating "You could look at people with like 10 people in their families and just double that amount that they use."

   How would you respond to this student? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.

B. The class is convinced that Mary's strategy is the best one to take. However, the curriculum writers designed this question to be an introduction to the line of best fit, which is to be the focus of the day's lesson. What would you do in this situation? Things to address include a way to convince the students that Mary's strategy is not the best one; how to transition from this "scaling" method to the "best fit line" method; how to help the students look at the data from an aggregate view (looking at data set as a whole) as opposed to a case view (looking at one data point at a time).

5. A. Go to the following website. Practice placing the line of best fit on the scatterplot by eye (without any buttons activated), then compare it to a best fit line. Do this for multiple scatterplots until you feel confident in your ability to accurately place a line of best fit by eye.

http://hspm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html

Now that you're an expert at fitting a line by eye, why is a least squares regression line still a useful tool?

B. Students in your class have been doing an activity where they place the line of best fit on scatterplots by eye then come up with their own criteria for determining where to place the line. Below are some of the student responses. For each response, determine if it is generally a good criteria (works in all reasonable cases), create counter examples of scatterplots where that criteria would create a poor line of best fit (if possible), and provide your detailed response to the student.

JOE: Goes through the most points

SUSAN: Equal number of points above and below the line

JONATHAN: Connect the first and last points

(optional) Anticipate other possible student criteria & formulate responses to those students.

Association of Quantitative Variables, Part II

1. A student in your class raises his hand and asks "What's the difference between correlation, association, and regression?". How would you respond to this student? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.

2. Create three distinctly differently patterned scatterplots which all have a correlation coefficient of approximately zero. Explain why the correlation coefficient is near zero for each of the cases.

3. Page 273 of the textbook (Walpole et al.) presents some common transformations used to linearize data. Pick one of them and show why that transformation works to linearize the data. For example, if the original data set has an exponential relationship between x and y, why is the relationship between x and ln(y) linear?

4. You have been teaching your class about finding the best model for a data set. A student says "So I just try all of these different transformations on the data set and whichever one gives me the biggest value of r is the best one, right?". How would you respond to this student? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.

5. One way to calculate the correlation coefficient r is through the following formula:

$$r = \frac{\Sigma z_x z_y}{n-1}$$

   A. Define what z_x and $z_y$represent in this formula & how to find each.

   B. Explain why the correlation coefficient is calculated this way, keeping in mind that the purpose of this statistic is to measure the strength and direction of the linear association between the x- and y-variables (How is it calculating the strength? How is it calculating the direction?). Include reasoning for every aspect of this formula, including the z-scores, the product of the z-scores, the summation of the product of the z-scores, and division by n-1.

   C. Use this formula to explain why the correlation coefficient is unaffected by linear transformations of the data (e.g., if all of the x-value have 5 added to them, the value of r doesn't change. Why?)
   D. Use this formula to explain why the correlation coefficient is unit-less (which gives it the advantage to be compared across various data sets!)
   E. Use this formula to explain why the correlation coefficient of y on x is the same as the correlation coefficient of x on y.
   F. Use this formula to explain why the correlation coefficient is sensitive to influential points.

Correlation vs. Causation

Part of talking about a data set is graphing it. When you (or someone else) graphs a data set, a choice must be made about which variable to put on which axis. The choice is guided by some traditions, but the choice also influences the way that people think about the data set and any possible causation. As an impartial analyst, you need to think about these influences. Or, if you are trying to convey your point of view, you might deliberately decide to use these influences to convince people your are right—surely, those with an opposing point of view are doing the same thing.

First, the names we give the variables or axes show how they are often interpreted:
Common Names for Variables:

horizontal	vertical
x	Y
independent	dependent
free	outcome
predictor	predicted
stimulus	response
cause	effect
controlled	uncontrolled
explanatory	explained
regressor	Result

Note that “independent” and “dependent” are often used in algebra and calculus classes, and in science/engineering experiments where the experimenters can indeed control one variable and see how the other depends on it. However, in observational studies, it is perhaps misleading to label one variable as depending on the other, since the influence could easily go the other way, or perhaps each depends on the other in a feedback loop. Statistically speaking, “dependence” is a two-way road even when the causation is clear. You might want to bring up this issue with your future students.

Here are the traditions about what kinds of variable go on the horizontal axis. Put a variable on the horizontal axis:

* If we directly controlled or could directly control it (e.g. price, if we run the store, though economists do the opposite)
* If we think it is the “cause”, and the other variable is the “result”
* If it is a date/time value, e.g. 2007 or 12:43pm (a “when”)
  but “3 hours” or “78 years of age” (a “how long”) could be either x or y.
* If it is the first in a sequence of measurements (e.g. 5th grade test score vs. 8th grade test score)

If neither variable is a clear choice to go on the horizontal axis, it might help to try the graph both ways and see if one interpretation makes more sense than another (“try it both ways”)

If there is no clear preference for one variable vs. the other, then maybe “it doesn't matter” (# people at one university, #people at another university)

There is a difference between "try it both ways" and "it doesn't matter".

Speaking of vocabulary, here are some other common "synonyms" for association:
link, relationship, correspondence, trend

Should “regression” be included in that list? Is it indeed a synonym for association?
Should “correlation” be included in that list? Is it indeed a synonym for association?

Correlation vs. Causation Worksheet 1

For each of the following studies, fill in "x=" or "y=" where appropriate. If neither way seems right, write either "try it both ways" or "doesn't matter". Also sketch what the graph might look like.

a. __ =Price charged, __=number of this book bought today on Amazon

b. __=weight, __=height of a sample of American 20-year-olds

c. __=Average murder sentence duration by state, __=murder rate, by state

d. __=Sales of Toyota minivans, __=Sales of Chrysler minivans, by year

e. __=Year and __=U.S. national debt

f. __=A person's height, __=salary

g. __=A person's salary, __=years of education

h. __=Production costs, __=gross box office income, movie by movie

i. __=Student standardized test scores, __=avg teacher salary, school by school

j. __=Age of first cancer, __=packs/day smoked, by person

k. __=Cancer rate, __=distance to high-voltage power lines, by street/block

l. __=Student standardized math test scores, 5^{th} grade and __= score in 8^{th} grade

m. __=Average income, __=unemployment rate, by year

Correlation vs. Causation Worksheet 2

Lurking variable worksheet Name:_____

Suppose you read an article that says that people who eat at least one carrot a day tend to spend less on health care than those that don't. Does this mean you should eat more carrots to stay healthier? Give an argument based on what we've discussed in class.

Here's a good argument: Perhaps a hidden variable is a person's attitude toward health. People who try to take good care of themselves probably eat a lot of veggies. They probably also have better health than those who don't care what they eat.

Notice what is good about this argument: it proposes a specific hidden (also known as lurking) variable (rather than saying "there is a lurking variable" with no further explanation), and it says how that variable affects both of the variables already mentioned, to produce the observed correlation. It doesn't argue that the link doesn't exist. Another good way to argue is to give a plausible reason why the causation might go the other way than was suggested. Now it's your turn to practice. Write down good arguments about (usually against) these statements:

1. There is a positive link between consumption of tobacco and non-use of seatbelts. So if you want to cut down on smoking, buckle up!

2. There is a link between the # of years of math someone takes in high school and their future income. So, Michigan should require high school students to take at least Algebra 2.

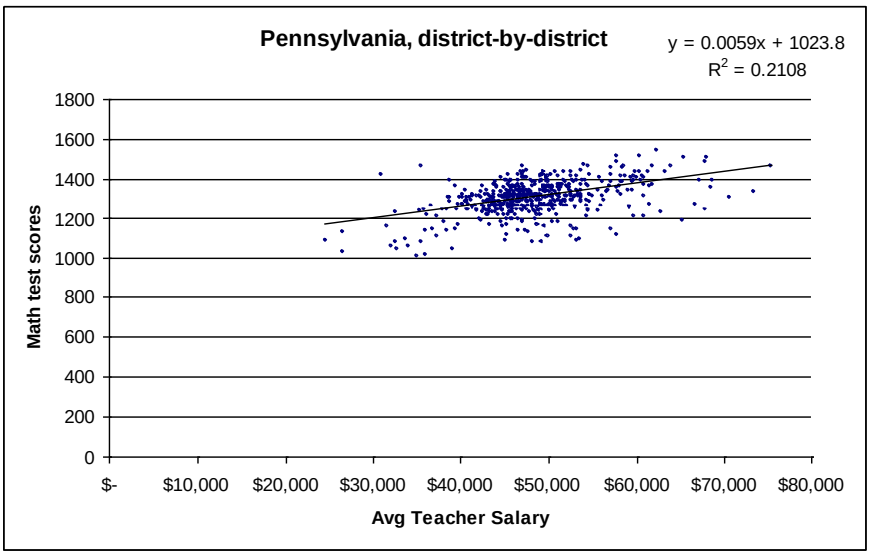
3. An actual article said something like: there is a link between credit card debt and health problems. So, to make yourself healthier, pay down your credit cards, since credit card debt causes stress which can cause health problems.

4. There is a link between the presence of computers in K-12 schools and their standardized test scores. Therefore, we should spend more money on computers in schools.

5. A study shows that couples that go to church more often have a lower divorce rate than other couples. So, to keep the divorce rate low, couples should go to church more often. (PS: is this a negative correlation?)

Correlation vs. Causation Worksheet 3

Consider this real data set of average teacher salary vs. math test scores, district-by-district in Pennsylvania for one year (around 2005):



For each statement, mark True or False. Write an explanation of why it is true or false.

1. T/F: If you give a raise of $10,000 to each teacher in a particular district, that district's score will go up 59 points.

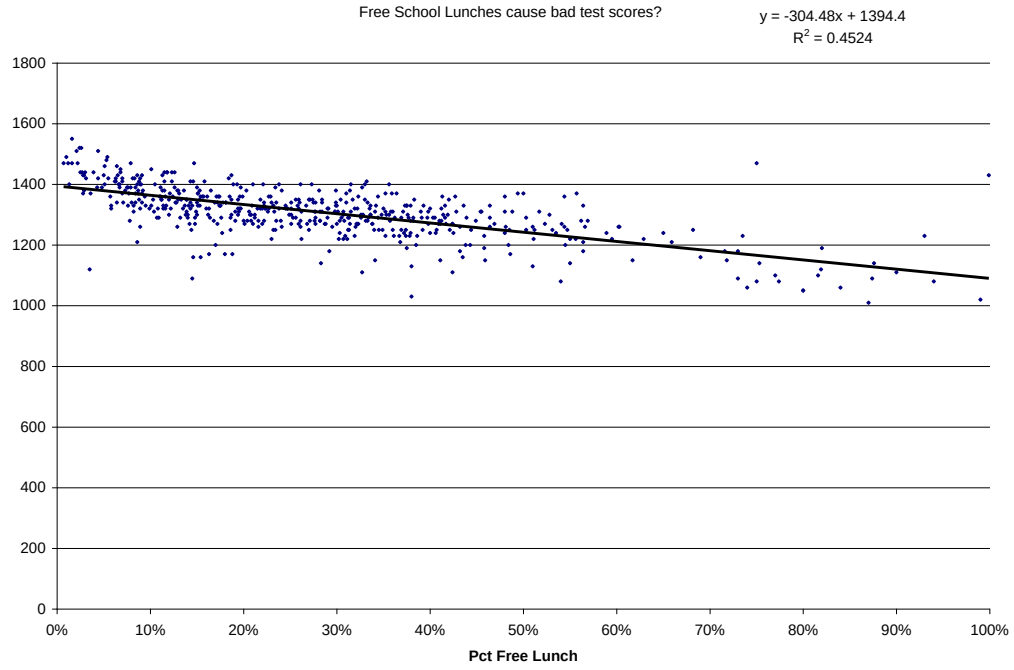
2. T/F: If you live in a district with a $50,000 average teacher salary and move to a district with a $60,000 average, your child's test scores will go up, on average, by 59 points.

3. T/F: If you live in a district with a $50,000 average and move to a district with a $60,000 average, the scores in that district will be 59 points higher.

4. T/F: If you live in a district with a $50,000 average and move to a district with a $60,000 average, then on average the scores in that district will be 59 points higher.

5. T/F: Since there is a district with a salary of about $30,000 with test scores above 1400, and another district around $65,000 with test scores below 1200, we can see that there's no correlation between salary and test scores.

6. Now write some true statements and some false statements about this graph (from the same real data set); x=% of kids in the district who get free or reduced lunch, y=math test scores in the district



Correlation vs. Causation Worksheet 4

The book "Born to Buy" by Juliet B. Schor (2004) talks about how companies are marketing toys and foods to children.

1. On page 125 it says "In the 1970s, Marvin Goldberg studied differences between children who saw and did not see television advertising and found that sugared cereals were more likely to be present in the homes of the former." Is this evidence that food marketing to children is effective? Explain.

2. On page 128, paraphrased the way a reporter might paraphrase it: "A two-year study of eleven and twelve year olds, discovered that each additional soft drink consumed per day increased the likelihood of a child's becoming obese by 60 percent." What do you think of this study, the way it is summarized here?

3. Here is what it actually said on page 128: "A two-year study of eleven and twelve year olds, which controlled for exercise, other food intake, and media viewing, discovered that each additional soft drink consumed per day increased the likelihood of a child's becoming obese by 60 percent." Now what do you think?

4. Newsweek reported on July 12, 2010: "Five experiments by Northwestern's Adam Galinsky showed that those who have lived abroad outperform others on creativity tasks." Does this mean that living abroad increases creativity? Explain.

5. Newsweek reported on July 12, 2010: "According to University of Texas professor Elizabeth Vandewater, for every hour a kid regularly watches television, his overall time in creative activities—from fantasy play to arts projects—drops as much as 11 percent." Does this mean:

a) The R^2 value in this study is 0.11? b) The slope in this study is 0.11? c) The slope in this study is -0.11?
d) The intercept in this study is 0.11? e) something else? What?

Journal Article Pre-Reading Assignment

The following newspaper article extract comes from page 2 of the June 11, 1991 edition of the Hobart Mercury

Family car is killing us, says Tasmanian researcher

Twenty years of research has convinced Mr. Robinson that motoring is a health hazard. Mr. Robinson has graphs which show quite dramatically an almost perfect relationship between the increase in heart deaths and the increase in use of motor vehicles. Similar relationships are shown to exist between lung cancer, leukaemia, stroke and diabetes.

1. Draw and label a sketch of what one of Mr. Robinson's graphs might look like.

2. What questions would you ask Mr. Robinson about his research?

Journal Article Reading assignment

Please read the article Watson, J. (2000). "Statistics in context". Mathematics Teacher, 93(1), 54-58 a*nd respond to the following prompts.*

1) The first sentence in the article states "Judging statistical claims in social contexts is fundamental to statistical literacy." What do *you* think someone should be able to do to be deemed statistically literate?
2) This article focuses on the topics of graphing relationships and interpreting associations found among variables (2nd paragraph p.54). What are the CCSS-M (http://www.corestandards.org/the-standards/mathematics) relevant to these goals?
3) Investigate the curriculum recommendations in other disciplines here in the United States to see where these topics are covered and what the relevant standards are.

   Next Generation Science Standards http://www.nextgenscience.org/
   National Curriculum Standards for Social Studies http://www.socialstudies.org/standards
   National Health Education Standards http://www.cdc.gov/healthyyouth/sher/standards/

4) Add at least 2 more bullet points to the list at the top of the first column on page 55.
5) Do you think Mr. Robinson's numbers (Figure 1) are raw counts or rates? Which would be more helpful in looking for things that explain heart disease?
6) Pick one of the graphs in (b) or (c) of Figure 4. What would you say to the student who created that graph to assist him/her in moving to the next step of the hierarchy?
7) What are distinct advantages & disadvantages of each of the 3 types of representations shown in Figures 5, 6, and 7? Which representation type do you prefer and why?
8) How would you design a study on driving and heart deaths?
9) Evaluate your own answers from the pre-reading assignment to the two tasks, using the structure outlined in the article.

Journal Article Post-Reading Assignment

Design a lesson or cross-curricular activity around an example of a statistical claim from the media (you are welcome to find your own or you can use one of the references below). The goal of your lesson is to help your class progress from step 2 of the hierarchy to step 3.

Article title	Source
Rethinking the economics of traffic congestion	http://www.theatlanticcities.com/commute/2012/06/defense-congestion/2118/#
Delaying kindergarten: Effects on test scores and childcare costs	http://house.michigan.gov/SessionDocs/2011-2012/Testimony/Committee5-3-21-2012-4.pdf
Sleeping pills linked with early death	http://healthland.time.com/2012/02/28/study-sleeping-pills-linked-with-early-death/
Why video games are good for you	http://www.straightdope.com/columns/read/3034/why-video-games-are-good-for-you
Fear-mongering from the bench [does prison time reduce violent crime?]	http://www.commonwealmagazine.org/fear-mongering-bench
Why being thin can actually translate into a bigger paycheck for women	http://www.forbes.com/sites/lisaquast/2011/06/06/can-being-thin-actually-translate-into-a-bigger-paycheck-for-women/