

# **AP Statistics**

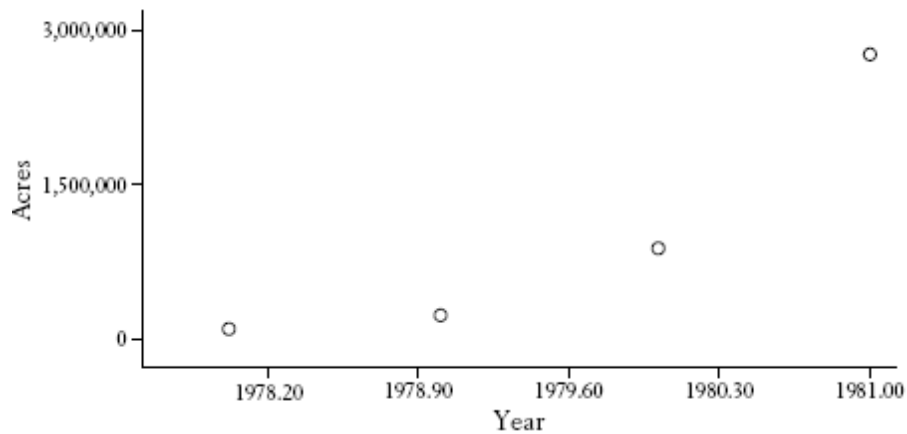
## **Solutions to Packet 4**

More on Two-Variable Data  
Transforming Relationships  
Cautions about Correlation and Regression  
Relations in Categorical Data

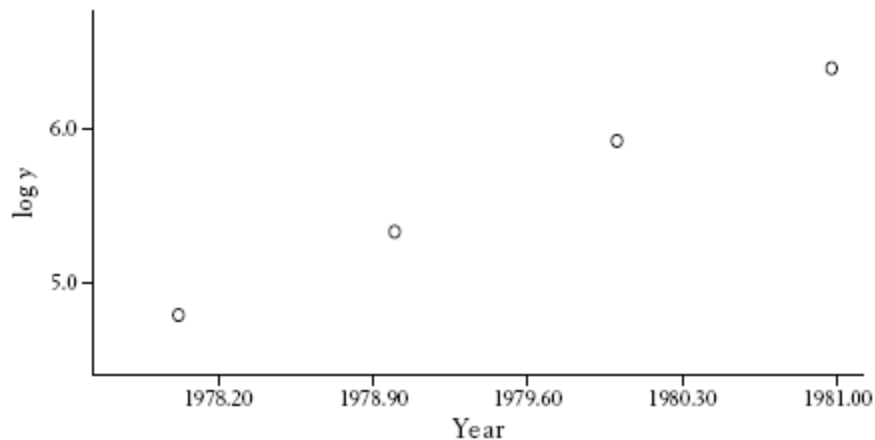
4.6 **GYPSY MOTHS** Biological populations can grow exponentially if not restrained by predators or lack of food. The gypsy moth outbreaks that occasionally devastate the forests of the Northeast illustrate approximate exponential growth. It is easier to count the number of acres defoliated by the moths than to count the moths themselves. Here are data on an outbreak in Massachusetts:

Year	Acres
1978	63,042
1979	226,260
1980	907,075
1981	2,826,095

- Let L1 = year, L2 = acres
- Look at scatterplot: year vs acres, scatterplot looks exponential

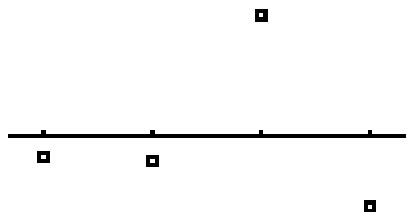


- Verify that y is being multiplied by about 4 each year by calculating the ratio of acres defoliated each year to the previous year. (start with 1979 to 1978, when the ratio is  $226,260/63,042 = 3.59$ ) **The ratios are  $226,260/63,042 = 3.59$ ,  $907,075/226,260 = 4.01$ , and  $2,826,095/907,075 = 3.12$ .**
- Let L3 = log (acres) = log (L2) **log y yields 4.7996, 5.3546, 5.9576, and 6.4512.**
- Look at scatterplot, year vs log (acres), L1 vs L3



- Calculate LSRL for this scatterplot.

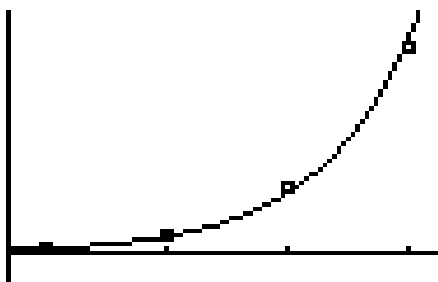
$$\log \hat{y} = -1094.5 + 0.5558x$$

- $r = 0.999$
- Look at residual plot. How does it look?  residual plot shows random scatter.

- Perform the inverse transformation to express  $\hat{y}$  as an exponential equation.  
 eqt:  $\hat{y} = 10^{(-1094.5 + 0.5558x)}$  Sketch a scatterplot of the original data with the exponential curve model superimposed. Is your exponential function a satisfactory model for the data? **YES!** Use your model to predict the # of acres defoliated in 1982. **10,719,965**

```

2011 Plot2 Plot3
\Y1 -1094.507174
8963+.5557706197
851X
\Y2 10^Y1
\Y3 =
\Y4 =
\Y5 =
  
```



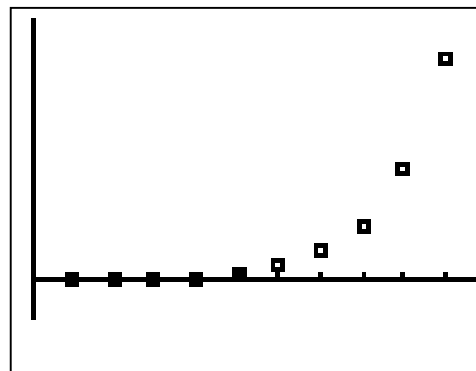
(Postscript: A viral disease reduced the gypsy moth population between the readings in 1981 and 1982. The actual count of defoliated acres in 1982 was 1,383,265.)

4.10 **GUN VIOLENCE** (Exact exponential growth) A paper in a scholarly journal once claimed (I am not making this up), “Every year since 1950, the number of American children gunned down has doubled.” To see that this is silly, suppose that in 1950 just 1 child was “gunned down” and suppose that the paper’s claim is right.

(a) Fill in table with # of children killed in each of the next 10 years.

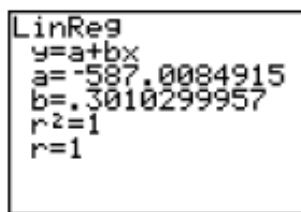
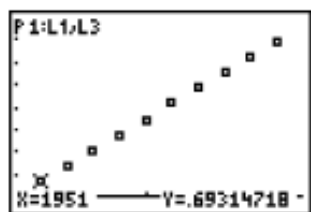
year	# children killed
1950	1
1951	2
1952	4
1953	8
1954	16
1955	32
1956	64
1957	128
1958	256
1959	512
1960	1024

(b) Plot the number of deaths against the year and connect the points with a smooth curve. This is an exponential curve.

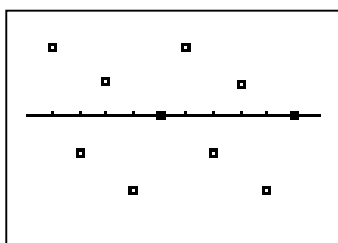


(c) The paper appeared in 1995, 45 years after 1950. How many children were killed in 1995, according to the paper? *If  $x$  = number of years after 1950, then  $y$  = the number of children killed  $x$  years after 1950 =  $2^x$ . At  $x = 45$ ,  $y = 2^{45} = 3.52 \times 10^{13}$ , or 35,200,000,000,000 (35 trillion!)*

(d) Let  $L3 = \log(\# \text{ children}) = \log(L2)$ . Look at scatterplot of year vs  $\log(\# \text{ of children})$  (L1 vs L3). This should appear linear.

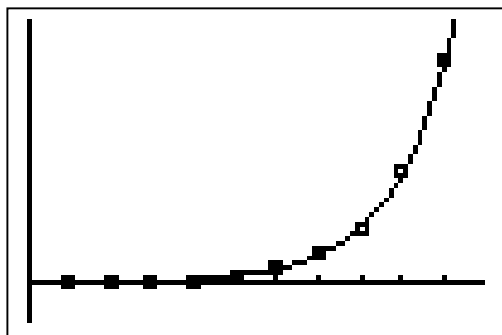


(e) Find LSRL of scatterplot above.  $\log \hat{y} = -587.0085 + .30103x$   $r = 1$   
Make sure to look at residual plot.



(f) Perform the inverse transformation to express  $\hat{y}$  as an exponential equation.

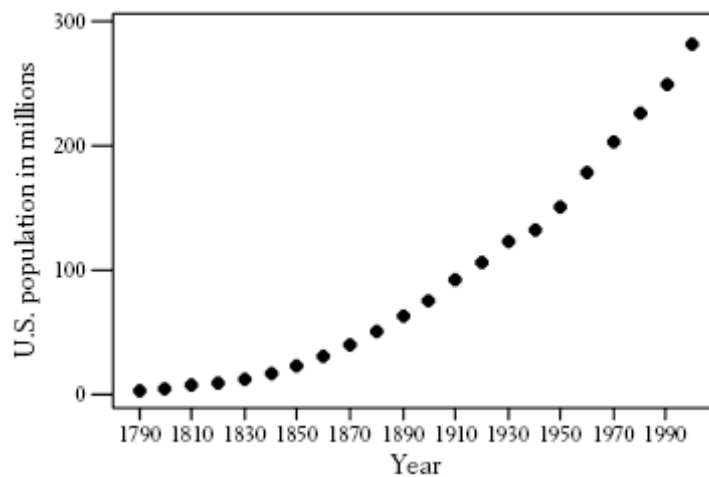
eqt:  $\hat{y} = 10^{\wedge} y_1$  Sketch a scatterplot of the original data with the exponential curve model superimposed. Is your exponential function a satisfactory model for the data? **YES!**



4.11 **US POPULATION** The following table gives the resident population of the United States from 1790 to 2000, in millions of persons.

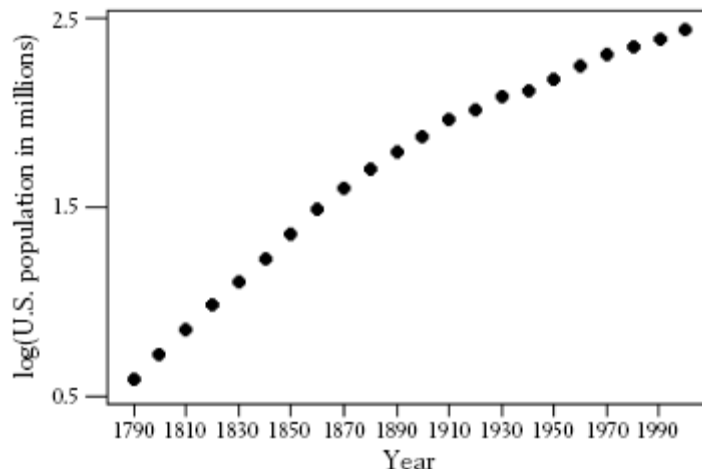
Date	Pop.	Date	Pop.	Date	Pop.	Date	Pop.
1790	3.9	1850	23.2	1910	92.0	1970	203.3
1800	5.3	1860	31.4	1920	105.7	1980	226.5
1810	7.2	1870	39.8	1930	122.8	1990	248.7
1820	9.6	1880	50.2	1940	131.7	2000	281.4
1830	12.9	1890	62.9	1950	151.3		
1840	17.1	1900	76.0	1960	179.3		

(a) Plot year (L1) vs population (L2). The growth of the American population appears roughly exponential.



(b) Let  $L3 = \log(L2)$ . Plot year (L1) vs  $\log(\text{population})$  (L3). The pattern of growth is now clear. An expert says that “the population of the US increased exponentially from 1790 to about 1880. After 1880 growth was approximately exponential, but at a slower rate.” Explain how this description is obtained from the graph.

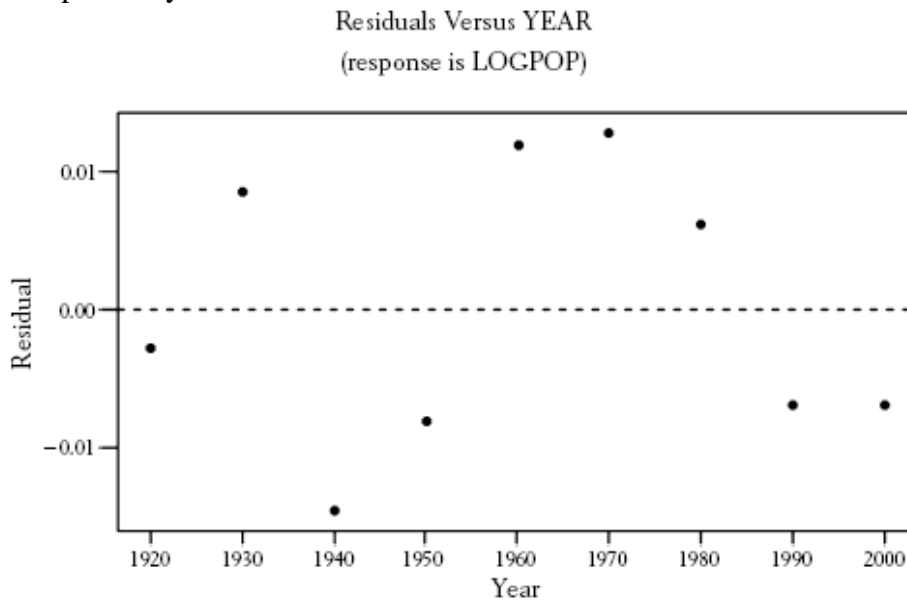
The transformed data appear to be linear from 1790 to about 1880, and then linear again, but with a smaller slope. This reflects a lower exponential growth rate after 1880.



(c) Use part or all of the data to construct an exponential model for the purpose of predicting the population in the year 2010.

- Years used to construct model: 1920 to 2000
- LSRL equation:  $\log(\text{pop}) = -8.2761 + 0.005366(\text{year})$   $r = 0.998$
- Exponential model:  $\text{pop} = 10^{(-8.2761 + 0.005366(\text{year}))}$
- Prediction for 2010: 323,531,803
- Do you think your prediction will be too high or too low? too high why? because the rate of growth slowed after 2000. So the actual growth will be smaller than that predicted by our model.

Construct a residual plot for your transformed data. What is the value of  $r^2$ ? 0.995



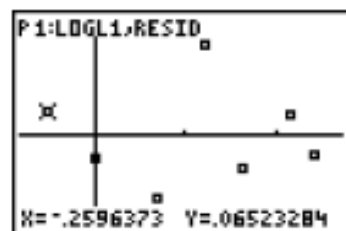
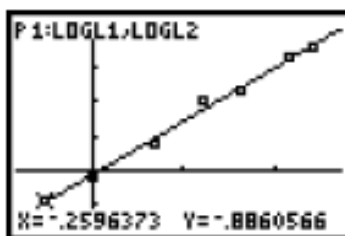
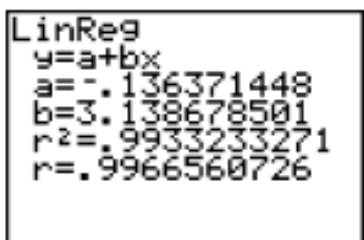
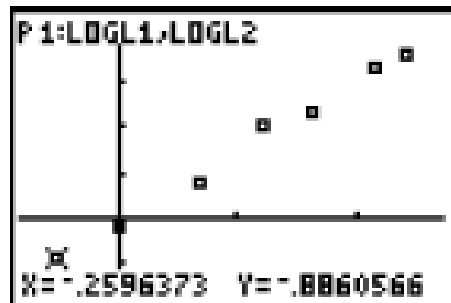
4.14 **HEART WEIGHTS OF MAMMALS** Use the methods discussed in this section to analyze the following data on the hearts of various mammals.

Mammal	Heart weight (grams)	Length of cavity of left ventricle (centimeters)
Mouse	0.13	0.55
Rat	0.64	1.0
Rabbit	5.8	2.2
Dog	102	4.0
Sheep	210	6.5
Ox	2030	12.0
Horse	2900	16.0

Let L1 = length, L2 = weight, L3 = log(L1) = log(length), L4 = log(L2) = log(weight)

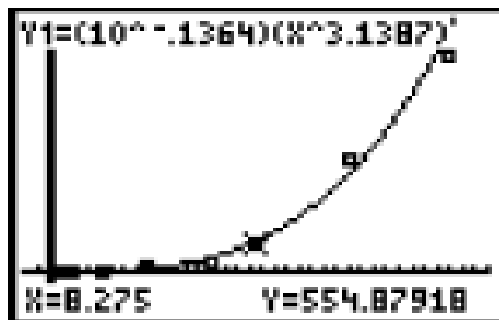
- Look at four scatterplots:
  - Ø length vs weight L1 v L2
  - Ø length vs log (weight) L1 v L4
  - Ø log (length) vs weight L3 v L2
  - Ø log (length) vs log (weight) L3 v L4

- Best scatterplot **L3 v L4** →
- LSRL for this scatterplot  $\log \hat{y} = -0.1364 + 3.1387 \log x$   
 $r = 0.996656$  Look at residual plot



- Back transformed equation  $\log \hat{y} = -0.1364 + 3.1387 \log x$   
 $\hat{y} = 10^{(-0.1364 + 3.1387 \log x)}$

- Look at scatterplot L1 vs L2 with power model superimposed over it!! **YEA!**



4.15 **HEIGHT AND OBESITY** The U.S. Department of Health and Human Services characterizes adults as “seriously overweight” if they meet certain criterion for their heights as shown in the table below (only a portion of the chart is reproduced here).

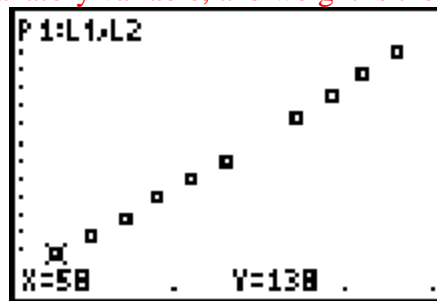
Height (ft, in)	Height (in)	Severely overweight (lb)	Height (ft, in)	Height (in)	Severely overweight (lb)
4'10"	58	138	5'8"	68	190
5'0"	60	148	6'0"	72	213
5'2"	62	158	6'2"	74	225
5'4"	64	169	6'4"	76	238
5'6"	66	179	6'6"	78	250

Weights are given in pounds, without clothes. Height is measured without shoes. There is no distinction between men and women; a note accompanying the table states, “The higher weights apply to people with more muscle and bone, such as many men.” Despite any reservations you may have about the department’s standards for both genders, do the following:

(a) Without looking at the data, hypothesize a relationship between height and weight of U.S. adults. That is, write a general form of an equation that you believe will model the relationship.

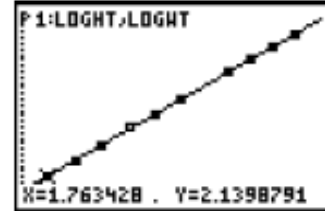
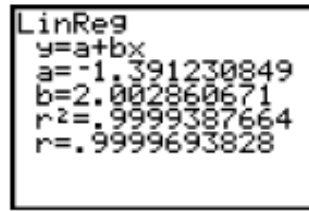
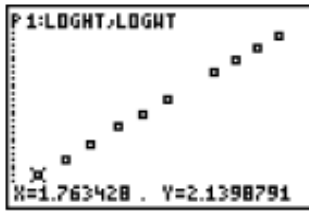
As height increases, weight increases. Since weight is a 3-dimensional attribute and height is 1-dimensional, weight should be proportional to the cube of the height. One might speculate a model of the form  $WEIGHT = a \times HEIGHT^b$  where a and b are constants. This is a power function.

(b) Which variable would you select as explanatory and which would be the response? Plot the data from the table. Height is the explanatory variable, and weight is the response variable.

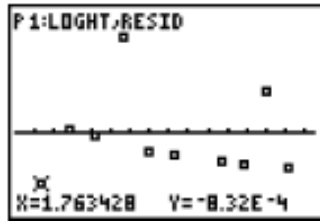


(c) Perform a transformation to linearize the data. Do a least-squares regression on the transformed data and check the correlation coefficient.

(Let L1 = height, L2 = weight, L3 = log (L1) = log(height), L4 = log (L2) = log(weight)) Plot log (L1) v. log (L2). The plot appears to be very linear, so least-squares regression is performed on the transformed data. The correlation is 0.99997. The regression line fits the transformed data extremely well.



(d) Construct a residual plot of the transformed data. Interpret the residual lot.

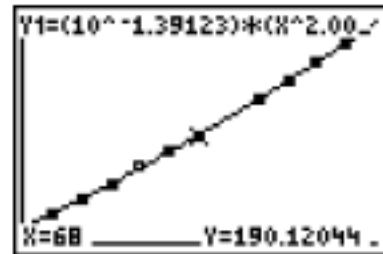


There is no discernable pattern in the residual plot. The line clearly fits the transformed data well.

(e) Perform the inverse transformation and write the equation for your model. Use your model to predict how many pounds a 5'10" adult would have to weigh in order to be classified by the department as "seriously overweight." Do the same for a 7-foot individual.

$$\hat{y} = 10^{(-1.39123 + 2.00286 \log x)}$$

The predicted severely overweight value for a 5'10" adult = 201.5 pounds. The predicted severely overweight value for a 7' adult = 290.3 pounds.



4.18 **EXACT EXPONENTIAL GROWTH, II** Fred and Alice were born the same year, and each began life with \$500. Fred added \$100 each year, but earned no interest. Alice added nothing, but earned interest at 7.5% annually. After 25 years, Fred and Alice are getting married. Who has more money?

Fred's equation  $f = 500 + 100x$

After 25 years, Fred has  $500 + 100(25) = \$3000.00$

Alice's equation  $a = 500 (1.075)^x$

After 25 years, Alice has  $500 (1.075)^{25} = \$3049.17$ . Alice has more money!!!

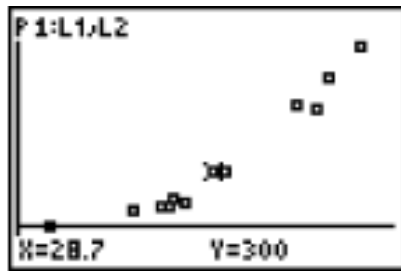
4.19 **FISH IN FINLAND, I** Here are the data for 12 perch caught in a lake in Finland.

Weight (grams)	Length (cm)	Width (cm)	Weight (grams)	Length (cm)	Width (cm)
5.9	8.8	1.4	300.0	28.7	5.1
100.0	19.2	3.3	300.0	30.1	4.6
110.0	22.5	3.6	685.0	39.0	6.9
120.0	23.5	3.5	650.0	41.4	6.0
150.0	24.0	3.6	820.0	42.5	6.6
145.0	25.5	3.8	1000.0	46.6	7.6

(a) Let L1 = length, L2 = weight

Look at scatterplot of length vs weight. What do you see?

The association is positive: as the length of the fish increases the weight increases. The association is strong, and the pattern is clearly curved.



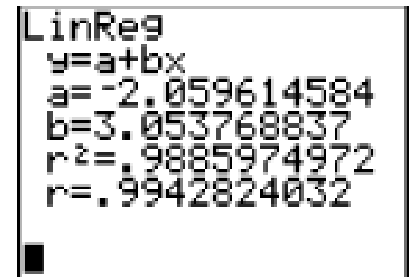
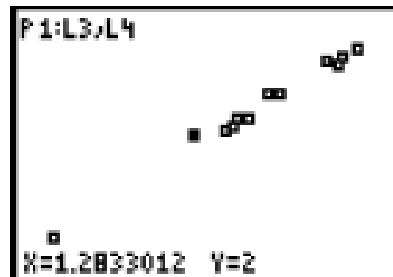
b) How do you expect the weight of animals of the same species to change as their length increases?

This is a power law relationship, so using logarithms to transform both length and weight should straighten the scatterplot.

(length is one-dimensional and weight is 3-dimensional, so we expect the weight of the fish to change proportional to the cube of the length.)

Let L3 = log (length) and  
L4 = log (weight)  
Look at scatterplot of  
log (length) vs log(weight)  
Is this plot now roughly linear?

Yes



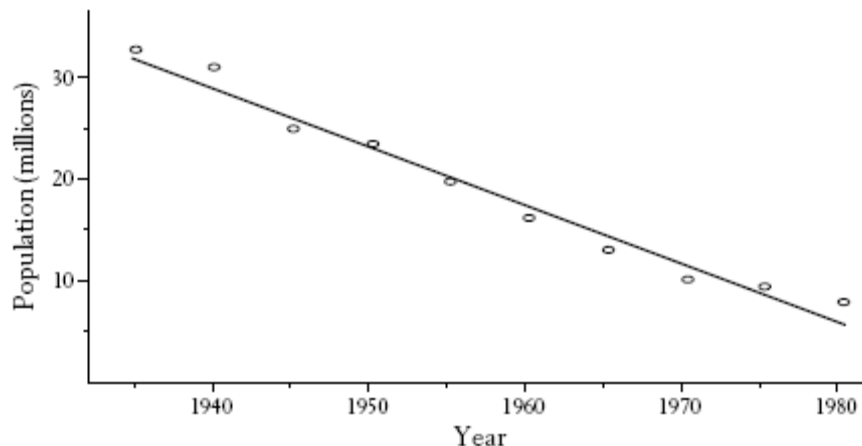
LSRL  $\log \hat{y} = -2.0596 + 3.0538 \log x$ ,  $r = .9943$

Back transformed equation  $\hat{y} = 10^{(-2.0596 + 3.0538 \log x)}$

4.27 **THE SIZE OF AMERICAN FARMS** The number of people living on American farms has declined steadily during this century. Here are data on the farm population (millions of persons) from 1935 to 1980.

Year:	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population:	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

(a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.  $\hat{y} = 1166.93 - 0.58679x$



(b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?

Based on the slope, the farm population decreased about 590 thousand (0.59 million) people per year. 97.7% of the variation in farm population is explained by linear regression on the year.

(c) Use the regression equation to predict the number of people living on farms in 1990. Is this result reasonable? Why?

-782,100 — clearly a ridiculous answer, since a population must be greater than or equal to 0.

4.28 **THE POWER OF HERBAL TEA** A group of college students believes that herbal tea has remarkable powers. To test this belief, they make weekly visits to a local nursing home, where they visit with the residents and serve them herbal tea. The nursing home staff reports that after several months many of the residents are more cheerful and healthy. A skeptical sociologist commends the students for their good deeds but scoffs at the idea that herbal tea helped the residents. Identify the explanatory and response variables in this informal study. Then explain what lurking variables account for the observed association.

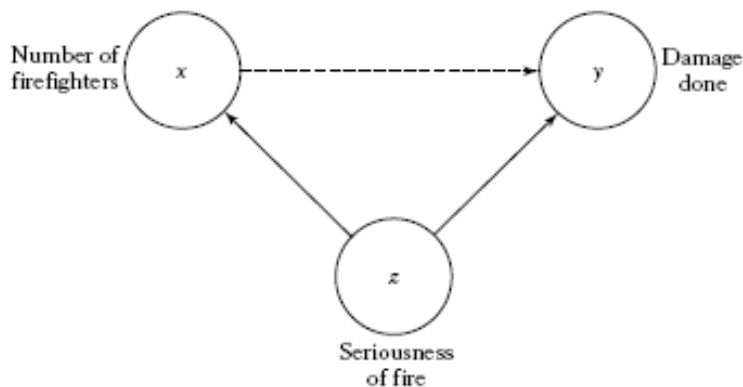
The explanatory and response variables were “consumption of herbal tea” and “cheerfulness.” The most important lurking variable is social interaction—many of the nursing home residents may have been lonely before the students started visiting.

**4.30 HOW TO SHORTEN A HOSPITAL STAY** A study shows that there is a positive correlation between the size of a hospital (measured by its numbers of beds  $x$ ) and the median number of days  $y$  that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital?

No; more likely it means that patients with more serious conditions (which require longer stays) tend to go to larger hospitals, which are more likely to have the facilities to treat those problems.

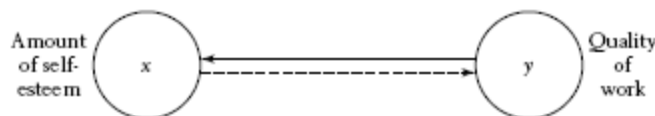
For Exercises 4.33 – 4.36, answer the question. State whether the relationship between the two variables involves causation, common response, or confounding. Identify possible lurking variable(s). Draw a diagram of the relationship in which each circle represents a variable. Write a brief description of the variable by each circle.

**4.33 FIGHTING FIRES** Someone says, “There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage.” Why is this reasoning wrong? Seriousness of the fire is a lurking (common-response) variable: more serious fires require more attention and do more damage. It would be more accurate to say that a large fire “causes” more firefighters to be sent, rather than vice versa.



**4.34 HOW’S YOUR SELF-ESTEEM** People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help them do better in school and life. Raising self-esteem became for a time a goal in many schools. California even created a state commission to advance the cause. Can you think of explanation for the association between high self-esteem and good school performance other than “Self-esteem causes better work in school”?

A reasonable explanation is that the cause-and-effect relationship goes in the other direction: Doing well makes students feel good about themselves, rather than vice versa.



4.35 **SAT MATH AND VERBAL SCORES** On page 70 in your text there is a table giving education data for the states. The correlation between the average SAT math scores and the average SAT verbal scores for the states is  $r = 0.962$ .

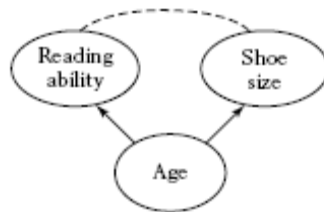
(a) Find  $r^2$  and explain in simple language what this number tells us.

$r^2 = 0.925$  —more than 90% of the variation in one SAT score can be explained through a linear regression on the other score.

(b) If you calculated the correlation between the SAT math and verbal scores of a large number of individual students, would you expect the correlation to be about 0.96 or quite different? Explain your answer.

The correlation would be much smaller, since individual students have much more variation between their scores. Some may have greater verbal skills and low scores in math (or vice versa); some will be strong in both areas, and some will be weak in both areas. By averaging—or, as in this case, taking the median of—the scores of large groups of students, we muffle the effects of these individual variations.

4.36 **BETTER GRADES** A study of elementary school children, ages 6 to 11, finds a high positive correlation between shoe size  $x$  and score  $y$  of a test of reading comprehension. What explains this correlation? Age is the lurking (common–response) variable here: we would expect both quantities—shoe size and reading level—to increase as a child ages.

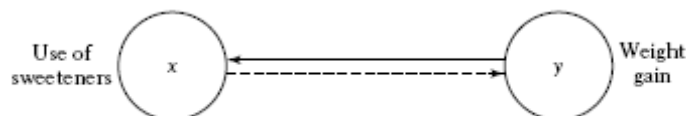


HW #25 38 – 45, 52, 53, 55

For Exercises 4.38 – 4.45, carry out the instructions. Then state whether the relationship between the two variables involves causation, common response, or confounding. Identify possible lurking variable(s). Draw a diagram of the relationship in which each circle represents a variable. Write a brief description of the variable by each circle.

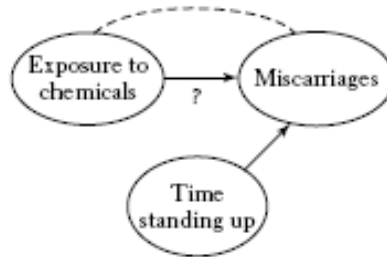
4.38 **DO ARTIFICIAL SWEETENERS CAUSE WEIGHT GAIN?** People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

In this case, there may be a causative effect, but in the direction opposite to the one suggested: People who are overweight are more likely to be on diets, and so choose artificial sweeteners over sugar. (Also, heavier people are at a higher risk to develop diabetes; if they do, they are likely to switch to artificial sweeteners.)



**4.39 DOES EXPOSURE TO INDUSTRIAL CHEMICALS CAUSE MISCARRIAGES?** A study showed that women who work in the production of computer chips have abnormally high numbers of miscarriages. The union claimed that exposure to chemicals used in production causes the miscarriages. Another possible explanation is that these workers spend most of their time standing up.

Time standing up is a confounding variable in this case. The diagram below illustrates the confounding between exposure to chemicals and standing up.



**4.40 IS MATH THE KEY TO SUCCESS IN COLLEGE?** Here is the opening of a newspaper account of a College Board study on 15,941 high school graduates:

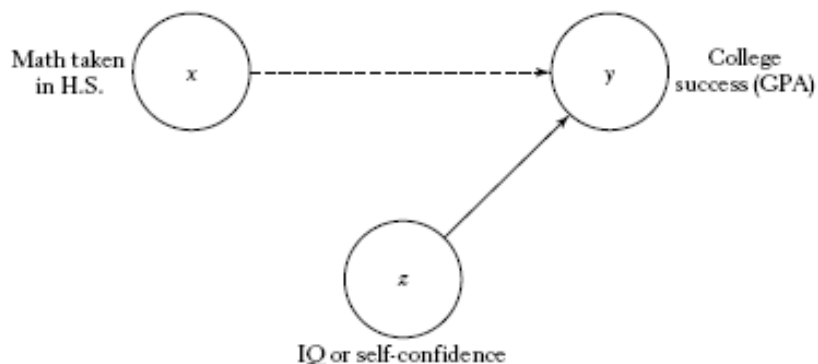
*Minority students who take high school algebra and geometry succeed in college at almost the same rate as whites, a new study shows.*

*The link between high school math and college graduation is “almost magical,” says College Board President Donald Stewart, suggesting “math is the gatekeeper for success in college.”*

*“These findings,” he says, “justify serious consideration of a national policy to ensure that all students take algebra and geometry.”*

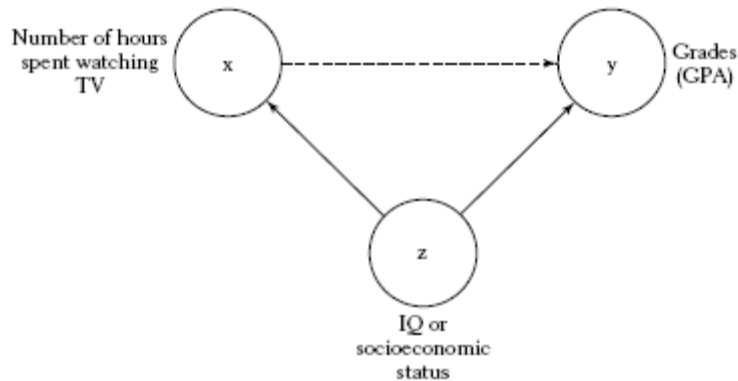
What lurking variables might explain the association between taking several math courses in high school and success in college? Explain why requiring algebra and geometry may have little effect on who succeeds in college.

A student’s intelligence may be a lurking (confounding) variable: Stronger students (who are more likely to succeed once they get to college) are more likely to choose to take these math courses, while weaker students may avoid them. Other possible answers might be variations on this idea; for example, if we believe that success in college depends on a student’s self-confidence, we might suppose that confident students are more likely to choose math courses.



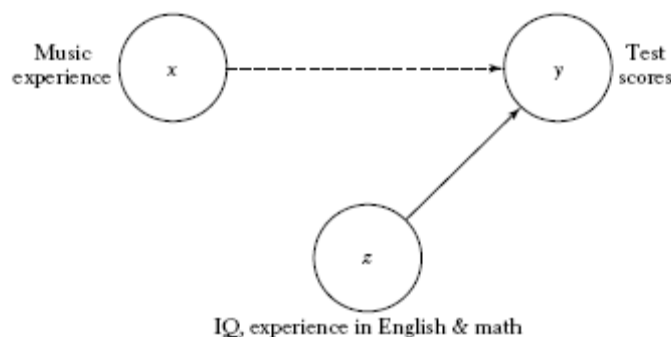
**4.41 ARE GRADES AND TV WATCHING LINKED?** Children who watch many hours of television get lower grades in school on the average than those who watch less TV. Explain clearly why this fact does not show that TV *causes* poor grades. In particular, suggest some other variables that may be confounded with heavy TV viewing and may contribute to poor grades.

This is a case of common response. It could be that children with lower intelligence watch many hours of TV and get lower grades as well. Or children from lower socioeconomic households where the parent(s) are less likely to limit television viewing and be unable to help their children with their schoolwork because the parents themselves lack education may watch more TV and get lower grades as a result.



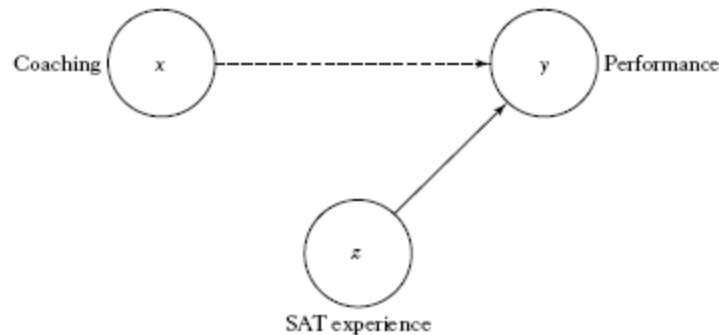
**4.42 MOZART FOR MINORS** In 1998, the Kalamazoo (Michigan) Symphony advertised a “Mozart for Minors” program with this statement: “Question: Which students scored 51 points higher in verbal skills and 39 points higher in math? Answer: Students who had experience in music.” What do you think of the claim that “experience in music” causes higher test scores?

Several possible lurking variables could help explain this supposed “causal” relationship. Factors such as innate intelligence and experience in English and mathematics classes are hopelessly confounded with the effects of music experience on test scores. Both music experience and test scores could also be responding to the socioeconomic status of the student. Students from affluent families may be more likely to receive extensive exposure to music and may also do better on tests because they can afford test-taking courses and other costly means of preparation.



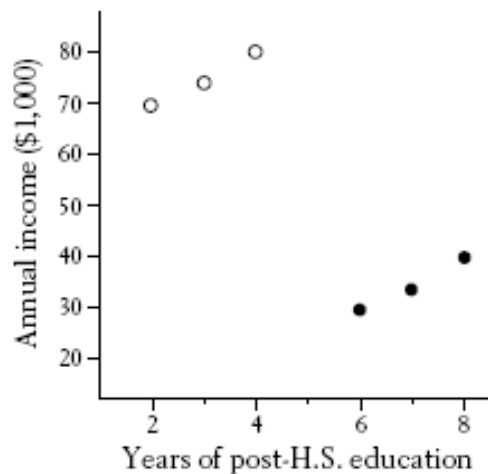
**4.43 RAISING SAT SCORES** A study finds that high school students who take the SAT, enroll in an SAT coaching course, and then take the SAT a second time raise their SAT mathematics scores from a mean of 521 to a mean of 561. What factors other than “taking the course causes higher scores” might explain this improvement?

The effects of coaching are confounded with those of experience. A student who has taken the SAT once may improve his or her score on the second attempt because of increased familiarity with the test.



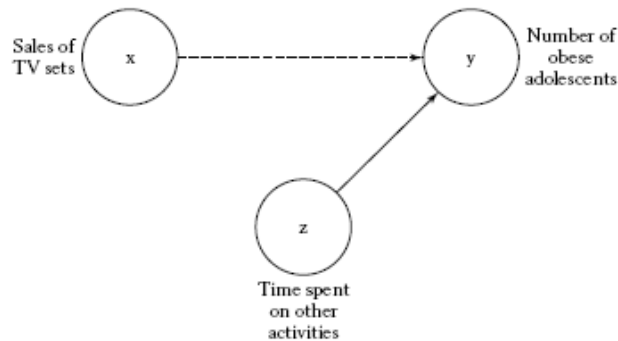
**4.44 ECONOMISTS’ EDUCATION AND INCOME** There is a strong positive correlation between years of education and income for economists employed by business firms. (In particular, economists with doctorates earn more than economists with only a bachelor’s degree.) There is also a strong correlation between years of education and income for economists employed by colleges and universities. But when all economists are considered, there is a negative correlation between education and income. The explanation for this is that business pays high salaries and employs mostly economists with bachelor’s degrees, while colleges pay lower salaries and employ mostly economists with doctorates. Sketch a scatterplot with two groups of cases (business and academic) that illustrates how a strong positive correlation within each group and a negative overall correlation can occur together.

The plot below is a very simplified (and not very realistic) example—open circles are economists in business; filled circles are teaching economists. The plot should show positive association when either set of circles is viewed separately, and should show a large number of bachelor’s degree economists in business and graduate degree economists in academia.



4.45 **TV AND OBESITY** Over the last 20 years there has developed a positive association between sales of television sets and the number of obese adolescents in the United States. Do more TVs cause more children to put on weight, or are there other factors involved? List some possible lurking variables.

Spending more time watching TV means that less time is spent on other activities: these may suggest lurking (confounding) variables. For example, perhaps the parents of heavy TV watchers do not spend as much time at home as other parents. Also, heavy TV watchers would typically not get as much exercise. Another factor could be the economy; as the economy has grown over the past 20 years, more families can afford TV sets (even multiple TV sets), and as a result, TV viewing has increased and children have less physical work to do in order to make ends meet

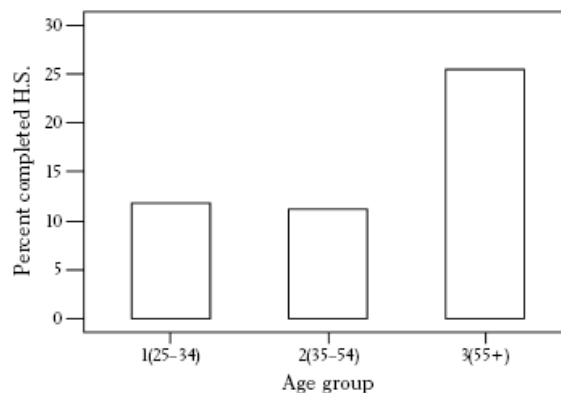


4.52 Using the counts in the table below, find the percent of people in each age group who did not complete high school. Draw a bar graph that compares these percents. State briefly what the data show.

**TABLE 4.6** Years of school completed, by age, 2000 (thousands of persons)

Education	Age group			Total
	25 to 34	35 to 54	55+	
Did not complete high school	4,474	9,155	14,224	27,853
Completed high school	11,546	26,481	20,060	58,087
1 to 3 years of college	10,700	22,618	11,127	44,445
4 or more years of college	11,066	23,183	10,596	44,845
Total	37,786	81,435	56,008	175,230

11.8%, 11.2%, and 25.4%. The percentage of people who did not finish high school is about the same for the 25–34 and 35–54 age groups, but more than doubles for the 55 and over age group.



4.53 **SMOKING BY STUDENTS AND THEIR PARENTS** Here are data from eight high schools on smoking among students and among their parents.

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400

- (a) How many students do these data describe? **5375**
- (b) What percent of these students smoke?  **$1004/5375 = 18.7\%$**
- (c) Give the marginal distribution of parents' smoking behavior, both in counts and in percents.  
**Both parents smoke: 1780 (33.1%)**  
**One parent smokes: 2239 (41.7%)**  
**Neither parent smokes: 1356 (25.2%)**

4.55 **IS HIGH BLOOD PRESSURE DANGEROUS?** Medical researchers classified each of a group of men as "high" and "low" blood pressure, then watched them for 5 years. (Men with systolic blood pressure 140 mm Hg or higher were "high"; the others "low.") The following two-way table gives the results of the study:

	Died	Survived
Low blood pressure	21	2655
High blood pressure	55	3283

- (a) How many men took part in the study? What percent of these men died during the 5 years of the study? **6014; 1.26%**
- (b) The two categorical variables in the table are blood pressure (high or low) and outcome (died or survived). Which is the explanatory variable? **Blood pressure is explanatory.**
- (c) Is high blood pressure associated with a higher death rate? Calculate and compare percents to answer this question. **Yes, among those with low blood pressure, 0.785% died; the death rate in the high blood pressure group was 1.65% - about twice as high as the other group.**

4.59 **MAJORS FOR MEN AND WOMEN IN BUSINESS** A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded:

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

(a) Find the two conditional distributions of major, one for women and one for men. Based on your calculations, describe the differences between women and men with a graph and in words.

Use column percentages (68/225 = 30.2% of females are in accounting, etc.) The biggest difference between women and men is in Administration, a higher percentage of women chose this major. Meanwhile a greater proportion of men chose other fields, especially finance.

	Female	Male
Accounting	30.2%	34.8%
Admin.	40.4%	24.8%
Econ.	2.2%	3.7%
Fin.	27.1%	36.6%

(b) What percent of the students did not respond to the questionnaire? The nonresponse weakens conclusions drawn from these data. There were 386 responses; 336/722 = 46.5% did not respond.

4.61 **RACE AND THE DEATH PENALTY** Whether a convicted murderer gets the death penalty seems to be influenced by the race of the victim. Here are data on 326 cases in which the defendant was convicted of murder:

	White defendant		Black defendant	
	White victim	Black victim	White victim	Black victim
Death	19	0	11	6
Not	132	9	52	97

(a) Use these data to make a two-way table of defendant's race (white or black) versus death penalty (yes or no).

	Black	White
Death	17	19
Not	149	141

(b) Show that Simpson's paradox holds: a higher percentage of white defendants are sentenced to death overall, but for both black and white victims a higher percent of black defendants are sentenced to death. Overall, 11.9% of white defendants and 10.2% of black defendants get the death penalty. However, for white victims, the percentages are 12.6% and 17.5% (respectively); when the victim is black, they are 0% and 5.8%.

(c) Use the data to explain why the paradox holds in language that a judge could understand.

In cases involving white victims, 14% of defendants got the death penalty; when the victim was black, only 5.4% of defendants were sentenced to death. White defendants killed whites 94.3% of the time—but are less likely to get the death penalty than blacks who killed whites.

4.70 **BASEBALL PARADOX** Most baseball hitters perform differently against right-handed and left-handed pitching. Consider two players, Joe and Moe, both of whom bat right-handed. The table below records their performance against right-handed and left-handed pitchers.

Player	Pitcher	Hits	At bats
Joe	Right	40	100
	Left	80	400
Moe	Right	120	400
	Left	10	100

(a) Make a two-way table of player (Joe or Moe) versus outcome (hit or no hit) by summing over both kinds of pitcher.

	Hit	No hit
<i>All pitchers</i>		
Joe	120	380
Moe	130	370
<i>Right-handed</i>		
Joe	40	60
Moe	120	280
<i>Left-handed</i>		
Joe	80	320
Moe	10	90

(b) Find the overall batting average (hits divided by total times at bat) for each player. Who has the higher batting average? Joe: .240, Moe: .260. Moe has the best overall batting average.

(c) Make a separate two-way table of player versus outcome for each kind of pitcher. From these tables, find the batting averages of Joe and Moe against right-handed pitching. Who does better? Do the same for left-handed pitching. Who does better?

Against right-handed pitchers: Joe: .400, Moe: .300.

Against left-handed pitchers: Joe: .200, Moe: .100.

Joe is better against both kinds of pitchers.

(d) The manager doesn't believe that one player can hit better against both left-handers and right-handers yet have a lower overall batting average. Explain in simple language why this happens to Joe and Moe.

Both players do better against right-handed pitchers than against left-handed pitchers.

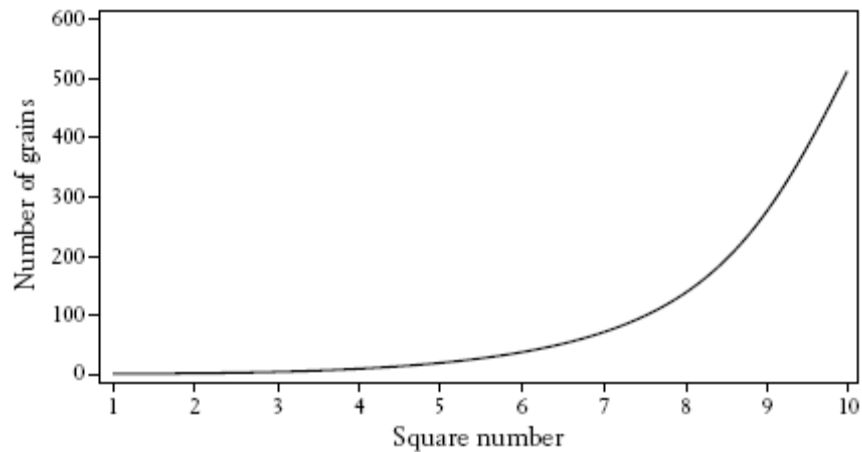
But Joe faced left-handers 80% of the time, while Moe only faced left-handers 20% of the time, so Moe's overall average is better!

4.74 **EXACT EXPONENTIAL GROWTH, I** A clever courtier, offered a reward by an ancient king of Persia, asked for a grain of rice on the first square of a chess board, 2 grains on the second square, then 4, 8, 16, and so on.

(a) Make a table of the number of grains on each of the first 10 squares of the board.

1, 2, 4, 8, 16, 32, 64, 128, 256, 512.

(b) Plot the number of grains on each square against the number of the square for squares 1 to 10, and connect the points with a smooth curve. This is an exponential curve.

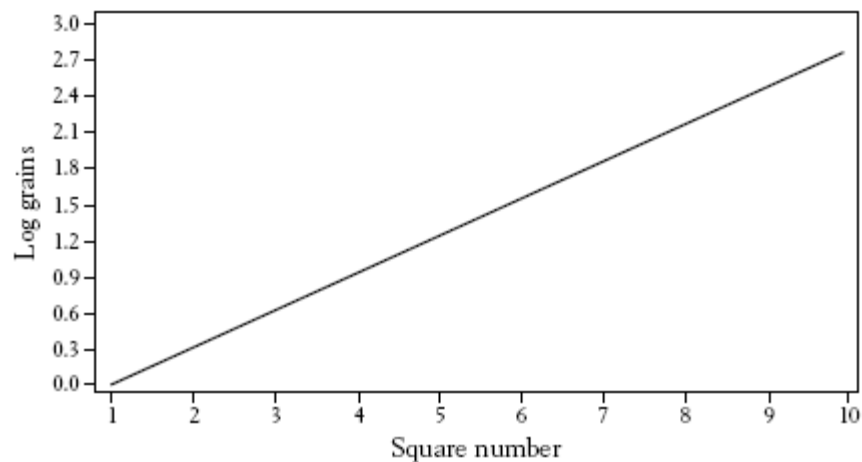


(c) How many grains of rice should the king deliver for the 64<sup>th</sup> (and final) square?

Approximately 9,000,000,000,000,000.

(d) Take the logarithm of each of your numbers of grains from (a). Plot these logarithms against the number of squares 1 to 10. You should get a straight line.

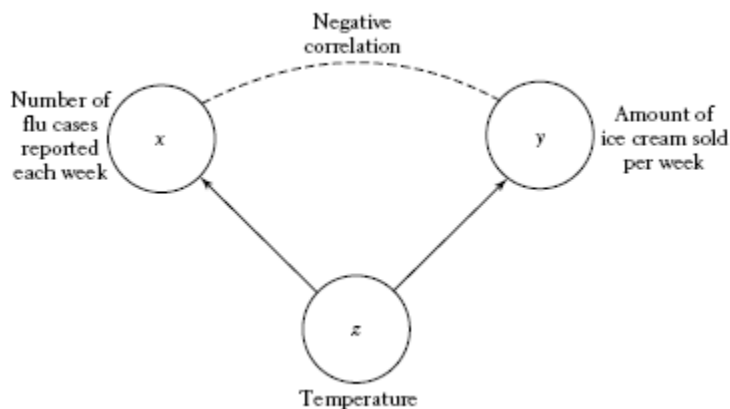
0.00, 0.30, 0.60, 0.90, 1.20, 1.51, 1.81, 2.11, 2.41, 2.71.



(e) From your graph in (d) find the approximate values of the slope  $b$  and the intercept  $a$  for the line. Use the equation  $y = a + bx$  to predict the logarithm of the amount for the 64<sup>th</sup> square. Check your result by taking the logarithm of the amount you find in (c).

$b = .3, a = -.3, \quad -.3 + .3(64) = 18.9. \quad \text{The log of the answer in part (c) is 18.95.}$

4.79 **ICE CREAM AND FLU** There is a negative correlation between the number of flu cases reported each week throughout the year and the amount of ice cream sold in that particular week. It's unlikely that ice cream prevents flu. What is a more plausible explanation for this observed correlation?



Both  $x$  and  $y$  are responding to the variable  $z = \text{temperature}$ . In the winter, when temperatures are low, there are many flu cases ( $x$ ) but relatively little ice cream sold ( $y$ ). In the summer, when temperatures are high, there are few cases of flu, but a large amount of ice cream is sold. This is an example of common response.

4.83 **SMOKING AND STAYING ALIVE** In the mid-1970s, a medical study contacted randomly chosen people in a district in England. Here are data on the 1314 women contacted who were either current smokers or who had never smoked. The table classifies these women by their smoking status and age at the time of the survey and whether they were still alive 20 years later.

	Age 18 to 44	
	Smoker	Not
Dead	19	13
Alive	269	327

	Age 45 to 64	
	Smoker	Not
Dead	78	52
Alive	167	147

	Age 65+	
	Smoker	Not
Dead	42	165
Alive	7	28

(a) Make a two-way table of smoking (yes or no) by dead or alive. What percent of the smokers stayed alive for 20 years? What percent of the nonsmokers survived? It seems surprising that a higher percent of smokers stayed alive.

	Smoker	Not	Total
Dead	139	230	369
Alive	443	502	945
Total	582	732	1314

76.1% of the smokers stayed alive for 20 years; 68.6% of the nonsmokers did so.

(b) The age of women at the time of the study is a lurking variable. Show that within each of the three age groups in the data, a higher percent of nonsmokers remained alive 20 years later. This is another example of Simpson's paradox.

For the youngest group, 96.2% of the nonsmokers and 93.4% of the smokers were alive 20 years later. For the middle group, 73.9% of the nonsmokers and 68.2% of the smokers were alive 20 years later. For the oldest group, 14.5% of the nonsmokers and 14.3% of the smokers were alive 20 years later. So the results are reversed when the data for the three age groups are combined.

(c) The study authors give this explanation: "Few of the older women (over 65 at the original survey) were smokers, but many of them had died by the time of follow-up." Compare the percent of smokers in the three age groups to verify the explanation.

45.9% of the youngest group, 55.2% of the middle aged group, but only 20.2% of the oldest group were smokers.