

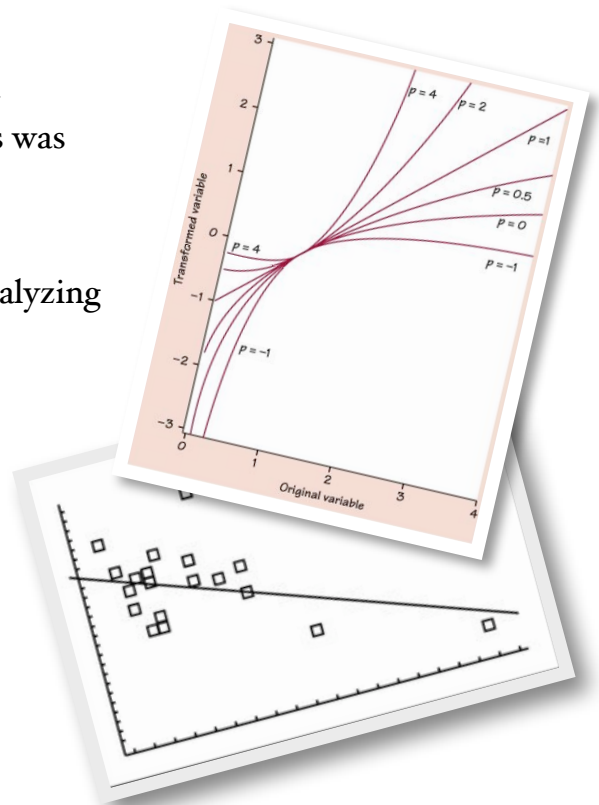
CHAPTER 4

MORE ON BIVARIATE DATA

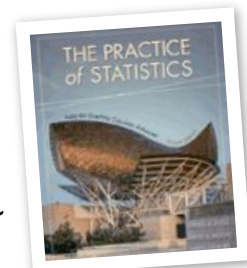
Chapter 3 discussed methods for describing and summarizing bivariate data. However, the focus was on *linear* relationships. In this chapter, we are introduced to methods for describing *nonlinear* bivariate relationships as well as methods for analyzing categorical data.

MORE ON BIVARIATE DATA:

- 4.1: Transforming Relationships
- 4.2: Cautions about Correlation
- 4.3: Relationships in Categorical Data



AP STATS CHAPTER 4: MORE ON BIVARIATE DATA



"NUMBERS ARE LIKE PEOPLE...TORTURE THEM LONG ENOUGH AND THEY'LL TELL YOU ANYTHING..."

Tentative Lesson Guide					
Date	Stats	Lesson	Assignment	Done	
Mon	10/23	4.1	Transforming Relationships	Rd 192-205 Do 3-5	
Tues	10/24	4.1	Power and Exponential Models	Rd 214-219 Do 13-16	
Wed	10/25	4.1	Modeling Nonlinear Data	Nonlinear Modeling Practice	
Thu	10/26	Rev	Review 4.1	Rd 149-156 Do 38-41	
Fri	10/27	Quiz	Quiz 4.1	Rd 157-165 Do 42-43, 45	
Mon	10/30	4.2	Cautions about Correlation	Rd 225-230 Do 27-31	
Tues	10/31	4.2	The Question of Causation	Rd 231-237 Do 33-37	
Wed	11/1	4.2	Causation Practice	Rd 238 Do 38-49	
Thu	11/2	4.3	Categorical Relationships	Rd 241-245 Do 53-55	
Fri	11/3	4.3	Simpson's Paradox	Rd 246-250 Do 59-61	
Mon	11/6	Rev	Review	Do 72-73, 81-83	
Tues	11/7	Exam	Exam Chapter 4	Online Quiz Due	

Note:

The purpose of this guide is to help you organize your studies for this chapter.

The schedule and assignments may change slightly.

Keep your homework organized and refer to this when you turn in your assignments at the end of the chapter.



Class Website:

Be sure to log on to the class website for notes, worksheets, links to our text companion site, etc.

<http://web.mac.com/statsmonkey>

Don't forget to take your online quiz!. Be sure to enter my email address correctly!

<http://bcs.whfreeman.com/yates2e>

My email address is:

jmmolesky@isd194.k12.mn.us

Chapter 4 Objectives and Skills:

These are the expectations for this chapter. You should be able to answer these questions and perform these tasks accurately and thoroughly. Although this is not an exhaustive review sheet, it gives a good idea of the "big picture" skills that you should have after completing this chapter. The more thoroughly and accurately you can complete these tasks, the better your preparation.

Transforming Data:

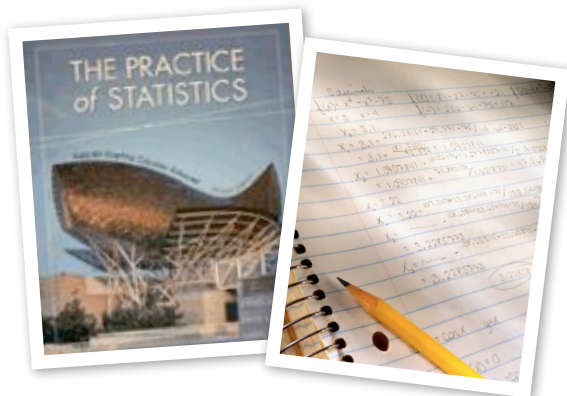
- If a scatterplot shows a curved pattern, it can perhaps be conveniently modeled by an exponential growth or decay function of the form $y = ab^x$ or a power function of the form $y = ax^b$. In these situations, we can linearize the data by making use of logarithms.
- If a scatterplot suggests an exponential or power function, we should check to see which model is appropriate by transforming x , y , or both variables and checking the linearity of the resulting scatterplot. If $(x, \log y)$ appears linear, an exponential model may work best. If $(\log x, \log y)$ appears linear, we may want to try a power model. Find the LSRL for the transformed data and then perform the reverse transformation to find the regression model. Know how to transform back to a power/exponential model

CAUTIONS ABOUT CAUSATION:

- Correlation and regression need to be interpreted with caution. Two variables may be strongly associated, but this does not mean that one causes the other. Bottom line: High correlation does not imply causation. Among other things, lurking variables and common response should always be considered:
 - Extrapolation: Using LSRL to predict values outside the domain of the explanatory variable used to create the line. (It is generally best to avoid extrapolating.)
 - Lurking variable: A variable that effects the relationship of the variables in a study, but is not included among the variables studied.
 - Correlation based on averages are frequently higher than correlation based on all of the numbers that make up the averages.
- If there is a strong association between two variables x and y , any one of the following statements could be true.
 - x causes y
 - both x and y are responding to changes in some unobserved variable or variables.
 - the effect of x on y is hopelessly mixed up with the effects of other variables on y . Because of this, it is impossible to directly determine the effects of x on y .

RELATIONSHIPS IN CATEGORICAL DATA:

- A two-way table of counts describes the relationship between two categorical variables... the row variable and the column variable. The row totals and column totals give the marginal distributions of the two variables separately, but do not give any information about the relationships between the variables. Probabilities, including conditional probabilities, can be calculated from two-way tables.
- In some cases, lurking variables can cause a reversal in the direction of a comparison or association when data from several groups are combined to form a single group. This phenomenon is called "Simpson's Paradox." Mmmmmmmmm...



4.1: Introduction to Nonlinear Relationships

Data that displays a curved pattern can be modeled by a number of different functions. Two of the most common nonlinear models are exponential ($y=AB^x$) and power ($y=Ax^B$). We will focus on these two models in this chapter. Our goal is to fit a model to curved data so that we can make predictions as we did in Chapter 3. However, the only statistical tool we have to fit a model is the least-squares regression model. Therefore, in order to find a model for curved data, we must first “straighten it out”...

Mmmmm...Theory.

- Consider an Exponential Model
 $y = A B^x$
- Take the log of both sides

- Consider a Power Model
 $y = A x^B$
- Take the log of both sides

Conclusion:

- If data is **exponential** in nature, the graph of $(x, \log y)$ will be linear.
- If data has a **power** relationship, the graph of $(\log x, \log y)$ will be linear.

In order to make predictions with our model, we need to convert the LSRL of the transformed data back to the original terms of the problem. We can do this by using inverse functions...

Predicting with an Exponential Model

LSRL of transformed data:

Exponential Model:

Predicting with a Power Model

LSRL of transformed data:

Power Model:

So, in a nutshell, to find an Exponential or Power Model:

- 1)
- 2)
- 3)
- 4)

“Moleskium” Decay Activity

The decay of radioactive isotopes typically creates a nonlinear relationship between time and amount remaining. “Moleskium” is a little-known element found in a short, highly caffeinated math teacher at LSHS. Without proper caffeination, “Moleskium” can decay quite rapidly...trust me, it’s not a pretty site when that happens.

We will be simulating the decay of “Moleskium” through the use of Skittles® bite-sized candies. {Actually depriving Mr.M. of caffeine to observe the real-life decay of “Moleskium” would be unethical, so we’ll rely on Skittles to simulate the phenomenon.} As we proceed through this simulation, be sure to dispose of the “decayed” Skittles in a nearby oral cavity or waste-disposal container.

Disclaimer: DO NOT eat *actual* radioactive isotopes. These are *pretend* so it’s ok.

“Moleskium” Decay Simulation:

- Start with a sizeable ($n > 60$) quantity of Skittles in a covered container.
- Shake the container to simulate typical Mr.M. activity.
- Open the container and remove all Skittles that have “decayed” {ie, no “S” showing}
- Count the remaining Skittles and record in the table below.
- Repeat until all Skittles have “decayed”.

Round Number	Isotopes Remaining
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	

Enter the “Round” into L1 and “Isotopes” into L2.

Display the relationship with a scatterplot and interpret in context:



Use your calculator to find the LSRL for (round, isotopes):

LSRL: _____

Sketch the residual plot for this LSRL and use it to discuss the appropriateness of using a linear model for this relationship.

“Moleskium” Decay Activity - Continued

We can tell by the scatterplot of (round, isotopes) and by the residual plot of (round, residual) that a nonlinear model may be a more appropriate model for this situation. However, the question that remains is, “Which model is more appropriate? Exponential or Power?” Our eyes are fairly good at judging whether or not points lie on a straight line. However, they are not very good at spotting the difference between power and exponential curves. Therefore, we’ll use the fact that certain transformations of power and exponential data produce linear relationships. If we can determine which transformation does a better job of “straightening out the data” then we’ll have a better idea which model is best.

Use your list editor to define $L3 = \log(\text{round})$ and $L4 = \log(\text{isotopes})$.

Remember, if $(x, \log y)$ is linear, an **exponential** model may be best. If $(\log x, \log y)$ is linear, a **power** model may be best. All we have to do is plot the two and determine which is “more linear”.

<input checked="" type="checkbox"/> Sketch $(x, \log y)$ and interpret	<input checked="" type="checkbox"/> Sketch $(\log x, \log y)$ and interpret
--	---

Find the LSRL for the transformed data that exhibits the most linear relationship. Justify “most linear” by considering the correlation and residual plots for each set of transformed data.

Since the LSRL is written in terms of transformed data, transform it back into the original terms of the problem. That is, find a prediction model that will convert “round number” into “isotopes remaining”.

Final Prediction Model for “Moleskium” Decay: _____

4.1: Nonlinear Transformation Practice

1. Do aircraft with a higher typical speed also have greater average flight lengths? Use the following data to create a model that predicts the average flight length (miles) for a plane based on its typical speed (mph). Show all scatterplots, residual plots, and analyses used to determine your model.

mph	518	539	529	498	513	498	504	516	467	524	523	495	464	486
mi	2882	5063	3321	1363	2451	1493	1963	2379	1126	3253	2995	2331	1167	2135
mph	441	440	458	414	432	416	374	388	412	387	389	384	380	
mi	782	742	1101	702	798	602	345	442	570	487	468	500	413	

Use your model to predict an average flight length for a plane that typically flies at a speed of 525mph.

2. The following data describe the number of police officers (thousands) and the violent crime rate (per 100,00 population) in a sample of states. Use these data to determine a model for predicting violent crime rate based on number of police officers employed. Show all appropriate plots and work.

Police	86.2	9.2	45	39.9	6	11.8	2.9	14.6	30.5	12.3	46.2	15.2	10.9
Crime	1090	559	1184	1039	303	951	132	763	635	726	840	373	523

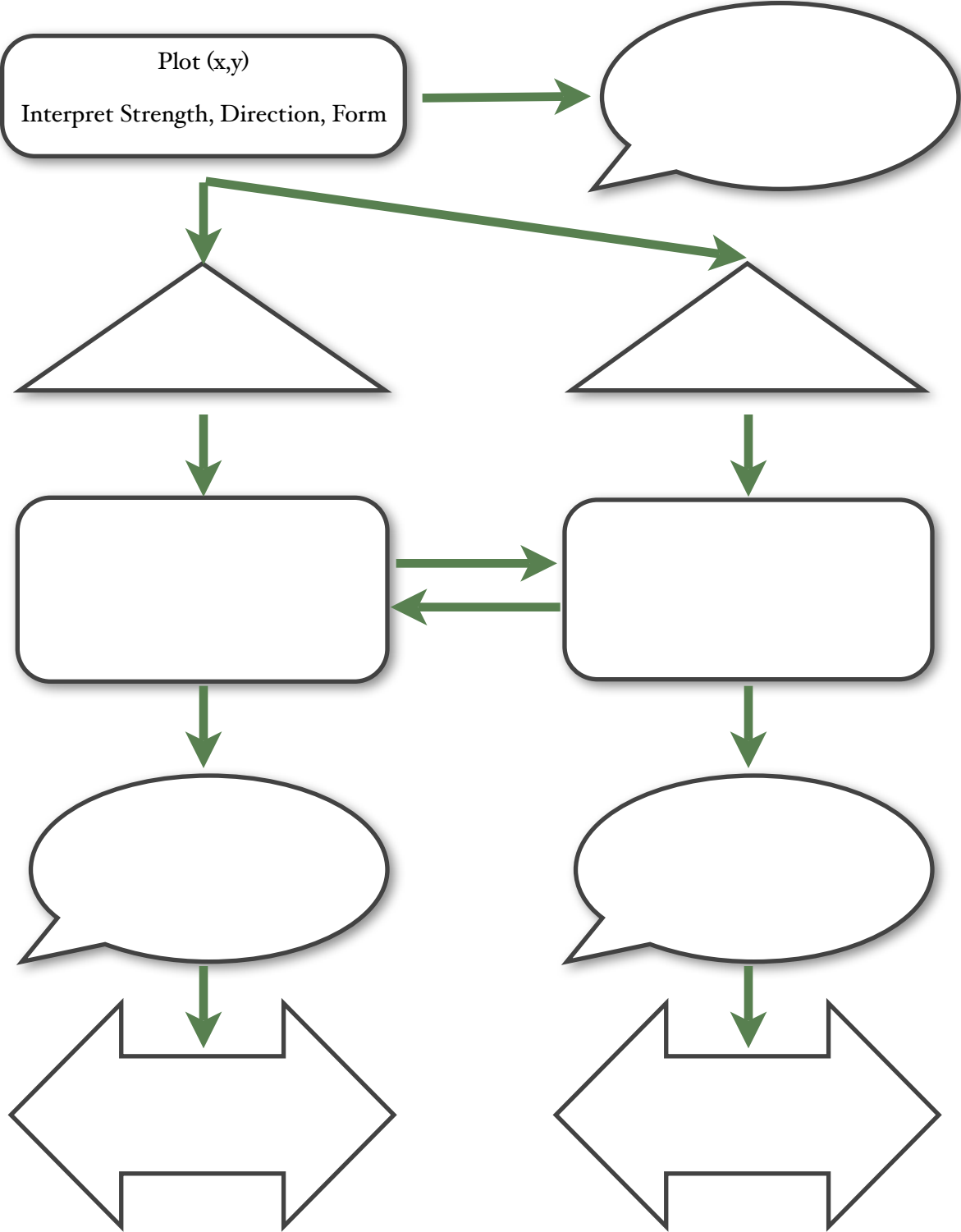
Use your model to predict the violent crime rate for a state with 25,400 police officers.

3. Consider the following data on population density of the United States (Statistical Abstract of the United States, 1996). Find a prediction model for population density based on year. Show all work.

Year	Density	Year	Density	Year	Density	Year	Density	Year	Density
1800	6.1	1840	9.8	1880	16.9	1920	35.6	1960	50.6
1810	4.3	1850	7.9	1890	21.2	1930	41.2	1970	57.4
1820	5.5	1860	10.6	1900	25.6	1940	44.2	1980	64.0
1830	7.4	1870	13.4	1910	31.0	1950	50.7	1990	70.3

Use your model to predict the population density in 2006.

Modeling NonLinear Data



The Question of Causation

When building prediction models for bivariate data, you should always start with a plot of your explanatory vs. response variables. Analysis of the residual plots and correlation coefficient can help you assess the appropriateness of your model for prediction. However, you should always use caution when interpreting exactly what the cause of the relationship between two variables is.

Consider the following historical data. Construct a prediction model based on the data and discuss the strength of the observed relationship. We will discuss what x and y are after we build the model:

Year	$x = ?$	$y = ?$
1860	63	8376
1865	48	6406
1870	53	7005
1875	64	8486
1880	72	9595
1885	80	10643
1890	85	11265
1895	76	10071
1900	80	10547
1905	83	11008
1910	105	13885
1915	140	18559
1920	175	23024
1925	183	24185
1930	192	25434
1935	221	29238
1940	262	34705



LSRL: _____

Correlation Coefficient $r =$ _____

What does the correlation imply about the relationship?

Note: $x =$ _____ and $y =$ _____

Important Note for the Day:

Possible Reasons for Strong Correlation:



The Unusual Episode - Relationships in Categorical Data

Consider the following categorical data. Take some time to read the tables and make a guess as to what could be described here. You may ask “yes” or “no” questions to help figure out what the “episode” was.

By Economic Status and Sex									
Econ Status	Population Exposed to Risk			Number of Deaths			Deaths per 100 Exposed to Risk		
	Male	Female	Both	Male	Female	Both	Male	Female	Both
I (high)	180	145	325	118	4	122	65	3	37
II	179	106	285	154	13	167	87	12	59
III	510	196	706	422	106	528	83	54	73
Other	862	23	885	670	3	673	78	13	76
Total	1731	470	2201	1364	126	1490	80	27	67

By Economic Status and Age									
Econ Status	Population Exposed to Risk			Number of Deaths			Deaths per 100 Exposed to Risk		
	Adult	Child	Both	Adult	Child	Both	Adult	Child	Both
I (high)	319	6	325	122	0	122	38	0	37
II	261	24	285	167	0	167	64	0	59
III	627	79	706	476	52	528	76	66	73
Other	885	0	885	673	0	673	76	0	76
Total	2092	109	2201	1438	52	1490	69	48	67

What was the “unusual episode” that was responsible for the categorical data above?

What are some of the conclusions you can reach about the risk, who was exposed, differences in groups, etc.? What are some ways you could display the relationships? What types of plots could you use?

Categorical Analyses - Simpson's Paradox

Discrimination in Admissions Policies

The University of California at Berkeley was charged with having discriminated against women in their graduate admissions process for the fall quarter of 1973. The table below identifies the number of acceptances and denials for both men and women applicants in each of the six largest graduate programs at the institution at the time:

Program	Men Accepted/ Applied	Women Accepted/ Applied
A	511/825	89/108
B	352/560	17/25
C	120/325	202/593
D	137/407	132/375
E	53/191	95/393
F	22/373	24/341
Total	1195/2681	559/1835

Ignore the "Programs" and collapse the data into a two-way table of gender by admission. Record your results below:

	Accepted	Denied	Total
Male			
Female			
Total			

Of men who applied to these programs, what percent were admitted? _____

Of women who applied to these programs, what percent were admitted? _____

What would you conclude about the admissions policies at Berkeley based on these proportions?

Program	% Men Accepted	% Women Accepted
A		
B		
C		
D		
E		
F		

Using the data in the first table, calculate the proportion of males and females accepted in each program.

Simpson's Paradox:

Categorical Analyses - Simpson's Paradox

Who's the Better Batter?

Occasionally, the effect of a confounding factor is strong enough to produce a paradox known as Simpson's Paradox. The paradox is that the relationship appears to be in a different direction when the confounding variable is not considered than when the data are separated into the categories of the confounding variable itself.

Here are the batting averages for two players for each half of the season:

	First Half of the Season			Second Half of the Season		
	Hits	At Bat	Average	Hits	At Bat	Average
Homer	60	200		50	200	
Bart	29	100		1	5	

What is Homer's Batting Average for the entire season? $\frac{\quad}{\quad} = \frac{\quad}{\quad}$

What is Bart's Batting Average for the entire season? $\frac{\quad}{\quad} = \frac{\quad}{\quad}$

Explain Simpson's Paradox in the context of this situation.

Race and the Death Penalty

The following data from "Racial Characteristics and Imposition of the Death Penalty," American Psychological Review, 46 (1981), pp. 918-927 refers to the Supreme Court Case "Furman v. Georgia, 408 U.S. 238 (1972)" in which it was argued that the imposition and carrying out of the death penalty held to constitute cruel and unusual punishment in violation of the Eighth and Fourteenth Amendments.

	White Defendant			Black Defendant		
Victim	White	Black	Total	White	Black	Total
Death	19	0		11	6	
Not	132	9		52	97	
Total						

Explain how Simpson's Paradox can lead you to two different conclusions regarding race and the death penalty.