Chapter 3: Describing Bivariate Data

Illustration: Gender vs Housework Study

Suppose a sociologist is interested in investigating the relationship between gender and amount of time spent on housework for married couples. Assuming she has a sample of subjects who are known to be married, what data might she collect from each subject to help in her investigation?

Recall that _____ *data* occurs when we observe two variables on each subject. Here, the bivariate data consists of one _____ and one _____ variable. It is also possible to have bivariate data with both variables quantitative or both variables qualitative.

Why do some studies require bivariate as opposed to univariate data? Bivariate data is needed to investigate the _____ between two variables

Bivariate data would be used to answer questions such as: Is there a _____ between

- The household income of a child and the number of years of school he will complete?
- Treatment A and amount of pain?
- Income and happiness? Education and income?
- A customer's income and their likelihood of buying a widget?

Summaries and Graphs for Bivariate Data

The techniques available depend on the _____ of the two variables.

Both Variables Qualitative

Contingency Table (also called cross-tabulation), side-by-side pie or bar charts, or stacked bar charts may be used. These summaries display _____ for every possible combination of the two variables.

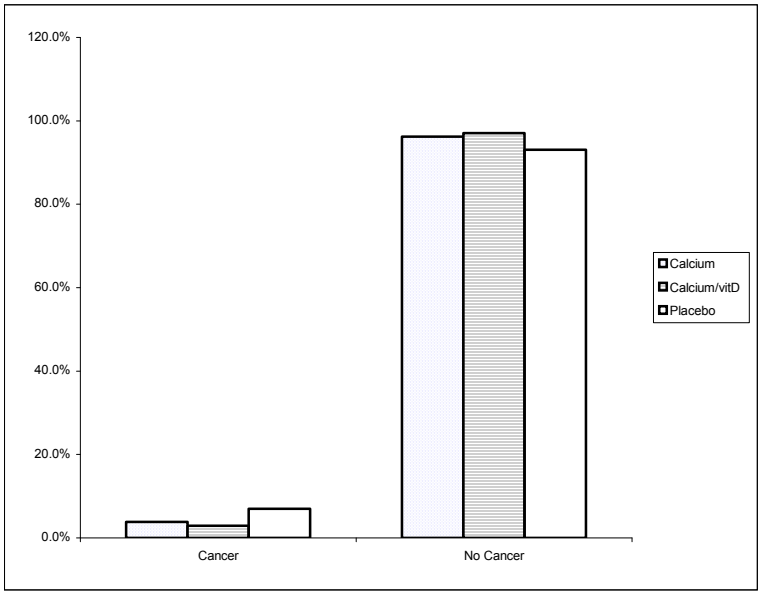
Example: Vitamin D vs Cancer

For each subject, we record the treatment and whether or not cancer occurred. We summarize the data using a _____ below.

	Cancer	No Cancer	Total
Calcium	17	429	446
Calcium/Vitamin D	13	433	446
Placebo	20	268	288

It's hard to tell if either treatment is effective due to differing sample sizes across treatments. Conditional data distributions help. These are the percent of cancer and no cancer *within* each treatment group (or conditional on the treatment group) and are shown below.

	Cancer	No Cancer
Calcium	17/446=0.04	429/446=0.96
Calcium/Vitamin D	??/446=0.03	433/446=.97
Placebo	20/288=0.07	268/???=0.93



One Qualitative/One Quantitative Variable

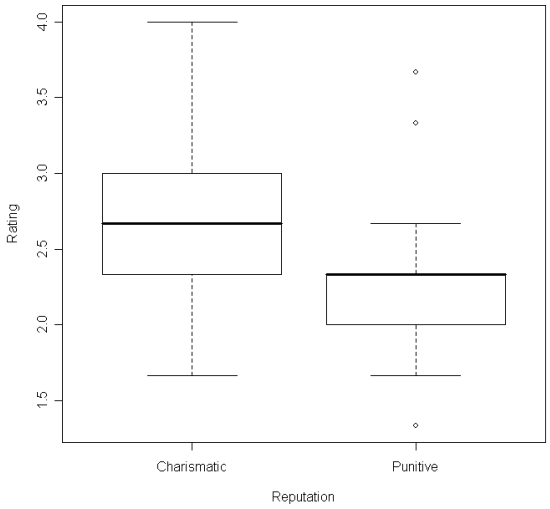
We may group by the _____ variable and calculate a five-number summary for each group. These may be displayed graphically using side-by-side _____.

Grouping on the qualitative variable, we may also summarize each group with the mean and standard deviation. Side-by-side histograms may be helpful (use the same scale on the x-axis).

Example: Instructor Reputation and Teacher Rating

Study to determine if an instructor's reputation affected student evaluations of her teaching ability. Random assignment to groups. One group told the instructor had a reputation for being good and charismatic; the other group told the instructor had a bad reputation (punitive). Both watched the same 20 minute lecture and numerically evaluated the instructor's teaching.

Five-number summaries: Charismatic 1.7, 2.3, 2.7, ___, 4.0
Punitive 1.3, 2.0, ___, 2.3, 3.7



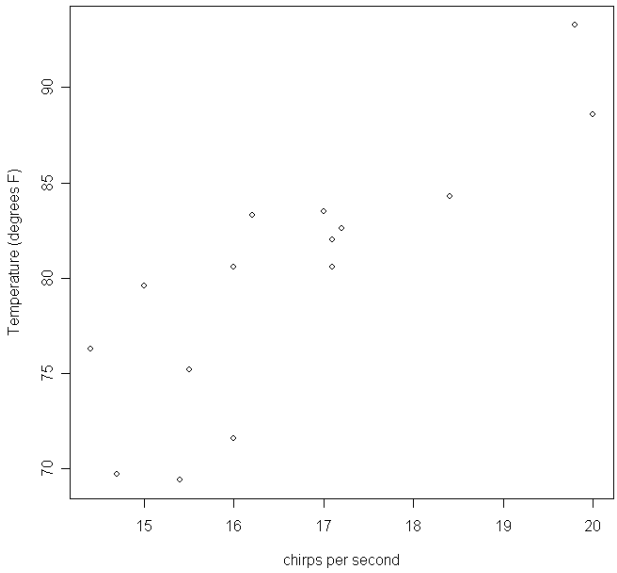
Both Variables Quantitative

Data are graphed using a _____. We will learn about some important summary numbers for this type of data next.

Example: Temperature and Cricket Chirps

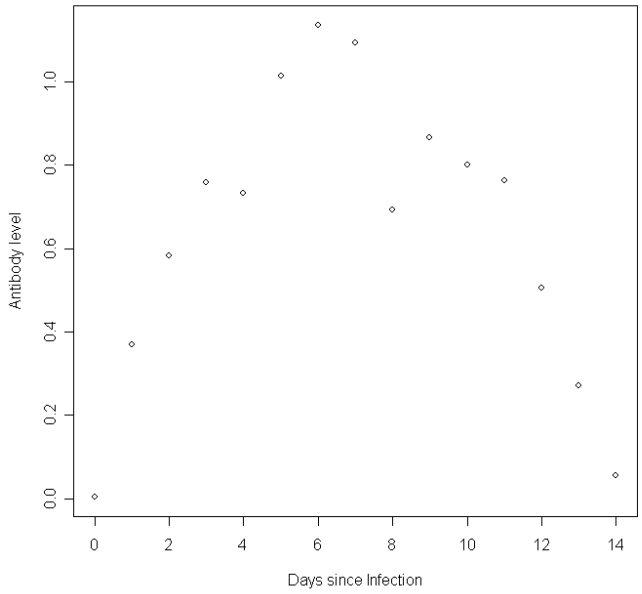
What can we tell from the scatterplot? Three features of interest:

_____ of the scatterplot – does it roughly resemble a line, parabola or something else
_____ of the relationship – how closely does the scatterplot fit to the pattern
_____



Linear Regression and Correlation

Consider the scatterplot for *x*=days since infection with a virus and *y*=antibody level. How could you describe the average pattern of the data?



We note that not all scatterplots exhibit a linear pattern. However, for this class we will concentrate on modeling _____ ***relationships only*** (i.e., chirps vs temperature data). Before applying the following methods, check the scatterplot to ensure the two variables follow a _____ pattern.

For bivariate data that follow a linear pattern, we wish to:

- measure the strength of the linear relationship
- calculate the line that best fits the data
- given a value of x, use the best-fit line to predict y; x = explanatory or independent variable; y = response or dependent variable

Correlation Coefficient

The correlation coefficient, denoted r, is a measure of the _____ of the linear relationship between two variables observed for a sample of subjects.

$$\text{covariance of x and y} = s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

An equivalent, but easier to use, formula for computing the covariance is

$$s_{xy} = \frac{\left(\sum x_i y_i\right) - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{n-1}$$

Facts about r:

- $-1 \le r \le 1$
- If r>0, x and y are positively correlated, which means that as x increases, y also _____
- If r<0, x and y are negatively correlated, which means that as x increases, y _____
- r=0 means there is ____ linear correlation between x and y. However, they may be correlated in a non-linear manner like the antibody data. Plots of how this can happen:

- If r = 1, all points in the scatterplot will fall on a line with positive slope; if r = -1, all points will fall on a line with negative slope. As r moves away from either ± 1 toward 0, the points will become more spread out around the best-fitting line. More plots:

Example: Use the applet to show the value of r for different scatterplots.

Exercise: Match each dataset with the most likely correlation coefficient selected from: -0.8, ,0, 0.8

a) x=energy consumption of a residence in a month,y=electricity bill for the same month
b) x=last digit of student's SSN, y=number of CD's student owns

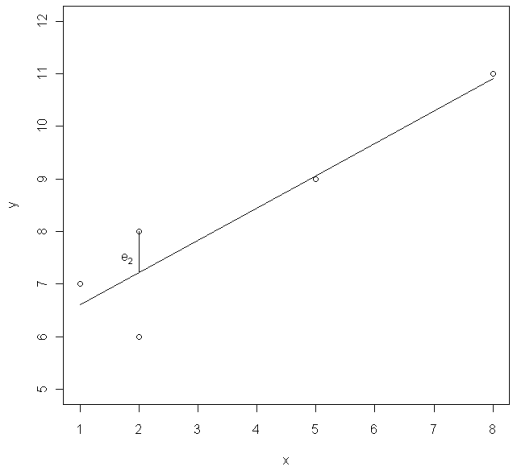
c) x=age of a car, y= resale value of the car

+Example: Calculate and interpret the correlation coefficient for the given data.

x	1	2	2	5	8
y	7	8	6	9	11

The Equation of the Best Fit Line

This line is the "model" for our data and will be used to calculate predicted value of y for a given value of x. The best fit line can be written in the form $y = a + bx$ where we will give a formula for the y-intercept, *a*, and the slope, *b*, shortly. The line that "best fits" the data will minimize the squares of the residuals, $\sum e_i^2$. One residual, e_2, is shown in the graph below. Draw the remaining residuals on the plot.



The slope and intercept of the best fit line are calculated as follows:

$$b = \frac{s_{xy}}{s_x^2} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}$$

Example: Calculate the best fit line for the data from the last example.