Lecture Notes & Examples 3.1

Section 3.1 – Scatterplots and Correlation (pp. 143-163)

Most statistical studies examine data on more than one variable. We will continue to use tools we have already learned as well as adding others to assist us in analysis.

- Plot the data, add numerical summaries
- Look for overall patterns and deviations from those patterns
- If there is a regular pattern, use a simplified model to describe it

1. Explanatory and Response Variables

Definition: A **response variable** measures the outcome of a study. An **explanatory variable** *may* help explain or influence changes in a response variable.

This means that the *explanatory variable* "accounts for" or "predicts" changes in the response variable.

Examples:

Explanatory	Response
Amt of Rain	Weed Growth
Win % of BB Team	Attendance @ games
Amt of Exercise	Resting Pulse Rate

CHECK YOUR UNDERSTANDING Identify the explanatory and response variables in each setting.

1. How does drinking beer affect the level of alcohol in our blood? The legal limit for driving in all states is 0.08%. In a study, adult volunteers drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol levels.

   Explanatory: # of cans of beer

   Response: blood alcohol level

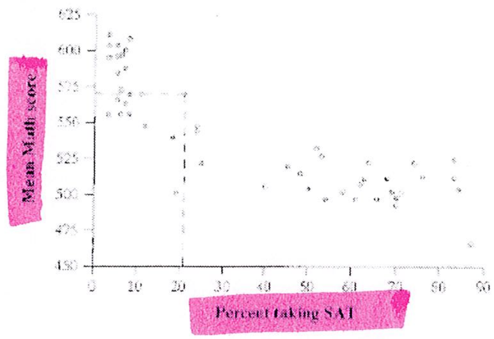
2. The National Student Loan Survey provides data on the amount of debt for recent college graduates, their current income, and how stressed they feel about college debt. A sociologist looks at the data with the goal of using amount of debt and income to explain the stress caused by college debt.

   Explanatory: amount of debt and income

   Response: stress caused by college debt

2. Displaying Relationships: Scatterplots

Definition: A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each individual in the data set appears as a point on the graph.



If there is an explanatory variable, it is plotted on the x-axis and the response variable is on the y-axis.

If there is no explanatory-response distinction, either variable can go on the x-axis.

Discuss Association → States with higher % taking SAT typically have lower mean math scores.

How to make a scatterplot:

1. Decide which variable should go on which axis.
2. Label and scale your axes
3. Plot individual data values

(Common error on AP Exam – failing to label axes.)

Calculator:

TI-84 pg 149 Put data in L_1 and L_2

Plot 1

ON

Type: [✓ ⁙] ⩘ ▥

XList: L_1

Y List: L_2

Mark: [✓ □] ÷ .

Zoom ZoomStat (#9)

3. Interpreting Scatterplots

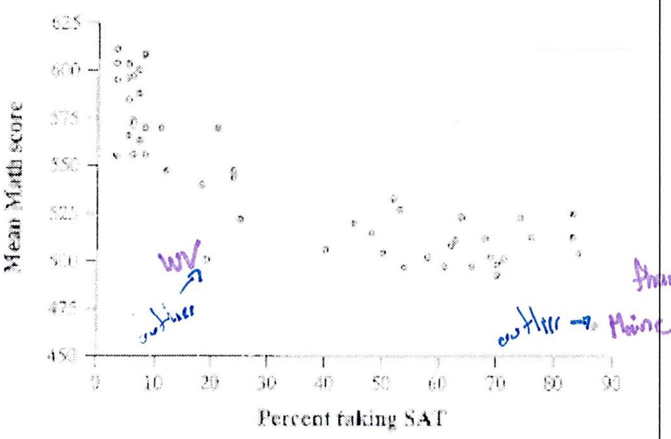
How to examine a scatterplot

Look for *overall pattern* and for striking *departures* from that pattern

- Overall pattern is described by the **direction**, **form**, and **strength** of the relationship.
- An important type of departure is an **outlier**, an individual pattern that falls outside the overall pattern of the relationship.

Influential Points!

DOFS In Context!!



D: In general, it appears that the higher % taking SAT, The lower the Mean Math Score. (pattern moves from upper left to lower right) Negative Association between 2 variables.

F: The Form of the relationship is slightly curved with 2 distinct clusters. In about ½ of states less than 25% took SAT, in other ½ more than 40% took SAT.

S: Moderately Strong: States with similar % taking SAT tend to have roughly similar mean Math scores.

O: Two states standout. WV at (19, 501) and Maine at (87, 466). They fall outside overall Pattern.

Definition:
Two variables have a **positive association** when the above average values of one tend to accompany above average values of the other and when below average values also tend to occur together.

Two variables have a **negative association** when the above average values of one tend to accompany below average values of the other.

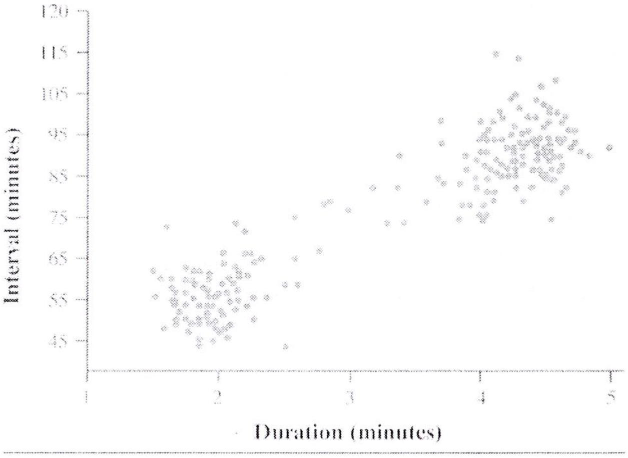
******Causation and Association******

Association does not imply causation!!!!!

Examples:

- AP Calculus causes Global Warming.
As # of AP Calc Exams has grown over the years, so has avg global warming temp.

- months with higher ice cream sales imply more drownings.



CHECK YOUR UNDERSTANDING

Here is a scatterplot that plots the interval between consecutive eruptions of Old Faithful (a geyser) against the duration of the previous eruption.

1. Describe the direction of the relationship. Explain why this makes sense.

Relationship is positive. The longer the duration of the erruption, the longer the wait between eruptions. This might be due to when expending more energy, it takes longer to build up the energy to erupt again

2. What form does the relationship take? Why are there two clusters of points?

The form is roughly linear with 2 clusters. Clusters indicate that there are 2 types of erruptions, one shorter, the other somewhat longer.

3. How strong is the relationship? Justify your answer.

The relationship is fairly strong. Two points define a line, and in this case we could think of each cluster as a point, so the 2 clusters seem to define a line.

4. Are there any outliers?

There are a few outliers around the clusters, but not many and not very distant from the main grouping of points

5. What information does the Starnes family need to predict when the next eruption will occur?

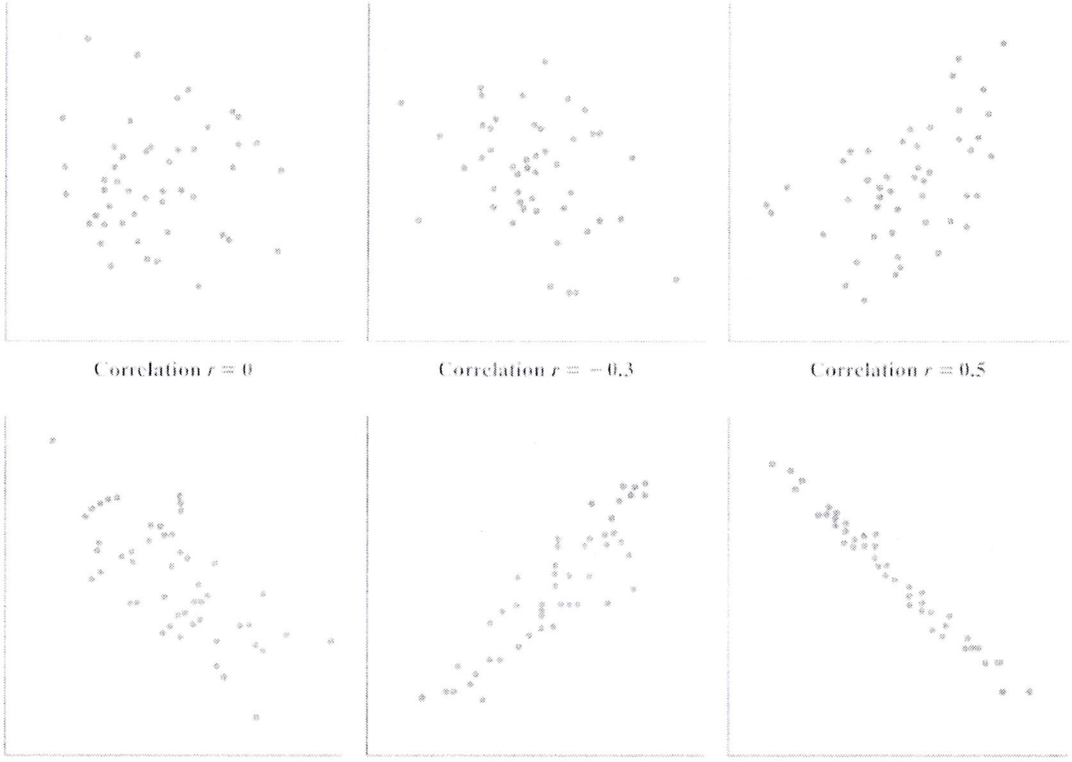
They need to know how long the last eruption lasted in order to predict how long until the next one.

4. Measuring Linear Association: Correlation

A linear relationship may appear in a scatterplot. The linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line. We are going to use a statistic called **correlation** to measure linearity in a scatterplot. **Correlation *r*** measures the *direction* and *strength* of the linear relationship between two quantitative variables.

The correlation r is always a number between -1 and 1. The sign indicates the direction of the association. Values close to 0 indicate a weak linear relationship. As r approaches -1 or 1, the strength of the relationship increases. -1 and 1 only occur if the values lie *exactly* on a straight line.

Refer to figure 3.6 on page 151 for examples of different values of r.



Correlation $r = 0$ Correlation $r = -0.3$ Correlation $r = 0.5$

Correlation $r = -0.7$ Correlation $r = 0.9$ Correlation $r = -0.99$

Calculator TI-84

Linear Reg → r

Catalog

Diagnostics ON

$\boxed{\text{STAT}}$ → CALC #8: LinearReg L₁, L₂, y

$\boxed{\text{VARS}}$ → Y-VARS #1: Function #1: Y₁

[illegible]

Weak → 0 → 0.5

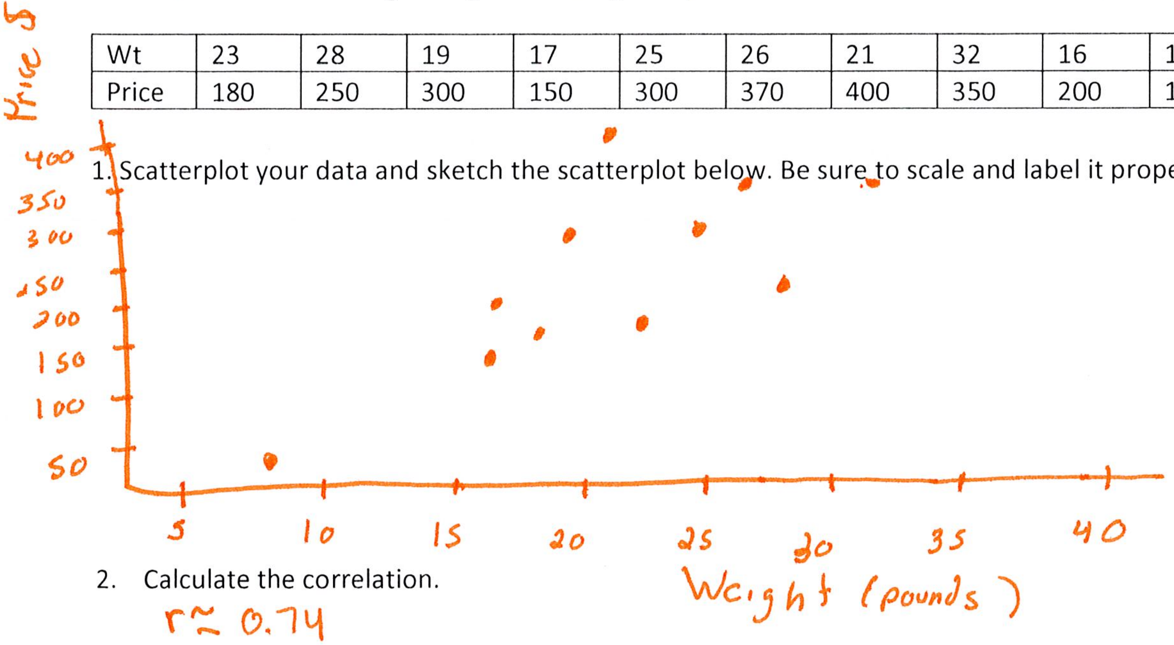
Moderate → 0.5 → 0.8

Strong → 0.8 - 1.0

Team work: The following data give the weight in pounds and cost in dollars of a sample of 11 stand mixers.

Wt	23	28	19	17	25	26	21	32	16	17	8
Price	180	250	300	150	300	370	400	350	200	150	30

1. Scatterplot your data and sketch the scatterplot below. Be sure to scale and label it properly.



2. Calculate the correlation.

$r \approx 0.74$

3. The last mixer in the table is from Walmart. What happens to the correlation when you remove this point?

$r \approx 0.56$

4. What happens to the correlation if the Walmart mixer weighs 25 pounds instead of 8 pounds? Add the point (25, 30) and recalculate the correlation.

$r \approx 0.34$

5. Suppose a new titanium mixer was introduced that weighed 8 points, but the cost was $500. Remove the point (25, 30) and add the point (8, 500). Recalculate the correlation.

$r \approx -.08$

6. Summarize what you learned about the effect of a single point on the correlation.

A single point can affect correlation greatly.
(Correlation is not resistant)

How to calculate r

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

AP Formula Sheet

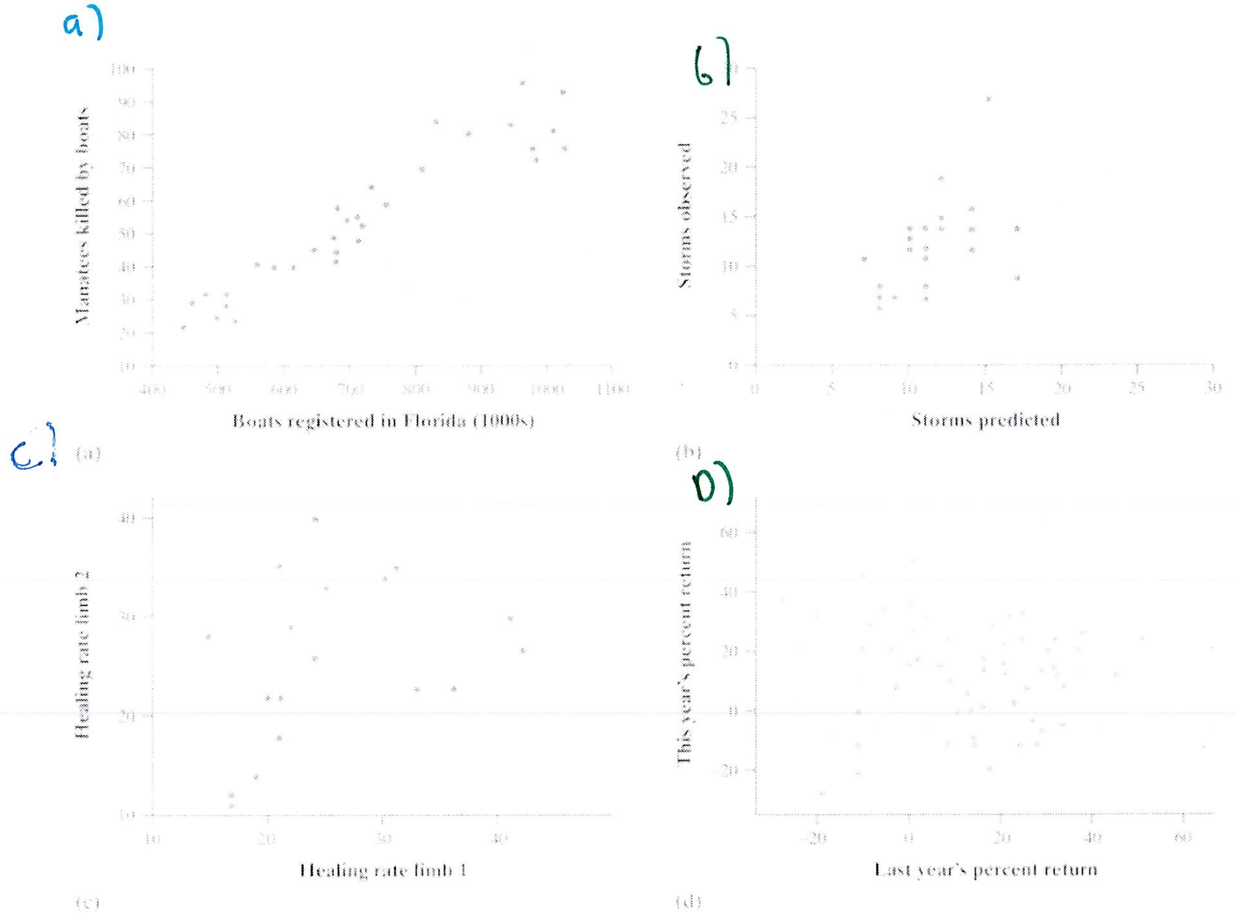
What does this mean? $r = \frac{1}{n-1} \sum z_x z_y$

Begins by standardizing observations. Multiply z scores. Average Products
(2-Dimensional z-score)

Note: A value of r close to 1 or -1 *does not guarantee a linear relationship between two variables*. A scatterplot with a clear curved form can have a correlation that is near -1 or 1. **Always plot your data!** *

CHECK YOUR UNDERSTANDING

The scatterplots below show four sets of real data: (a) repeats the manatee plot; (b) shows the number of named tropical storms and the number predicted before the start of hurricane season each year between 1984 and 2007 by William Gray of Colorado State University; (c) plots the healing rate in micrometers (millionths of a meter) per hour for the two front limbs of several newts in an experiment; and (d) shows stock market performance in consecutive years over a 56-year period.



1. For each graph, estimate the correlation r. Then interpret the value of r in context.

a) r≈0.9 There is a strong, positive linear relationship between # of boats registered in FL and # of manatees killed.

b) r≈0.5 There is a moderate, positive linear relationship, between the # of named storms predicted and actual # of named storms.

c) r≈0.3 There is a weak positive linear relationship between the healing rate of the 2 front limbs of the newts.

2. The scatterplot in (b) contains an outlier: the disastrous 2005 season, which had 27 named storms, including Hurricane Katrina. What effect would removing this point have on the correlation? Explain.

d) r≈-0.1. There is a weak, neg linear relationship between last year's % return and this years % return in the stock market.

2. The correalation would decrease. This point has the effect of strengthening the observed linear relationship we see.

5. Facts about Correlation

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

1. Correlation makes no distinction between explanatory and response variables.

Which variable is x and which is y doesn't matter

2. Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x, y, or both.

3. The correlation r itself has no unit of measurement. It's just a #

4. Correlation requires that both variables be quantitative.

5. Correlation measures the strength of only the linear relationship between ~~tow~~ two variables. It does not describe curved relationships between variables.

6. The correlation is not *resistant*: it is strongly affected by a few outlying observations.

7. Correlation is not a complete summary of two-variable data. You should always give means and standard deviations of both x and y along with the correlation.

Team work. Read and discuss the example on p. 156

Example – Scoring Figure Skaters Why Correlation doesn't tell the whole story

Until a scandal at the 2002 Olympics brought change, figure skating was scored by judges on a scale from 0.0 to 6.0. The scores were often controversial. We have the scores awarded by two judges, Pierre and Elena,for many skaters. How well do they agree? We calculate that the correlation between their scores is r = 0.9. But the mean of Pierre's scores is 0.8 point lower than Elena's mean.

These facts don't contradict each other. They simply give different kinds of information. The mean scores show that Pierre awards lower scores than Elena. But because Pierre gives every skater a score about 0.8 point lower than Elena does, the correlation remains high. Adding the same number to all values of either x or y does not change the correlation. If both judges score the same skaters, the competition is scored consistently because Pierre and Elena agree on which performances are better than others. The high r shows their agreement. But if Pierre scores some skaters and Elena others, we should add 0.8 point to Pierre's scores to arrive at a fair comparison.

.