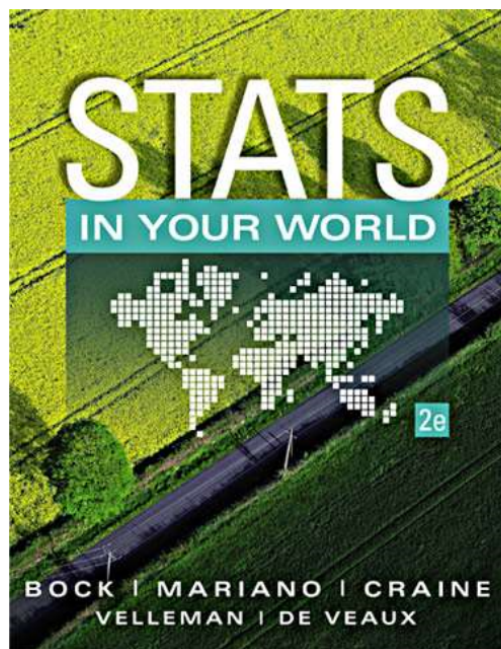


Chapter 6

A Tale of Two Variables

Quantitative



Looking at Scatterplots

- **Scatterplots** may be the most common and most effective display for data.
 - In a scatterplot, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others.



Scatterplots are the best way to start observing the relationship and the ideal way to picture associations between two *quantitative* variables.

Looking at Scatterplots (cont.)

When looking at scatterplots, we will look for direction, form, strength, and unusual features.

Handwritten notes: \pm slope, linear, curved, weak, moderate, strong, ~~strong~~, outliers

■ Direction:

- A pattern that runs from the upper left to the

lower right is said to have a negative direction. (association)

- A trend running the other way
- has a positive direction. (association)

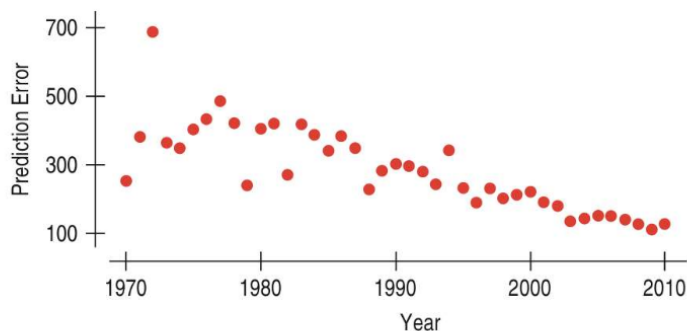


as $x \uparrow, y \downarrow$

as $x \uparrow, y \uparrow$

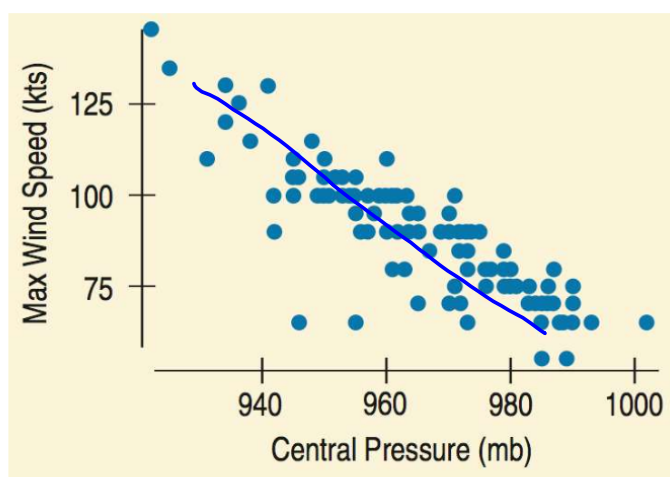
Looking at Scatterplots (cont.)

Can the NOAA predict where a hurricane will go?



- The figure shows a negative direction between the year since 1970 and the prediction errors made by NOAA.
- As the years have passed, the predictions have improved (errors have decreased).

Looking at Scatterplots (cont.)

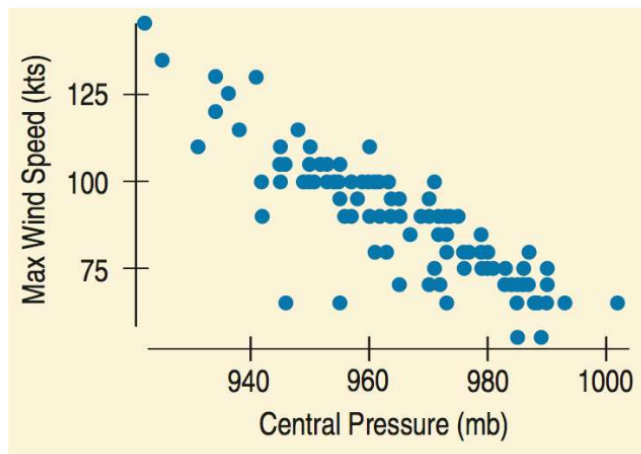


- The example in the text shows a negative association between central pressure and maximum wind speed
- As the central pressure increases, the maximum wind speed decreases.

Looking at Scatterplots (cont.)

- Form:

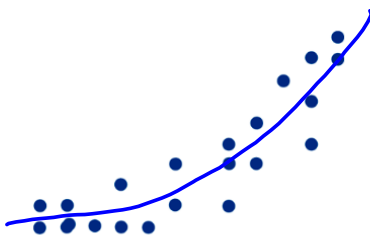
- If there is a straight line (linear) relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form.



Looking at Scatterplots (cont.)

- Form:

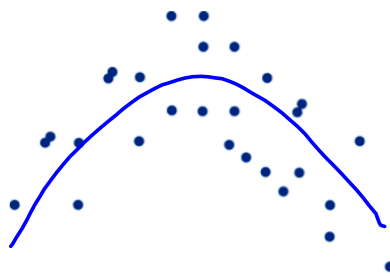
- If the relationship isn't straight, but curves gently, while still increasing or decreasing steadily,



we can often find ways to make it more nearly straight (chapter 9!).

Looking at Scatterplots (cont.)

- Form:
 - If the relationship curves sharply,

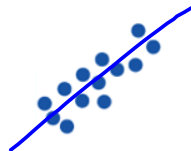


a linear method is not helpful.

Looking at Scatterplots (cont.)

■ Strength:

- The points can appear to follow a single stream



(whether straight, curved, or bending all over the place).

- We say this is relatively strong association.

Looking at Scatterplots (cont.)

- Strength:
 - If the points appear as a vague cloud with no discernible trend or pattern:




- We say this a weak association. (No association)
- Note: we will quantify the amount of scatter soon.

Looking at Scatterplots (cont.)

- Unusual features:

- Look for the unexpected.

 Often the most interesting thing to see in a scatterplot is the thing you never thought to look for.

- One example of such a surprise is an outlier standing away from the overall pattern of the scatterplot.

Roles for Variables

- It is important to determine which of the two quantitative variables goes on the x-axis and which on the y-axis.
- This determination is made based on the roles played by the variables.
- When the roles are clear, the **explanatory** or **predictor variable** goes on the x-axis, and the **response variable** (variable of interest) goes on the y-axis.

Roles for Variables (cont.)

- The roles that we choose for variables are more about how we think about them rather than about the variables themselves.
- Just placing a variable on the x-axis doesn't necessarily mean that it explains or predicts *anything*. And the variable on the y-axis may not respond to it in any way.

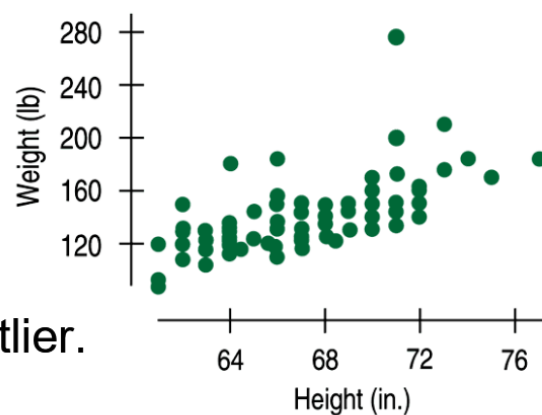
Correlation \leftarrow linear association

- Data collected from students in Statistics classes included their heights (in inches) and weights (in pounds):



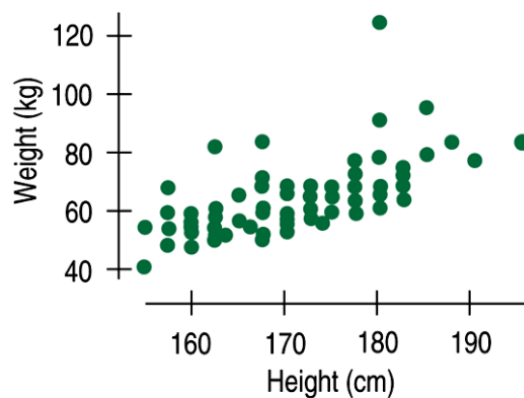
Good description:
moderately positive
association with a
fairly straight (linear)
form, although there
seems to be a high outlier.

- form
- dir
- strength
- unusual



Correlation (cont.)

- How strong is the association between weight and height of Statistics students?
- If we had to put a number on the strength, we would not want it to depend on the units we used.
- A scatterplot of heights (in centimeters) and weights (in kilograms) doesn't change the shape of the pattern:



Correlation (cont.)

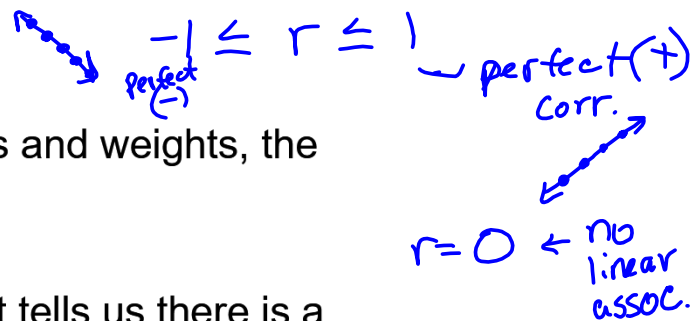
- The correlation coefficient (r) gives us a numerical measurement of the strength of the linear relationship between the explanatory and response variables.

- Don't let this formula scare you! Technology will provide us with the correlation.

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

- The formula depends on z-scores which have no units.

Correlation (cont.)






- For the students' heights and weights, the correlation is 0.644.
- What does this mean?
- The sign is positive, so it tells us there is a positive association.
- 0.644 is moderate in strength.
- So we say a correlation of 0.644 tells us there is a positive, moderate, linear relationship between height and weight.
- Of course, we'd also like to see a scatterplot! — look for outliers

Correlation Properties


- The sign of a correlation coefficient gives the direction of the association.
- Correlation is always between -1 and $+1$.
 - Correlation *can* be exactly equal to -1 or $+1$, but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.
 - A correlation near zero corresponds to a weak (or no) linear association.

Correlation Properties (cont.)

-  Correlation treats x and y symmetrically:
 - The correlation of x with y is the same as the correlation of y with x .
-  Correlation has no units.
-  Correlation is not affected by changes in the center or scale of either variable.
 - Correlation depends only on the z -scores, and they are unaffected by shifting or rescaling.

Correlation Properties (cont.)

- Correlation measures the strength of the linear association between the two quantitative variables.
 - Variables can have a strong association but still have a small correlation if the association isn't linear. As always, make a picture.
- Correlation is sensitive to outliers. A single outlying value can make a small correlation large or make a large one small. As always, *beware of outliers*.

ex:

 $r \sim 0$
But strong
association
(No corr.)

Correlation Conditions

- **Correlation** measures the strength of the linear association between two quantitative variables.

* Before you use correlation, you must check several conditions:

- Quantitative Variables Condition
- Straight Enough Condition
- No Outliers Condition

look at
Scatterplot

Correlation Conditions (cont.)



- Quantitative Variables Condition:
 - Correlation applies only to quantitative variables.
 - Don't apply correlation to categorical data masquerading as quantitative.
 - Check that you know the variables' units and what they measure.

Correlation Conditions (cont.)

- **Straight Enough Condition:**
 - Correlation measures the strength only of the *linear* association, and will be misleading if the relationship is not linear.
 - Thus we *only* calculate and use the correlation coefficient for linear data.

Correlation Conditions (cont.)

■ No Outliers Condition:

-  Outliers can distort the correlation dramatically (like it would the mean or standard deviation).
 - An outlier can make an otherwise small correlation look big or hide a large correlation.
 - It can even give an otherwise positive association a negative correlation coefficient (and vice versa).
-  When you see an outlier, it's often a good idea to report the correlations with and without the point.

Review Question!

- Which statistics have we studied so far this year that are resistant to outliers?

*Center = Median
Spread = IQR*

- Which statistics are not resistant?

*Center = Mean
Spread = SD*

"Correlation" vs. "Association"

✓ The word "correlation" refers specifically to an association between two linear and quantitative variables.

✓ Also, correlation does not prove causation.
Ex. Ice cream sales vs. A/C sales

Lurking Variable

- a hidden variable that stands behind a relationship and simultaneously affects the other two variables

Ex. Ice cream sales and A/C sales are have a moderately strong linear correlation, however, both are affected by the lurking variable, temperature.

Read Pg. 137 - 146

Homework 1:

pg. 154-155 #1, 2, 3, 7, 10

For #7 write GOOD, detailed answers, a few sentences each. Be sure to answer all parts of the question.

#7) $x = \text{explanatory}$
 $y = \text{Response}$

Packet Page 8

Homework 2:

pg. 155-156 #6, 11, 13, 14, 17, 18

#6) ~~like~~ like #7
also form, strength, dir.

Correlation:

- Quant. Var
- ~~is~~ straight enough
- No outliers

Packet Pg. 16