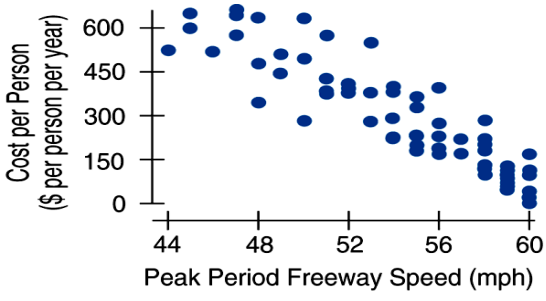
Chapter 7 I can describe the association (direction, form, strength) of a scatterplot that relates two quantitative variables. I can then find the correlation, when the association appears linear. Date:_____

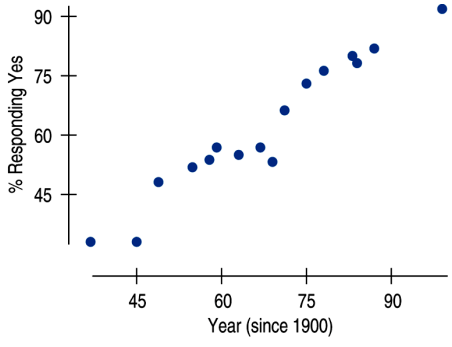
Scatterplots may be the most common and most effective display for data. In a scatterplot, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others. Scatterplots are the best way to start observing the relationship and the ideal way to picture associations between two ***quantitative* variables.**

When looking at a scatterplot, we will look for its associations (**direction** , **form** , **strength**), and for any unusual features.

Direction:

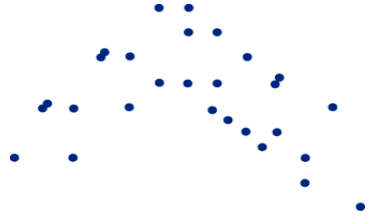
- A pattern that runs from the upper left to the lower right is said to have a negative direction.
- A trend running the other way has a positive direction.





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.
- 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 (re-expressing the data).
- If the relationship curves sharply, the methods of this book cannot really help us.



Strength:

- At one extreme, the points appear to follow a single stream (whether straight, curved, or bending all over the place).
- At the other extreme, the points appear as a vague cloud with no discernable trend or pattern:

Unusual Features:

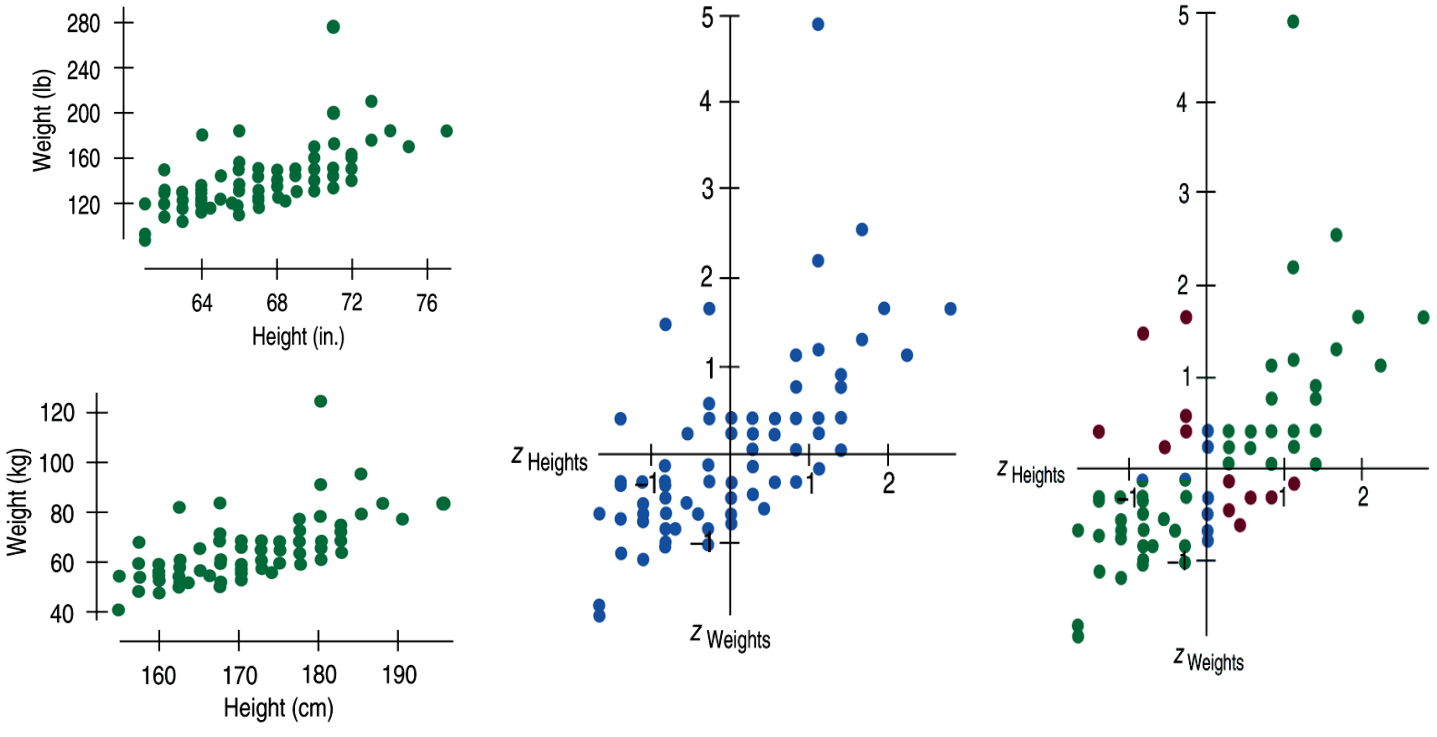
- Look for the unexpected.
- One example is an outlier standing away from the overall pattern of the scatterplot.
- Clusters or subgroups should also raise questions. They may be a clue that you should split the data into subgroups instead of looking at it all together.

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 goes on the *y*-axis.
- 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

- Data collected from students in Statistics classes included their heights (in inches) and weights (in pounds). The graph below shows a positive association and a fairly straight form, although there seems to be a high outlier.
- 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.
- Since the units don't matter, why not remove them altogether? We could standardize both variables and write the coordinates of a point as (z_x, z_y). Below is a scatterplot of the standardized weights and heights. Note that the underlying linear pattern seems steeper in the standardized plot than in the original scatterplot. That's because we made the scales of the axes the same. Equal scaling gives a neutral way of drawing the scatterplot and a fairer impression of the strength of the association.
- The 4th scatterplot shows some points (those in green) that strengthen the impression of a positive association between height and weight. Other points (those in red) tend to weaken the positive association. Points with *z*-scores of zero (those in blue) don't vote either way.



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

$$r = \frac{\sum z_x z_y}{n-1} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}$$

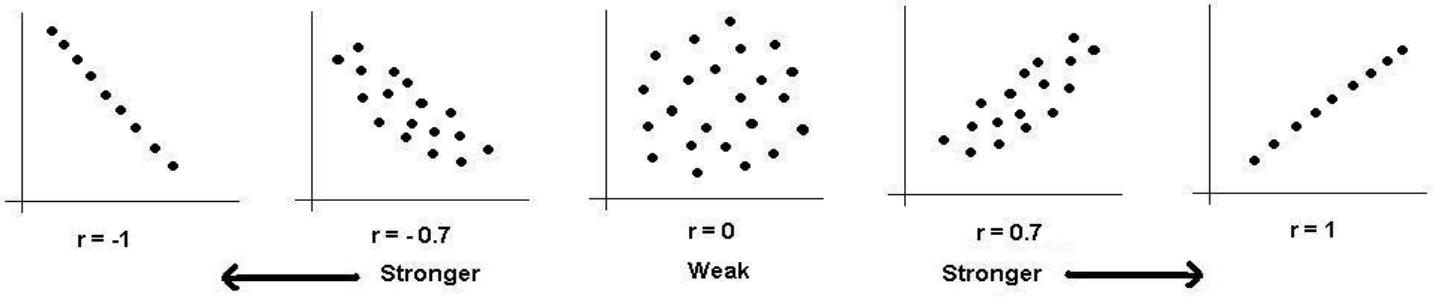
- For the students' heights and weights, the correlation is 0.644. What does this mean in terms of strength? We'll address this shortly.

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 and Outlier Condition

- 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.
- Straight Enough Condition:
  You can *calculate* a correlation coefficient for any pair of variables. But correlation measures the strength only of the *linear* association, and will be misleading if the relationship is not linear.
- Outlier Condition:
  Outliers can distort the correlation dramatically. 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.

Correlation Properties

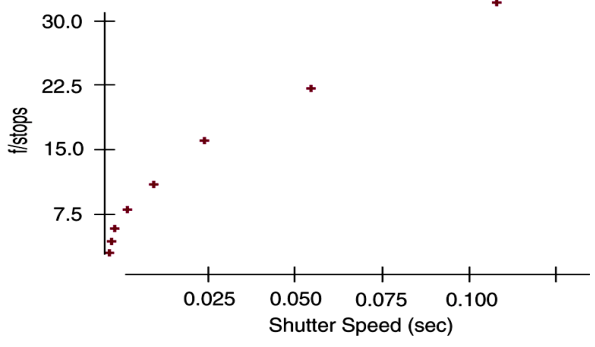
- The sign of a correlation coefficient gives the direction of the association.
- Correlation is always between -1 and +1. Correlation *can* be exactly equal to -1 or +1, but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line. A correlation near zero corresponds to a weak linear association.
- 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 changes in center or scale.
- Correlation measures the strength of the *linear* association between the two variables. Variables can have a strong association but still have a small correlation if the association isn't linear.
- Correlation is sensitive to outliers. A single outlying value can make a small correlation large or make a large one small.



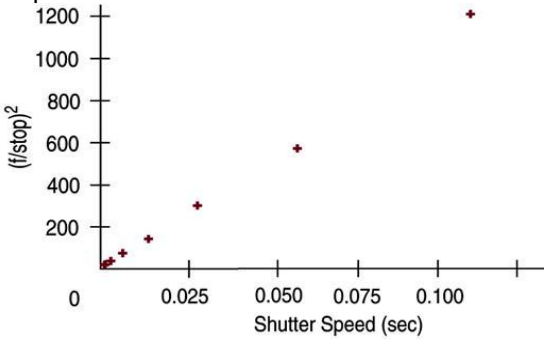
Straightening Scatterplots

- Straight line relationships are the ones that we can measure with correlation.
- When a scatterplot shows a bent form that consistently increases or decreases, we can often straighten the form of the plot by re-expressing one or both variables.

A scatterplot of f/stop vs. shutter speed shows a bent relationship:

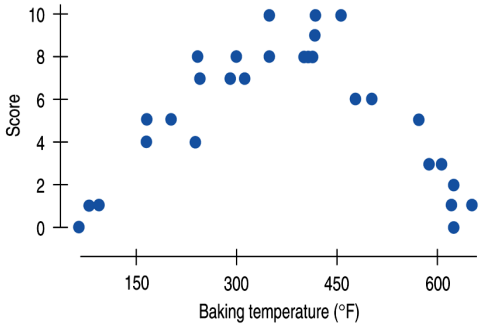


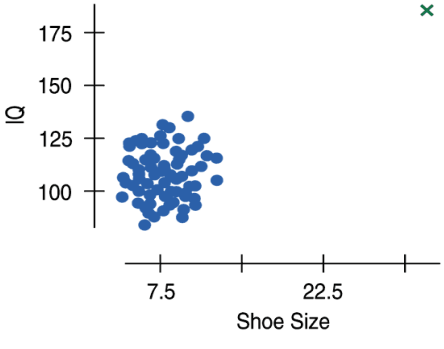
Re-expressing f/stop speed by squaring straightens the relationship:

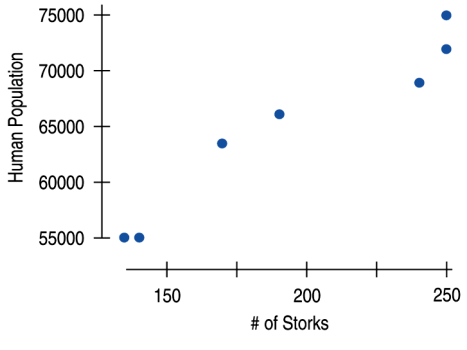


What Can Go Wrong?

- Don't say "correlation" when you mean "association." More often than not, people say correlation when they mean association. The word "correlation" should be reserved for measuring the strength and direction of the linear relationship between two quantitative variables.
- Don't correlate categorical variables. Be sure to check the Quantitative Variables Condition.
- Be sure the association is linear. There may be a strong association between two variables that have a nonlinear association. (Baking temp vs. Score scatterplot)
- Beware of outliers. Even a single outlier can dominate the correlation value. (Shoe size vs IQ scatterplot)
- Don't confuse correlation with causation. Not every relationship is one of cause and effect. (Storks vs Human Population plot)
- Watch out for lurking variables. A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a lurking variable.







1. Describe what you think a scatterplot might look like for the following variables:

a) Drug dosage and degree of pain relief

b) Calories consumed and weight loss

c) Hours of sleep and score on a test

d) Shoe size and grade point average

e) Time for a mile run and age

f) Age of car and cost of repairs

2. a) Below are heights & weights data for male & female AP Stat students. Create TI lists MHT, MWT, FHT, FWT, and enter the data.

Males

HT (in)	WT (lbs)	HT (in)	WT (lbs)
67	140	71	132
71	165	70	140
73	168	71	140
71	142	70	140
74	200	69	130
74	175	70	150
68	135	74	170
73	145	71	175
71	150	74	180
72	155	72	150
69	168	70	150
66	106	73	190
70	144		

Females

HT (in)	WT (lbs)	HT (in)	WT (lbs)
63	117	64	110
62	107	63	123
75	170	64	110
61	91	71	134
62	118	64	129
63	130	62	129
66	135	65	123
63	120	64.5	115
67	125	68.5	122
67	117	65.5	120
64	135	64	111
61	88	64	115

b) Is it reasonable to assume these data are drawn from populations that are normally distributed? Check summary statistics and histograms for each variable.

c) Make a scatterplot for each gender. Which is the explanatory variable? Describe the relationship (form, strength, direction, outliers, etc.) for each gender

d) Calculate and interpret *r* for each gender.

e) What would you predict about the weight of i) someone of average height?
ii) a male 2 standard deviations above average in height. iii) a female 1 standard deviation below average in height?

f) Write an equation of the least squares line of best fit for each gender.

g) Explain what the slope of each line means in the context of this relationship.

h) Predict weights for : a 60" male; a 60" female

i) Predict the weight of a 7'2" male; of a 20" newborn baby girl. Comment on these results.