Math 140
Introductory Statistics

Professor Bernardo Ábrego
Lecture 35
Sections 3.1, 3.2

Scatterplots: How to make them.

Here is the scatterplot with all the 50 employees.
(data on page 5)

In your calculator, you can put the x-coordinates on L1 and the y-coordinates on L2. Then you can see the scatterplot by selecting STAT PLOT -> ON and using L1 for XList and L2 for YList.

Comparison:
One variable vs. Two variables

<table>
<thead>
<tr>
<th>One Variable</th>
<th>Two Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape (of Distribution)</td>
<td>Shape (of Scatterplot)</td>
</tr>
<tr>
<td>Center</td>
<td>Trend</td>
</tr>
<tr>
<td>Spread</td>
<td>Strength (correlation)</td>
</tr>
</tbody>
</table>

Describing the Pattern of a Scatterplot

1. Identify the variables and cases.
   - What exactly does each point represent?
   - Describe the scale (units of measurement) and range of each variable.
2. Describe the overall shape of the relationship:
   - linearity: is the pattern linear (scattered about a line) or curved?
   - clusters: is there just one, or more than one?
   - outliers: are there any striking exceptions to the overall pattern?
3. Describe the trend.
   - If, as x gets larger, y tends to get larger, there is a positive trend. (The cloud of points tends to slope up as you go from left to right.)
   - If, as x gets larger, y tends to get smaller, there is a negative trend. (The cloud of points tends to slope down as you go from left to right.)
Describing the Pattern of a Scatterplot

4. Describe the strength of the relationship.
   - If the points cluster closely around an imaginary line, the association is strong. If the points are scattered farther from the line, the association is weak. The strength could vary on different places of the histogram or it could be constant.

5. Does the pattern generalize to other cases, or is the relationship an instance of “what you see is all there is”?

6. Are there plausible explanations for the pattern?
   - Is it reasonable to conclude that one variable causes the other?
   - Is there a third or lurking variable that might be causing both?

Example Dormitory Populations

The plot below, for the 50 United States, the number of people living in college dormitories versus the number of people living in cities, in thousands. Describe the pattern in the plot.

Example: Dormitory Populations

1. Variables and cases. The scatterplot shows dormitory population versus urban population, for the 50 U.S. states. Dormitory population ranges from near zero to a high of more than 165,000 in New York. The urban population ranges from near zero to about 15 million in New York and 28 million in California.

2. Shape. While most states follow a linear trend, the three states with the highest urban population suggest curvature in the plot because for those states, the number of people in dormitories is proportionately lower than in the smaller states. California can be considered an outlier with respect to its urban population, which is much higher than in other states. It is also an outlier with respect to the overall pattern because it lies far below the generally linear trend.
Example: Dormitory Populations

3. **Trend.** The trend is positive—states with higher urban populations tend to have higher dormitory populations, and vice versa.

4. **Strength.** The relationship varies in strength. As $x$ increases, $y$ becomes more variable—the cloud of points fans out.
   - A tight cluster in the lower left corresponds to a large number of states with small urban populations and small numbers of people living in dormitories.
   - Toward the upper right, the points are much more spread out. There are comparatively fewer states with large urban populations, and these states show much more variation in the number of people living in college dormitories.
   - The variation is roughly proportional to population size. For the states with smallest urban population, the points cluster rather closely to a line. For the states with the largest urban population, the states are more scattered from the line.
   - Overall, the strength of the relationship is moderate.

3.2 Getting a Line on the Pattern

- **Lines Review:**
  - Equation: $y = y$-intercept + slope $x$
  - (In statistics the independent term is written first.)
  - Recall that $\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{\text{change in } y}{\text{change in } x}$

Using Lines for Predictions

- We want to find a line that ‘fits best’ the points in a scatterplot.
- **Reasons to do this:**
  - to find a summary, or model, that describes the relationship between the two variables
  - to use the line to predict the value of $y$ when you know the value of $x$. In cases where it makes sense to do this, the variable on the $x$-axis is called the predictor or explanatory variable and the variable on the $y$-axis is called the response variable.
  - The line is called fitted line or regression line.