9.1 Well-Designed Statistical Pictures

Basic Characteristics:
1. Data should stand out clearly from background.
2. Clear labeling that indicates
   a. title or purpose of picture.
   b. what each axis, bar, pie segment, …, denotes.
   c. scale of each axis, including starting points.
3. Source for the data.
4. As little “chart junk” (extraneous material) as possible.
9.2 Pictures of Data

Pie Charts

Show what percentage of the whole fall into each category for a single variable.

Pie chart of hair colors of white American children.

Source: Krantz, 1992, p. 188.
Bar Graphs
Show what percentage or frequency of the whole fall into each category – can be used for two or three variables simultaneously.

Percentage of men and women 16 and over in the labor force

Pictograms
Bar graph that uses pictures related to topic.

Percentage of Ph.D.s earned by women.

Left pictogram: Misleading because eye focuses on area rather than just height.

Right pictogram: Visually more accurate, but less appealing.

9.3 Pictures of Measurement Variables

Single Variable Pictures:
- Stemplots
- Histograms

Displaying Relationships:
- Line Graphs: displays a variable over time
- Scatterplots: displays relationship between two measurement variables
Scatterplots

Displays relationship between two measurement variables.

Scatterplot of GPA and verbal SAT score.

Overall increasing trend but still variability in GPAs at each level of verbal SAT scores.

9.4 Difficulties and Disasters in Plots, Graphs, and Pictures

Most Common Problems:
1. No labeling on one or more axes
2. Not starting at zero as a way to exaggerate trends
3. Change(s) in labeling on one or more axes
4. Misleading units of measurement
5. Using poor information
9.5 A Checklist for Statistical Pictures

Ten questions to ask before interpreting:

1. Does the message of interest stand out clearly?
2. Is the purpose or title of the picture evident?
3. Is a source given for the data, either with the picture or in an accompanying article?
4. Did the information in the picture come from a reliable, believable source?
5. Is everything clearly labeled, leaving no ambiguity?
9.5 A Checklist for Statistical Pictures

Ten questions to ask before interpreting:

6. Do the axes start at zero or not?
7. Do the axes maintain a constant scale?
8. Are there any breaks in the numbers on the axes that may be easy to miss?
9. For financial data, have the numbers been adjusted for inflation?
10. Is there information cluttering the picture or misleading the eye?
Thought Question 1:

Do you think each of the following pairs of variables would have a **positive correlation**, a **negative correlation**, or **no correlation**?

a. Calories eaten per day and weight
b. Calories eaten per day and IQ
c. Amount of alcohol consumed and accuracy on a manual dexterity test
d. Number of ministers and number of liquor stores in cities in Pennsylvania
e. Height of husband and height of wife
10.1 Statistical Relationships

**Correlation:** measures the *strength* of a certain type of relationship between two measurement variables.

**Regression:** gives a numerical method for trying to *predict* one measurement variable from another.
Statistical Relationships versus Deterministic Relationships

**Deterministic:** if we know the value of one variable, we can determine the value of the other *exactly*. e.g. relationship between volume and weight of water.

**Statistical:** *natural variability* exists in both measurements. Useful for describing what happens to a population or aggregate.
10.2 Strength versus Statistical Significance

A relationship is \textit{statistically significant} if the chances of observing the relationship in the sample when actually nothing is going on in the population are less than 5%.

A relationship is statistically significant if that relationship is stronger than 95% of the relationships we would expect to see just by chance.
Two Warnings about Statistical Significance

• Even a minor relationship will achieve “statistical significance” if the sample is very large.

• A very strong relationship won’t necessarily achieve “statistical significance” if the sample is very small.
10.3 Measuring Strength Through Correlation

A Linear Relationship

Correlation (or the Pearson product-moment correlation or the correlation coefficient) represented by the letter $r$:

- Indicator of how closely the values fall to a straight line.
- Measures linear relationships only; that is, it measures how close the individual points in a scatterplot are to a straight line.
Other Features of Correlations

1. Correlation of +1 indicates a perfect linear relationship between the two variables; as one increases, so does the other. All individuals fall on the same straight line (a deterministic linear relationship).

2. Correlation of –1 also indicates a perfect linear relationship between the two variables; however, as one increases, the other decreases.

3. Correlation of zero could indicate no linear relationship between the two variables, or that the best straight line through the data on a scatterplot is exactly horizontal.
Other Features of Correlations

4. A **positive correlation** indicates that the variables increase together.

5. A **negative correlation** indicates that as one variable increases, the other decreases.

6. Correlations are unaffected if the units of measurement are changed. For example, the correlation between weight and height remains the same regardless of whether height is expressed in inches, feet or millimeters (as long as it isn’t rounded off).
Example 3: Verbal SAT and GPA

Scatterplot of GPA and verbal SAT score.

The correlation is 0.485, indicating a moderate positive relationship.

Higher verbal SAT scores tend to indicate higher GPAs as well, but the relationship is nowhere close to being exact.
Example 4: Husbands’ and Wifes’ Ages and Heights

Husbands’ and wives’ ages are likely to be closely related, whereas their heights are less likely to be so.

Source: Marsh (1988, p. 315) and Hand et al. (1994, pp. 179-183)
Example 6: Professional Golfers’ Putting Success

Scatterplot of distance of putt and putting success rates.

Correlation $r = -0.94$. Negative sign indicates that as distance goes up, success rate goes down.

Source: Iman (1994, p. 507)
10.4 Specifying Linear Relationships with Regression

Goal: Find a straight line that comes as close as possible to the points in a scatterplot.

- Procedure to find the line is called regression.
- Resulting line is called the regression line.
- Formula that describes the line is called the regression equation.
- Most common procedure used gives the least squares regression line.
The Equation of the Line

\[ y = a + bx \]

- \( a = \text{intercept} \) – where the line crosses the vertical axis when \( x = 0 \).
- \( b = \text{slope} \) – how much of an increase there is in \( y \) when \( x \) increases by one unit.

\[ y = \text{temperature in Fahrenheit} \]
\[ x = \text{temperature in Celsius} \]

\[ y = 32 + 1.8x \]

Intercept of 32 = temperature in F when C temperature is zero. Slope of 1.8 = amount by which F temperature increases when C temperature increases by one unit.
Example 7: Husbands’ and Wifes’ Ages, Revisited

Scatterplot of British husbands’ and wives’ ages with regression equation: \( y = 3.6 + 0.97x \)

**Intercept:** has no meaning.

**Slope:** for every year of difference in two wives ages, there is a difference of about 0.97 years in their husbands ages.

<table>
<thead>
<tr>
<th>Wife’s Age</th>
<th>Predicted Age of Husband</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 years</td>
<td>3.6 + (.97)(20) = 23.0 years</td>
</tr>
<tr>
<td>25 years</td>
<td>3.6 + (.97)(25) = 27.9 years</td>
</tr>
<tr>
<td>40 years</td>
<td>3.6 + (.97)(40) = 42.4 years</td>
</tr>
<tr>
<td>55 years</td>
<td>3.6 + (.97)(55) = 57.0 years</td>
</tr>
</tbody>
</table>

husband’s age = 3.6 + (.97)(wife’s age)
Extrapolation

Not a good idea to *use a regression equation to predict values far outside the range where the original data fell.*

No guarantee that the relationship will continue beyond the range for which we have data. Use the equation only for a *minor extrapolation* beyond the range of the original data.

**Final Cautionary Note:**
Easy to be misled by inappropriate interpretations and uses of correlation and regression. Chapter 11: how that can happen, and how you can avoid it.
Case Study 10.1: Are Attitudes about Love and Romance hereditary?

Study Details:

• 342 pairs of monozygotic (MZ) twins (share 100% of genes); 100 pairs of dizygotic (DZ) twins (share about 50% of genes); 172 spouse pairs (a twin and his or her spouse)

• Each filled out “Love Attitudes Scale” (LAS) questionnaire; 42 statements (7 questions on each of 6 love styles); Respondents ranked 1 (strongly agree) to 5 (strongly disagree).

• Six scores (1 for each love type) determined for each person. Correlations were computed for each of three types of pairs.

Case Study 10.1: Are Attitudes about Love and Romance hereditary?

**Key:** If love styles are genetic then matches between MZ twins should be much higher than those between DZ twins.

**Results:** Correlations are not higher for the MZ twins than they are for the DZ twins.

<table>
<thead>
<tr>
<th></th>
<th>Correlation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monozygotic Twins</td>
<td>Dizygotic Twins</td>
<td>Spouses</td>
</tr>
<tr>
<td><strong>Love Style</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eros</td>
<td>.16</td>
<td>.14</td>
<td>.36</td>
</tr>
<tr>
<td>Ludus</td>
<td>.18</td>
<td>.30</td>
<td>.08</td>
</tr>
<tr>
<td>Storge</td>
<td>.18</td>
<td>.12</td>
<td>.22</td>
</tr>
<tr>
<td>Pragma</td>
<td>.40</td>
<td>.32</td>
<td>.29</td>
</tr>
<tr>
<td>Mania</td>
<td>.35</td>
<td>.27</td>
<td>-.01</td>
</tr>
<tr>
<td>Agape</td>
<td>.30</td>
<td>.37</td>
<td>.28</td>
</tr>
<tr>
<td><strong>Personality Trait</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well-being</td>
<td>.38</td>
<td>.13</td>
<td>.04</td>
</tr>
<tr>
<td>Achievement</td>
<td>.43</td>
<td>.16</td>
<td>.08</td>
</tr>
<tr>
<td>Social closeness</td>
<td>.38</td>
<td>.01</td>
<td>-.04</td>
</tr>
</tbody>
</table>

*This surprising, and very unusual, finding suggests that genes are not important determinants of attitudes toward romantic love. Rather, the common environment appears to play the cardinal role in shaping familial resemblance on these dimensions.* (p. 271)
Case Study 10.2:  A Weighty Issue: Women Want Less, Men Want More

Ideal versus actual weight for females  Ideal versus actual weight for males

Equation: $\text{ideal} = 43.9 + 0.6(\text{actual})$

Equation: $\text{ideal} = 52.5 + 0.7(\text{actual})$

- If everyone at their ideal weight, all points fall on line $\text{Ideal} = \text{Actual}$.
- Most women fall below that line.
- Men under 175 pounds would prefer to weight same or more, while men over 175 pounds would prefer to weight same or less.
For Those Who Like Formulas

The Data

\( n \) pairs of observations, \((x_i, y_i), \ i = 1, 2, \ldots, n\), where \( x_i \) is plotted on the horizontal axis and \( y_i \) on the vertical axis.

Summaries of the Data, Useful for Correlation and Regression

\[
SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}
\]

\[
SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}
\]

\[
SXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}
\]

Correlation for a Sample of \( n \) Pairs

\[
r = \frac{SXY}{\sqrt{SSX \cdot SSY}}
\]

The Regression Slope and Intercept

\[
slope = b = \frac{SXY}{SSX}
\]

\[
intercept = a = \bar{y} - b \bar{x}
\]