Chapter 7

- Scatterplots, Association, and Correlation
Scatterplots & Correlation

- Here, we see a **positive** relationship between a bear’s age and its neck diameter.

As a bear gets older, it tends to have a larger neck.
Scatterplots & Correlation

- Statistics is about \textit{variation}.

- Recognize, quantify and try to explain variation.
  - Variation in neck measurements can be explained, at least in part, by the age of the bear.
  - Older bear \rightarrow Larger neck
Positive Association

Cell phone usage per 100 people vs. Percent of individuals between 15 & 64

- Data from 2008.
- These variables have a positive correlation…
  - A country with a larger percentage of people between 15-64 tends to have more cell phone users.
Negative Association

- Outside temperature and amount of natural gas used.
- These variables have a **negative** correlation...
  - Days with higher temperature tend to use less natural gas.
  - Higher temperature $\rightarrow$ Less gas used
Scatterplots & Correlation

- When the two variables of interest are **continuous variables**, we can plot their relationship with a **scatterplot** (or scatter diagram).

  - A scatterplot gives you a quick look at the general relationship between the variables.
  - Each observation provides one point on the plot.
- **Response variable** – plotted on the vertical axis.
  - Also called the dependent variable.

- **Explanatory variable** – plotted on the horizontal axis.
  - Used to try to explain variation in the response variable.
  - Also called the independent variable.

Here, we use Engine HPW to explain the variability in HWY-mpg.
Correlation and Association

Definition

A correlation exists between two variables when higher values of one variable consistently go with higher values of another variable or when higher values of one variable consistently go with lower values of another variable.

When describing relationships, we use the terms correlation and association interchangeably. If variables are correlated, we say they are associated.
Positive Association (correlation)

- Above average values of Age are associated with above average values of Neck Measure (age-high goes with neck-high)

- Below average values of Age are associated with below average values of Neck Measure (age-low goes with neck-low)
Negative Association (correlation)

- Below average values of Engine HPW are associated with above average values of HWY-mpg (HPW-low goes with MPG-high).

- Above average values of Engine HPW are associated with below average values of HWY-mpg (HPW-high goes with MPG-low).
Strength of Association

- Correlation applies only to **quantitative (continuous) variables**.
- Correlation measures the strength of **linear association**.
- The **correlation coefficient** \((r)\) gives the direction of the linear association and quantifies the strength of the linear association between two quantitative variables.
- Correlation is a `unitless` quantity (not in `feet` or `inches`… no units)
**Strength of Association**

Correlation Coefficient \((r)\) will be between -1 and 1.

- **Strong Negative Linear Relationship**
  - Correlation Coefficient: -1.0
- **Very Weak or No Linear Relationship**
  - Correlation Coefficient: 0.0
- **Strong Positive Linear Relationship**
  - Correlation Coefficient: 1.0
$r = 0.3$

weak (fuzzy)

(a) Weak positive correlation between $x$ and $y$

$r = 0.7$

stronger (more clear)

(b) Strong positive correlation between $x$ and $y$

$r = 1$

super strong

(c) Perfect positive correlation between $x$ and $y$

$r = 0.0$

none

(g) No correlation between $x$ and $y$

$r = ?$

$r$ not meaningful, this is non-linear

(h) Nonlinear correlation between $x$ and $y$

$r = -0.3$

weak (fuzzy)

(d) Weak negative correlation between $x$ and $y$

$r = -0.7$

stronger (more clear)

(e) Strong negative correlation between $x$ and $y$

$r = -1$

super strong

(f) Perfect negative correlation between $x$ and $y$
Things to look for in a scatterplot

1. Direction of association
   - Positive or negative.

2. Form of association
   - Linear, curved, clustered, scattered (no relationship).

3. Strength of association
   - How closely the points follow a clear form.

4. Outliers
   - A point that lies outside of the general pattern.
Example

Direction _______________
Form _________________
Strength _______________
Outliers? _______________
Association vs. Causation

- The existence of an association does not equate to causation.
- To imply that a change in one variable causes a change in another is a very strong statement – use ‘association’ for our relationships in this class.
Correlation Cautions

- Don’t confuse correlation with causation.
  - There is a strong positive correlation between shoe size and intelligence.

- Beware of lurking variables.
Beware of lurking variables

- **Lurking variable** – a hidden variable that stands behind a relationship and affects the other two variables.

![Graph showing the relationship between fire damage (dollars) and number of firefighters at scene.](image)
Association vs. Causation

- Increasing the size of the fire will cause greater damage.

- Increasing the number of firefighters at the fire will not cause greater damage, but we do tend to see more firefighters at larger fires.

- Correlation does NOT imply causality.