Bivariate Data
Univariate (one variable) data

- Involves only a single variable
  - So cannot describe associations or relationships
- Descriptive Statistics
  - Central tendencies: mean, median, mode
  - Dispersion: range, max, min, quartiles, variance, standard deviation
- Visualizations
  - Pie charts, Bar charts, Line charts, Histograms, Box plots
Bivariate (two variables) data

● Involves two variables
  ○ So *can* describe associations or relationships

● Descriptive Statistics
  ○ Central tendencies: mean, median, mode
  ○ Dispersion: variance, standard deviation, covariance, correlation

● Visualizations
  ○ Scatter plots
Bivariate data and scatter plots
Bivariate data and scatter plots

- A scatter plot of bivariate data shows one variable vs. the other on a 2-dimensional graph.
- If there is an explanatory variable, it is plotted it on the horizontal ($x$) axis, and the response variable is plotted on the vertical ($y$) axis.
  - If there is no explanatory-response distinction either variable can be plotted on either axis.
- $x$ is also known as the independent variable, and $y$ the dependent.
Covariance & Correlation
Hybrid cars sold in the U.S. from 1997 to 2013

The variables:

- Model of the car
- Year of manufacture
- MSRP (manufacturer's suggested retail price) in 2013 dollars
- Acceleration rate in km per second
- Fuel economy in miles per gallon
- Model’s class
Positive association

- The points are scattered in an upward direction, indicating that cars with greater acceleration tend to cost more.
- Conversely, cars that cost more tend to have greater acceleration.
- This is an example of **positive association**: above average values of one variable tend to be associated with above average values of the other.
Negative association

- There is a clear downward trend: i.e., a negative association
- Hybrid cars with higher mpg tend to cost less; conversely, cars that cost more tend to have lower mpg
- This might seem confusing at first, until we consider that cars that accelerate faster tend to be less fuel efficient and have lower mpg
Sample Covariance

c is the average product, across all observations, of deviations (i.e., the differences between the measurements and their sample means)

\[ \gamma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \]

\[ c_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \]
Standardized Units

- Cov(GNI per Capita, Expected Years of Schooling) $\sim 32898.63$
- Cov(Mean Years of Schooling, Expected Years of Schooling) $\sim 7.232066$
- This is not intuitive: Covariance with Mean Years of Schooling should be higher.

Data Source: 2014 HDI Index
Standard Units

**Problem:** scales differ (e.g., Fahrenheit degrees $x$ dollars vs. Fahrenheit degrees $x$ euros vs. Celsius degrees $x$ euros)

**Solution:** normalize the deviations
I.e., divide by a measure of spread

**Interpretation:**
The measurement is \( \frac{x_i - \mu_x}{\sigma_x} \) many standard deviations from the mean

Correlation = normalized covariance
Sample Correlation

\( r \) is the average product, across all observations, of deviations, measured in standard units

\[
\rho_{XY} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i - \mu_x}{\sigma_x} \right) \left( \frac{x_i - \mu_y}{\sigma_y} \right)
\]

\[
r_{XY} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i - \bar{x}}{s_{xx}} \right) \left( \frac{x_i - \bar{x}}{s_{yy}} \right)
\]
Standardized Units

- Cor(GNI per Capita, Expected Years of Schooling) \( \sim 0.610305 \)
- Cor(Mean Years of Schooling, Expected Years of Schooling) \( \sim 0.8152542 \)
- This is intuitive: Covariance with Mean Years of Schooling is now higher!

Data Source: 2014 HDI Index
Properties of Covariance

- Symmetric measure: does not distinguish between the explanatory and response variables
- Both variables must be quantitative
- If two variables are independent, then their covariance is 0.
- But if the covariance of two variables is zero, they are not necessarily independent, because covariance captures only linear associations.
  - The following data sets exhibit zero or near-zero covariance:
Properties of Correlation

- Symmetric measure: does not distinguish between the explanatory and response variables
- Both variables must be quantitative
- If two variables are independent, then their correlation is 0.
- But if the correlation of two variables is zero, they are not necessarily independent, because correlation captures only linear associations.
- Is standardized, so is invariant to change of units
- Is a number between -1 and 1
A toy example: Ice cream sales

Based on the scatter plot of temperature vs. ice cream sales, we expect $r$ to be positive and close to 1, as there seems to be a strong linear relationship.
Step 1: Convert values to standard units

- Let’s calculate temperature in standard units, starting with 16.4
- Mean of temperature: 19.08182
- Standard deviation of temperature: 3.755437

Value in standard units = \( \frac{\text{value} - \text{average}}{\text{SD}} \)

Temperature 16.4 in standard units = \( \frac{16.4 - 19.08182}{3.755437} = -0.714116 \)

All the values in the columns of temperatures and sales in standard units were calculated this way.
Step 2: Multiply corresponding pairs of values in standard units

<table>
<thead>
<tr>
<th>temperature</th>
<th>sales</th>
<th>temperature_standard_units</th>
<th>sales_standard_units</th>
<th>product_of_standard_units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.4</td>
<td>-0.71411617</td>
<td>-0.8481268</td>
<td>0.605650985</td>
</tr>
<tr>
<td>2</td>
<td>11.9</td>
<td>-1.91237891</td>
<td>-2.10518057</td>
<td>4.025902917</td>
</tr>
<tr>
<td>3</td>
<td>15.2</td>
<td>-1.03365290</td>
<td>-0.78525929</td>
<td>0.811685544</td>
</tr>
<tr>
<td>4</td>
<td>18.5</td>
<td>-0.15492690</td>
<td>-0.12080912</td>
<td>0.018716583</td>
</tr>
<tr>
<td>5</td>
<td>22.1</td>
<td>0.80368329</td>
<td>0.92076141</td>
<td>0.740000559</td>
</tr>
<tr>
<td>6</td>
<td>19.4</td>
<td>0.08472565</td>
<td>-0.06693478</td>
<td>-0.005671093</td>
</tr>
<tr>
<td>7</td>
<td>25.1</td>
<td>1.60252511</td>
<td>1.74683459</td>
<td>2.799346300</td>
</tr>
<tr>
<td>8</td>
<td>23.4</td>
<td>1.14984808</td>
<td>1.11830065</td>
<td>1.285875854</td>
</tr>
<tr>
<td>9</td>
<td>18.1</td>
<td>-0.26143914</td>
<td>0.01387672</td>
<td>-0.003627919</td>
</tr>
<tr>
<td>10</td>
<td>22.6</td>
<td>0.93682359</td>
<td>0.22937408</td>
<td>0.214883045</td>
</tr>
<tr>
<td>11</td>
<td>17.2</td>
<td>-0.50109169</td>
<td>-0.10285101</td>
<td>0.051537786</td>
</tr>
</tbody>
</table>
Step 3: \( r \) is the average of the products of the values in standard units

- We compute the mean of the products in standard units
- In this case \( r \) is 0.9585728
- This confirms our conjecture that \( r \) is positive and close to 1
Computing $r$ in R

> hybrid <- read.csv("hybrid.csv")
> cor(hybrid$msrp, hybrid$accelerate)
[1] 0.6955779
Computing $r$ in R

```r
> hybrid <- read.csv("hybrid.csv")
> cor(hybrid$msrp, hybrid$mpg)
[1] -0.5318264
```
Correlation Matrix

```r
> hybrid <- read.csv("hybrid.csv")
> x <- select(hybrid, msrp : mpgmpge)
> corr(x, x)
```

<table>
<thead>
<tr>
<th></th>
<th>msrp</th>
<th>accelerate</th>
<th>mpg</th>
<th>mpgmpge</th>
</tr>
</thead>
<tbody>
<tr>
<td>msrp</td>
<td>1.00000000</td>
<td>0.6955779</td>
<td>-0.5318264</td>
<td>-0.3722185</td>
</tr>
<tr>
<td>accelerate</td>
<td>0.6955779</td>
<td>1.00000000</td>
<td>-0.5060704</td>
<td>-0.3988673</td>
</tr>
<tr>
<td>mpg</td>
<td>-0.5318264</td>
<td>-0.5060704</td>
<td>1.00000000</td>
<td>0.6677531</td>
</tr>
<tr>
<td>mpgmpge</td>
<td>-0.3722185</td>
<td>-0.3988673</td>
<td>0.6677531</td>
<td>1.00000000</td>
</tr>
</tbody>
</table>
Pairs

> pairs(x)
Summary

Correlation measures the **direction** and **strength** of a **linear** relationship between two quantitative variables.
Correlation is powerful and simple but easy to misinterpret:

- *Correlation does not imply causation!*
  - Correlation only measures association.
- Correlation only measures **linear association**.
- **Outliers** can have a significant effect on correlation.
- Correlation can be misleading when data are **aggregated**.
Correlation does not imply causation

Intuitive Example:

- Imagine a positive correlation between math abilities and kids’ weight.
- Does this imply that students who are better at math gain weight easily?
- Or that gaining weight can improve a student’s math abilities?
- No! Of course not!
- Age is a confounding variable, which explains the correlation.

Older children both weigh more and are better at math than younger children.
Correlation measures linear associations

```r
> x <- seq(-3, 3, by = 0.5)
> y <- x ** 2
> cor(x, y)
[1] 0
> plot(x, y, xlab = "x", ylab = "y")
```
Outliers can gravely impact correlation

\begin{verbatim}
> cor(line$x, line$y)
[1] 1

> cor(outlier$x, outlier$y)
[1] 0
\end{verbatim}
Correlations across aggregated data

- When the data for each country are collapsed to a single point, the variables seem strongly correlated.
- But individuals vary, so there is really a cloud of points per country.
- Thus, the true correlation is lower than the value calculated using aggregate data.
- Aggregate correlations like these are known as ecological correlations.
- When the correlations are spurious, it is called an ecological fallacy.
A 2012 paper in the New England Journal of Medicine

Some responded harshly:
http://blogs.scientificamerican.com/the-curious-wavefunction/chocolate-consumption-and-nobel-prizes-a-bizarre-juxtaposition-if-there-ever-was-one/

Other responses were more nuanced:
http://www.reuters.com/article/us-eat-chocolate-win-the-nobel-prize-idUSBRE8991MS20121010#vFdfFkbPVilSjsB.97
Simpson’s Paradox

A Related Phenomenon
Simpson’s Paradox

- A phenomenon in which a trend that appears in different groups of data disappears or reverses when the groups are aggregated
Example: Hospital Surgeons

- In a certain hospital, there are two surgeons.
  - Surgeon A operates on 100 patients, and 95 survive.
  - Surgeon B operates on 80 patients and 72 survive.

- We want to choose the better of the two surgeons.
  - We calculate the percentage of surgeon A's patients who survived: $\frac{95}{100} = 95\%$
  - And we do the same for surgeon B: $\frac{72}{80} = 90\%$

- Is it decided? Should we choose surgeon A.

This example was borrowed from Overview of Simpson’s Paradox in Statistics.
Example: Hospital Surgeons (cont’d)

- Not so fast: there is a latent variable. Some surgeries are high-risk.
- Of the 100 patients that surgeon A treated:
  - 50 were high risk, three of whom died: \( \frac{47}{50} = 94\% \) survival rate for high-risk surgeries.
  - The other 50 were routine, and two died: \( \frac{48}{50} = 96\% \) survival rate for routine surgeries.
- Of the 80 patients that surgeon B treated:
  - 40 were high risk, seven of whom died: \( \frac{7}{40} = 82.5\% \) survival rate for high-risk surgeries.
  - The other 40 were routine, and one died: \( \frac{1}{40} = 97/5\% \) survival rate for routine surgeries.

This example was borrowed from *Overview of Simpson’s Paradox in Statistics*. 
Explaining the paradox

- Incorrectly assuming the relationships among the variables in a sample mirror those of the population from which the sample was drawn, due to an incorrect characterization of the population.
- Failure to identify uncontrolled or unobserved (i.e., latent) variables.
Extras
Correlations across aggregated data

First 15 rows of average SAT scores by state in 2015

<table>
<thead>
<tr>
<th>State</th>
<th>Critical.Reading</th>
<th>Mathematics</th>
<th>Writing</th>
<th>Total.SAT.score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Alabama</td>
<td>545</td>
<td>538</td>
<td>533</td>
<td>1616</td>
</tr>
<tr>
<td>2 Alaska</td>
<td>509</td>
<td>503</td>
<td>482</td>
<td>1494</td>
</tr>
<tr>
<td>3 Arizona</td>
<td>523</td>
<td>527</td>
<td>502</td>
<td>1552</td>
</tr>
<tr>
<td>4 Arkansas</td>
<td>568</td>
<td>569</td>
<td>551</td>
<td>1688</td>
</tr>
<tr>
<td>5 California</td>
<td>495</td>
<td>506</td>
<td>491</td>
<td>1492</td>
</tr>
<tr>
<td>6 Colorado</td>
<td>582</td>
<td>587</td>
<td>567</td>
<td>1736</td>
</tr>
<tr>
<td>7 Connecticut</td>
<td>504</td>
<td>506</td>
<td>504</td>
<td>1514</td>
</tr>
<tr>
<td>8 Delaware</td>
<td>462</td>
<td>461</td>
<td>445</td>
<td>1368</td>
</tr>
<tr>
<td>9 District.of.Columbia</td>
<td>441</td>
<td>440</td>
<td>432</td>
<td>1313</td>
</tr>
<tr>
<td>10 Florida</td>
<td>486</td>
<td>480</td>
<td>468</td>
<td>1434</td>
</tr>
<tr>
<td>11 Georgia</td>
<td>490</td>
<td>485</td>
<td>475</td>
<td>1450</td>
</tr>
<tr>
<td>12 Hawaii</td>
<td>487</td>
<td>508</td>
<td>477</td>
<td>1472</td>
</tr>
<tr>
<td>13 Idaho</td>
<td>467</td>
<td>463</td>
<td>442</td>
<td>1372</td>
</tr>
<tr>
<td>14 Illinois</td>
<td>599</td>
<td>616</td>
<td>587</td>
<td>1802</td>
</tr>
<tr>
<td>15 Indiana</td>
<td>496</td>
<td>499</td>
<td>478</td>
<td>1473</td>
</tr>
</tbody>
</table>
Correlations across aggregated data

> cor(sat$Critical.Reading, sat$Mathematics)
[1] 0.9843772

- This correlation does not reflect the true relationship between students’ Math and Critical Reading scores
- States do not take tests; students do