Introduction to Statistics
Xiamen Academic Program

Dr. Jason A. Osborne

Monday, Tuesday, July 16-17, 2012
What is Statistics?
Statistics is the name for the discipline that considers the analysis (and collection) of data. Often the purpose of this analysis is to make inferences about a population based on a sample.

- In agriculture, which treatments help yields?
- How much resemblance between parents and children?

Galton’s jittered heights data

midparent ht(inches)
son height(inches)
What is Statistics (continued)?

- Is a new medical treatments safe? Effective?

- How can a patient’s molecular profile be used to prescribe medication?
What is Statistics (continued)?

- Why does the casino make a profit on roulette?

(www.propunttingsystems.com) - a website with articles on “winning strategies” for roulette.
What is Statistics (continued)?

▶ Who will win the next election? By how much?

(Picture of President Harry S. Truman published in 1948.)

▶ How many people are employed? Unemployed?

▶ What would the general population think of this movie?
Designed experiments

and **observational studies**

- Designed experiments
  - To infer *causation*
  - Randomization
  - Inclusion of *controls* (e.g. *placebo*)
  - (Double)-blindness

- **Observational studies**
  - To infer *association*
  - *Confounding*, lurking variables
  - Multicollinearity
  - Simultaneity

- Some examples
  - Polio and the Salk vaccine trials
  - Berkeley admissions data
  - The effects of ultrasound exams on pregnant women
1916 - First polio epidemic hits the U.S.
- Claims hundreds of thousands of victims, mostly children
1950s - several vaccines discovered
- one by Jonas Salk
  - proved safe in the lab
  - caused the production of
    antibodies against polio.
Large-scale field trial needed to establish effectiveness outside the lab.

1954 Public Health Service organizes an experiment in which the subjects are children in most vulnerable age groups, grades 1-3.

Q: How to assess effectiveness?
Q: Give vaccine to a group of kids and compare to 1953?
A: No. Incidence varies from year to year (confounded w/ time).
Polio - Salk vaccine trials

A method of comparison is needed:

▶ control group - receives placebo
▶ treatment group - receives vaccine

(Placebos have been shown to have effects in some cases.)

To eliminate other unforeseeable differences between groups which may affect response, (confounding factors), subjects are randomized to the two groups.
For the same reason, it is best if the experiment is double-blind; neither subjects nor evaluators know who is in the treatment group.
Polio (continued)

What happened: children could only be vaccinated with parental permission. Among the 400,000 children whose parents gave permission, half were randomized to the control group, half to the treatment group with the following results:

<table>
<thead>
<tr>
<th>Group</th>
<th>Size</th>
<th>(cases/100,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>200,000</td>
<td>28</td>
</tr>
<tr>
<td>Control</td>
<td>200,000</td>
<td>71</td>
</tr>
<tr>
<td>No Consent</td>
<td>350,000</td>
<td>46</td>
</tr>
<tr>
<td>Grade 2 (vaccine)</td>
<td>225,000</td>
<td>25</td>
</tr>
<tr>
<td>Grades 1,3</td>
<td>725,000</td>
<td>54</td>
</tr>
<tr>
<td>Grade 2 - no consent</td>
<td>125,000</td>
<td>44</td>
</tr>
</tbody>
</table>

Conclusion: Estimated effect of polio vaccine among children in grades 1-3 with parents granting permission is

\[ \Delta \text{rate} = 43 \text{ cases/100,000}^* \]

* It can be shown that this figure is **statistically significant.**
Polio (continued)

Another design that was used:

- **NFIP**
  - Treatment Group: Grade 2 with consent
  - Control Group: Grades 1 and 3

Conclusion: estimated effect is

\[
\Delta \text{ rate } = \frac{29}{100,000}
\]

but this estimate is biased by parental consent.

Q: Why is consent a **confounding factor**?

Q: Children with parental consent more susceptible to polio?

Would you believe . . .

that children from households with less income are less susceptible?

that children from less hygienic surroundings are less susceptible?

these children more likely to contract polio very early, while still protected by antibodies from their mothers.

- The NFIP study is called an **observational study**.
Berkeley Admissions Data (grad school)

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Deny</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>3738</td>
<td>4704</td>
<td>8442</td>
</tr>
<tr>
<td>Women</td>
<td>1494</td>
<td>2827</td>
<td>4321</td>
</tr>
</tbody>
</table>

▶ Fact: More men are being admitted than women.
▶ Question: Is there a gender bias?
▶ An observational study (randomization not possible)
<table>
<thead>
<tr>
<th>Major</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>% admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28</td>
</tr>
<tr>
<td>F</td>
<td>393</td>
<td>6</td>
</tr>
</tbody>
</table>

- Gender is not the only way in which the grps differ.
- Majors A & B have much higher rates of admission than C,D,E, or F and more than half of the men considered here applied to major A or B.
- Majors C,D,E and F have lower rates of admission. 90% of the women applied to one of these majors.
Berkeley admissions data

A simplification:

- 1200 applicants (550 Men, 600 women)
- Two Majors:
  - Machismatics (admission rate is 50%)
  - Social Welfare (admission rate is 33%)
Simplified admissions data, continued

<table>
<thead>
<tr>
<th>Machismatics</th>
<th>Admit</th>
<th>Deny</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>200</td>
<td>200</td>
<td>50%</td>
</tr>
<tr>
<td>Women</td>
<td>100</td>
<td>100</td>
<td>50%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social Welfare</th>
<th>Admit</th>
<th>Deny</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>50</td>
<td>100</td>
<td>33%</td>
</tr>
<tr>
<td>Women</td>
<td>150</td>
<td>300</td>
<td>33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>Admit</th>
<th>Deny</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>250</td>
<td>300</td>
<td>45%</td>
</tr>
<tr>
<td>Women</td>
<td>250</td>
<td>400</td>
<td>38%</td>
</tr>
</tbody>
</table>

This is known as *Simpson’s Paradox*. 
Ultrasound

- Experiments with lab animals → ultrasound exams cause low birthweight
- Investigators from JHU in Baltimore conducted an observational study, making many msmts on many mothers:

\[ Y_i = \beta_0 + \beta_1 M_i + \beta_2 F_i + \beta_3 DOB_i + \beta_U U_i + E_i \]

concluded that mothers with exams had lighter babies \((\hat{\beta}_U < 0)\)

- A randomized controlled experiment run subsequently and found the opposite \((\hat{\beta}_U >> 0)\)
Descriptive statistics

▶ A statistic is a function of data that we use for inference about a population.
▶ “Descriptive statistics” is often used to mean quantitative summaries of a sample of data.
▶ These summaries are often used to convey
  ▶ Location (e.g. “average” or “typical” or “middle” values)
  ▶ Spread or variability or dispersion
  ▶ Shape (bell-curved?, flat?, unimodal? bimodal?, symmetric?)
Types of variables

- **Qualitative**: classified without meaningful numerical scale.
  - Nominal: Cannot be meaningfully ordered
    - Sex of an organism
    - Color of a flower
  - Ordinal: Can be meaningfully ordered (but no scale)
    - Color of a flower
    - Likert responses (Disagree strongly, . . . , Agree Strongly)
    - Grades - appearance of produce, 5 point scale

- **Quantitative**: measured on a meaningful numerical scale (sometimes classified as ratio or not)
  - Continuous
    - Age, Weight, or Length
  - Discrete
    - Counting variables such as # of offspring, # of offspring, progeny, attendance at a game, # of security faults in code.
See lib.stat.cmu.edu/DASL/

<table>
<thead>
<tr>
<th>OBS</th>
<th>NAME</th>
<th>SHELF</th>
<th>SUGARS</th>
<th>CARBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Honey-comb</td>
<td>1</td>
<td>11</td>
<td>14.0</td>
</tr>
<tr>
<td>2</td>
<td>Apple_Cinnamon_Cheerios</td>
<td>1</td>
<td>10</td>
<td>10.5</td>
</tr>
<tr>
<td>3</td>
<td>Honey_Nut_Cheerios</td>
<td>1</td>
<td>10</td>
<td>11.5</td>
</tr>
<tr>
<td>4</td>
<td>Frosted_Flakes</td>
<td>1</td>
<td>11</td>
<td>14.0</td>
</tr>
<tr>
<td>5</td>
<td>Golden_Crisp</td>
<td>1</td>
<td>15</td>
<td>11.0</td>
</tr>
<tr>
<td>6</td>
<td>Wheaties_Honey_Gold</td>
<td>1</td>
<td>8</td>
<td>16.0</td>
</tr>
<tr>
<td>7</td>
<td>Multi-Grain_Cheerios</td>
<td>1</td>
<td>6</td>
<td>15.0</td>
</tr>
<tr>
<td>8</td>
<td>Rice_Krispies</td>
<td>1</td>
<td>3</td>
<td>22.0</td>
</tr>
<tr>
<td>9</td>
<td>Corn_Cheex</td>
<td>1</td>
<td>3</td>
<td>22.0</td>
</tr>
<tr>
<td>10</td>
<td>Rice_Cheex</td>
<td>1</td>
<td>2</td>
<td>23.0</td>
</tr>
<tr>
<td>11</td>
<td>Corn_Flakes</td>
<td>1</td>
<td>2</td>
<td>21.0</td>
</tr>
<tr>
<td>12</td>
<td>Bran_Cheex</td>
<td>1</td>
<td>6</td>
<td>15.0</td>
</tr>
<tr>
<td>13</td>
<td>Wheat_Cheex</td>
<td>1</td>
<td>3</td>
<td>17.0</td>
</tr>
<tr>
<td>14</td>
<td>Cheerios</td>
<td>1</td>
<td>1</td>
<td>17.0</td>
</tr>
<tr>
<td>15</td>
<td>Wheaties</td>
<td>1</td>
<td>3</td>
<td>17.0</td>
</tr>
<tr>
<td>16</td>
<td>Special_K</td>
<td>1</td>
<td>3</td>
<td>16.0</td>
</tr>
</tbody>
</table>
2\textsuperscript{nd} (middle) shelf

<table>
<thead>
<tr>
<th>OBS</th>
<th>NAME</th>
<th>SHELF</th>
<th>SUGARS</th>
<th>CARBO</th>
<th>RATING</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>Cap’n’Crunch</td>
<td>2</td>
<td>12</td>
<td>12.0</td>
<td>18.0429</td>
</tr>
<tr>
<td>18</td>
<td>Cinnamon Toast Crunch</td>
<td>2</td>
<td>9</td>
<td>13.0</td>
<td>19.8236</td>
</tr>
<tr>
<td>19</td>
<td>Honey Graham Ohs</td>
<td>2</td>
<td>11</td>
<td>12.0</td>
<td>21.8713</td>
</tr>
<tr>
<td>20</td>
<td>Count Chocula</td>
<td>2</td>
<td>13</td>
<td>12.0</td>
<td>22.3965</td>
</tr>
<tr>
<td>21</td>
<td>Cocoa Puffs</td>
<td>2</td>
<td>13</td>
<td>12.0</td>
<td>22.7364</td>
</tr>
<tr>
<td>22</td>
<td>Golden Grahams</td>
<td>2</td>
<td>9</td>
<td>15.0</td>
<td>23.8040</td>
</tr>
<tr>
<td>23</td>
<td>Lucky Charms</td>
<td>2</td>
<td>12</td>
<td>12.0</td>
<td>26.7345</td>
</tr>
<tr>
<td>24</td>
<td>Trix</td>
<td>2</td>
<td>12</td>
<td>13.0</td>
<td>27.7533</td>
</tr>
<tr>
<td>25</td>
<td>Fruity Pebbles</td>
<td>2</td>
<td>12</td>
<td>13.0</td>
<td>28.0258</td>
</tr>
<tr>
<td>26</td>
<td>Nut &amp; Honey Crunch</td>
<td>2</td>
<td>9</td>
<td>15.0</td>
<td>29.9243</td>
</tr>
<tr>
<td>27</td>
<td>Smacks</td>
<td>2</td>
<td>15</td>
<td>9.0</td>
<td>31.2301</td>
</tr>
<tr>
<td>28</td>
<td>Froot Loops</td>
<td>2</td>
<td>13</td>
<td>11.0</td>
<td>32.2076</td>
</tr>
<tr>
<td>29</td>
<td>Apple Jacks</td>
<td>2</td>
<td>14</td>
<td>11.0</td>
<td>33.1741</td>
</tr>
<tr>
<td>30</td>
<td>Corn Pops</td>
<td>2</td>
<td>12</td>
<td>13.0</td>
<td>35.7828</td>
</tr>
<tr>
<td>31</td>
<td>Kix</td>
<td>2</td>
<td>3</td>
<td>21.0</td>
<td>39.2411</td>
</tr>
<tr>
<td>32</td>
<td>Raisin Bran</td>
<td>2</td>
<td>12</td>
<td>14.0</td>
<td>39.2592</td>
</tr>
<tr>
<td>33</td>
<td>Life</td>
<td>2</td>
<td>6</td>
<td>12.0</td>
<td>45.3281</td>
</tr>
<tr>
<td>34</td>
<td>Maypo</td>
<td>2</td>
<td>3</td>
<td>16.0</td>
<td>54.8509</td>
</tr>
<tr>
<td>35</td>
<td>Frosted Mini Wheats</td>
<td>2</td>
<td>7</td>
<td>14.0</td>
<td>58.3451</td>
</tr>
<tr>
<td>36</td>
<td>Strawberry Fruit Wheats</td>
<td>2</td>
<td>5</td>
<td>15.0</td>
<td>59.3640</td>
</tr>
</tbody>
</table>
Histories

To get a histogram for a sample, \( y_1, y_2, \ldots, y_n \), first order it so that

\[
y(1) \leq y(2) \leq y(3) \cdots \leq y(n)
\]

(Parentheses indicate order statistics.)

Then, put into “measurement classes” or bins to indicate frequency. For the bottom shelf cereal sugars, we have

\[
y(1) = 1, y(2) = y(3) = 2, \ldots, y(16) = 15.
\]

<table>
<thead>
<tr>
<th>Frequency</th>
<th>( \leq 2 )</th>
<th>3-4</th>
<th>5-6</th>
<th>7-8</th>
<th>9-10</th>
<th>11-12</th>
<th>( \geq 13 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bottom shelf

Middle shelf
Another histogram:

Q: What is a “typical” baseball salary?
Numerical descriptive statistics:
For qualitative data - report class frequencies
For quantitative data $y_1, y_2, \ldots, y_n$, we’re interested in
▶ Measures of **central** tendency
  ▶ mean, $\bar{y}$
  ▶ median $\tilde{y}$ or 50$^{th}$ percentile
  ▶ mode
▶ measures of **spread**
  ▶ sample variance $s^2$
  ▶ sample standard deviation $s$
  ▶ range $r$
  ▶ interquartile range, $IQR = 75^{th}$ percentile $- 25^{th}$ percentile.
▶ Shape (histograms, boxplots)
Measures of central tendency

Winglengths (in mm) for $n = 45$ female hook-billed kites:

<table>
<thead>
<tr>
<th>Winglengths (in mm) for $n = 45$ hook-billed kites</th>
</tr>
</thead>
<tbody>
<tr>
<td>284 266 271 280 285 286 289 280 262</td>
</tr>
<tr>
<td>285 285 280 283 297 285 285 277 245</td>
</tr>
<tr>
<td>288 295 300 288 268 286 272 310 250</td>
</tr>
<tr>
<td>273 282 272 271 303 282 305 262</td>
</tr>
<tr>
<td>275 305 292 257 285 261 280 274 258</td>
</tr>
</tbody>
</table>

Ordered values:

Hook-billed kites

© 2006 Kevin T. Karlson
Measures of central tendency

1. The (sample) **mean** is the arithmetic mean or average:

   \[
   \bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = n^{-1} \sum_{i=1}^{i=n} y_i
   \]

2. The (sample) **median** divides the ranked sample into halves:

   \[
   \frac{\text{(proportion of sample } \leq \tilde{y})}{\text{proportion of sample } \geq \tilde{y}} \geq 0.50 \quad (1)
   \]

   if \( n \) is odd, take the \((n + 1)/2\) ordered observation,

   \[
   \tilde{y} = y_{(n+1)/2}
   \]

   if \( n \) is even, take the midpoint of the two middle observations:

   \[
   \tilde{y} = \frac{y_{(n/2)} + y_{(n/2+1)}}{2}.
   \]

3. The (sample) **mode** is the observation which occurs the most frequently.
For the female hook-billed kites data,

\[(y(1) = 245, 250, 257, 258, 261, 262, 262, 266, 268, 271, 271, 271, 272, \]
\[272, 273, 274, 275, 277, 280, 280, 280, 282, 282, 283, 284, \]
\[297, 300, 303, 305, 305, 310 = y(45)) \]

1. Mean

\[\bar{y} = \sum \frac{y_i}{n} = \frac{12590}{45} = 279.8\]

2. Since \( n = 45 \) is odd, the median, or 50\(^{th}\) percentile is the 23\(^{th}\) ordered length:

\[y(23) = ?\]

3. The mode is ?, since it occurred more frequently than any other value.
Measures of Spread

The range, $R$ of the sample $y_1, \ldots, y_n$:

$$R = \max\{y_1, \ldots, y_n\} - \min\{y_1, \ldots, y_n\}$$

The variance ($s^2$) of the sample $y_1, \ldots, y_n$ is the sum of the squared deviations from the $\bar{y}$, divided by $(n - 1)$.

$$s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}$$

$$= (n - 1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

A little algebra will yield the alternative formulas:

$$s^2 = \frac{\left( \sum y_i^2 \right) - n\bar{y}^2}{n - 1} = \frac{\left( \sum y_i^2 \right) - \left( \sum y_i \right)^2}{n - 1}$$
More about the **variance**

Q: What does the sample variance quantify?
A: A typical squared distance from the mean.

Q: Why *squared* distance? Why not just take distances?
A: Because \[ \sum (y_i - \bar{y}) = \]

Q: Why \( n - 1 \)?
A: This is known as the degrees of freedom (d.f.). It is the number of independent pieces of information we have with which to describe the spread.

For the kites, \( \sum y_i = 12590, \sum y_i^2 = 3531578 \), and \( n = 45 \) so that

\[
s^2 = \frac{3531578 - (12590)^2/45}{44} = \frac{9175.8}{44} = 208.5\text{mm}^2
\]
The standard deviation

Q: What does $s^2 = 208.5 \text{mm}^2$ mean?
A: A little bit difficult to interpret.

The **standard deviation** ($s$) of a sample $y_1, \ldots, y_n$ is the positive square root of $s^2$:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}.$$

Note that the units of $s$ are always the same as the units of $y_j$. For the kites data, the **standard deviation** is $s = \sqrt{208.5} = 14.4$.

Q: Again, why didn’t we just use $(\sum(y_i - \bar{y}))/n$?
Chebyshev’s Rule: For \( k \geq 2 \), the interval from \( \bar{y} - k \cdot s \) to \( \bar{y} + k \cdot s \) contains at least \( 100(1 - 1/k^2)\% \) of the population. This rule works no matter what the distribution looks like. Take \( k = 2 \).

Chebyshev’s Rule implies that at least 75% of the observations \( y_1, \ldots, y_n \) will be contained in the interval \( (\bar{y} - 2s, \bar{y} + 2s) \). For the kites data, \( \bar{y} = 279.8 \) and \( s = 14.4 \). The \( \bar{y} \pm 2s \) interval is then \( (251.0, 308.6) \), which contains 42 of the 45 observations. (This is a proportion of \( 42/45 = 93\% \).)

The Empirical Rule says that for normal (bell-curve) populations,

\[
\text{(proportion in } \bar{y} \pm s) \approx 68\%
\]
\[
\text{(proportion in } \bar{y} \pm 2s) \approx 95\%
\]
Quantiles/Percentiles

For $0 < p < 1$, the $100 \times p$ percentile, $y_p$ is any number satisfying

\[
\left( \text{proportion of sample} \leq y_p \right) \geq p \\
\left( \text{proportion of sample} \geq y_p \right) \geq 1 - p.
\]

The lower and upper quartiles, denoted by $q_1$ and $q_3$ are the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles. The \textbf{interquartile range}, is

\[
IQR = q_3 - q_1 = y_{.75} - y_{.25}.
\]

The idea is that this is how long of an interval you need to contain the middle half of the data. It is another measure of dispersion.
Quantile/percentiles

For the kites data,


\[\text{(proportion of sample} \leq 271) = \frac{12}{45} = 0.27 \geq 0.25\]
\[\text{(proportion of sample} \geq 271) = \frac{36}{45} = 0.80 \geq 1 - 0.25\]

and

\[\text{(proportion of sample} \leq 286) = \frac{34}{45} = 0.76 \geq 0.75\]
\[\text{(proportion of sample} \geq 286) = \frac{13}{45} = 0.29 \geq 1 - 0.75\]

\[IQR = q_3 - q_1 = 286 - 271 = 15\]
Boxplots

- simple graph which highlights a sample’s location and spread.
- a box with lines at the three quartiles
- whiskers which extend to most extreme observations that aren’t farther out than $1.5 \times IQR$ beyond quartiles.
- observations beyond this range are outliers and appear as individual lines or points
Boxplots

Recall the kites data, where
$q_1 = 271\, mm$,  $q_2 = 282\, mm$,  $q_3 = 286\, mm$,  $IQR = 15\, mm$.
and also the baseball data
Are male spiders the same size as female spiders?
A picture of a dahlia with a green lynx spider:
Side-by-side boxplots to visually compare distributions
More about the baseball data ($n = 840$ players)

Q: What is a typical salary in 2011? (1994 given in parentheses)

A1: $\bar{x} = $3,222,000 (1994 - $1,264,000)
A2: $\tilde{x} = $1,100,000 (1994 - $525,000)
A3: Mode = $414,000 (55 players) (1994 - $109,000)

For right-tailed distributions, mean $>$ median
For left-tailed distributions, mean $<$ median

The mean is the balancing point of histogram
the median breaks the histogram up into equal halves
mean and median are equal for symmetric distributions.

Q: What is the spread of the salaries?

A1: $s^2 = 18,656,290,000,000$ squared dollars
A2: $s = $4,319,292 (proportion within 2$s$ of $\bar{x}$?)
A3: $IQR = q_3 - q_1 = $4,290,000 - $430100 = $3,860,000
A4: $Range = $25,773,500
**Linear Transformations**

A linear transformation, $w_1, w_2, \ldots$ of a sample $y_1, y_2, \ldots$

$$w_i = a + b \ y_i$$

where $a$ and $b$ are fixed and known. Two useful identities are

$$\bar{y} = a + b \bar{x}$$

$$s_y = |b| s_x$$

**Example.** A sample of Fahrenheit temperatures:

$$f_1 = 81, 73, 85, 71, 78, 83, 80, 80, 82, f_{10} = 77^\circ F$$

$$\bar{f} = 79^\circ F, \quad s_F = \sqrt{19.1} = 4.4^\circ F$$

$$c_i = 5/9(f_i - 32) \quad \text{(in Celcius)}$$

$$c_1 = 27.2, \ldots, c_{10} = ?$$

$$\bar{c} = 5/9(\bar{f} - 32) = 5/9(79 - 32) = 26.1^\circ C$$

$$s_c = \left|5/9\right| s_f = 5/9(4.4) = 2.4^\circ C$$
Elements of probability

- **experiment** - a procedure with *outcome* which can’t be predicted with certainty.
- **sample space** - set of all possible outcomes
- **event** - a collection of outcomes.

Q: What is the probability of an event $A$? (Write $\Pr(A)$.)

A: more than 8 male offspring in a group of $n = 10$

B: a dog afflicted with lymphoma lives at least 10 years

Q: What do we mean by probability?

1. Long-run relative frequency for repeatable experiments.
2. Subjective assessment of likelihood of an event (for experiments which cannot be repeated.) What is the probability that President Barack Obama will be re-elected?
Introductions, Designing Experiments
Descriptive Statistics
Elements of probability

Probability continued

Probabilities are between 0 and 1: $0 \leq \Pr(A) \leq 1$.

The probability of an event $A$ is the sum of the probabilities of all the outcomes in $A$. If $A = \{s_1, \ldots, s_n\}$, then

$$\Pr(A) = \sum_{s \in A} \Pr(s) = \Pr(s_1) + \Pr(s_2) + \cdots + \Pr(s_n).$$

E.g. let $Y$ be the number of male offspring out of $n = 10$ then $S = \{0, 1, 2, \ldots, 10\}$ and

$$\Pr(A) = \Pr(Y > 8) = \Pr(Y = 9) + \Pr(Y = 10)$$

The intersection of $A$ and $B$ - the event that occurs if both $A$ and $B$ occur, or the set of all sample points in both $A$ and in $B$. Two events $A$ and $B$ are called mutually exclusive if the intersection of $A$ and $B$ contains no sample points ($\Pr(A \cap B) = 0$). The complement of $A$ (denoted $A'$ or $\bar{A}$ or $A^c$) is the event that $A$ does not occur:

$$\Pr(\bar{A}) = 1 - \Pr(A).$$
Conditional Probability

The conditional probability of event $A$ given $B$ occurs:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$ 

Example: 91 married couples answered the question “My partner and I are happily married.”
One couple is randomly sampled.
Let $H$ denote the events where the husband replies yes
Let $W$ denote the events where the wife replies yes respectively.

<table>
<thead>
<tr>
<th>freq.</th>
<th>$W$</th>
<th>$\bar{W}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>67</td>
<td>5</td>
</tr>
<tr>
<td>$\bar{H}$</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conditional probability continued

Note that

\[ \Pr(W|H) = 0.93 \]
\[ \Pr(\bar{W}|H) = 0.07 \]
\[ \Pr(W|\bar{H}) = 0.63 \]
\[ \Pr(W) = 0.87 \]

So, the probability of \( W \) given \( H \) depends on \( H \):

\[ \Pr(W|H) \neq \Pr(W|\bar{H}) \]

and

\[ \Pr(W|H) \neq \Pr(W) \]
**Definition:** Two events $A$ and $B$ are said to be **independent** if

$$\Pr(A|B) = \Pr(A). \text{ (write } A \perp B)$$

The following statements are equivalent:

1. $A \perp B$
2. $\Pr(B|A) = \Pr(B)$
3. $\Pr(A \cap B) = \Pr(A) \Pr(B)$.

U.S. population by region and opinion about marijuana:

<table>
<thead>
<tr>
<th></th>
<th>rel. freq.</th>
<th>$F$</th>
<th>$\bar{F}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>7.8%</td>
<td>22.2%</td>
<td>22.2%</td>
</tr>
<tr>
<td>$\bar{E}$</td>
<td>18.2%</td>
<td>51.8%</td>
<td>51.8%</td>
</tr>
</tbody>
</table>

One person *randomly sampled.*

$F$ : person favors the legalization of marijuana and
$E$ : person is from the East.

Q: Are region and opinion about marijuana **independent**?

A: Yes, since $\Pr(F|E) = \Pr(F)$.

Q: Are **mutually exclusive** and **independent** the same?