Statistics

- Statistics is the ‘science’ of drawing inferences about data in the face of uncertainty
- A direct outgrowth from probability theory
- In early astronomy, it was noted that observations are subject to random errors
- Many branches and variations on statistics, depending on the assumptions about distributions that get made (frequentists, Bayesians, etc)
A *population* is a complete set of individuals (e.g., US voters, sentences of English). A *parameter* is a numeric value characterizing a population.

In general, the true value of a parameter is unknowable.

A *sample* is a subset of a population, and a *statistic* is a numeric quantity calculated from a sample.

The aim of *descriptive statistics* is to summarize a sample in a clear and accessible manner.

*Inferential statistics* draws conclusions about a population from the properties of a sample (parameter estimation).
Types of data

- **Nominal scale**
  Discrete, qualitative data with no inherent order.
  Examples: eye color, religion, grammaticality judgements

- **Ordinal scale**
  Discrete data on an ordered scale. Intervals between ranks are not necessarily equal, no zero point.
  Examples: course grades, graded grammaticality judgements

- **Interval scale**
  Continuous data on an ordered scale. Intervals are equal, but no zero.
  Examples: temperature in Fahrenheit

- **Ratio scale**
  Continuous data on an ordered scale. Intervals are equal, true zero point.
  Examples: duration, income
Location statistics

Measures of central tendency identify the “center” of a distribution.

- **Mode**
  Most frequently occurring element(s). The only measure that can be used with nominal data.

- **Median**
  The middle score, such that half the scores are above the median and half are below. Useful for ordinal data, or highly skewed distributions (e.g., income).

- **Mean**
  Arithmetic average. Useful for interval and ratio data. For large samples, the mean approaches the expected value.

- **Trimmed mean**
  Arithmetic average with extreme values omitted. E.g., Olympic scoring.
Sample Mean

- Despite its limitations, the sample mean is the most widely used measure of central tendency.

- Strong Law of Large Numbers
  If $X_1, X_2, \ldots$ is an infinite sequence of independent random variables, all of which have the same expected value $\mu$ and the same finite variance $\sigma^2$, then the sample average $m_n$

  $$m_n = (X_1 + \cdots + X_n)/n$$

  converges almost surely to $\mu$:

  $$P(\lim_{n \to \infty} m_n = \mu) = 1$$

- This means that the sample mean tells us something about distribution of the population.
Sample Mean

- The sample mean is also an unbiased estimator of the population mean.

- The bias of an estimator $\hat{\theta}$ for a population parameter $\theta$ is $E[\hat{\theta}] - \theta$

- The expected value of the sample mean is:

\[
E[m] = E\left[\frac{1}{n}\sum X_i\right]
\]
\[
= \frac{1}{n}\sum E[X_i]
\]
\[
= \frac{1}{n}\sum \mu
\]
\[
= \mu
\]
Because of the Strong Law of Large Numbers (and other theorems yet to come), the sample mean $m$ is a good estimator for the population mean $\mu$ (i.e., the expected value).

The population mean is a parameter (in the statistical sense), and is usually related to a parameter (in the probability distribution sense).

Even when the sample mean is strictly speaking not appropriate, it is often used, hopefully without too much harm (Likert scales, GPA).
Dispersion statistics

Measures of dispersion indicate how spread out a distribution is.

- **Range**
  Largest and smallest values. Simple, but very sensitive to outliers.

- **Interquartile range**
  Difference between the 75th percentile and the 25th percentile. Like
  the median, useful for skewed distributions.

- **Variance**
  Average squared deviation from the sample mean:

  \[
  V = \frac{1}{n} \sum_{i} (x_i - m)^2
  \]

  The standard deviation is the square root of the variance.

Other shape statistics include *skew* and *kurtosis*. 
Sample variance

- Unlike the sample mean, the sample variance $V$ is a biased estimator of the population variance $\sigma^2$ when $\mu$ is unknown.

- With a little bit of algebra, we can show that:

$$E[V] = \frac{n-1}{n} \sigma^2$$

and that:

$$E[V] - \sigma^2 = -\frac{\sigma^2}{n}$$

- An unbiased estimator of the population variance is:

$$s^2 = \frac{1}{n-1} \sum (x_i - m)^2$$
Graphical methods

- Numerical statistics are precise, but graphical methods are superior for finding patterns in data.

- Visualizations may indicate problems with data that would be obscured by a purely numeric approach.

- Bar graphs, pie charts, scatter plots, histograms

- Stem and leaf plots, box plots, quantile-quantile (QQ) plots
Graphical methods

Stem and leaf plots are enhanced histograms.

```
3 : 5
4 : 35
5 : 46778
6 : 12445556678899
7 : 01223346677899
8 : 0233467
9 : 223
```
Graphical methods

Box plots summarize the shape of a distribution.
Graphical methods

Scatter plots show relationships between variables

Petal and Sepal Dimensions in Iris Blossoms
Nominal data

- Graphical methods work best for ordinal, interval, and ratio data
- Most linguistic data is nominal
- Multiple variables can be presented as a contingency table

<table>
<thead>
<tr>
<th></th>
<th>NP</th>
<th>to</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP mod.</td>
<td>88</td>
<td>32</td>
</tr>
<tr>
<td>VP mod.</td>
<td>56</td>
<td>30</td>
</tr>
<tr>
<td>Pred.</td>
<td>17</td>
<td>105</td>
</tr>
</tbody>
</table>
Mosiac plot
Graphical methods