Probability and Statistics
Thought experiment
Know the population
Will sample $n$ members at random
What can we say about probability of sample statistics?
$p(M = x)$ for some value $x$
$p(M_1 > M_2)$
etc.
All hypothetical
Based on assumptions about population
Worked out mathematically
Not based on any real data
Logic behind all inferential statistics
IF the population has a certain distribution, what will the probabilities be?

Marbles
Half red, half blue. $p(\text{red}) =$
2/3 red, 1/3 blue. $p(\text{red}) =$
Probability is just a bag of marbles
Population = bag
Sample = drawing from bag
Sample four marbles
$p(2 \text{ reds, 2 blues}) =$

Probability Distributions
Probability distribution
Distribution defined only by probabilities (not frequencies)
Can think of as infinite population
Discrete vs. continuous probability distributions
Discrete: like histogram, but $p(x)$ instead of $f(x)$
Probability is height of each bar
Continuous: density function
Probability is area under the curve
Random variable
Variable defined by probability distribution
Can take on one of many values
Each value (or range of values) has a probability
Expected Value

Expected value
Mean of a probability distribution
\[ E(R) = \sum_x x \cdot p(R = x) \]

<table>
<thead>
<tr>
<th>x</th>
<th>p(x)</th>
<th>x \cdot p(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.0625</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>.25</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>.375</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>.25</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>.0625</td>
<td>0.25</td>
</tr>
</tbody>
</table>

\[ E(R) = 2.00 \]

Properties of Expected Value
Multiplying by a fixed number
I give you $3 per red: \[ E(3 \cdot R) = 3 \cdot E(R) \]
\[ E(c \cdot R) = c \cdot E(R) \]
Adding a fixed number
I give you $1 per red, plus a bonus $1: \[ E(R + 1) = E(R) + 1 \]
\[ E(R + c) = E(R) + c \]
Sum of two random variables
\[ E(R + G) = E(R) + E(G) \]
Relationship to mean
Expected value of any random variable equals its mean

Sampling from a Population
Sampling a single item \((n = 1)\)
What is \(E(X)\)?
\[ E(X) = \sum_x x \cdot p(x) = \mu \]
Sampling many items
What is \(E\left(\frac{\sum X}{n}\right)\)?
\[ E\left(\frac{\sum X}{n}\right) = E\left(\frac{X[1] + ... + X[n]}{n}\right) \]
\[ = E(X[1]) + ... + E(X[n]) \]
\[ = \mu + \mu + ... + \mu = n \cdot \mu \]
What is \(E(\text{mean}(X))\)?
\[ E\left(\frac{\sum X}{n}\right) = \frac{E\left(\sum X\right)}{n} = \frac{n \cdot \mu}{n} = \mu \]
\[ \rightarrow E(M) = \mu \]

Unbiased estimators
Statistic \(a\) is unbiased estimator of parameter \(a\) if \(E(a) = a\)
Sample mean is unbiased estimator of population mean
Biased Estimators: The Example of Variance

Population variance

$$\sigma^2 = \frac{\sum_{pop} (X - \mu)^2}{N}$$

$$= \text{mean}((X - \mu)^2)$$

$$= E((X - \mu)^2)$$

$(X-\mu)^2$ is unbiased estimator of $\sigma^2$

Many observations

$$E\left(\frac{\sum_{sample} (X-M)^2}{n}\right) = \sigma^2$$

Problem: We don't know $\mu$

$$\frac{\sum_{sample} (X-M)^2}{n}$$?

Sample mean always shifted toward sample

$$\frac{\sum_{sample} (X-M)^2}{n} < \frac{\sum_{sample} (X-\mu)^2}{n}$$

$$E\left(\frac{\sum_{sample} (X-M)^2}{n}\right) < \sigma^2$$

$$\frac{\sum_{sample} (X-M)^2}{n}$$ is biased estimate of $\sigma^2$

Not a good definition for sample variance

Sample Variance

Goal: Define sample variance to be unbiased estimator of population variance

$$E(s^2) = \sigma^2$$

Problem: Obvious answer is biased

$$E\left(\frac{\sum_{sample} (X-M)^2}{n}\right) < \sigma^2$$

$M$ is always closer to $X$ than $\mu$ is

Solution:

$$s^2 = \frac{\sum_{sample} (X-M)^2}{n-1}$$

Unbiased: $E(s^2) = \sigma^2$

Sample standard deviation

$$s = \sqrt{\frac{\sum_{sample} (X-M)^2}{n-1}}$$
Take-away Points
Random variables and probability distributions
Expected value
Formula: \( E(R) = \sum_x x \cdot p(R = x) \)
Relationship to mean
Adding and multiplying
Samples and sample statistics as random variables
Biased and unbiased estimators
\( M \) is an unbiased estimator of \( \mu \): \( E(M) = \mu \)
\[ \frac{\sum_{\text{sample}} (X - M)^2}{n} \] is a biased estimator of \( s^2 \)
Sample variance
Formula: \( s^2 = \frac{\sum_{\text{sample}} (X - M)^2}{n - 1} \)
Unbiased estimator of population variance
Sample standard deviation
\[ s = \sqrt{\frac{\sum_{\text{sample}} (X - M)^2}{n - 1}} \]