iClicker Question

How much do you care of the theory of data science vs. practice?
A. I love the theory
B. Everything in moderation
C. I only care about the practice
Estimation & Inference
The Signal vs. the Noise

As Nate Silver* will tell you, the difficulty in statistical inference is separating the signal from the noise.

- If we flip a fair coin, it is possible that the outcome will be a sequence of all heads, just due to random chance.
- If this is the only sample that we see, then how can we separate the signal (that the coin is fair, say) from the noise (the sequence of all heads that we observe).
Paul the Octopus

- A German animal oracle who successfully predicted the outcome of all 7 of Germany’s matches in the 2010 World Cup
- He also predicted Spain to win the Cup final
- How likely is his success rate?

```
> dbinom(7, 7, 1/2)
[1] 0.0078125
```

- Other factors: He chose Germany 11/14 times. And Spain (twice) and Serbia otherwise. Similar flags. Likely color blind, but perhaps has preference for horizontal shapes.
Statistical Estimation & Inference

- Design and build a parametric model intended to explain data
  - about a population, or a sample of a population
- Estimate its parameters
- Quantify the uncertainty in those estimates
- Use the model to draw inferences about the entire population

“All models are wrong, but some are useful.” -- George Box
Properties of Estimators
What is an Estimator?

- A point estimator is a **function** that takes data (i.e., a sample) as input, and produces point estimates as output.
  - The sample mean **function** outputs the mean of its input.
  - Likewise, the sample variance function outputs the variance of its input.
- Note the nomenclature: a **point estimator** is a rule for generating **point estimates**.
  - “Average all the values in the sample” is a rule/function.
  - The average if all the values in a particular sample is an estimate.
Example

- Assume $n$ i.i.d. (independent and identically distributed) random variables $X_1, X_2, \ldots, X_n$ with mean $\mu$ and standard deviation $\sigma$

- $f(X_1, X_2, \ldots, X_n) = (1/n) \sum X_i$ (i.e., the sample mean function) is an estimator of the mean

- $\bar{x} = (1/n) \sum x_i$ (i.e., a sample mean) is an estimate of $\mu$
Evaluating Estimators (and Estimates)

- Any function of the data is an estimator!
- So how do we know we’ve got a good one?
- Desiderata:
  - In the limit, as the sample size tends to $\infty$, a consistent estimator converges to the model parameter it is estimating
  - An estimator is called unbiased if its expected value is the model parameter it is estimating
  - The efficiency of an estimator measures the quantity of data necessary to produce a certain quality estimate
Consistency

- An estimator is consistent if its value approaches its true value as the sample size tends to $\infty$.
- Consistent estimators become more accurate as the sample size increases.
- Is the sample mean a consistent estimator? Why or why not?
Bias

- Suppose $\theta^*$ is our model parameter, and $\theta$ is our estimator.
- The function $\theta$ applied to data yields a point estimate.
- $E[\theta]$ is the expected value of the estimator, where the randomness comes from the randomness in the data.
- $\text{Bias}[\theta] = E[\theta] - \theta^*$
- If $\text{Bias}[\theta] = 0$, then $\theta$ is unbiased.
- If an estimator is unbiased, then on average it yields an accurate prediction of the model parameter.
Example

- Let $\bar{x} = (1/n) \sum x_i$ represent the sample mean.
- $\text{Bias}[\bar{X}] = E[\bar{X}] - \mu = E[X_i] - \mu = \mu - \mu = 0$.
- The sample mean is unbiased.
Example, continued

- But $X_1$ and $X_2$ and so on are also unbiased.
- So why is $\bar{X}$ a better estimator than $X_1$ (or $X_2$ and so on)?
- Given two unbiased estimators, the preferred one is the one with lower variance:
  - $\text{Var}(X_1) = \sigma$
  - $\text{Var}(\bar{X}) = \text{Var}(1/n \sum X_i) = 1 / n^2 \sum \text{Var}(X_i) = \sigma / n$
  - $\text{Var}(\bar{X}) < \text{Var}(X_1)$
Mean Squared Error

- $\text{MSE}(\theta, \theta^*) = \mathbb{E}[(\theta - \theta^*)^2]$
- **Fact**: For an unbiased estimator, the MSE is just the variance!
  - $\text{MSE}(X_1) = \text{Var}(X_1)$
  - $\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) / n$
  - $\text{MSE}(\bar{X}) < \text{MSE}(X_1)$
Bias-Variance Decomposition

- **Theorem**: \( \text{MSE}(\theta, \theta^*) = B(\theta, \theta^*)^2 + \text{Var}(\theta) \)
- So error is a combination of bias and variance in our estimator.
- Ideally, we would reduce both, but this is often impossible.
- Instead, we usually trade off one against the other.

*This theorem implies our earlier fact.*
Proof of Bias-Variance Decomposition

- $\text{MSE}(\theta, \theta^*) = \mathbb{E}[(\theta - \theta^*)^2] = \mathbb{E}[\theta^2 - 2\theta^*\theta + (\theta^*)^2] = \mathbb{E}[\theta^2] - 2\theta^*\mathbb{E}[\theta] + (\theta^*)^2$
- $\text{B}(\theta, \theta^*)^2 = (\mathbb{E}[\theta] - \theta^*)^2 = (\mathbb{E}[\theta] - \theta^*)(\mathbb{E}[\theta] - \theta^*) = (\mathbb{E}[\theta])^2 - 2\theta^*\mathbb{E}[\theta] + (\theta^*)^2$
- $\text{Var}[\theta] = \mathbb{E}[(\theta - \mathbb{E}[\theta])^2] = \mathbb{E}[\theta^2] - (\mathbb{E}[\theta])^2$
- $\text{B}(\theta, \theta^*)^2 + \text{Var}[\theta] = (\mathbb{E}[\theta])^2 - 2\theta^*\mathbb{E}[\theta] + (\theta^*)^2 + \mathbb{E}[\theta^2] - (\mathbb{E}[\theta])^2$
Bias-Variance Tradeoff

Low Bias

Low Variance

High Variance

High Bias

Image Source
Bias-Variance Tradeoff

- Simple (e.g., linear) models are highly biased; as such, they often underfit, meaning they fail to capture regularities in the data.
- Otoh, they are not sensitive to noise, so are comparatively low variance (i.e., they have so much bias that they don’t change much with the data).
- More complicated models are less biased. Because of their flexibility, they end up modeling noise (as well as signal), and consequently overfit.
- Flexible models have high variance, b/c the models themselves can vary enormously with the data.
Linear Regression, Revisited
Statistical Models

A statistical model is a set of assumptions about how sample data are generated, characterized by parameters (e.g., $\mu$, $\beta_0$, and $\beta_1$).

Examples:

- Here’s a simple one: $Y = \mu + \varepsilon$, where $\mu$ is a model parameter representing the mean of a population, and $\varepsilon$ is a random error term (a.k.a. noise).
- Here’s another one: $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon$ represents noise/error.
Linear Model

- The distribution of $X$ is arbitrary (perhaps even non-random).
- The distribution of $Y$ depends on that of $X$ in a linear fashion:
  - The distribution of $Y$ given $X$ is $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon$ represents noise/error.
Linear Model

- The distribution of $X$ is arbitrary (perhaps even non-random).
- The distribution of $Y$ depends on that of $X$ in a linear fashion:
  - The distribution of $Y$ given $X$ is $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon$ represents noise/error.
- Find $\beta_0$ and $\beta_1$ that minimize the mean squared error:
  - $(\beta_0, \beta_1)$ s.t $E[(Y - \beta_0 - \beta_1 X)^2]$ is minimized
- Solve as usual
  - Take partial derivatives, and set them equal to zero.
- Out pops:
  - $\beta_0 = E[Y] - \beta_1 E[X]$
  - $\beta_1 = \text{Cov}(X, Y) / \text{Var}(X) = \text{Corr}(X, Y) \sigma_Y / \sigma_X$
- The same answer as before (in expectation)!

Linear Model: Additional Assumptions

- The distribution of $X$ is arbitrary (perhaps even non-random).
- The distribution of $Y$ depends on that of $X$ in a linear fashion:
  - The distribution of $Y$ given $X$ is $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon$ represents noise/error.
  - Alternatively, for all $1 \leq i \leq n$, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.
- Both the $X_i$ and $\varepsilon_i$ (the error) are random variables.
  - The error terms have an expectation of 0: $E[\varepsilon_i \mid X = x] = 0$.
  - The error terms have constant variance: $\text{Var}[\varepsilon_i \mid X = x] = \sigma^2$.
  - The error terms are uncorrelated with themselves: (i.e., no time-series data):
    $\text{Cov}[\varepsilon_i = \varepsilon_j] = 0$, for all $i \neq j$.
- Under these assumptions, $\beta_0$ and $\beta_1$ are unbiased and consistent estimators.
  - Unbiased, because the error terms have an expectation of 0.
  - Consistent, by the law of large numbers (and the model assumptions).