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).
Population vs. Sample

- A **population** is a set of individuals under study, chosen by the statistician, or nowadays, the data scientist.
- Typically, the entirety of the population is unobservable.
- A **sample** is an observed subset of the population.
- Our goal is to study the sample, and then draw inferences about the population from what we observe about the sample.
- E.g., **polling** (ask only a sample of college students their favorite ice cream flavor; then infer the preferences of all students).
Statistical Models

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. (This, by the way, is a linear model.)

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

“All models are wrong, but some are useful.” -- George Box
Statistical Estimation & Inference

- Statistical (machine) **learning**, or **estimation**, involves fitting parameters of a statistical model based on **in-sample** data.

- Given a statistical model, and learned parameters, we apply **inference** to draw conclusions about **out-of-sample** data.
  - E.g., We can infer that a jury panel is or is not representative of the population from which the jurors were selected.
Estimation
Statistical Estimation

- Classical (a.k.a. frequentist) approach
  - The unknown quantity $\theta^*$ is deterministic.
  - It is estimated by a random variable called an estimator $\theta$.
  - A goal of learning is to learn parameters that minimize expected loss, $E_x[\mathcal{L}(\theta(x), \theta^*)]$, where $x$ denotes (random) measurements.

- Bayesian approach
  - The unknown quantity $\Theta$ is a random variable, about which we have some prior distributional knowledge.
  - After observing the data, we update our prior to build a posterior, using Bayes’ rule.
  - A goal of learning is to learn parameters that minimize expected loss, $E_{\theta,x}[\mathcal{L}(\theta(x), \theta^*)]$, assuming both $x$ and $\theta$ are random.
What is a Point Estimate?

- Given a statistical model, a **point estimate** is a single value (as opposed to an interval) used to estimate a model parameter.

Examples:

  - The sample mean is usually used to estimate a model parameter representing the mean (i.e., the “true” mean).
  - Likewise, the sample variance is used to estimate a model parameter representing the variance (i.e., the “true” variance).
Confidence Intervals
A Confidence Interval

- Goal is to find not just a single point estimate, but an interval estimate
- Delimited by an upper and lower bound, such that the probability that the interval includes $\theta$ is $1 - \propto$
- $1 - \propto$ is called the confidence level, and preferably, $\propto$ is small
  - For example: $\propto = 0.05$ implies a 95% confidence level
  - Likewise: $\propto = 0.1$ implies a 90% confidence level

Formally: find $\theta_{lo}$ and $\theta_{hi}$ such that $\Pr[\theta_{lo} \leq \theta \leq \theta_{hi}] \geq 1 - \propto$
A Motivational Story

- Let’s say you are running for mayor.
  - You hire a polling agency to determine if you are likely to win or not.
  - After sampling 100 likely voters, the polling agency reports that 55 of those 100 support you.
- But more than 100 people live in your town! Are you going to win?
  - The polling agency also reports: “with 95% confidence, the true proportion of the (entire) population who support you is between .45 and .65.”
  - Here’s what this means:
    - If the polling agency takes this poll for you 100 times, then (on average) 95% of the confidence intervals they report will contain the true proportion of the population who support you.
  - Here’s what this doesn’t mean:
    - There is a 95% chance the proportion of the population who support you lies in this interval.
- But wait…how did the polling agency come up with the interval?
Let $X$ be the number of people who support you in a poll, and let $\hat{p} = X / n$.

By the central limit theorem (YAY!), for large $n$, $\hat{p}$ is approximately normal.

Perfect. Let’s apply a $z$-transformation: $\Pr(z_{\text{lo}} \leq Z \leq z_{\text{hi}}) = .95$.

- $\Pr(z_{\text{lo}} \leq (\hat{p} - \mu) / \sigma \leq z_{\text{hi}}) = .95$
- $\Pr(\mu - z_{\text{lo}} \sigma \leq \hat{p} \leq \mu + z_{\text{hi}} \sigma) = .95$
- $\Pr(\hat{p} - z_{\text{lo}} \sigma \leq \mu \leq \hat{p} + z_{\text{hi}} \sigma) = .95$
- N.B. $z_{\text{lo}} = -1.96$ and $z_{\text{hi}} = 1.96$, since $\alpha = 0.05$. 

Building a Confidence Interval
Building a Confidence Interval (cont’d)

- Let $X$ be the number of people who support you in a poll, and let $p_{\text{hat}} = X / n$.
- By the central limit theorem (YAY!), for large $n$, $p_{\text{hat}}$ is approximately normal.
- Perfect. Let’s apply a $z$-transformation: $\Pr(z_{\text{lo}} \leq Z \leq z_{\text{hi}}) = .95$.
  - $\Pr(z_{\text{lo}} \leq (p_{\text{hat}} - \mu)/\sigma \leq z_{\text{hi}}) = .95$
  - $\Pr(\mu + z_{\text{lo}} \sigma \leq p_{\text{hat}} \leq \mu + z_{\text{hi}} \sigma) = .95$
  - $\Pr(p_{\text{hat}} + z_{\text{lo}} \sigma \leq \mu \leq p_{\text{hat}} + z_{\text{hi}} \sigma) = .95$
  - N.B. $z_{\text{lo}} = -1.96$ and $z_{\text{hi}} = 1.96$, since $\alpha = 0.05$.
- Sounds good, but what are the mean $\mu$ and variance $\sigma^2$ of $p_{\text{hat}}$?
  - $E[p_{\text{hat}}] = E[X / n] = 1/n \ E[X]$
  - $\text{Var}[p_{\text{hat}}] = \text{Var}[X / n] = 1/n^2 \ \text{Var}[X]$
- This begs the question: what are the mean and variance of $X$?
The mean and variance of $p_{\text{hat}}$

- Recall that $X$ is the number of people who support you in a poll.
- $X$ is the result of a random experiment (the poll), so it is a random variable.
- Great! How is it distributed? Easy: $X$ is binomially distributed:
  - Mean = $np$
  - Variance = $np(1-p)$
- The expected value of $p_{\text{hat}}$ is $p$.
  - $E[p_{\text{hat}}] = E[X / n] = np / n = p$
- The variance of $p_{\text{hat}}$ is $(p)(1 - p) / n$.
  - $\text{Var}[p_{\text{hat}}] = \text{Var}[X / n] = 1/n^2 \text{Var}[X] = 1/n^2 (np)(1 - p)$
Z-transformation

- Recall: $\Pr(-1.96 \leq Z \leq 1.96) = .95$
- In other words:
  - $\Pr(z_{lo} \leq (\hat{p} - p)/\sigma \leq z_{hi}) = .95$
  - $\Pr(p + z_{lo} \sigma \leq \hat{p} \leq p + z_{hi} \sigma) = .95$
  - $\Pr(\hat{p} + z_{lo} \sigma \leq p \leq \hat{p} + z_{hi} \sigma) = .95$

- By definition, $\hat{p} = X / n = 0.55$
- We estimate the variance $\sigma^2$ using $\hat{p}$ instead of $p$ (since we don’t know $p$):
  - $\sigma^2 = (.55)(.45) / 100 = 0.002475$
  - $\sigma = \sqrt{0.002475} = 0.05$
An Interval Around Proportions

Hence, we build our 95% confidence interval as follows:

- $\Pr(\hat{p} - z_{lo} \sigma \leq p \leq \hat{p} + z_{hi} \sigma) = .95$
  - Lower Bound: $0.55 + (-1.96)(0.05) = .45$
  - Upper Bound: $0.55 + (1.96)(0.05) = .65$

- The value $(1.96)(0.05) \approx 0.1$ is called the margin of error.

Reminder, restated in terms of margin of error:

- If we conduct this poll repeatedly, the true proportion of your supporters would be expected to fall within in the margin of error 95 times out of 100.
Deriving Standard Error
Standard Error

- The **sample mean** is the average of all the sample values.
- The **sample variance** is the average of the squared deviations from the mean.
- The **sample standard deviation** is the square root of the sample variance.
- The **standard error** is the standard deviation of the sampling distribution.

- In this class, we are learning to build confidence intervals for sample means and sample proportions.
- This means we need to be able to calculate the standard error of sample means and sample proportions.
The Standard Error of the Sample Mean

- Let $X_M$ represents the sample mean:
  - $X_M = (X_1 + \ldots + X_n) / n$

- First, we need the variance of the sample mean:
  - $\text{Var}[X_M] = \text{Var}[(X_1 + \ldots + X_n) / n] = (1 / n^2) (n) \text{Var}[X] = \sigma^2 / n$

- We now have a formula for the standard error of the sample mean:
  - $\text{SE}[X_M] = \sigma / \sqrt{n}$
Let $P_{\text{hat}}$ represent the sample proportion:

- $P_{\text{hat}} = X / n$, where $X$ is the binomial random variable distributed according to $(n, p)$

First, we need the variance of the sample proportion:

- $\text{Var}[P_{\text{hat}}] = \text{Var}[X / n] = (1 / n^2) (np)(1 - p) = p(1 - p) / n$

We now have a formula for the standard error of the sample proportion:

- $\text{SE}[P_{\text{hat}}] = \sqrt{p(1 - p) / n}$
Difference of Two Sample Means

- Let’s find the standard error of the difference between two sample means.
  - Let $X_M - Y_M$ represent the difference between two sample means.

- $X_M - Y_M = (X_1 + \ldots + X_N) / n - (Y_1 + \ldots + Y_M) / m$
  - $= V[(X_1 + \ldots + X_N) / n] + V[(Y_1 + \ldots + Y_M) / m]$
  - $= (1/n^2) nV[X_1] + (1/m^2) mV[Y_1]$
  - $= \sigma_x^2/n + \sigma_y^2/m$

- We now have a formula for the standard error of the difference between two sample means:
  - $SE[X_M] = \sigma_x / \sqrt{n} + \sigma_y / \sqrt{m}$
Difference of Two Sample Proportions

- Let’s find the standard error of the difference between two sample proportions.
  - Let \( P_1 \) represent one proportion and \( P_2 \) represent a second proportion.
- \( P_1 = X / n \) and \( P_2 = Y / m \)
  - \( X \sim B(n, p_1) \) and \( Y \sim B(m, p_2) \)
  - \( \text{V}[P_1 - P_2] = \text{V}[X / n] + \text{V}[Y / m] \)
  - \( = (1 / n^2) (np_1)(1 - p_1) + (1 / m^2) (mp_2)(1-p_2) \)
  - \( = p_1(1-p_1) / n + p_2(1 - p_2) / m \)
- We now have a formula for the standard error of the difference between two sample proportions:
  - \( SE[P_1 - P_2] = \sqrt{p_1(1-p_1) / n + p_2(1 - p_2) / n} \)
Favorite Ice Cream Flavor among U.S. Adults

- A Harris Poll surveyed 2242 adults in the United States about their favorite flavor of ice cream.
  - 27% of people said chocolate was their favorite flavor.
  - 23% of people said vanilla was their favorite flavor.
- Is the preference for chocolate over vanilla significant? That is, do people really prefer chocolate?
- Plan of attack:
  - Look at the difference in the proportions. It is 4%.
  - Calculate the confidence interval around this difference.
  - If it contains 0, then we cannot conclude that there is a difference.
We need to calculate the standard error to find a confidence interval:

- \((.27)(2242) = 605\) people preferred chocolate
- \((.23)(2242) = 516\) people preferred vanilla

\[
\text{Var}[p_1 - p_2] = \frac{(0.27)(1 - 0.27)}{605} + \frac{(0.23)(1 - 0.23)}{516} = 0.00067
\]

The standard error is the square root of the variance: \(\sqrt{0.00067} = 0.026\)

So, the confidence interval, at the 95% level, is:

\[
[0.04 - (1.96)(0.026), 0.04 + (1.96)(0.026)]
\]

\([-0.011, 0.09]\)

\([-1.1\%, 9\%]\)

0% is included in this interval!

- This means we cannot conclude, at the 95% confidence level, that people prefer chocolate to vanilla.
- But maybe people prefer chocolate to vanilla at the 90% confidence level. (Check this!)
Some words of warning!

- Our methods make use of the central limit theorem.
  - In the example, we do not assume anything about how ice cream preferences are distributed.
  - However, by the CLT, we do assume the sampling distribution is approximately normal.
  - The survey was large enough sample for the CLT to apply. However, the CLT might not have applied if the sample size were smaller.

- Very common Mistake: if I have a 95% confidence interval between 1 and 2, people often interpret it as a 95% chance the parameter lies between 1 and 2.
  - This is a misconception! A 95% interval means that if we repeated the experiment 100 times, 95 of the resulting 100 intervals would contain the true population parameter.
  - These two interpretations are not the same. Be careful to avoid this potential pitfall!
Hypothesis Testing
A Motivating Example

- Between 1960 and 1980, there were many lawsuits in the south claiming racial bias in jury selection.
- Here’s some made up* (but similar) supporting data:
  - 50% of citizens in the local area are African American
  - On an 80 person panel, only 4 were African American
- Can this outcome be explained as the result of pure chance?
- Unlikely: If $X \sim B(80, .5)$, then $P[X \leq 4] = 1.8 \times 10^{-18}$
- N.B.: Statistics can never prove anything!

*This example was borrowed from The Cartoon Guide to Statistics.
Steps in Hypothesis Testing

- **Step 1: Formulate the hypothesis**
  - Assume a binomial random variable, with $n = 80$.
  - The **null** hypothesis: There is no bias in jury selection (i.e., $p = 0.5$).
  - The **alternative** hypothesis: $p < 0.5$.

- **Step 2: The test statistic**
  - Z-score for a binomial random variable, with $n = 80$ and $p = 0.5$, and $X = 4$

- **Step 3: Calculate the $p$-value**
  - If the null hypothesis were true, what is the probability of observing a test statistic as extreme as the one we observed?

- **Step 4: Perform a significance test at the $\alpha$-level**
  - If $p$-value $\leq \alpha$-level, then the effect is significant.
  - We reject the null hypothesis and search for alternative explanations.
Calculations

- Numerator: \( \hat{p} - p = 0.05 - 0.5 = -0.45 \)
  - By subtracting \( p \), we are assuming that the null hypothesis is true.

- Denominator: Standard Error
  - \( \text{Var}(\hat{p}) = (0.5)(1 - 0.5) / 80 = 0.003125 \)
  - The standard error is the square root of this variance: \( \sqrt{0.003125} = 0.056 \)

- Z-score: \( -0.45 / 0.056 \approx -8.0 \)
  - That’s a whole lot of standard deviations below the mean!

- If the null hypothesis were true, the probability of observing a test statistic as extreme as the one we observed is essentially 0.
- We reject the null hypothesis and search for alternative explanations.
Who is going to be the next US president?

- YouGov surveyed 1300 people in the United States and asked “Who will you vote for in the election for President in November?”
- The poll was conducted on October 7th and 8th.
  - 44% of people planned to vote for Clinton.
  - 38% of people planned to vote for Trump.
Step 1: Formulate Hypotheses

- Our null hypothesis, $H_0$, is that the proportion of people that plan to vote for Trump is identical to the proportion of people that plan to vote for Clinton.
  - $p_H = p_T$
- Our alternative hypothesis, $H_A$, is that the proportion of people who plan to vote for Clinton is higher than the proportion who plan to vote for Trump.
  - $p_H > p_T$
Step 2: Calculate the Test Statistic

- **Z-score for the difference between two sample proportions**
  
  Numerator: \((.44 - .38) - 0 = .06\)
  
  By subtracting 0, we are assuming that the null hypothesis is true.

- **Denominator: Standard Error**
  
  \((.44)(1300) = 572\) people preferred Clinton
  
  \((.38)(1300) = 494\) people preferred Trump
  
  \(\text{Var}[p_H - p_T] = (0.44)(1 - 0.44) / 572 + (0.38)(1 - 0.38) / 494 = 0.0009\)
  
  The standard error is the square root of this variance: \(\sqrt{0.0009} = 0.03\)

- **Z-score**: \(0.06 / 0.03 = 2\)
Step 3: Calculate the $p$-value

- Given that the null hypothesis is true, what is the probability that we would see data as extreme as what we observed?
  - $\text{pnorm}(2, \text{lower.tail} = \text{FALSE}) = 0.023$
Step 4: Perform a Significance Test

- If we assume the null hypothesis is true, then there is only a 2% chance we would see data that favors Clinton as much or more than what we saw in the YouGov poll.
- This means one of two things:
  - We witnessed something really rare.
  - The assumption that the null hypothesis is true is incorrect.
- Typically, we set a benchmark threshold ($\alpha$-level) before we run the test.
  - Often, the threshold is 5% (corresponding to a 95% confidence interval).
  - If the $p$-value is below this threshold, then we reject the null hypothesis.
“Innocent until proven guilty”

- Hypothesis testing is a statistical implementation of this maxim
- Null hypothesis: the defendant is innocent
- Alternative hypothesis: the defendant is guilty
  - A type 1 error (false positive) occurs when we put an innocent person in jail
  - A type 2 error (false negative) occurs when we do not jail a guilty person
- Another example:
  - Type 1 error: false alarm (fire alarm when there is no fire)
  - Type 2 error: fire but no fire alarm
Type I vs. Type II Errors

- Cancer screening
  - Null: no cancer
  - Type I: cancer suspected where there is none (not good, but not terrible)
  - Type II: cancer goes undetected (very very bad)

- Err on the side of type I errors (choose higher significance level)
Type I vs. Type II Errors

- Spam Filters
  - Null: an email is legitimate
  - Type I: filter a legitimate email (could be very bad)
  - Type II: don’t filter spam (not so bad)

- Err on the side of type II errors (choose lower significance level)
Type I vs. Type II Errors

- Suspected terrorists
  - Null hypothesis: person is not a terrorist
  - Type I: send an innocent person to Guantánamo Bay
  - Type II: let a terrorist (who intends to commit mass murder) free

- US is erring on the side of type I errors, which explains why people are held at Guantánamo Bay without a fair trial
Extras
The Presidential Election

- YouGov surveyed 1300 people in the United States and asked “Who will you vote for in the election for President in November?”
  - The poll was conducted on October 7th and 8th
  - 44% of people planned to vote for Clinton
  - 38% of people planned to vote for Trump

- The difference in proportions in this survey is 44% - 38% = 6%.

- Given these polls results, how confident are you that Clinton will win the election?
The Presidential Election

- We need to calculate the standard error to find a confidence interval:
  - \((.44)(1300) = 572\) people preferred Clinton
  - \((.38)(1300) = 494\) people preferred Trump
  - \(\text{Var}[p1 - p2] = (0.44)(1-0.44) / 572 + (0.38)(1-0.38) / 494 = 0.0009\)
  - The standard error is the square root of the variance: \(\sqrt{0.0009} = 0.03\)

- So, the confidence interval, at the 95% level, is:
  - \(0.06 - (1.96)(0.03)\) to \(0.06 + (1.96)(0.03)\)
  - \(0.01\) to \(0.12\)
  - \(1\%\) to \(12\%\)

- 0\% is **not** included in this interval!
  - This means we can conclude, at the 95% confidence level, that people prefer Clinton to Trump.
  - But people probably don’t prefer Clinton to Trump at the 99% confidence level. (Check this!)

- Will Clinton win the election?
Central Limit Theorem Visualizations

- [http://blog.vctr.me/posts/central-limit-theorem.html](http://blog.vctr.me/posts/central-limit-theorem.html)
A Brief Aside
The Linearity of Expectation

- We introduced the concept of expectation in the last lecture.
- Let $X_1$ be a random variable represented the first die roll.
  - $E[X_1] = 3.5$
- Now, let $X_2$, $X_3$, $X_4$, and $X_5$ represent additional die rolls.
  - $E[X_1 + X_2 + X_3 + X_4 + X_5] = E[5X_1] = 5E[X_1] = 17.5$
- More generally:
  - $E[aY_1 + bY_2] = E[aY_1] + E[bY_2] = aE[Y_1] + bE[Y_2]$
  - This rule is called the linearity of expectation.
The Linearity of Standard Deviations

- The same property of linearity holds for standard deviation as well.
  - \( \text{SD}[aY_1 + bY_2] = \text{SD}[aY_1] + \text{SD}[bY_2] = a\text{SD}[Y_1] + b\text{SD}[Y_2] \)

- Remember that variance is the square of standard deviation.
  - \( \text{Var}[aY_1 + bY_2] = \text{Var}[aY_1] + \text{Var}[bY_2] = a^2\text{Var}[Y_1] + b^2\text{Var}[Y_2] \)

- Summary:
  - You can pull constants out of expectation and standard deviation formulas.
  - You can pull constants out of variance formulas, but then you must square the constant when you pull it out.

- One more important note:
  - \( \text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y] \)
  - Why? \( E[(X-Y)^2] - (E[X-Y])^2 \Rightarrow \text{simplify and you’ll get this result!} \)