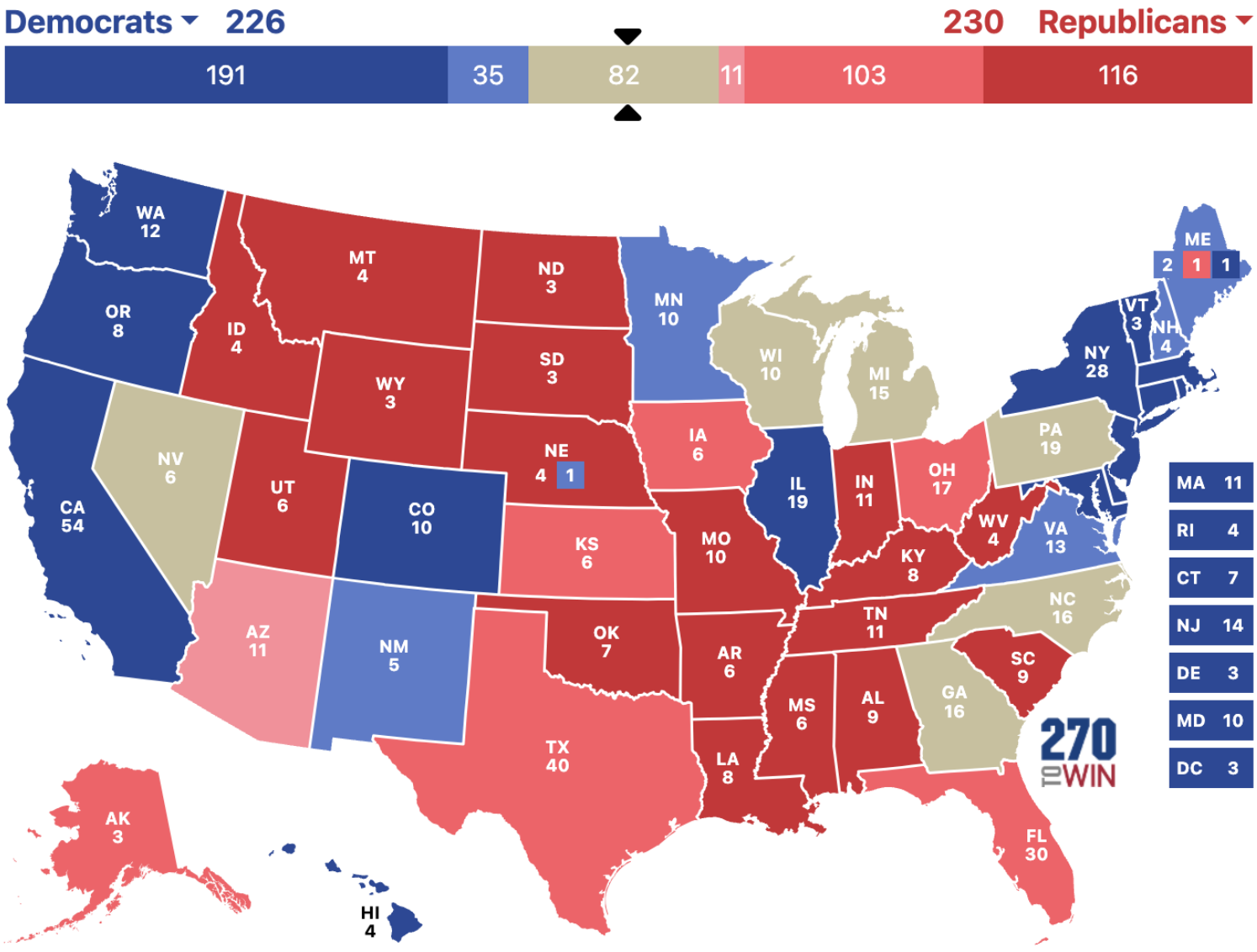
S&DS 242/542: Theory of Statistics

Lecture 1: Course introduction, survey sampling

2024 presidential election forecast



(Source: consensus forecast from 270towin.com, 4 November 2024)

Estimating the support for a candidate

What percentage of registered voters in Connecticut support Kamala Harris?

Let

- $N = 2{,}317{,}657$ (number of registered voters)
- $\theta =$ fraction who support Harris
- $1 - \theta =$ fraction who support a different candidate

$\theta \in [0, 1]$ is an unknown **parameter**.

Possible questions of interest:
1.) Is $\theta > 0.5$?
2.) What is our best estimate of θ?
3.) How much uncertainty is there in this estimate?
4.) Given a new voter, can we predict if they support Harris?

Survey sampling

These questions are typically answered by *survey sampling* or polling. Possible sampling methods:

- Survey every registered voter in Connecticut. This is very expensive and usually infeasible.
- Survey a subset of "representative" voters in Connecticut, as determined and selected by the investigator.
- Survey a subset of voters in Connecticut selected at **random**.

Random sampling was popularized by British statistician Arthur Lyon Bowley (1869 – 1957), who recognized several advantages:

- Guards against conscious or unconscious biases that may be introduced by the investigator. (What if the selected "representative" voters are not representative in some way?)
- Enables mathematical quantification of the uncertainty and magnitudes of error in inferences drawn from the data.

Simple random sample

Suppose we poll a **simple random sample** of $n = 1000$ people from Connecticut. This means:

- Person 1 is chosen at random (equally likely) from all N registered voters in Connecticut. Then person 2 is chosen at random from the remaining $N - 1$ people. Then person 3 is chosen at random from the remaining $N - 2$ people, etc.
- Or equivalently, all $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ possible sets of n people are equally likely to be chosen as our sample.

Then we can estimate

$$\theta = \frac{\#\text{ voters who support Harris}}{N}$$

by

$$\hat{\theta} = \frac{\#\textit{sampled}\text{ voters who support Harris}}{n}$$

Simple random sample

Say 540 of the 1000 people surveyed support Harris, so $\hat{\theta} = 0.54$.

What can we infer about θ?

Let's call our **data** $X_1, \ldots, X_n$:

$$X_i = \begin{cases} 1 & \text{if person } i \text{ supports Harris} \\ 0 & \text{if person } i \text{ does not support Harris} \end{cases}$$

Then $\hat{\theta} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$.

The data $X_1, \ldots, X_n$ are random, because we took a random sample. Therefore $\hat{\theta}$ is a random variable.

The sampling distribution of $\hat{\theta}$

$\hat{\theta}$ is a value computed from our observed data. Such a quantity is called a **statistic**. The probability distribution of $\hat{\theta}$ is called its **sampling distribution**.

We may be interested in its following properties:

- What is the mean of $\hat{\theta}$? If the mean is equal to θ, then $\hat{\theta}$ is an **unbiased** estimator for θ. Otherwise, $\hat{\theta}$ is **biased**.
- What is the standard deviation of $\hat{\theta}$? This quantifies the variability of our estimate $\hat{\theta}$, known as the **standard error**.
- What are the quantiles of $\hat{\theta}$? This information can allow us to create a **confidence interval** for θ.

Understanding the bias

For each $i = 1, \ldots, n$, by symmetry, the i^{th} person sampled is equally likely to be each of the N individuals.

Then the probability that this person supports Harris is $\frac{N\theta}{N} = \theta$. So $X_i \sim \text{Bernoulli}(\theta)$ and $\mathbb{E}[X_i] = \theta$.

Recall that expectation is *linear*:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y], \quad \mathbb{E}[cX] = c\mathbb{E}[X] \text{ if } c \text{ is constant.}$$

So

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{1}{n}(\mathbb{E}X_1 + \ldots + \mathbb{E}X_n) = \theta$$

The mean of $\hat{\theta}$ is θ, so $\hat{\theta}$ is **unbiased**. (If we were to repeat the survey many times, on average the value of $\hat{\theta}$ would be θ.)

Understanding the variability

Unbiasedness doesn't tell us how far off $\hat{\theta}$ is from θ, for a single survey. For this, let's consider the standard deviation $\sqrt{\text{Var}[\hat{\theta}]}$.

Consider first a simpler setting, where we sample $n = 1000$ individuals *with replacement* from the population of size N.

Then the random variables $X_1, \ldots, X_n$ would be **independent**: The event that the i^{th} sampled individual supports Harris is independent of all other samples.

Understanding the variability

Recall that

$$\text{Var}[cX] = c^2 \,\text{Var}[X] \text{ if } c \text{ is a constant.}$$

If X and Y are independent, then also

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

So $\text{Var}[\hat{\theta}] = \text{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right]$

$$= \frac{1}{n^2}\left(\text{Var}\, X_1 + \cdots + \text{Var}\, X_n\right) = \frac{1}{n}\text{Var}\, X_1$$

$$\text{Var}\, X_1 = \mathbb{E}[X_1^2] - (\mathbb{E}X_1)^2 = \theta - \theta^2 = \theta(1-\theta)$$

$$\Rightarrow \text{Var}[\hat{\theta}] = \frac{\theta(1-\theta)}{n}.$$

Understanding the variability

In our setting, the $n = 1000$ individuals are sampled *without replacement*. Then the previous calculation is not exactly correct, because sampling without replacement introduces dependence between the X_i's:

Suppose $X_1 = 1$, i.e. person 1 supports Harris. Conditional on $X_1 = 1$, the probability that $X_2 = 1$ is now $\frac{N\theta - 1}{N-1}$, instead of θ. So X_1 and X_2 are dependent.

Let's compute $\text{Var}[\hat{\theta}]$ a different way: Recall the definition of variance,

$$\text{Var}[\hat{\theta}] = \mathbb{E}[\hat{\theta}^2] - (\mathbb{E}\hat{\theta})^2$$

Here, the second term is $(\mathbb{E}\hat{\theta})^2 = \theta^2$.

Understanding the variability

For the first term,

$$\mathbb{E}[\hat{\theta}^2] = \mathbb{E}\left[\left(\frac{X_1 + \dots + X_n}{n}\right)^2\right]$$

$$= \frac{1}{n^2} \cdot \mathbb{E}\left[X_1^2 + \dots + X_n^2 + 2X_1X_2 + 2X_1X_3 + \dots + 2X_{n-1}X_n\right]$$

$$= \frac{1}{n^2} \cdot \left(n \cdot \mathbb{E}X_1^2 + 2 \cdot \underbrace{\binom{n}{2}}_{\frac{n(n-1)}{2}} \cdot \mathbb{E}X_1X_2\right)$$

$$= \frac{1}{n}\underbrace{\mathbb{E}[X_1^2]}_{=X_1} + \frac{n-1}{n} \cdot \underbrace{\mathbb{E}[X_1X_2]}_{= \begin{cases} 1 & \text{if both } X_1, X_2 = 1 \\ 0 & \text{otherwise} \end{cases}}$$

Understanding the variability

$$\mathbb{E}[X_1 X_2] = \mathbb{P}[X_1 = 1, X_2 = 1]$$

$$= \mathbb{P}[X_1 = 1] \cdot \mathbb{P}[X_2 = 1 \mid X_1 = 1]$$

$$= \theta \cdot \frac{N\theta - 1}{N - 1}$$

$$\Rightarrow \mathrm{Var}[\hat{\theta}] = \mathbb{E}[\hat{\theta}^2] - (\mathbb{E}\hat{\theta})^2$$

$$= \frac{1}{n} \cdot \mathbb{E}[X_1^2] + \frac{n-1}{n} \cdot \mathbb{E}[X_1 X_2] - (\mathbb{E}\hat{\theta})^2$$

$$= \frac{1}{n}\theta + \frac{n-1}{n} \cdot \theta \cdot \frac{N\theta - 1}{N - 1} - \theta^2$$

$$= \frac{\theta(1-\theta)}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Understanding the variability

Putting all of this together,

$$\text{Var}[\hat{\theta}] = \frac{\theta(1-\theta)}{n}\left(1 - \frac{n-1}{N-1}\right).$$

The factor $1 - \frac{n-1}{N-1}$ corrects for sampling without replacement.

For $N = 2{,}317{,}657$, $n = 1000$, and $\hat{\theta} = 0.54$, the standard error is

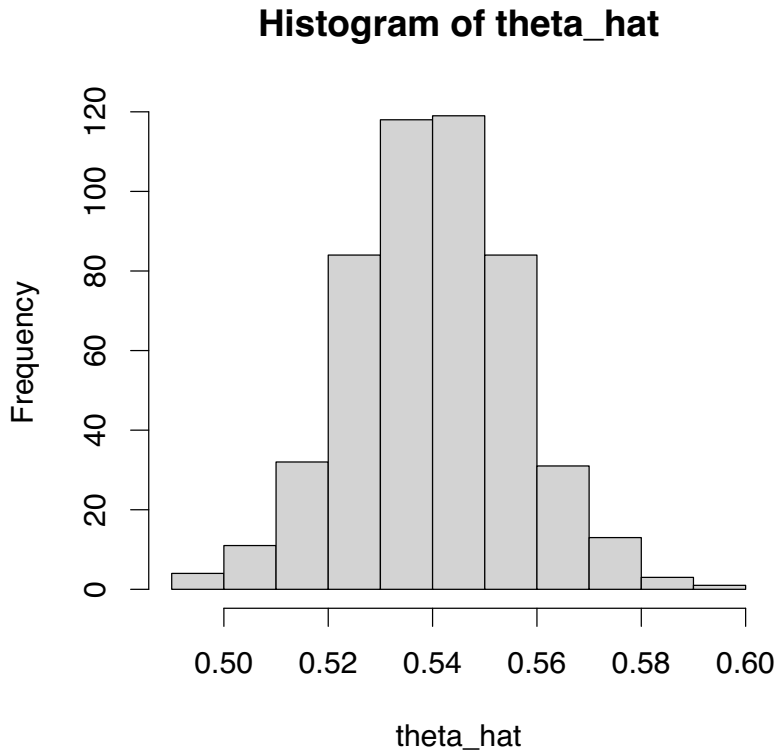
$$\sqrt{\text{Var}[\hat{\theta}]} \approx 0.016$$

It's "not unlikely" for $\theta \approx 0.52$, and "highly unlikely" that $\theta \approx 0.45$.

Understanding the sampling distribution

To make "not unlikely" and "highly unlikely" more precise, let's look at the distribution of $\hat{\theta}$. We can *simulate* $X_1, \ldots, X_n$ from a population of N people, $N\theta$ of whom support Harris (supposing $\theta = 0.54$) and then compute $\hat{\theta}$.

Here's a histogram of the values of $\hat{\theta}$ that we obtain:



Understanding the sampling distribution

The histogram of $\hat{\theta}$ looks like a normal bell curve, with mean 0.54 and standard deviation 0.016. Why?

Again, first suppose that we sampled *with replacement*. Then $X_1, \ldots, X_n$ are independent. By the **Central Limit Theorem**, if n is large, then the distribution of

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\left(\frac{X_1 + \ldots + X_n}{n} - \theta\right) = \frac{(X_1 - \theta) + \ldots + (X_n - \theta)}{\sqrt{n}}$$

is approximately $\mathcal{N}(0, \theta(1-\theta))$.

So the distribution of $\hat{\theta} - \theta$ is approximately $\mathcal{N}(0, \frac{\theta(1-\theta)}{n})$, and the distribution of $\hat{\theta}$ is approximately $\mathcal{N}(\theta, \frac{\theta(1-\theta)}{n})$.

A confidence statement

For sampling *without* replacement, the same normal approximation is still correct, provided that the sample size n is large but much smaller than the population size N.

Recall that roughly 95% of the probability density of a normal distribution is within 2 standard deviations of its mean. Then, applying this normal approximation,

$$(0.54 - 2 \times 0.016,\ 0.54 + 2 \times 0.016) = (0.508, 0.572)$$

is a **95% confidence interval** for θ. In particular, we are more than 95% confident that $\theta > 0.5$.

A confidence statement

"It can be shown that if quantities are distributed according to almost any curve of frequency satisfying simple and common conditions, the average of successive groups of say, 10, 20, 100, ... n of these conform to a normal curve (the more and more closely as n is increased) whose standard deviation diminishes in inverse ratio to the number in each sample.... If we can apply this method—and for clearness I give an example immediately—we are able to give not only a numerical average, but a reasoned estimate for the real physical quantity of which the average is a local or temporary instance."

—A. L. Bowley, 1906.

Some extensions

- Today, usually a combination of targeted and random sampling called **stratified sampling** is used.

  Suppose there are two demographics of people in Connecticut, each of population size $N/2$, who have very different tendencies to support Harris vs. Trump. Using a random sample of 500 individuals from each demographic can yield smaller variability for $\hat{\theta}$.

- Our calculations were based on the strong assumption that we have a simple random sample. If sampling is not uniform across the population, small sampling biases in large samples can yield highly misleading confidence statements.

The randomized view of data

In much of statistical analysis, it is assumed that

Data is a realization of a random process

Why? Possible reasons:

1. We introduced randomness in our experimental design (for example: polling, clinical trials, A/B testing)
2. We are studying an actually random phenomenon (for example: coin tosses or dice rolls)
3. Randomness is a modeling assumption for something we don't understand (for example: errors or "noise" in measurements)

Treating our data as random allows us to perform statistical inference and quantify uncertainty.

Statistical inference

Statistical inference $=$ Probability^{-1}

Probability: For a specified probability distribution, what are the properties of data from this distribution?

Example: $X_1, \ldots, X_{10} \overset{IID}{\sim} \mathcal{N}(2.3, 1)$. What is $\mathbb{P}[X_1 > 5]$? What is the distribution of $\frac{1}{10}(X_1 + \ldots + X_{10})$?

Statistical inference: From a realization of random data, what can we learn about its probability distribution?

Example: $X_1, \ldots, X_{10} \overset{IID}{\sim} \mathcal{N}(\theta, 1)$ for some θ. We observe $X_1 = 3.67$, $X_2 = 2.24$, etc. What is θ?

Inference questions

In this course, we'll focus on the following inferential questions:

- Hypothesis testing: Asking a "yes" or "no" question about the distribution. (Is $\theta > 0.5$?)
- Estimation: Determining the distribution, or some parameter about the distribution. (What is θ?)
- Uncertainty quantification: Understanding the possible errors of our estimates. (What is a range of values to which we're reasonably confident θ belongs?)
- Prediction: Predicting the outcome for a new sample. (Does a new voter support Harris?)

Goals

In statistical inference, there is usually not a single right approach or a single right answer.

- For inferential questions that commonly arise in applications: What are the statistical methods that are often used to answer these questions? *Why* are these the methods of choice?
- In what ways can we compare different methods for answering the same question? In what settings should we prefer one method over another?
- How can we quantify the errors/uncertainties in our answers, and how do they depend on our modeling assumptions?
- For new inferential questions, when there are not existing statistical methods and tools, what are some principles/ideas that can guide us in developing new methods?

Intended audience and pre-requisites

Intended audience:

- Students studying statistics, data science, machine learning, and AI, who want to learn about the foundations and principles of statistical inference.
- Students/researchers in areas where statistics is commonly applied, who want a mathematical one-semester course on statistical methods and ideas.

Pre-requisites:

- Probability theory (S&DS 241/541 or equivalent)
- Multivariable calculus, with some matrix algebra (Math 120 or equivalent)
- Willingness to learn a little bit of computer programming

Differences with other courses at Yale

- We'll provide a more mathematical treatment of statistical methods and inference than S&DS 220/520 and 230/530.
- S&DS 238 covers both statistics and probability in one semester—our course is designed to be taken after a separate course in probability (S&DS 241/541). We'll have less emphasis on Bayesian methods.
- We'll cover a broader set of topics, with less depth, than S&DS 410/610. I'll often emphasize conceptual ideas over mathematical rigor, and provide heuristic explanations instead of formal proofs.

Course website

(seems that this is important!)

www.stat.yale.edu/~zf59/sds242

All course information (syllabus, office hours), lecture notes/slides, and homeworks will be posted here.

Homework solutions, practice exams, and restricted content will be posted to Canvas.

Homework

Approximately weekly, due Wednesdays 1PM on Gradescope. First homework is posted, and due next Wednesday, January 22.

Homework assignments will include computing exercises asking you to perform small simulations, create histograms and plots, and analyze data. Guidance will be provided in the programming language R, although you may choose to use any other language. You will be graded on your results, not on the quality of your code.

Policies

You are allowed a total of 8 late days over the semester, with at most 4 late days for a single assignment. Additional late assignments will incur a 20% penalty per day it is late. Assignments more than 4 days late will not be accepted.

(For homework due Wednesday 1PM, submission before Thursday 1PM is 1 late day, before Friday 1PM is 2 late days, etc.)

You are encouraged to discuss homework problems with your classmates, but you must submit your own individual write-up, **using your own code for the programming exercises.** Use of generative AI tools (e.g. ChatGPT, Claude, Gemini, Llama) is not permitted, unless otherwise noted.

Please indicate at the top of your assignment the number of late days used, and the names of your collaborators.

Exams and grading

There will be two (closed-book) in-class exams:

Midterm: An evening the week of Feb 24–28, date/time TBD
Final: Tuesday May 6, 9AM

Your final grade will be the maximum of the following two weightings:

$$30\% \times \text{homework} + 35\% \times \text{midterm} + 35\% \times \text{final}$$
$$30\% \times \text{homework} + 20\% \times \text{midterm} + 50\% \times \text{final}$$

Piazza

For all questions about course material, lectures, homework, exams, and logistics, Piazza is the fastest and best way to get an answer from me and the teaching staff.

You are encouraged to help answer other students' questions, and to use Piazza to find and form study groups.

Auditing

Auditors are welcome! Please register on Canvas and Piazza to receive our course announcements and gain access to course materials.

Due to demands on our course teaching staff, I ask that auditors please do not submit homeworks to Gradescope or participate in the course exams.

Course schedule

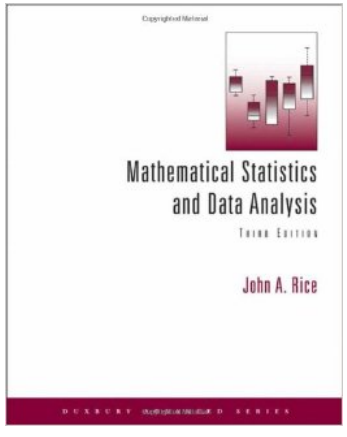
- Unit 0: Introduction and tools. Review of probability theory, limit theorems, introduction to computer simulation.
- Unit 1: Hypothesis testing. Test statistics, p-values, parameteric and nonparametric tests, power, experimental design, multiple testing.
- Unit 2: Parametric models. Method of moments, maximum likelihood estimation, Bayesian inference, confidence intervals, the "bootstrap".
- Unit 3: Predictive inference. Linear regression, logistic regression, classification models, cross-validation and conformal prediction.

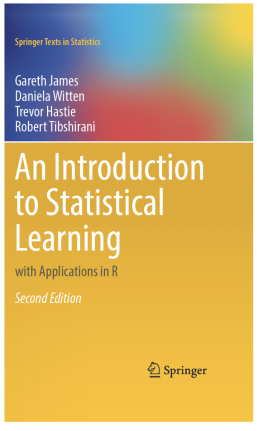
Lecture material and textbooks

Lecture slides will be posted to the course webpage.

Accompanying readings are from:

- John A. Rice, *Mathematical Statistics and Data Analysis*, 3rd edition. (You don't need the CD.)
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning (with Applications in R)*





A thought from Larry Wasserman

"Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid." — L. Wasserman

I hope this course will teach you how to use the band-aid.