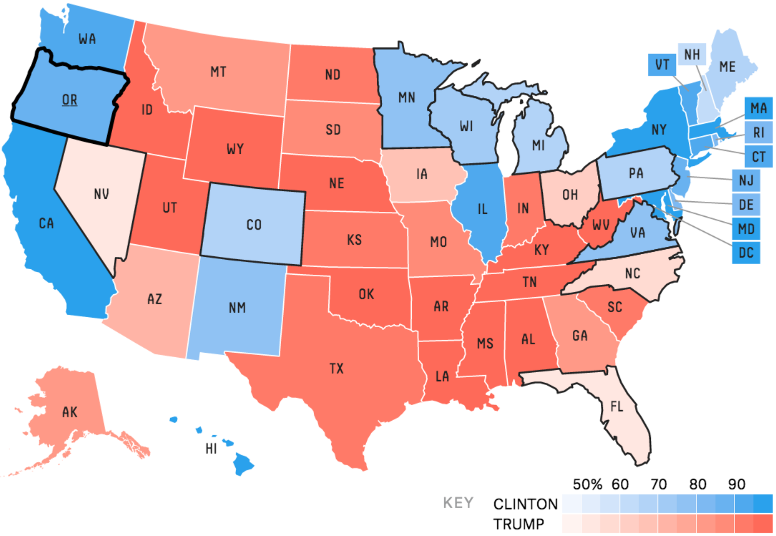
STATS 200: Introduction to Statistical Inference

Lecture 1: Course introduction and polling

U.S. presidential election projections by state



(Source: fivethirtyeight.com, 25 September 2016)

Polling

Let's try to understand how polling can be used to determine the popular support of a candidate in some state (say, Iowa).

Key quantities:

- $N = 3{,}046{,}355$ – population of Iowa
- $p = \frac{\#\text{ people who support Hillary Clinton}}{N}$
- $1 - p = \frac{\#\text{ people who support Donald Trump}}{N}$

We know N but we don't know p.

Polling

Let's try to understand how polling can be used to determine the popular support of a candidate in some state (say, Iowa).

Key quantities:

- $N = 3{,}046{,}355$ – population of Iowa
- $p = \frac{\text{\# people who support Hillary Clinton}}{N}$
- $1 - p = \frac{\text{\# people who support Donald Trump}}{N}$

We know N but we don't know p.

Question #1: What is p?

Polling

Let's try to understand how polling can be used to determine the popular support of a candidate in some state (say, Iowa).

Key quantities:

- $N = 3{,}046{,}355$ – population of Iowa
- $p = \frac{\#\text{ people who support Hillary Clinton}}{N}$
- $1 - p = \frac{\#\text{ people who support Donald Trump}}{N}$

We know N but we don't know p.

Question #1: What is p?
Question #2: Is $p > 0.5$?

Polling

Let's try to understand how polling can be used to determine the popular support of a candidate in some state (say, Iowa).

Key quantities:

- $N = 3{,}046{,}355$ – population of Iowa
- $p = \frac{\#\text{ people who support Hillary Clinton}}{N}$
- $1 - p = \frac{\#\text{ people who support Donald Trump}}{N}$

We know N but we don't know p.

Question $\#1$: What is p?
Question $\#2$: Is $p > 0.5$?
Question $\#3$: Are you sure?

Simple random sample

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

- Person 1 is chosen at random (equally likely) from all N people in Iowa. 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.

Then we can estimate p by

$$\hat{p} = \frac{\#\text{ sampled people who support Hillary Clinton}}{n}$$

Simple random sample

Say 540 out of the 1000 people we surveyed support Hillary, so $\hat{p} = 0.54$.

Does this mean $p = 0.54$? Does this mean $p > 0.5$?

Simple random sample

Say 540 out of the 1000 people we surveyed support Hillary, so $\hat{p} = 0.54$.

Does this mean $p = 0.54$? Does this mean $p > 0.5$?

No! Let's call our **data** $X_1, \ldots, X_n$:

$$X_i = \begin{cases} 1 & \text{if person } i \text{ supports Hillary} \\ 0 & \text{if person } i \text{ supports Donald} \end{cases}$$

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

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

Understanding the bias

$\hat{p}$ is a random variable—it has a probability distribution.

We can ask: What is $\mathbb{E}[\hat{p}]$? What is $\mathrm{Var}[\hat{p}]$? What is the distribution of $\hat{p}$?

Understanding the bias

$\hat{p}$ is a random variable—it has a probability distribution.

We can ask: What is $\mathbb{E}[\hat{p}]$? What is $\mathrm{Var}[\hat{p}]$? What is the distribution of $\hat{p}$?

Each of the N people of Iowa is equally likely to be the i^{th} person that we sampled. So each $X_i \sim \text{Bernoulli}(p)$, and $\mathbb{E}[X_i] = p$.

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

Interpretation: The "average value" of $\hat{p}$ is p.
We say that $\hat{p}$ is **unbiased**.

Understanding the variance

For the variance, recall that for any random variable X,

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

Let's compute $\mathbb{E}[\hat{p}^2]$:

$$\begin{aligned}\mathbb{E}[\hat{p}^2] &= \mathbb{E}\left[\left(\frac{X_1 + \ldots + X_n}{n}\right)^2\right] \\ &= \frac{1}{n^2}\mathbb{E}\left[X_1^2 + \ldots + X_n^2 + 2(X_1X_2 + X_1X_3 + \ldots + X_{n-1}X_n)\right]\end{aligned}$$

Understanding the variance

For the variance, recall that for any random variable X,

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

Let's compute $\mathbb{E}[\hat{p}^2]$:

$$\begin{aligned}
\mathbb{E}[\hat{p}^2] &= \mathbb{E}\left[\left(\frac{X_1 + \ldots + X_n}{n}\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[X_1^2 + \ldots + X_n^2 + 2(X_1X_2 + X_1X_3 + \ldots + X_{n-1}X_n)\right] \\
&= \frac{1}{n^2}\left(n\,\mathbb{E}[X_1^2] + 2\binom{n}{2}\mathbb{E}[X_1X_2]\right) \\
&= \frac{1}{n}\mathbb{E}[X_1^2] + \frac{n-1}{n}\mathbb{E}[X_1X_2]
\end{aligned}$$

Understanding the variance

From the previous slide:

$$\mathbb{E}[\hat{p}^2] = \frac{1}{n}\mathbb{E}[X_1^2] + \frac{n-1}{n}\mathbb{E}[X_1 X_2]$$

Since X_1 is 0 or 1, $X_1 = X_1^2$. Then $\mathbb{E}[X_1^2] = \mathbb{E}[X_1] = p$.

Q: Are X_1 and X_2 independent?

Understanding the variance

From the previous slide:

$$\mathbb{E}[\hat{p}^2] = \frac{1}{n}\mathbb{E}[X_1^2] + \frac{n-1}{n}\mathbb{E}[X_1 X_2]$$

Since X_1 is 0 or 1, $X_1 = X_1^2$. Then $\mathbb{E}[X_1^2] = \mathbb{E}[X_1] = p$.

Q: Are X_1 and X_2 independent?
A: No.

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

We have:

$$\mathbb{P}[X_1 = 1] = p, \quad \mathbb{P}[X_2 = 1 \mid X_1 = 1] = \frac{Np - 1}{N - 1}$$

Understanding the variance

$$\begin{aligned}\mathrm{Var}[\hat{p}] &= \mathbb{E}[\hat{p}^2] - (\mathbb{E}[\hat{p}])^2 \\ &= \frac{1}{n}p + \frac{n-1}{n}p\left(\frac{Np-1}{N-1}\right) - p^2\end{aligned}$$

Understanding the variance

$$
\begin{aligned}
\mathrm{Var}[\hat{p}] &= \mathbb{E}[\hat{p}^2] - (\mathbb{E}[\hat{p}])^2 \\
&= \frac{1}{n}p + \frac{n-1}{n}p\left(\frac{Np-1}{N-1}\right) - p^2 \\
&= \left(\frac{1}{n} - \frac{n-1}{n}\frac{1}{N-1}\right)p + \left(\frac{n-1}{n}\frac{N}{N-1} - 1\right)p^2
\end{aligned}
$$

Understanding the variance

$$
\begin{aligned}
\mathrm{Var}[\hat{p}] &= \mathbb{E}[\hat{p}^2] - (\mathbb{E}[\hat{p}])^2 \\
&= \frac{1}{n}p + \frac{n-1}{n}p\left(\frac{Np-1}{N-1}\right) - p^2 \\
&= \left(\frac{1}{n} - \frac{n-1}{n}\frac{1}{N-1}\right)p + \left(\frac{n-1}{n}\frac{N}{N-1} - 1\right)p^2 \\
&= \frac{N-n}{n(N-1)}p + \frac{n-N}{n(N-1)}p^2
\end{aligned}
$$

Understanding the variance

$$
\begin{aligned}
\mathrm{Var}[\hat{p}] &= \mathbb{E}[\hat{p}^2] - (\mathbb{E}[\hat{p}])^2 \\
&= \frac{1}{n}p + \frac{n-1}{n}p\left(\frac{Np-1}{N-1}\right) - p^2 \\
&= \left(\frac{1}{n} - \frac{n-1}{n}\frac{1}{N-1}\right)p + \left(\frac{n-1}{n}\frac{N}{N-1} - 1\right)p^2 \\
&= \frac{N-n}{n(N-1)}p + \frac{n-N}{n(N-1)}p^2 \\
&= \frac{p(1-p)}{n}\frac{N-n}{N-1} = \frac{p(1-p)}{n}\left(1 - \frac{n-1}{N-1}\right)
\end{aligned}
$$

Understanding the variance

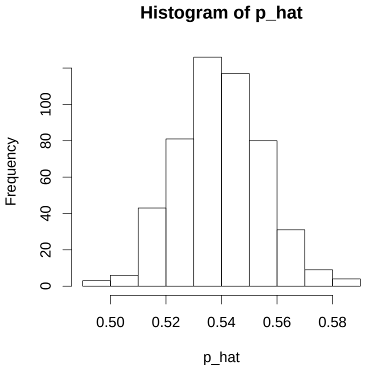
$$\mathrm{Var}[\hat{p}] = \frac{p(1-p)}{n}\left(1 - \frac{n-1}{N-1}\right)$$

When N is much bigger than n, this is approximately $\frac{p(1-p)}{n}$, which would be the variance if we sampled n people in Iowa with replacement. (In that case $\hat{p}$ would be a Binomial(n,p) random variable divided by n.) The factor $1 - \frac{n-1}{N-1}$ is the correction for sampling without replacement.

For $N = 3{,}046{,}355$, $n = 1000$, and $p \approx 0.54$, the standard deviation of $\hat{p}$ is $\sqrt{\mathrm{Var}[\hat{p}]} \approx 0.016$.

Understanding the sampling distribution

Finally, let's look at the distribution of $\hat{p}$. Suppose $p = 0.54$. We can use **simulation** to randomly sample $X_1, \ldots, X_n$ from Np people who support Hillary and $N(1-p)$ people who support Donald, and then compute $\hat{p}$. Doing this 500 times, here's a histogram of the 500 (random) values of $\hat{p}$ that we obtain:



Understanding the sampling distribution

$\hat{p}$ looks like it has a normal distribution, with mean 0.54 and standard deviation 0.016. Why?

Heuristically, if N is much larger than n, then $X_1, \ldots, X_n$ are "almost independent". If n is also reasonably large, then the distribution of

$$\sqrt{n}(\hat{p} - p) = \sqrt{n}\,\frac{(X_1 - p) + \ldots + (X_n - p)}{n}$$

is approximately $\mathcal{N}(0, p(1-p))$ by the Central Limit Theorem.

So $\hat{p}$ is approximately $\mathcal{N}(p, \frac{p(1-p)}{n})$.

A confidence statement

Recall that 95% of the probability density of a normal distribution is within 2 standard deviations of its mean.

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

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

Fundamental principle

We will assume throughout this course:

Data is a realization of a random process.

Fundamental principle

We will assume throughout this course:

Data is a realization of a random process.

Why? Possible reasons:

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

Statistical inference

$$\text{Statistical inference} = \text{Probability}^{-1}$$

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

$$\text{Statistical inference} = \text{Probability}^{-1}$$

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: For a specified set of data, what are properties of the distribution(s)?
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?

Goals

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

- For a given inferential question, what is a good (best?) method of answering that question using data? How do we compare different methods for answering the same question?
- How do we understand the error/uncertainty in our answer?
- How do we understand the dependence of our answer on our modeling assumptions?

Inference tasks

- Hypothesis testing: Asking a binary question about the distribution. (Is $p > 0.5$?)
- Estimation: Determining the distribution, or some characteristic of it. (What is our best guess for p?)
- Confidence intervals: Quantifying the uncertainty of our estimate. (What is a range of values to which we're reasonably sure p belongs?)

Course logistics

Webpage: stats200.stanford.edu

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

Grades and other restricted content will be posted on Stanford Canvas. (There's a link in the above page.)

Prerequisites

- Probability theory (STATS 116 or equivalent)
- Multivariable calculus (MATH 52 or equivalent)

Homework assignments will include simple computing exercises asking you to perform small simulations, create histograms and plots, and analyze data. You may use any language (e.g. R, Python, Matlab) and will be graded only on your results, not on the quality of your code.

Your TA Alex Chin will teach an Introduction to R section, time and place TBD. The first couple homework assignments will also walk you through how to do these things in R.

Requirements

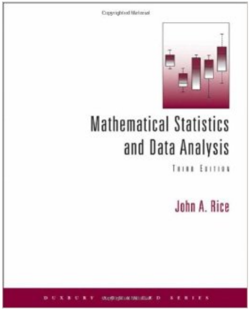
Homework: Due Wednesdays at the start of class. First homework due next Wednesday, October 5.

Collaboration: You can work together on homework, but you must submit your own write-up, **in your own words and using your own code for the programming exercises**. Please indicate at the top of your write-up the names of the students with whom you worked.

Exams: One midterm, one final (both closed-book).

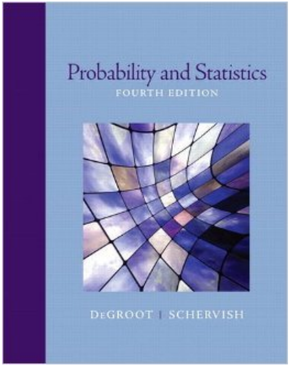
Notes and textbook

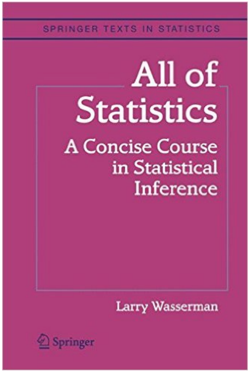
Lectures will switch between slides and blackboard; I'll post slides/notes online after class. Readings are assigned from John A. Rice, *Mathematical Statistics and Data Analysis*:



"Teaching two separate courses, one on theory and one on data analysis, seems to me artificial."—Rice

For reference:





Morris H. DeGroot and Mark J. Schervish, *Probability and Statistics*
Larry Wasserman, *All of Statistics: A concise course in statistical inference*

"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."—Wasserman