Welcome to Day - 1

MATH350 – Statistical Inference

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Dr. Tanujit Chakraborty, Ph.D. from ISI Kolkata.
Associate Professor of Statistics at Sorbonne University.
tanujit.chakraborty@sorbonne.ae
Course page: https://github.com/ctanujit/MATH350
Course for BSc Mathematics and Data Science Students.

Interesting Quotes

"All knowledge is, in final analysis, History.
All sciences are, in the abstract, Mathematics.
All judgements are, in their rationale, Statistics."

- C R Rao, Professor Emeritus at Pennsylvania State University.

"Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time."

- Bradley Efron, American Statistician and Professor at Stanford.

"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."

- Larry Wasserman, Canadian Statistician and Professor at CMU.

Course Prerequisites

Mathematics & Statistics:

- Descriptive Statistics
- Real Analysis
- Multivariate Analysis

Programming Skills:

- Basics of R and RStudio
- Basics of Python
- Basic Statistics using R and Python

Probability Theory:

- Probability Fundamentals
- Probability Distributions
- Sampling Distributions
- Basics of Measure Theory

Course Syllabus

Statistical Inference:

- Point Estimation
- Interval Estimation
- Hypothesis Testing
- Linear Models and ANOVA
- Nonparametric Inference
- Bayesian Inference

Lab Session & Projects:

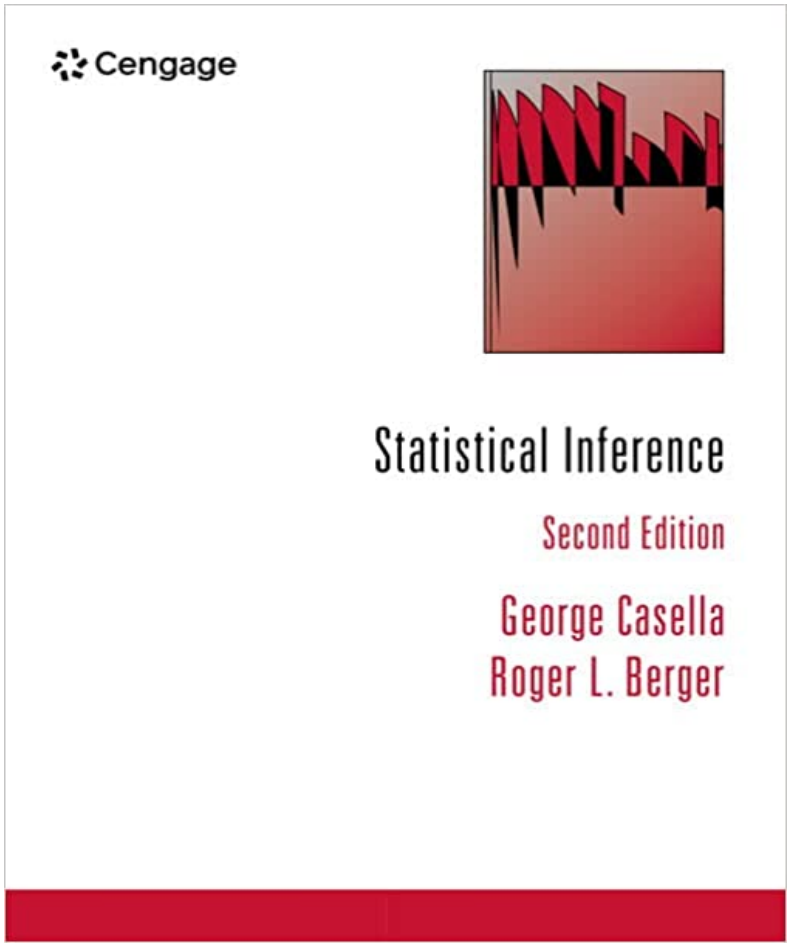
- Implementations in R
- Simulations in R
- MCMC in R
- Course Project using Real-world Data

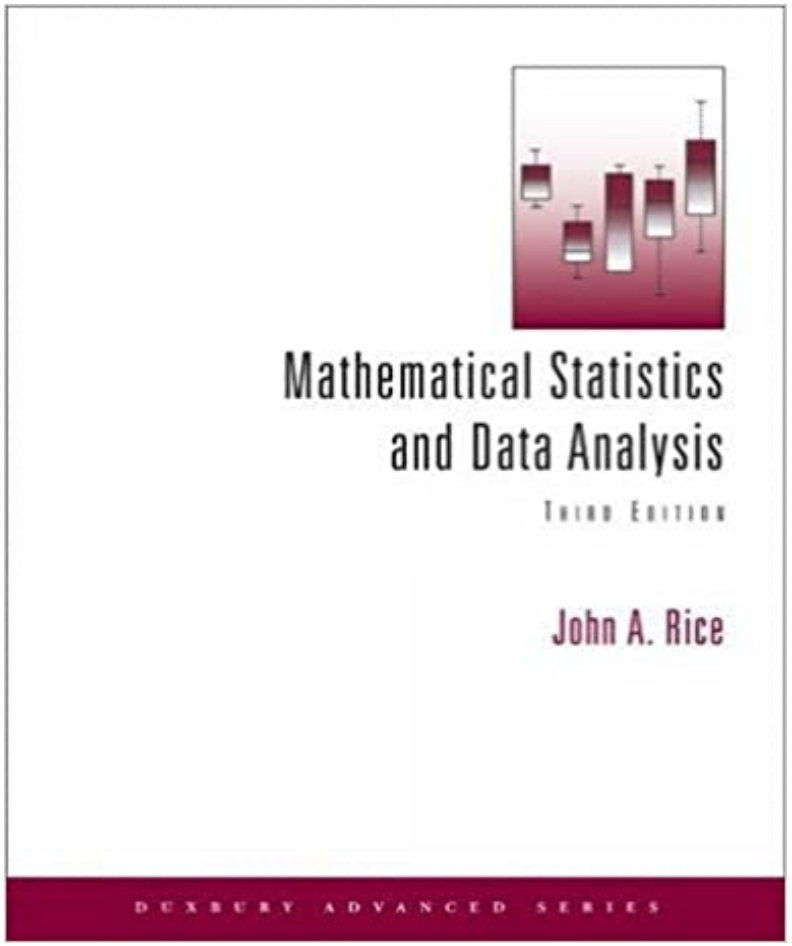
Evaluation:

- Course Project (50%)
- Final Exam (50%)

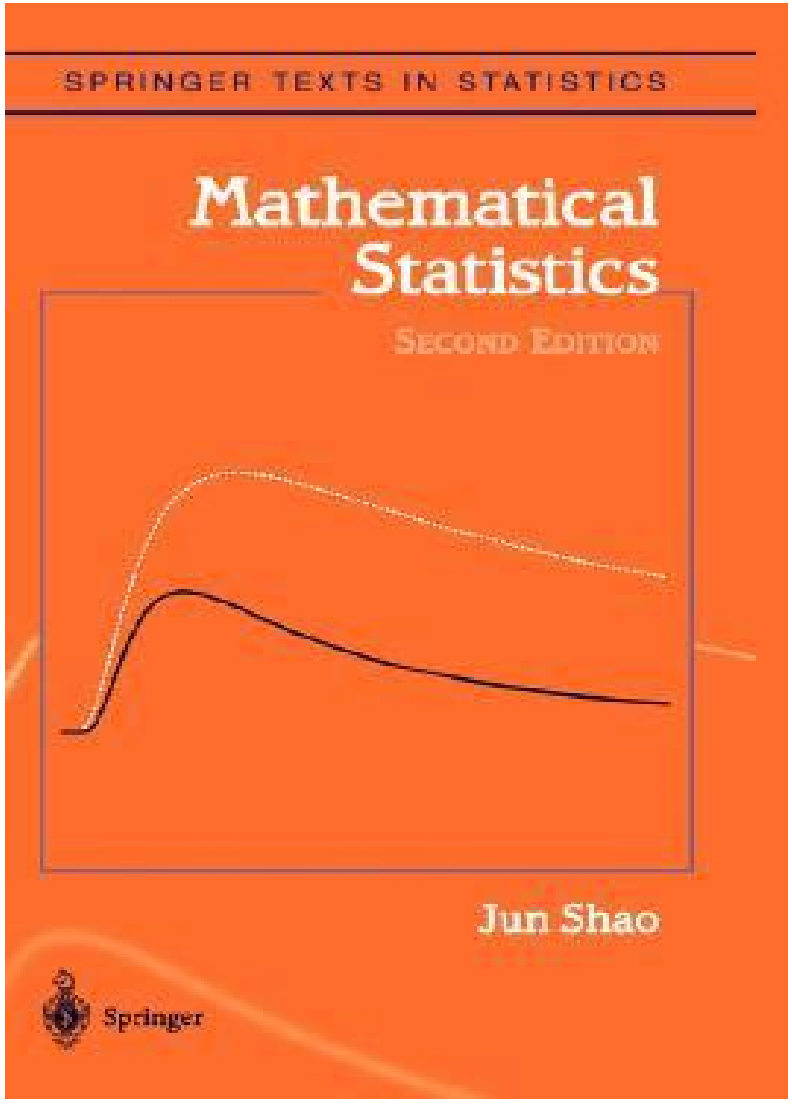
Textbooks

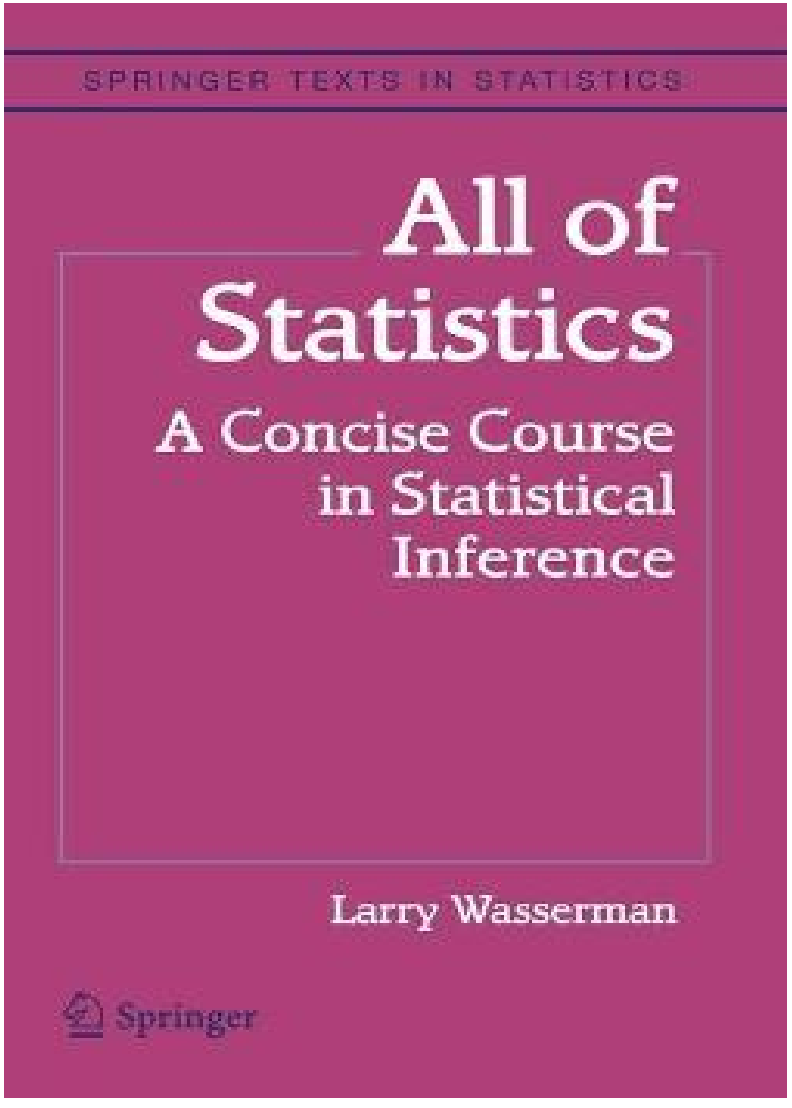
"Teaching two separate courses, one on theory and one on data analysis, seems to me artificial" - John A. Rice.





Important References and Exercises





Historical Background

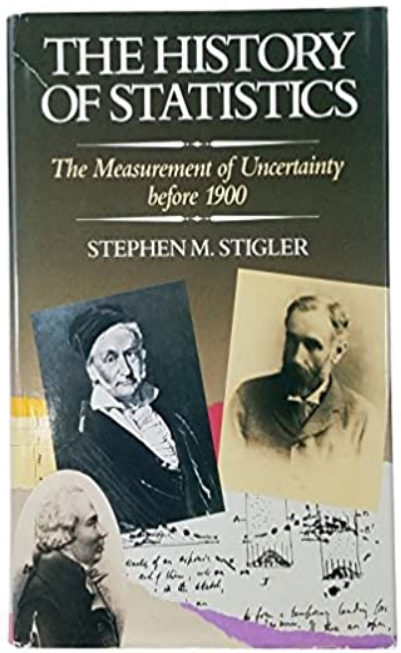
History of Normal Distribution

The normal distribution was first described by Carl Friedrich Gauss. Pierre-Simon Laplace was the first to determine the constant of proportionality $\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)$*, and hence was able to write the full probability density function.*

History of Probability & Statistics

Probability has been a subject of study for a long time (since 17th century). Statistics is a relatively young field. Linear regression due to Francis Galton (Natural Inheritance, 1880s), a cousin of Charles Darwin, become the basic technique of modern statistics. Karl Pearson's "goodness of fit" measure and correlation coefficient appeared in the early 19th century.

The science of Statistics blossomed in the 1920s and 1930s with the works of Ronald A Fisher, Jerzy Neyman, and Karl Pearson. A flurry of research was prompted by World War 2 which generated various difficult applied problems during the 1940s.



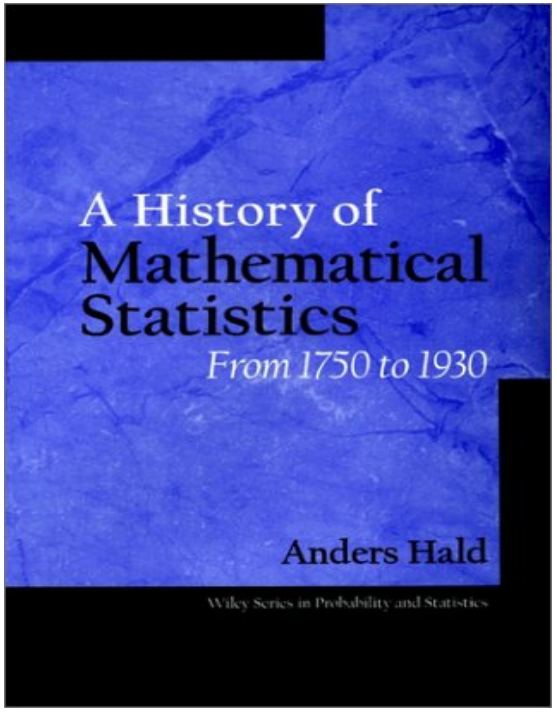
Three Revolutions in Parametric Inference

- The three revolutions in parametric statistical inference are due to Laplace, Gauss, and Fisher:

  - We use $p(.)$ generally to denote a frequency function (continuous or discontinuous) and $p(x|\theta)$ to denote a statistical model defined on a given sample space and parameter space.

  - Let $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$ denote a sample space of n independent observations. From the model we can find the sampling distribution of the statistics $T(\underset{\sim}{x})$ and from $p(T|\theta)$ we can find the probability limits for T for any given value of θ.

  - This is a problem in direct probability (as it was called in the 19^{th} century).



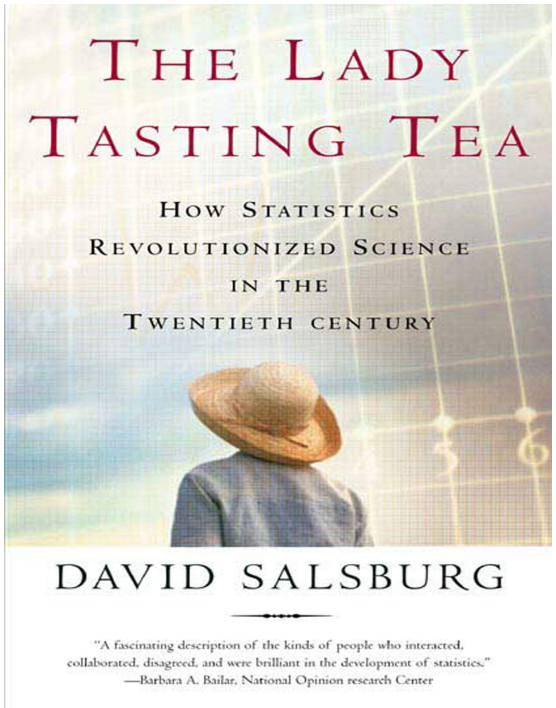
Three Revolutions in Parametric Inference

- In inverse probability, the problem is to find probability limits for θ for a given value of $\underset{\sim}{x}$.

  - Bayes was first to realize that a solution is possible only if θ itself is a random variable with probability density $p(\theta)$ and tried to find conditional distribution $p(\theta|\underset{\sim}{x})$ and $p(\theta|T)$.

  - Laplace also gave a similar theory independently of Bayes.

- Fisher introduced $L_x(\theta)$ (called the likelihood function), defined as proportional to $p(\theta|\underset{\sim}{x})$, to avoid the theory of inverse probability.



$$\underset{\text{Inverse Probability}}{p(\theta|\underset{\sim}{x})} \quad \underset{\text{Laplace}}{\propto} \quad \underset{\text{Direct Probability}}{p(\underset{\sim}{x}|\theta)} \quad \underset{\text{Fisher}}{\propto} \quad \underset{\text{Likelihood}}{L_x(\theta)}$$

Bayesian Approach

- Bayesian methods are much older, dating to the original paper of Thomas Bayes in 1760s. The area generated some interest in the works of Laplace, Gauss, and others in the 19^{th} century.
- However, the Bayesian approach was initially opposed and ignored by statisticians of the 20^{th} century.
- But several prominent non-statisticians, most notably Harold Jeffreys (a Physicist) and Arthur Bowley (an Econometrician) continued to lobby on behalf of the Bayesian ideas. They referred to it as "inverse probability" in their research works.
- The main tool for conducting inference in a Bayesian framework is the posterior distribution:

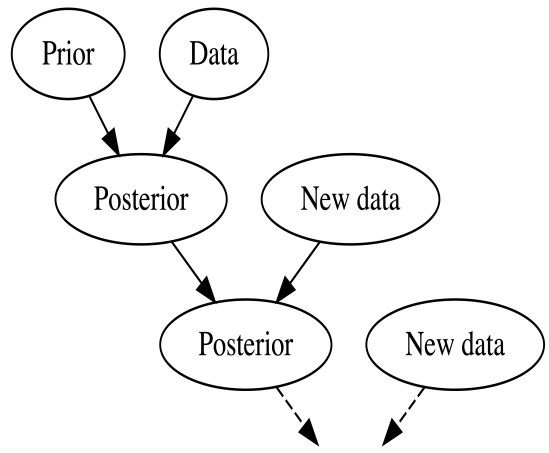
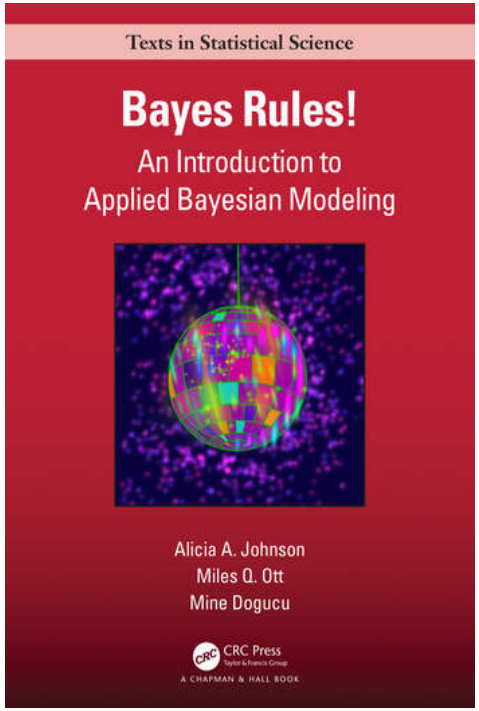


Figure: Bayesian knowledge-building diagram

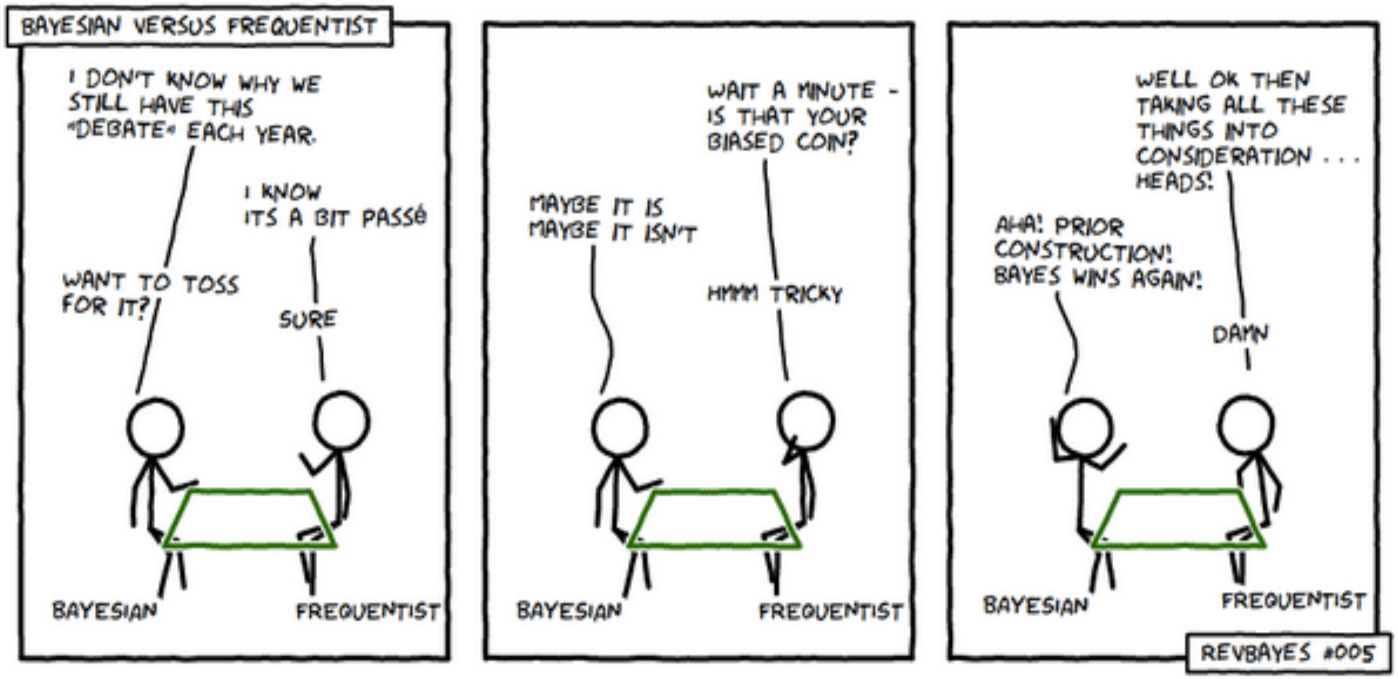
$$\text{Posterior } (\pi(\theta|\underset{\sim}{x})) \propto \text{Likelihood } (f(\underset{\sim}{x}|\theta)) \times \text{Prior } (\pi(\theta))$$

Bayesian Approach

- At the beginning of 1950s, statisticians such as Leonard Savage, Bruno de Finetti, and Dennis Lindley pointed out several deficiencies of the classical approach.
- The biggest impetus to Bayesian statistics came in 1990s after many computational algorithms and the modern computer started to surface. Markov Chain Monte Carlo (MCMC) methods arrived that comprise a class of algorithms for sampling a probability distribution.
- *Example:* Acceptance of paper in "Annals of Statistics".
  What is the probability of acceptance of a paper in Annals of Statistics with prior information that your two papers submitted to Annals got accepted?

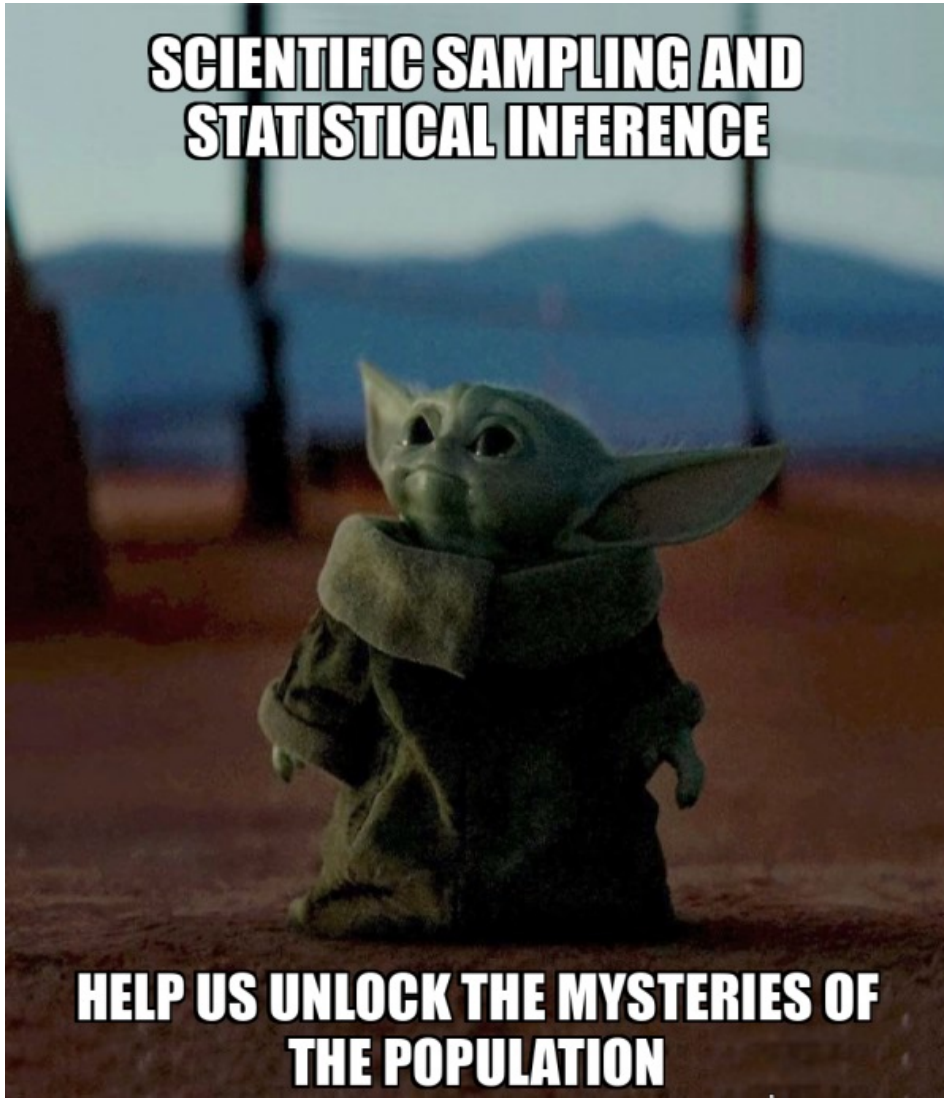


We all are Bayesian?

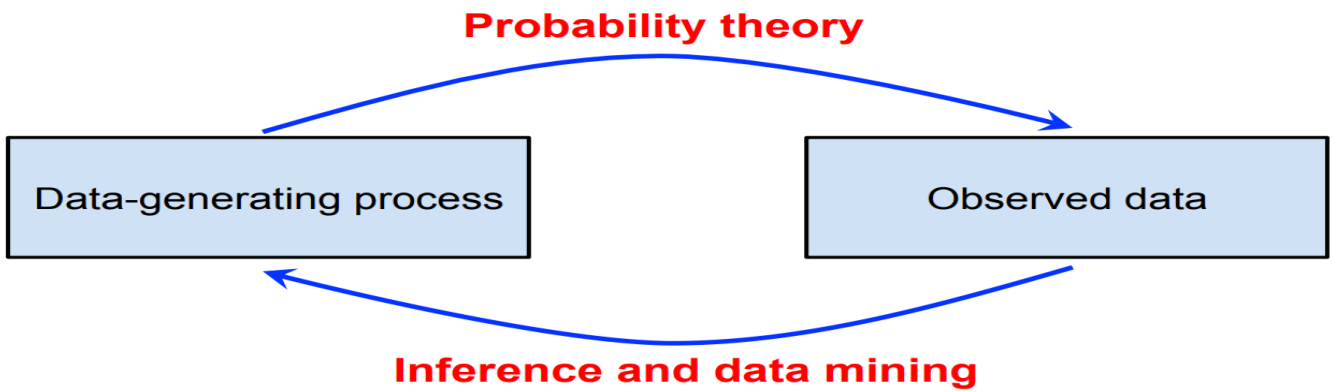


Introduction : Statistical Inference

Why Statistical Inference?



Probability & Inference



Probability theory

is a formalism to work with uncertainty. Given a data-generating process, what are properties of outcomes?

Statistical inference

deals with the inverse problem. Given outcomes, what can we say on the data-generating process?

Introduction

- The aim of Statistical Inference is to learn about the population using the observed data.

  This involves:

- computing something with the data
  - a statistic: function of data
- interpret the result
  - in probabilistic terms: sampling distribution of statistic
- Statistical inference $=$ Probability^{-1}

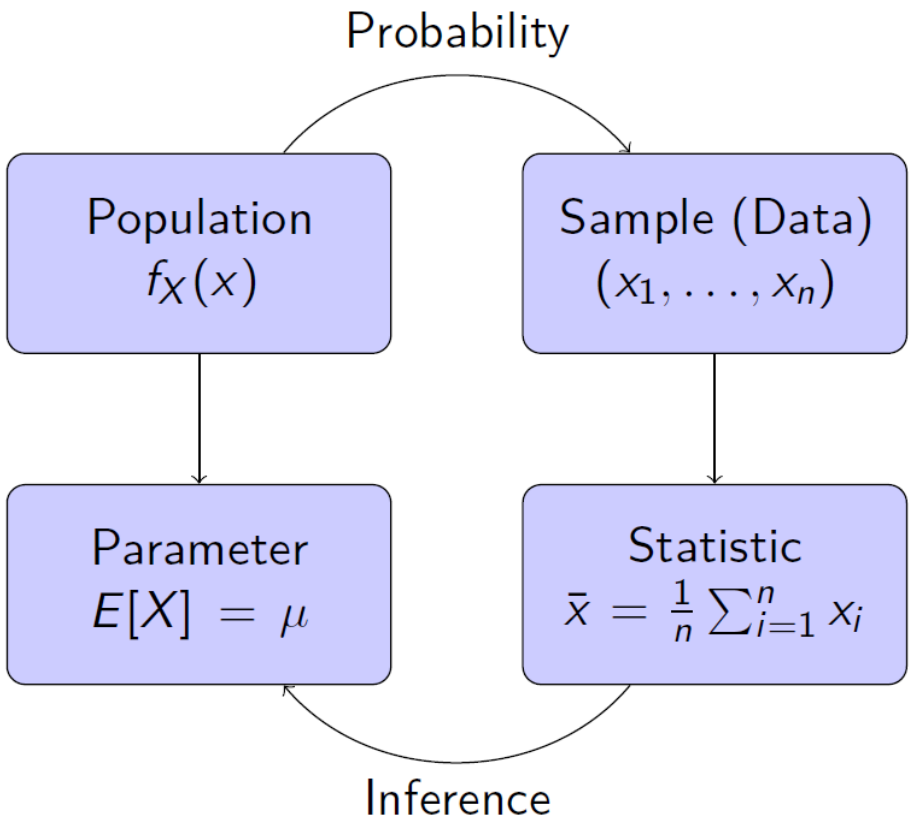
Introduction

- Statistical inference $=$ Probability^{-1}
- Probability: For a specified probability distribution, what are the properties of data from this distribution?
  - Example: Let $X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} \mathcal{N}(2.3, 1)$. 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: $X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$ for some θ. We observe $X_1 = 3.67, X_2 = 2.24$, etc. What is θ ?

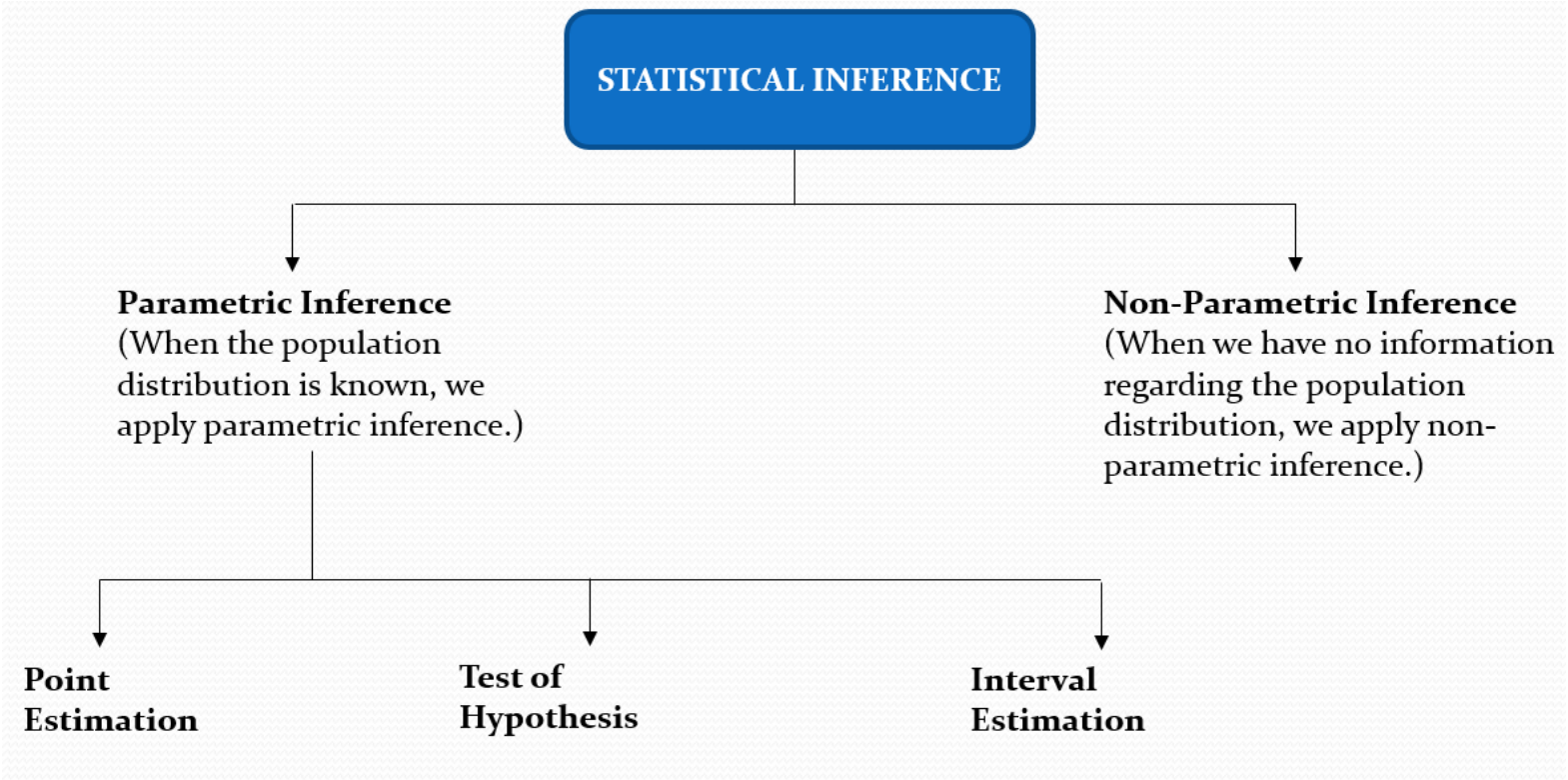
Introduction

$f(x|\theta) \Rightarrow x$: effect, θ : cause $\Rightarrow$ Probabilistic Thinking

$l(\theta|x) \Rightarrow$ Inversion Process $\Rightarrow$ Inferential Thinking



Taxonomy of Statistical Inference



Statistical Inference

Statistical Inference is based on probabilistic modeling of the observed phenomenon. There are two popular approaches:

- A first approach assumes that statistical inference must incorporate as much as possible of the phenomenon's complexity. Thus, it aims at estimating the distributions underlying the phenomenon under minimal assumptions, generally using functional estimation (density, regression, etc.). This is known as the NONPARAMETRIC approach.
- Conversely, the parametric approach represents the distribution through a density function $f(\underset{\sim}{x}|\theta)$ where only the parameter θ (of finite dimension) is unknown. This is called the PARAMETRIC approach. that parameter.

Point estimation

- We want to estimate a population parameter using the observed data, i.e., some measure of variation, an average, min, max, quantile, etc.
- Point estimation attempts to obtain a best guess for the value of that parameter.
- An estimator is a statistic (function of data) that produces such a guess.
- We usually mean by "best" an estimator whose sampling distribution is more concentrated about the population parameter value compared to other estimators.
- Hence, the choice of a specific statistic as an estimator depends on the probabilistic characteristics of this statistic in the context of the sampling distribution.

Confidence Interval

- We can also quantify the uncertainty (sampling distribution) of our point estimate.
- One way of doing this is by constructing an interval that is likely to contain the population parameter.
- One such an interval, which is computed on the basis of the data, is called a confidence interval.
- The sampling probability that the confidence interval will indeed contain the parameter value is called the confidence level.
- We construct confidence intervals for a given confidence level.

Hypothesis Testing

- The scientific paradigm involves the proposal of new theories that presumably provide a better description of the laws of Nature.
- If the empirical evidence is inconsistent with the predictions of the old theory but not with those of the new theory
  - then the old theory is rejected in favor of the new one.
  - otherwise, the old theory maintains its status.
- Statistical hypothesis testing is a formal method for determining which of the two hypothesis should prevail that uses this paradigm.

Statistical Hypothesis Testing

- Each of the two hypothesis, the old and the new, predicts a different distribution for the empirical measurements.
- In order to decide which of the distributions is more in tune with the data a statistic is computed.
- This statistic t is called the test statistic.
- A threshold c is set and the old theory is rejected if $t > c$.
- Hypothesis testing consists in asking a binary question about the sampling distribution of t.

Statistical Hypothesis Testing

- This decision rule is not error proof, since the test statistic may fall by chance on the wrong side of the threshold.
- Suppose we know the sampling distribution of the test statistic t.
- We can then set the probability of making an error to a given level by setting c.
- The probability of erroneously rejecting the currently accepted theory (the old one) is called the significance level of the test.
- The threshold is selected in order to assure a small enough significance level.

Multiple Measurements

- The method of testing hypothesis is also applied in other practical settings where it is required to make decisions.
- Consider a random trial of a new treatment to a medical condition where the
  - treated get the new treatment.
  - controls get the old treatment

  and measure their response.
- We now have 2 measurements that we can compare. We will use statistical inference to make a decision about whether the new treatment is better.

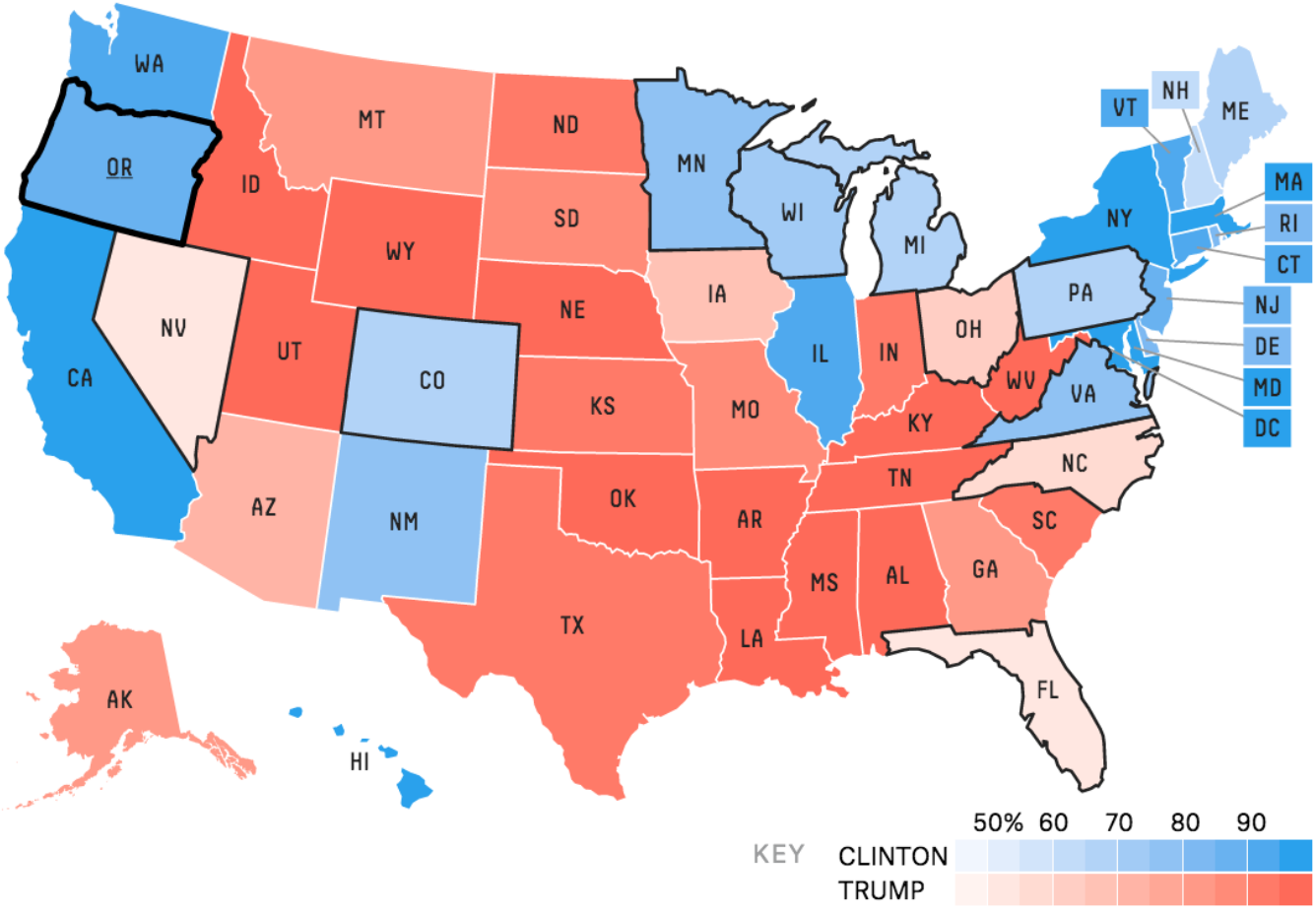
Statistic and Sampling Distributions

- Statistical inferences, be it point estimation, confidence intervals, or hypothesis tests, are based on statistics computed from the data.
- A statistic can be thought as a statistical summary of the data. Examples: the sample average and the sample standard deviation. where the
  - For a given dataset a statistic has a single numerical value.
  - The statistic is therefore a random variable.
- The distribution of the values we get when computing a statistic in (infinitely) many random samples is called the sample distribution of that statistic. The sampling distribution of a statistic gives the nature of this variability.

Motivating Example : Polling

U.S. Presidential Election (2016)

U.S. Presidential Election (2016)



(Source: fivethirtyeight.com, 25 September 2016)

Polling Problem (Ref: Stanford STAT-200)

- 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 (Population of Iowa) = 3.05 million
  - $p = \frac{\text{No. of people who support Hillary Clinton}}{N}$
  - $1 - p = \frac{\text{No. of people who support Donald Trump}}{N}$
- 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 $\hat{p} = \frac{\text{No. of sampled people who support Hillary Clinton}}{n}$.

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 $p > 0.5$?
- 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 $\hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n}$.
- 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 $\hat{p}$?
- Each of the N people of Iowa is equally likely to be the i^{th} person that we sampled. So each $X_i \sim \mathrm{Bernoulli}(p)$, and

$$\mathbb{E}[X_i] = p \times 1 + (1-p) \times 0 = p.$$

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

- Interpretation: The "average value" of $\hat{p}$ is p. We say that $\hat{p}$ is unbiased.
- Unbiasedness refers to the average error over repeated sampling, and not the error for the observed data!

Understanding the Variance

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

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

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

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

Understanding the Variance

From the previous slide:

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

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

Question: Are X_1 and X_2 independent?

Answer: No.

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

We have:

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

Understanding the Variance

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

Understanding the Variance

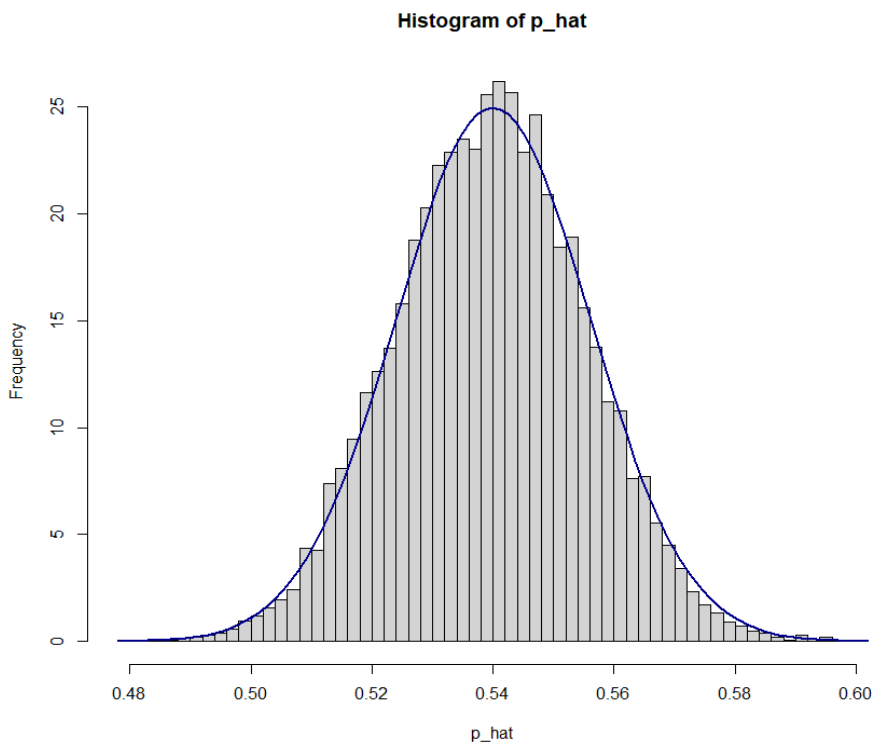
$$\mathrm{Var}[\hat{p}] = \frac{p(1-p)}{n}\left(1 - \frac{n-1}{N-1}\right)$$

- 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 $\mathrm{Binomial}(n, p)$ 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 $\sqrt{\mathrm{Var}[\hat{p}]} \approx 0.016$.

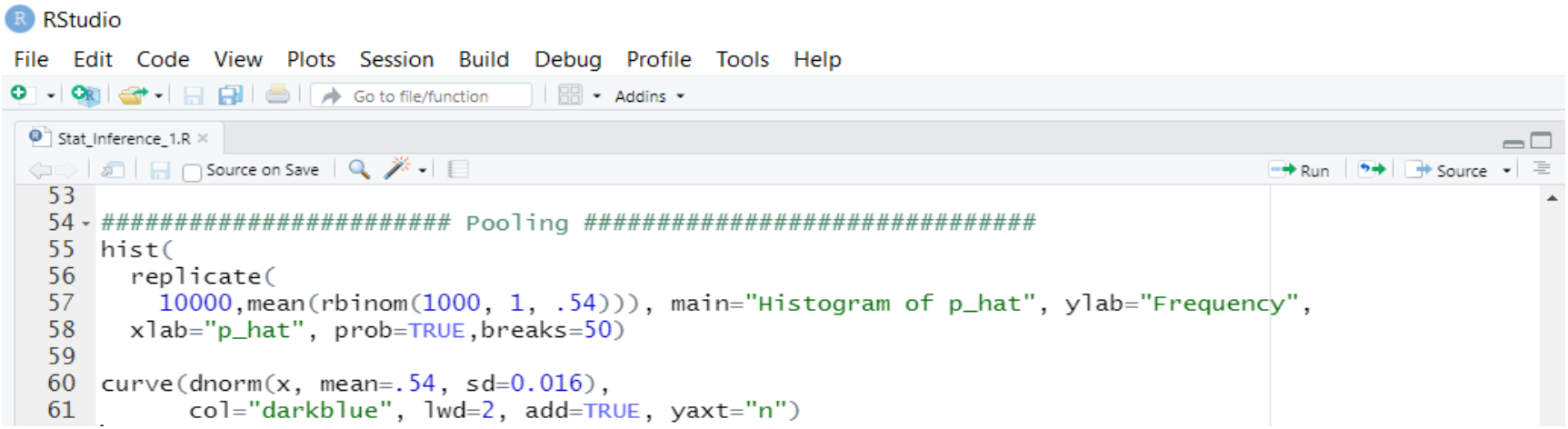
Understanding the Sampling Distribution

- 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:



R Code: Pooling Example



Code: https://github.com/tanujit123/MATH350

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 $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

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 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)

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?)

In the coming weeks, we will take a closer look at how this randomness affects what we can learn about the population from the data. The knowledge of Statistical Inference will make you a better Data Scientist.

Perspective | Open Access | Published: 10 November 2022

Statistical inference links data and theory in network science

Leto Peel, Tiago P. Peixoto & Manlio De Domenico

Nature Communications **13**, Article number: 6794 (2022) | Cite this article

38 Altmetric | Metrics

Abstract

The number of network science applications across many different fields has been rapidly increasing. Surprisingly, the development of theory and domain-specific applications often occur in isolation, risking an effective disconnect between theoretical and methodological advances and the way network science is employed in practice. Here we address this risk constructively, discussing good practices to guarantee more successful applications and reproducible results. We endorse designing statistically grounded methodologies to address challenges in network science. This approach allows one to explain observational data in terms of generative models, naturally deal with intrinsic uncertainties, and strengthen the link between theory and applications.

Figure: Read Online

Local

Women flocking to statistics, the newly hot, high-tech field of data science

A 11

Erin Blankeship, left, statistics professor at University of Nebraska- Lincoln, and Aimee Schwab, graduate teaching assistant and PhD student in statistics, in a classroom at Hardin Hall. Statistics is leading all other STEM fields in in attracting, retaining and promoting women. (Jake Crandall/For The Washington Post)

Figure: Read Online

Harvard Business Review

ARTWORK: TAHAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCR ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT

Big Data: The Management Revolution

5 Essential Principles for Understanding Analytics

Data Scientists Don't Scale

Figure: Read Online