COMP 680
Statistics for Computing and Data Science
Week 3: Statistical Inference Overview

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

1. Fundamental Concepts
2. Asymptotic Theory
3. Code Demo

Probability and Statistics

- Probability:
  - given a data generating process, what are the properties of the outcomes?
  - formal language of uncertainty
  - theoretical foundation of statistical inference

Probability and Statistics

- Probability:
  - given a data generating process, what are the properties of the outcomes?
  - formal language of uncertainty
  - theoretical foundation of statistical inference
- Statistics:
  - given the outcomes, what can we say about the process that generated the data?
  - "reverse process" of probability

Probability and Statistics

- Probability:
  - given a data generating process, what are the properties of the outcomes?
  - formal language of uncertainty
  - theoretical foundation of statistical inference
- Statistics:
  - given the outcomes, what can we say about the process that generated the data?
  - "reverse process" of probability
- Data mining and machine learning are close cousins of Statistics.

Population and Samples

- Population distribution: a census.
  - if the population distribution is known, no statistical inference needed!

Population and Samples

- Population distribution: a census.
  - if the population distribution is known, no statistical inference needed!
- Sample: data generated from the population distribution
  - most of the time, assume i.i.d sample

Population and Samples

- Population distribution: a census.
  - if the population distribution is known, no statistical inference needed!
- Sample: data generated from the population distribution
  - most of the time, assume i.i.d sample
- Statistical inference: make conclusion of population based on a random sample (the data you get to observe)

Inference

- Statistical inference:
  - given a sample $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} F$, how do we infer the data generating mechanism, i.e. the population distribution cdf $F_X(\cdot)$
  - may want to infer only some feature of F_X such as its mean

Inference

- Statistical inference:
  - given a sample $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} F$, how do we infer the data generating mechanism, i.e. the population distribution cdf $F_X(\cdot)$
  - may want to infer only some feature of F_X such as its mean
- Parametric models:
  - if F can be parameterized by a finite number of parameters
  - this often means we know the distribution family of F (or willing to make the assumption)

Inference

- Statistical inference:
  - given a sample $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} F$, how do we infer the data generating mechanism, i.e. the population distribution cdf $F_X(\cdot)$
  - may want to infer only some feature of F_X such as its mean
- Parametric models:
  - if F can be parameterized by a finite number of parameters
  - this often means we know the distribution family of F (or willing to make the assumption)
- Nonparametric models:
  - infinite number of parameters!
  - for example F can be any cdf that is continuous

Statistics and Sampling Distribution

- What is a statistic?
  - a function of data, a quantity that depends on data
  - examples: sample mean, median, variance, quantile, max and min...

Statistics and Sampling Distribution

- What is a statistic?
  - a function of data, a quantity that depends on data
  - examples: sample mean, median, variance, quantile, max and min...
- The sampling distribution of a statistic:
  - different values and associated probabilities of the statistic
  - based on ALL possible random samples from the population

Statistics and Sampling Distribution

- What is a statistic?
  - a function of data, a quantity that depends on data
  - examples: sample mean, median, variance, quantile, max and min...
- The sampling distribution of a statistic:
  - different values and associated probabilities of the statistic
  - based on ALL possible random samples from the population
- The sampling distribution quantify the uncertainty:
  - why is there uncertainty?
  - what if you get another random sample?

Outline

1. Fundamental Concepts
2. Asymptotic Theory
3. Code Demo

Theoretical Justification

- Asymptotic theory, or large sample theory, or limit theory
  - what happens if we have more and more data: $n \to \infty$

Theoretical Justification

- Asymptotic theory, or large sample theory, or limit theory
  - what happens if we have more and more data: $n \to \infty$
- Intuition: "more" data $\to$ better inference
  - i.e. we get closer to the "truth" – the population distribution
  - "more" in both quantity and quality

Theoretical Justification

- Asymptotic theory, or large sample theory, or limit theory
  - what happens if we have more and more data: $n \to \infty$
- Intuition: "more" data $\to$ better inference
  - i.e. we get closer to the "truth" – the population distribution
  - "more" in both quantity and quality
- Major theorems:
  - Law of Large Numbers
  - Central Limit Theorem

Convergence of Random Variables

- A sequence of random variables $X_1, X_2, \cdots$ converges to X
  - converge in probability $X_n \xrightarrow{\mathbb{P}} X$ if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ for any $\epsilon > 0$

Convergence of Random Variables

- A sequence of random variables $X_1, X_2, \cdots$ converges to X
  - converge in probability $X_n \xrightarrow{\mathbb{P}} X$ if $\mathbb{P}(|X_n - X| > \epsilon) \rightarrow 0$ for any $\epsilon > 0$
  - converge in distribution $X_n \xrightarrow{\mathbb{D}} X$ if $F_{X_n}(t) \rightarrow F_X(t)$ at all t for which $F_X(\cdot)$ is continuous

Convergence of Random Variables

- A sequence of random variables $X_1, X_2, \cdots$ converges to X
  - converge in probability $X_n \xrightarrow{\mathbb{P}} X$ if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ for any $\epsilon > 0$
  - converge in distribution $X_n \xrightarrow{\mathbb{D}} X$ if $F_{X_n}(t) \to F_X(t)$ at all t for which $F_X(\cdot)$ is continuous
  - converge in quadratic mean $X_n \xrightarrow{qm} X$ if $\mathbb{E}[(X_n - X)^2] \to 0$

Convergence of Random Variables

- A sequence of random variables $X_1, X_2, \cdots$ converges to X
  - converge in probability $X_n \xrightarrow{\mathbb{P}} X$ if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ for any $\epsilon > 0$
  - converge in distribution $X_n \xrightarrow{\mathbb{D}} X$ if $F_{X_n}(t) \to F_X(t)$ at all t for which $F_X(\cdot)$ is continuous
  - converge in quadratic mean $X_n \xrightarrow{qm} X$ if $\mathbb{E}[(X_n - X)^2] \to 0$

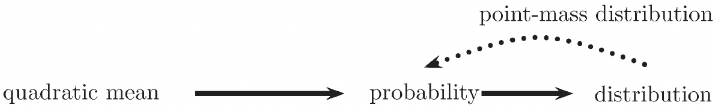

FIGURE 5.2. Relationship between types of convergence.

Law of Large Numbers

- **IF** $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(\cdot)$, then the sample mean converges to population mean in probability, i.e. $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.

Law of Large Numbers

- **IF** $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(\cdot)$, then the sample mean converges to population mean in probability, i.e. $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.
- Mild assumption about $f_X(\cdot)$ required:

Law of Large Numbers

- **IF** $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(\cdot)$, then the sample mean converges to population mean in probability, i.e. $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.
- Mild assumption about $f_X(\cdot)$ required:
- Proof: one line using Chebyshev's inequality:

Central Limit Theorem

- **IF** $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(\cdot)$ with mean μ and variance σ^2, then the sample mean converges in distribution to a normal distribution:

$$\bar{X}_n \xrightarrow{\mathbb{D}} N(\mu, \sigma^2/n)$$

Central Limit Theorem

- **IF** $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(\cdot)$ with mean μ and variance σ^2, then the sample mean converges in distribution to a normal distribution:

$$\bar{X}_n \xrightarrow{\mathbb{D}} N(\mu, \sigma^2/n)$$

- equivalently:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathbb{D}} Z \sim N(0, 1)$$

Central Limit Theorem

- known as the "asymptotic normality of sample mean":
  - **regardless** of the population distribution
  - even discrete population: Bernoulli, Poisson...

Central Limit Theorem

- known as the "asymptotic normality of sample mean ":
  - **regardless** of the population distribution
  - even discrete population: Bernoulli, Poisson...
- "asymptotic" means for any finite sample, it is an approximation:
  - for example, Bernoulli $0 \leq \bar{X}_n \leq 1$, can not be exact normal!

Central Limit Theorem

- known as the "asymptotic normality of sample mean":
  - **regardless** of the population distribution
  - even discrete population: Bernoulli, Poisson...
- "asymptotic" means for any finite sample, it is an approximation:
  - for example, Bernoulli $0 \leq \bar{X}_n \leq 1$, can not be exact normal!
- CLT still holds if replace σ^2 with sample variance s^2

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s} \xrightarrow{\mathbb{D}} Z \sim N(0,1)$$

Central Limit Theorem

- known as the "asymptotic normality of sample mean ":
  - **regardless** of the population distribution
  - even discrete population: Bernoulli, Poisson...
- "asymptotic" means for any finite sample, it is an approximation:
  - for example, Bernoulli $0 \leq \bar{X}_n \leq 1$, can not be exact normal!
- CLT still holds if replace σ^2 with sample variance s^2

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s} \xrightarrow{\mathbb{D}} Z \sim N(0, 1)$$

- only for sample mean, not any other statistics!
  - empirical demonstration of the CLT using simulation

Central Limit Theorem

- Special case when population is Normal, then normality holds exactly:
  - if $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N_X(x|\mu, \sigma^2)$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = Z$$

Central Limit Theorem

- Special case when population is Normal, then normality holds exactly:
  - if $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N_X(x|\mu, \sigma^2)$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = Z$$

- Replace σ^2 with s^2, you get the Student's-t distribution:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s} = t_{df=n-1}$$

Examples of Application

- If you toss a fair coin 100 times, what is the probability of getting 70 heads or more?

Examples of Application

- If you toss a fair coin 100 times, what is the probability of getting 70 heads or more?
    - remember we have some upper bounds using Markov and Chebyshev's inequality

Examples of Application

- If you toss a fair coin 100 times, what is the probability of getting 70 heads or more?
  - remember we have some upper bounds using Markov and Chebyshev's inequality
- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p = 0.5)$
  - $\mu = 0.5$
  - $\sigma^2 = p(1-p) = 0.25$

Examples of Application

- If you toss a fair coin 100 times, what is the probability of getting 70 heads or more?
  - remember we have some upper bounds using Markov and Chebyshev's inequality
- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p = 0.5)$
  - $\mu = 0.5$
  - $\sigma^2 = p(1-p) = 0.25$

$$
\begin{aligned}
&\mathbb{P}(\sum_{i=1}^{100} X_i \geq 70) = \mathbb{P}(\bar{X}_{100} \geq 0.7) \\
=&\mathbb{P}(\frac{\sqrt{100}(\bar{X}_{100} - 0.5)}{\sqrt{0.25}} \geq \frac{\sqrt{100}(0.7 - 0.5)}{\sqrt{0.25}}) \\
=&\mathbb{P}(Z \geq 4) \approx 0.003\%
\end{aligned}
$$

Examples of Application

- You would like to know the percentage p in general population that support certain legislation. You start a poll online to randomly survey 100 people for a Yes/No question and you estimate the proportion of Yes in the survey response.

Examples of Application

- You would like to know the percentage p in general population that support certain legislation. You start a poll online to randomly survey 100 people for a Yes/No question and you estimate the proportion of Yes in the survey response.
- How accurate is this estimate?
  - $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$
  - estimate p by $\bar{X}_n = \sum_{i=1}^{n} X_i / n$
  - apply CLT:

Outline