COMP 680
Statistics for Computing and Data Science
Week 5: Nonparametric Inference

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

Statistics vs. A Statistic

- Statistics as a subject

Statistics vs. A Statistic

- Statistics as a subject
- A statistic as a function of data, a quantity that depends on data

Statistics vs. A Statistic

- Statistics as a subject
- A statistic as a function of data, a quantity that depends on data
- Examples:
  - sample mean, median, variance, quantile, max and min...

Sampling Distribution

- The sampling distribution of a statistic:
  - different values and associated probabilities of the statistic
  - based on ALL possible random samples from the population

Sampling Distribution

- The sampling distribution of a statistic:
  - different values and associated probabilities of the statistic
  - based on ALL possible random samples from the population
- Why do we care?

Sampling Distribution

- The sampling distribution of a statistic:
  - different values and associated probabilities of the statistic
  - based on ALL possible random samples from the population
- Why do we care?
  - a statistic as an estimate of population parameter
  - sampling distribution can quantify the uncertainty

Sampling Distribution

- CLT: sampling distribution of sample mean $\bar{X}_n$ for i.i.d. samples

Sampling Distribution

- CLT: sampling distribution of sample mean $\bar{X}_n$ for i.i.d. samples
  - if population is $\mathrm{N}(\mu, \sigma^2)$:
  - $\bar{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$
  - $\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$
  - $\frac{(\bar{X}_n - \mu)}{s/\sqrt{n}} \sim \mathrm{t}_{n-1}$ where s is sample standard deviation

Sampling Distribution

- CLT: sampling distribution of sample mean $\bar{X}_n$ for i.i.d. samples
  - if population is $\mathrm{N}(\mu, \sigma^2)$:
  - $\bar{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$
  - $\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$
  - $\frac{(\bar{X}_n - \mu)}{s/\sqrt{n}} \sim \mathrm{t}_{n-1}$ where s is sample standard deviation
  - for any other population distribution with mean μ and variance σ^2:
  - $\frac{(\bar{X}_n - \mu)}{s/\sqrt{n}} \xrightarrow{\mathbb{D}} \mathrm{N}(0, 1)$ as $n \to \infty$

Sampling Distribution

- CLT: sampling distribution of sample mean $\bar{X}_n$ for i.i.d. samples
  - if population is $\mathrm{N}(\mu, \sigma^2)$:
  - $\bar{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$
  - $\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$
  - $\frac{(\bar{X}_n - \mu)}{s/\sqrt{n}} \sim \mathrm{t}_{n-1}$ where s is sample standard deviation
  - for any other population distribution with mean μ and variance σ^2:
  - $\frac{(\bar{X}_n - \mu)}{s/\sqrt{n}} \xrightarrow{\mathbb{D}} \mathrm{N}(0,1)$ as $n \to \infty$
- the more general result implies: $\mathrm{t}_{n-1} \xrightarrow{\mathbb{D}} \mathrm{N}(0,1)$ as $n \to \infty$

Sampling Distribution

- What about other statistics?

Sampling Distribution

- What about other statistics?
- Visualize a sampling distribution by simulation:

Sampling Distribution

- What about other statistics?
- Visualize a sampling distribution by simulation:
  1. define a population
  2. fix a sample size
  3. draw a random sample from the population
  4. calculate the specific statistic and save the value
  5. repeat step 3 - 4, many many...times

Sampling Distribution

- What about other statistics?
- Visualize a sampling distribution by simulation:
  1. define a population
  2. fix a sample size
  3. draw a random sample from the population
  4. calculate the specific statistic and save the value
  5. repeat step 3 - 4, many many...times
- Practical questions:

Sampling Distribution

- What about other statistics?
- Visualize a sampling distribution by simulation:
  1. define a population
  2. fix a sample size
  3. draw a random sample from the population
  4. calculate the specific statistic and save the value
  5. repeat step 3 - 4, many many...times
- Practical questions:
  - sample with or without replacement?
  - where is the probability?
  - sample how many times??

In Practice

- Do not know what the population distribution is

In Practice

- Do not know what the population distribution is
- Can not generate more random samples by simulation

In Practice

- Do not know what the population distribution is
- Can not generate more random samples by simulation
- May be interested in statistic other than sample mean

In Practice

- Do not know what the population distribution is
- Can not generate more random samples by simulation
- May be interested in statistic other than sample mean
- Stuck???

Outline

Empirical Distribution

- Empirical distribution of a random sample:

Empirical Distribution

- Empirical distribution of a random sample:
  - observe data $\boldsymbol{X} = X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x, \theta)$
  - a discrete uniform distribution
  - each data point X_i is associated with probability $1/n$

Empirical Distribution

- Empirical distribution of a random sample:
  - observe data $\boldsymbol{X} = X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x, \theta)$
  - a discrete uniform distribution
  - each data point X_i is associated with probability $1/n$
- Empirical distribution of a statistic:

Empirical Distribution

- Empirical distribution of a random sample:
  - observe data $\boldsymbol{X} = X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x, \theta)$
  - a discrete uniform distribution
  - each data point X_i is associated with probability $1/n$
- Empirical distribution of a statistic:
  - sampling distribution needs to exhaust "ALL" possible random samples
  - empirical distribution just take "enough" random samples
  - empirical distribution is what we get in simulation
  - intuitively, empirical $\rightarrow$ sampling distribution as rep $\rightarrow \infty$

From Empirical to Theoretical

- The “Fundamental Theorem of Statistics”
  - in plain English: the empirical distribution of a large random sample resembles the population distribution.

From Empirical to Theoretical

- The "Fundamental Theorem of Statistics"
  - in plain English: the empirical distribution of a large random sample resembles the population distribution.
  - somewhat technical: the empirical distribution converges to the population distribution as sample size increases.

From Empirical to Theoretical

- The "Fundamental Theorem of Statistics"
  - in plain English: the empirical distribution of a large random sample resembles the population distribution.
  - somewhat technical: the empirical distribution converges to the population distribution as sample size increases.
  - formally known as: Glivenko-Cantelli Theorem

From Empirical to Theoretical

- The "Fundamental Theorem of Statistics"
  - in plain English: the empirical distribution of a large random sample resembles the population distribution.
  - somewhat technical: the empirical distribution converges to the population distribution as sample size increases.
  - formally known as: Glivenko-Cantelli Theorem
- closely related to the"Law of Large Numbers"

From Empirical to Theoretical

- The "Fundamental Theorem of Statistics"
  - in plain English: the empirical distribution of a large random sample resembles the population distribution.
  - somewhat technical: the empirical distribution converges to the population distribution as sample size increases.
  - formally known as: Glivenko-Cantelli Theorem
- closely related to the"Law of Large Numbers"
- Intuition: large, random sample helps us understand the population

Outline

Motivation

- To quantify the sampling distribution, we need more samples

Motivation

- To quantify the sampling distribution, we need more samples
- In practice, we usually only have one sample

Motivation

- To quantify the sampling distribution, we need more samples
- In practice, we usually only have one sample
- How do we generate more random samples without explicitly access the population?

Motivation

- To quantify the sampling distribution, we need more samples
- In practice, we usually only have one sample
- How do we generate more random samples without explicitly access the population?
  - we need these samples look like as if they are from the population

Idea

- The "Fundamental Theorem of Statistics" says the empirical distribution of a large random sample resembles the population distribution.

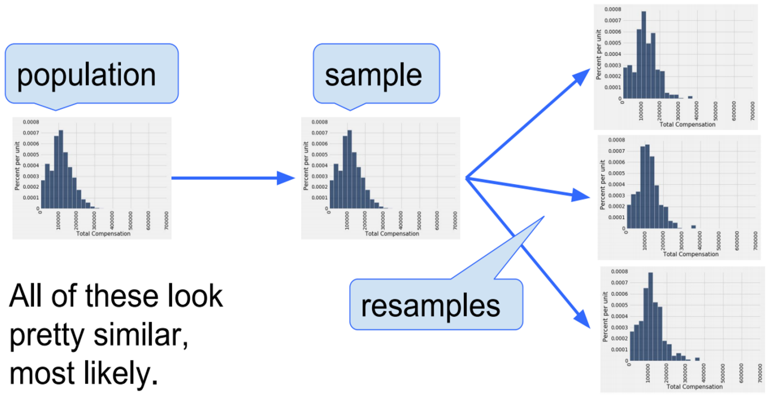
Idea

- The “Fundamental Theorem of Statistics” says the empirical distribution of a large random sample resembles the population distribution.
- The goal is to generate more random samples from the population distribution

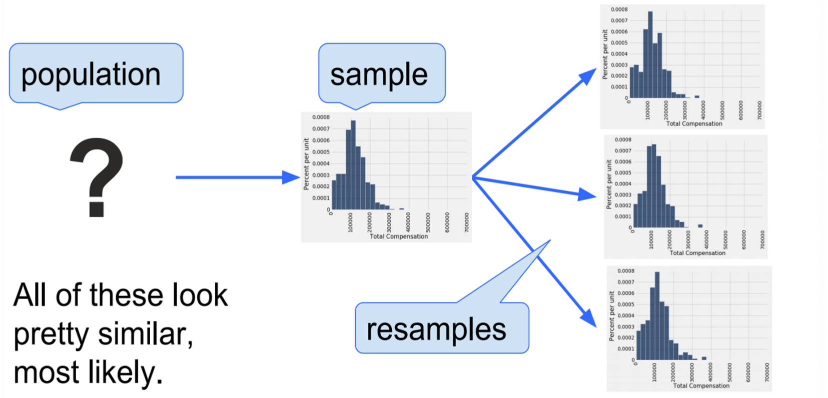
Idea

- The "Fundamental Theorem of Statistics" says the empirical distribution of a large random sample resembles the population distribution.
- The goal is to generate more random samples from the population distribution
- Instead, generate more random samples from the empirical distribution

Illustration



Illustration



Implementation

- Sample from the data

Implementation

- Sample from the data
  - one original sample
  - that's all you will sample from
- Sample randomly with replacement

Implementation

- Sample from the data
  - one original sample
  - that's all you will sample from
- Sample randomly with replacement
  - the empirical distribution = discrete uniform
  - how to sample from such a distribution?
- Sample the same number of observations as the data

Implementation

- Sample from the data
  - one original sample
  - that's all you will sample from
- Sample randomly with replacement
  - the empirical distribution = discrete uniform
  - how to sample from such a distribution?
- Sample the same number of observations as the data
  - fixed sample size
  - sampling distribution depends on n

Intuition

- A way to "shake up" your data to mimic different random samples

Intuition

- A way to "shake up" your data to mimic different random samples
- In each bootstrap resample:
    - some original observations may not appear
    - some may appear more than once
    - approximately 2/3 of original observations

Intuition

- A way to "shake up" your data to mimic different random samples
- In each bootstrap resample:
  - some original observations may not appear
  - some may appear more than once
  - approximately 2/3 of original observations
- How many resamples to generate?
  - the more, the better!
  - computation can be parallel

Parametric Bootstrap

- So far we have covered nonparametric bootstrap, where no assumption about population distribution is made.

Parametric Bootstrap

- So far we have covered nonparametric bootstrap, where no assumption about population distribution is made.
- For parametric bootstrap, you know the population distribution family with some unknown parameters:
  - $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x, \theta)$ where θ is the unknown parameter

Parametric Bootstrap

- So far we have covered nonparametric bootstrap, where no assumption about population distribution is made.
- For parametric bootstrap, you know the population distribution family with some unknown parameters:
  - $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x, \theta)$ where θ is the unknown parameter
- Parametric Bootstrap:
  - get a point estimate $\hat{\theta}$, for example MLE
  - generate bootstrap resamples from $X_1^b, X_2^b, \cdots X_n^b \overset{\text{i.i.d.}}{\sim} f_X(x, \hat{\theta})$

Outline

Point Estimate

- A point estimate is a statistic with a sampling distribution.

Point Estimate

- A point estimate is a statistic with a sampling distribution.
- We would like to estimate the variance of the point estimate, why?

Point Estimate

- A point estimate is a statistic with a sampling distribution.
- We would like to estimate the variance of the point estimate, why?
  - also known as the sampling variance
  - quantifies the uncertainty of the estimate
  - bias-variance decomposition

Point Estimate

- A point estimate is a statistic with a sampling distribution.
- We would like to estimate the variance of the point estimate, why?
  - also known as the sampling variance
  - quantifies the uncertainty of the estimate
  - bias-variance decomposition
- How to do this using bootstrap?

Bootstrap Variance Estimation

- Data: $\boldsymbol{X} = \{X_1, X_2, \cdots X_n\}$

Bootstrap Variance Estimation

- Data: $\boldsymbol{X} = \{X_1, X_2, \cdots X_n\}$
- Point estimate: $T_n(\boldsymbol{X})$

Bootstrap Variance Estimation

- Data: $\boldsymbol{X} = \{X_1, X_2, \cdots X_n\}$
- Point estimate: $T_n(\boldsymbol{X})$
- Bootstrap resample $\boldsymbol{X}^{(b)} = \{X_1^{(b)}, X_2^{(b)}, \cdots X_n^{(b)}\}$ for $b = 1, 2, \cdots B$ where B is the total number of resamples
  - for each b, calculate $T_n^{(b)} = T_n(\boldsymbol{X}^{(b)})$

Bootstrap Variance Estimation

- Data: $\boldsymbol{X} = \{X_1, X_2, \cdots X_n\}$
- Point estimate: $T_n(\boldsymbol{X})$
- Bootstrap resample $\boldsymbol{X^{(b)}} = \{X_1^{(b)}, X_2^{(b)}, \cdots X_n^{(b)}\}$ for $b = 1, 2, \cdots B$ where B is the total number of resamples
  - for each b, calculate $T_n^{(b)} = T_n(\boldsymbol{X^{(b)}})$
- Bootstrap variance and standard error estimation:

$$\text{Var}_{boot}(T_n) = \frac{1}{B}\sum_{b=1}^{B}\left(T_n^{(b)} - \frac{1}{B}\sum_{b=1}^{B} T_n^{(b)}\right)^2$$

Bootstrap CI

- Normal CI:

$$T_n \pm z_{\alpha/2} \cdot \widehat{\mathsf{se}}_{boot}$$

Bootstrap CI

- Normal CI:

$$T_n \pm z_{\alpha/2} \cdot \widehat{\text{se}}_{boot}$$

  - only use if the sampling distribution of T_n is approx normal
  - how do we know? sample mean, MLE, ...

Bootstrap CI

- Normal CI:

$$T_n \pm z_{\alpha/2} \cdot \widehat{\text{se}}_{boot}$$

  - only use if the sampling distribution of T_n is approx normal
  - how do we know? sample mean, MLE, ...
- Percentile CI:

$$\left(T^{(b)}_{n,\alpha/2}, \quad T^{(b)}_{n,1-\alpha/2}\right)$$

Bootstrap CI

- Normal CI:

$$T_n \pm z_{\alpha/2} \cdot \widehat{\mathrm{se}}_{boot}$$

  - only use if the sampling distribution of T_n is approx normal
  - how do we know? sample mean, MLE, ...
- Percentile CI:

$$\left(T^{(b)}_{n,\alpha/2}, \quad T^{(b)}_{n,1-\alpha/2}\right)$$

  - for a 95% CI, take the 2.5% and 97.5% percentile of $T_n^{(b)}$
  - intuition: the "middle" 95% of the empirical distribution

Implement Bootstrap

- Visualize an empirical distribution by bootstrap:
  1. have one large random sample
  2. generate one bootstrap resample
  3. calculate the specific statistic and save the value
  4. repeat step 2 - 3, many many... times

Implement Bootstrap

- Visualize an empirical distribution by bootstrap:
  1. have one large random sample
  2. generate one bootstrap resample
  3. calculate the specific statistic and save the value
  4. repeat step 2 - 3, many many... times
- No assumption about the population distribution – nonparametric

Implement Bootstrap

- Visualize an empirical distribution by bootstrap:
  1. have one large random sample
  2. generate one bootstrap resample
  3. calculate the specific statistic and save the value
  4. repeat step 2 - 3, many many... times
- No assumption about the population distribution – nonparametric
- Make assumption about the population distribution – parametric

Implement Bootstrap

- Visualize an empirical distribution by bootstrap:
  1. have one large random sample
  2. generate one bootstrap resample
  3. calculate the specific statistic and save the value
  4. repeat step 2 - 3, many many... times
- No assumption about the population distribution – nonparametric
- Make assumption about the population distribution – parametric
- The only difference in the step 2:
  - nonparametric: sample with replacement
  - parametric: sample from $f_X(x, \hat{\theta})$
  - other inference carried out exactly in the same way: sampling variance estimate and bootstrap CI...

When Not to Use Bootstrap

- When your original sample is NOT good
  - very small sample size
  - highly dependent samples

When Not to Use Bootstrap

- When your original sample is NOT good
  - very small sample size
  - highly dependent samples
- When you are inferring max, min values
  - extreme value theory
  - difficult problem

Outline