ECE560: Computer Systems Performance Evaluation

Lecture #8 –

Statistical Inference

Instructor: Dr. Liudong Xing
Spring 2024

Topics

- **Statistical Inference**
- Population *vs*. Sample
- Parameter estimators
- Confidence interval

Chapter 7 in Allen's book

This is a self-study lecture!

Statistical Inference (S.I.)

- Let X represent a performance measure of interest, e.g.
  - response time,
  - # of arrivals / completion per unit of time
- X can be regarded as a *r.v.* with an unknown distribution or with a known distribution form but with unknown parameters
- **Task/goal** is to estimate the unknown parameters based on the data collected from the experiments.

Basic Concepts (1)

- Population
  - A set of data points consists of all possible observations of a *r.v. X*
  - In general, a collection of elements under investigation
- Sample
  - A set of data points contains only part of all possible observations of a *r.v. X*
  - In general, a selected subset of a population

Basic Concepts (2)

- Methods of S.I. help us estimate the characteristics of the entire population based on the data collected from a sample.

- Population characteristics are called parameters
  - Mean (population/true mean): μ
  - Variance (population/true variance): σ^2
  - Standard deviation: σ

- Sample estimates are called statistics
  - Sample mean: $\bar{x}$
  - Sample variance: s^2
  - Sample standard deviation: s

Basic Concepts (3)

- Population parameters are **fixed**

- Statistics are **random**; different samples from the same population usually result in distinct estimates

Basic Concepts (4)

- Definition: Suppose we perform n measurements and observe a set of n experimental outcomes x_1, x_2, …, x_n. Each outcome x_i is a value of a *r.v.* X_i. Then, the set of *r.v.*s X_1, X_2, …, X_n is called a sample of size n from the population.

Basic Concepts (5)

- Definition: The set of *r.v.*s X_1, X_2, …, X_n is said to constitute a random sample of size n from the population that is determined by *r.v.* X (its c.d.f. is $F_X(x)$), provided that a) they are mutually independent and b) they are identically distributed with distribution $F_{Xi}(x) = F_X(x)$, for all i, for all x. →
- x_1, x_2, …, x_n are observed values of a random sample.
- Definition (Statistic): any function of a random sample X_1, X_2, …, X_n.
  - A statistic is also a r.v.,
  - Its distribution function -- "sampling distribution"

Basic Concepts (6)

- Definition (Estimator): any statistic $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots X_n)$ used to estimate the value of a parameter θ of the population is called an estimator of θ
  - An observed value of the statistics $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(x_1, x_2, \cdots x_n)$ is known as an estimate of θ
  - The two most common estimators are :
    - Sample mean (arithmetic mean) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
    - Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
    - $\overline{X}$ is an estimator of μ,
    - S^2 is an estimator of σ^2
  - Important properties
    - Unbiasedness
    - Consistency
    - Efficiency

Properties of Parameter Estimators (1)

- **Unbiasedness**: A statistic $\hat{\theta}$ is said to constitute an unbiased estimator of parameter θ if $E[\hat{\theta}] = \theta$
  - An estimator $\hat{\theta}$ is unbiased if $E[\hat{\theta}] = \theta$
  - In other words, on the average, the estimator is on target, or the estimated values of θ cluster about θ

Properties of Parameter Estimators (2)

- The sample mean $\overline{X}$ is an unbiased estimator
  - Proof (extra notes)
  - We can also compute the variance of $\overline{X}$

$$\mathrm{var}[\overline{X}] = \frac{\sigma^2}{n}$$

→ the accuracy of the sample mean as an estimator of the population mean μ increases with the sample size n when the population variance is finite

  - $\mathbf{SD}[\overline{X}] = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ is often called the "standard error"
    - Since usually we don't know the value of σ, The approximate value of the standard error is taken to be $s/\sqrt{n}$ where $s = \sqrt{s^2}$ is the sample SD

Properties of Parameter Estimators (3)

- The sample variance S^2 is an unbiased estimator
  - Proof (extra notes)
  - The proof shows why the denominator should be *n-1*

Properties of Parameter Estimators (4)

- **Consistency**: An estimator $\hat{\theta}$ of parameter θ is said to be consistent if $\hat{\theta}$ converges in probability to θ

$$\lim_{n\to\infty} P\{|\hat{\theta}-\theta| < e\} = 1$$

$$\lim_{n\to\infty} P\{|\hat{\theta}-\theta| \geq e\} = 0 \quad (\forall e > 0)$$

Properties of Parameter Estimators (5)

- **Efficiency**: An estimator $\hat{\theta}_1$ is said to be a more efficient estimator of the parameter θ than the estimator $\hat{\theta}_2$, if
  - $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ
  - $\text{var}[\hat{\theta}_1] < \text{var}[\hat{\theta}_2]$
- An unbiased estimator $\hat{\theta}$ is the *minimum variance unbiased estimator* of θ if

  $$\text{var}[\hat{\theta}] < \text{var}[\hat{\theta}_1]$$

  where $\hat{\theta}_1$ is any other unbiased estimator of θ

  → $\hat{\theta}$ is the most efficient estimator
- Sample mean is the minimum variance / the most efficient estimator of the population mean
- Sample variance is the minimum variance / the most efficient estimator of the population variance

Properties of Parameter Estimators (6)

- In summary, sample mean $\overline{X}$ (sample variance S^2) is the most efficient, unbiased and consistent estimator of population mean μ (population variance σ^2)
- The properties of estimators, such as being unbiased, consistent and the minimum-variance unbiased, etc., cannot help us make a probability judgment about the quality or accuracy of the estimates delivered
- The confidence intervals enable us to do that!

Confidence Intervals

- Probability statement: if we can ascertain that the estimator $\hat{\theta}$ satisfies the condition $\Pr\{\hat{\theta}-\varepsilon_1 < \theta < \hat{\theta}+\varepsilon_2\} = \alpha$ then, we can say that the interval $(\hat{\theta}-\varepsilon_1, \hat{\theta}+\varepsilon_2)$ is a $100\alpha\%$ confidence interval for parameter θ
  - α is called "confidence coefficient"
  - 100α is called "confidence level"
  - $1-\alpha$ is called "significance level"
  - $\varepsilon_1 + \varepsilon_2$ is the interval width ($\varepsilon_1 = \varepsilon_2 = \varepsilon$: Symmetric interval)

Confidence Intervals (2)

- 95% and 99% confidence interval are particularly popular, corresponding to α values of 0.95 and 0.99
  - e.g. $\Pr\{\overline{X} - e < \mu < \overline{X} + e\} = 0.99$
    Meaning the prob./confidence that the true mean μ lies within the interval $(\overline{X} - e, \overline{X} + e)$ is 0.99
- For a given confidence level, the shorter the interval $(\varepsilon_1 + \varepsilon_2 / 2\varepsilon)$, the more accurate the estimates!

Next Topics

- Stochastic Processes

Things to Do

- Project proposal
  - Due **February 23, Friday.**