工程概率统计

Probability and Statistics for Engineering

第五章 统计学基础

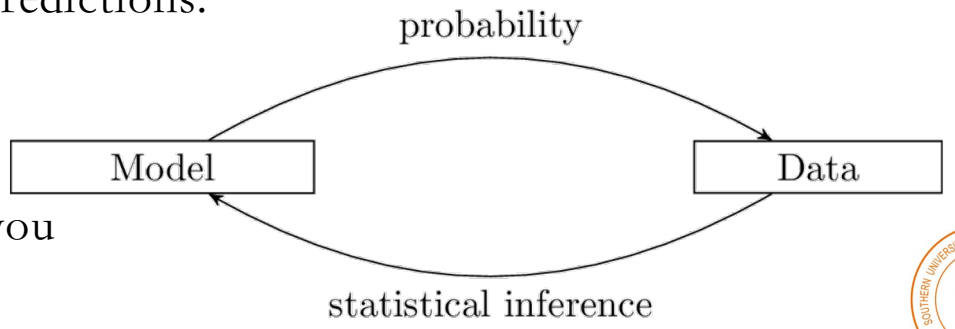
Chapter 5 Basic Concepts in Statistics

Chapter 5 Basic Concepts in Statistics

- 5.1 Population, Sample, Parameters and Statistics
- 5.2 Parameter Estimation – Point Estimation
  - 5.2.1 Properties of Point Estimation
  - 5.2.2 Method of Moments
  - 5.2.3 Method of Maximum Likelihood
- 5.3 Parameter Estimation – Confidence Interval

5.1 Population, Sample, Parameters and Statistics

- The first four chapters are about Probability Theory. In the following chapters, we will enter the field of Statistics.
- What's the relationship between Probability Theory and Statistics?
  - In Probability Theory, we are given the probability model and parameters of a random experiment, the question of interest is typically calculating the probability of a certain outcome.
  - In Statistics, the probability model or parameters of a random experiment are usually unknown, and we have collected some data from the experiment based on which the probability model and its parameters can be inferred, helping us understand the mechanism that generates the data.
  - Probability Theory provides the theoretical tools that Statistics builds on, and Statistics applies these tools to real-world data to draw conclusions and make predictions.
  - Probability Theory can be thought as the "theory of the game" where you know all the rules.
  - Statistics can be thought as the "detective work" where you figure out the rules by observing the game's outcomes.

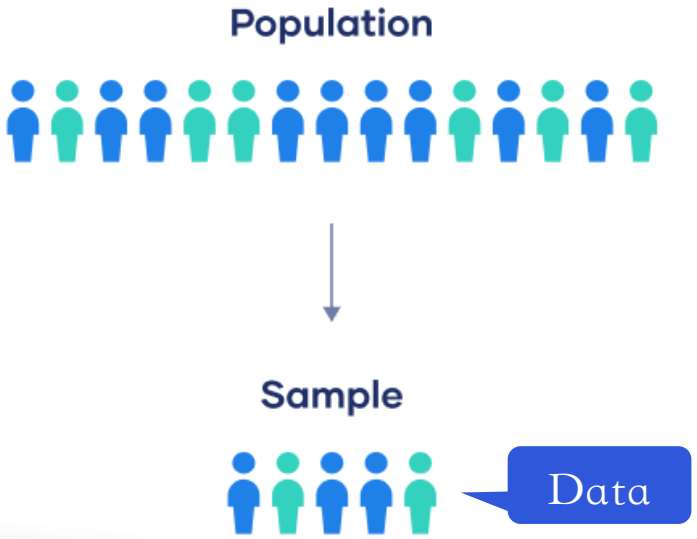


5.1 Population, Sample, Parameters and Statistics

- Data collection is a crucially important step in Statistics. We use the collected data (the sample) to make statements about a much larger set (the population).

Population, parameter, and Sample

- A population (总体) refer to the entire set of individuals, objects, events, or measurements that share a common characteristic and are of interest in a particular study.
- Any numerical characteristic of a population is a parameter (参数).
- A sample (样本) is a subset of the population, which is used to make inferences about the population.



- When the form of the population distribution is known but contains unknown parameters, inference about these parameters is referred to as parametric statistical inference (参数统计推断).

5.1 Population, Sample, Parameters and Statistics

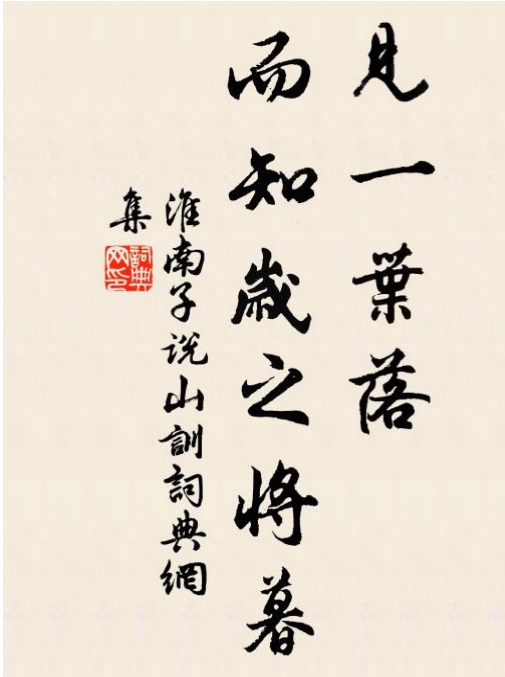
Example 5.1

- If you want to study the income of Shenzhen residents in 2024, you create a survey questionnaire and make it available online for people to fill out.
- The population in this context refers to all Shenzhen residents, or more specifically, the income of all Shenzhen residents in 2024. Mean/median/SD of the income can be considered parameters.
- A sample is this context refers to the Shenzhen residents who fill your questionnaire.

- Generally, we are interested in some **quantitative index (定量指标/数值指标)** X of the population.
- The value of X for each individual in the population may be different and these values collectively form a probability distribution.
- So, the population is usually expressed as a random variable X following a specific distribution, i.e., $X \sim F(x)$ (the CDF) or $X \sim f(x)$ (the PMF/PDF).
- In Statistics, the distribution of X is often unknown, but we can typically assume a specific form based on context, leaving only certain parameters $\boldsymbol{\theta}$ of the distribution unknown.
- So, the population can be expressed as $X \sim F(x;\boldsymbol{\theta})$ or $X \sim f(x;\boldsymbol{\theta})$. E.g., $X \sim \mathrm{N}(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)$.

5.1 Population, Sample, Parameters and Statistics

- n individuals are selected from the population, their quantitative index are recorded as $X_1, \ldots, X_n$, which is called **a sample of size n (一个容量为n的样本)**, n is the **sample size (样本容量)**.
- The process of selecting individuals from the population is called **sampling (抽样)**.
- Our goal is to make inference of the population based on a sample.
- The idea that a local perspective can reveal universal truths has long been recognized, there is a famous text: "From the small, we discern the large; seeing a single leaf fall, we know that the year is coming to an end; observing ice in a bottle, we understand the coldness of the world."
- However, sometimes a local perspective does not provide a complete understanding of the whole situation. For example, there is a proverb: "peering at a leopard through a tube, one sees only spots."
- **Question:** under what circumstances can conclusions drawn from a sample be generalized to the entire population?



5.1 Population, Sample, Parameters and Statistics

- This requires special attention during sampling to ensure that the samples drawn are sufficiently representative (样本要有足够的代表性).

Example 5.2

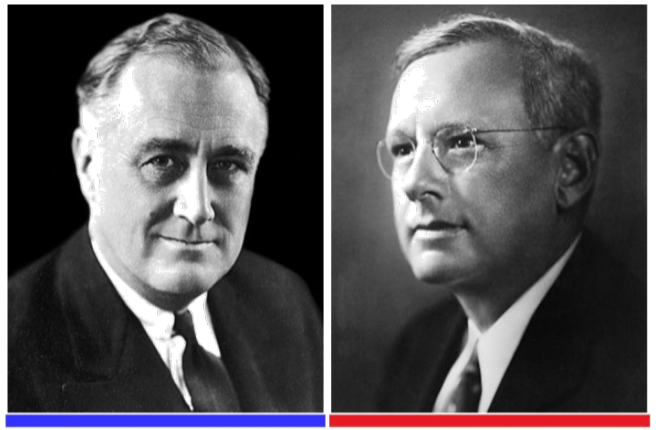
Public opinion polling in U.S. presidential elections

- The United States has numerous polling agencies, all striving for accuracy in their polls, as it is crucial for maintaining their reputation and even their survival and growth.
- From 1916 to 1932, the dominant player in polling was *The Literary Digest*. It accurately predicted the results of 5 consecutive presidential elections, earning widespread trust and acclaim.
- In 1932, *The Literary Digest* successfully predicted the election outcome, ushering the era of Roosevelt's New Deal.
- In the 1936 election, the two candidates were Roosevelt (Democratic) and Landon (Republican). Most polling organizations, media outlets, and political observers forecasted Roosevelt's victory.
- However, *The Literary Digest* offered a different prediction. It conducted a large-scale survey, mailing out a staggering 10 million questionnaires and receiving about 2.3 million responses.
- The results indicated that 57% of respondents supported Landon.

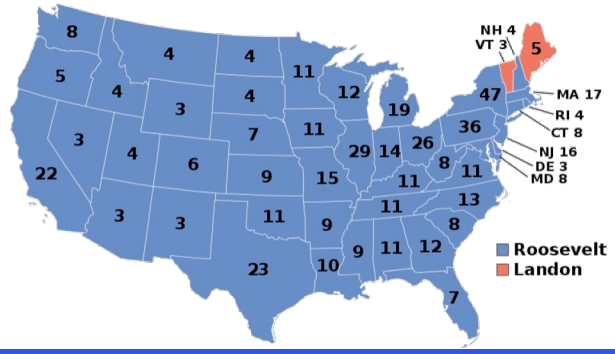
5.1 Population, Sample, Parameters and Statistics

Example 5.2

- Due to the significant influence of *The Literary Digest*, many believed that Landon would become the 33rd President of the U.S.
- However, it turned out that Roosevelt achieved a landslide victory, winning by the largest margin in the history of U.S. presidential elections.
- This grave error destroyed *The Literary Digest*'s credibility, leading to its demise within 18 months of the election.
- In contrast, a relatively unknown company, Gallup, correctly predicted the outcome with a sample of only 50,000 respondents, propelling it to fame.
- What went wrong with *The Literary Digest*'s polling?



Nominee	Franklin D. Roosevelt	Alf Landon
Party	Democratic	Republican
Home state	New York	Kansas
Electoral vote	523	8
States carried	46	2
Popular vote	27,747,636	16,679,543
Percentage	60.8%	36.5%



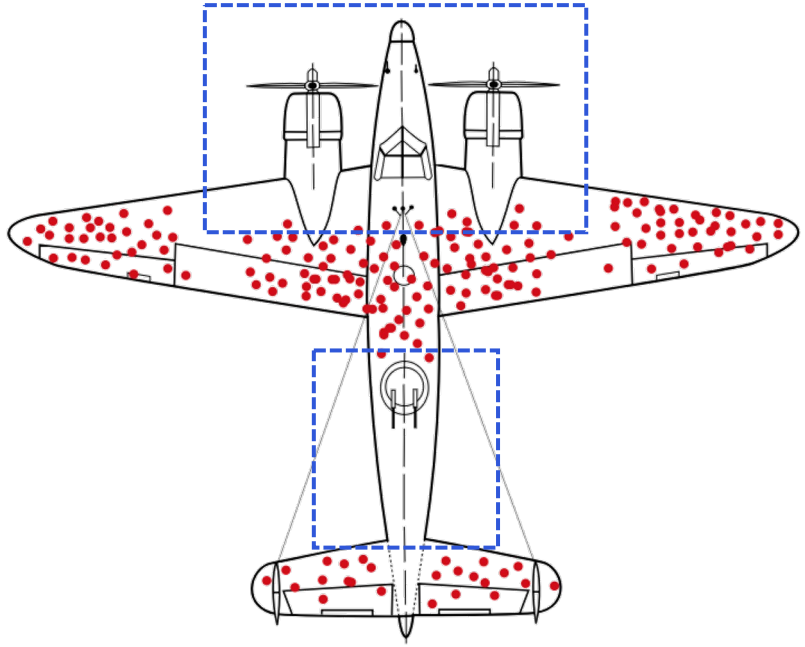
5.1 Population, Sample, Parameters and Statistics

Solution

5.1 Population, Sample, Parameters and Statistics

Example 5.3

- During World War II, the United States sought to reinforce specific areas of its bombers with additional armor.
- Analysts examined returning bombers and mapped the bullet holes and damage.
- Based on this analysis, they concluded that adding armor to the tail, fuselage, and wings would increase the pilots' chances of survival.
- What is your perspective on this approach?

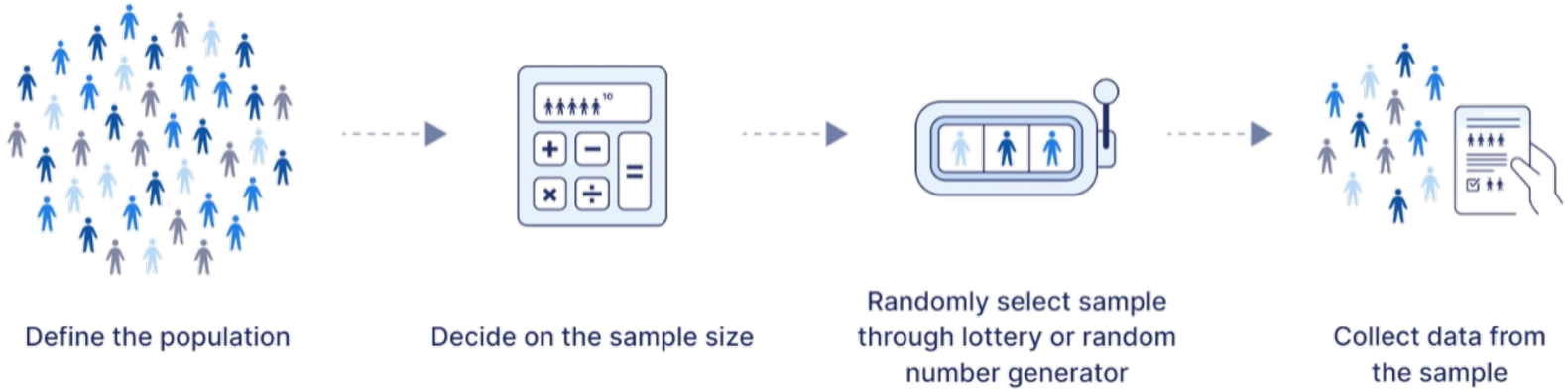


5.1 Population, Sample, Parameters and Statistics

Solution

5.1 Population, Sample, Parameters and Statistics

- The simplest method to ensure representativeness is **simple random sampling (简单随机抽样)**.



- The core of simple random sampling is to select samples in a completely random manner, ensuring that every individual in the population has an equal probability of being chosen.
- Expressed in statistical language, $X_1, \dots, X_n$ are independent, and follow the same distribution with X, i.e., $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} X$. This is a fundamental assumption of many statistical methods.
- **Question:** why $X_1, \dots, X_n$ are considered random variables?
- **Before** sampling, we have no idea who will be sampled; **after** sampling, the observed values $x_1, \dots, x_n$ of $X_1, \dots, X_n$ are obtained, called the **sample observed values (样本观测值)**.

5.1 Population, Sample, Parameters and Statistics

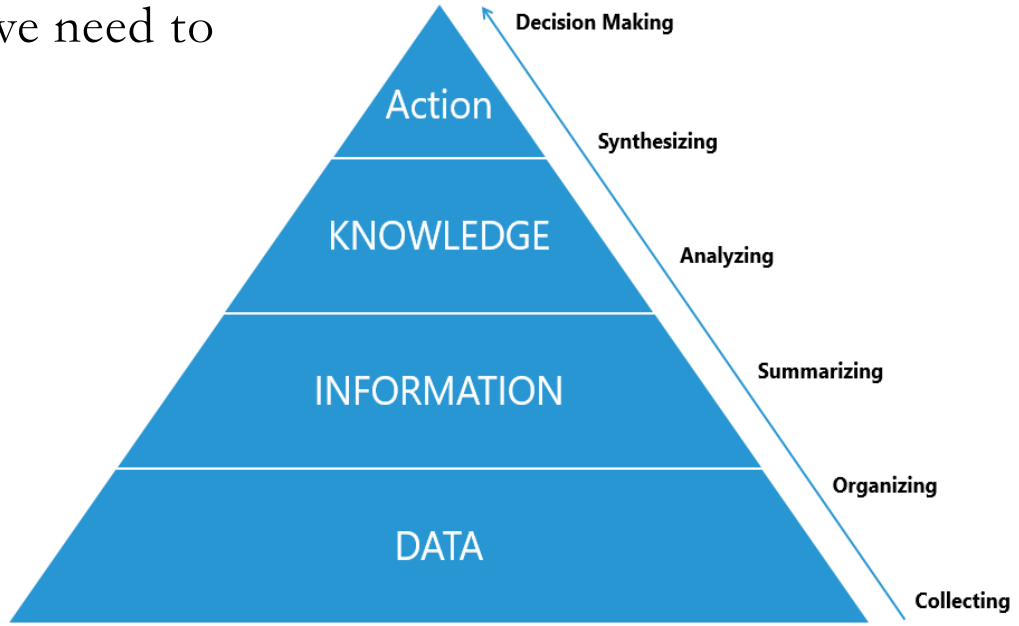
- Based on a simple random sample $X_1, \ldots, X_n$, what information can we get about the population?
- To extract information from the sample (i.e., data), we need to process the data, i.e., calculate $g(X_1, \ldots, X_n)$.

Statistic

- If $X_1, \ldots, X_n$ is a sample from population $X \sim f(x; \boldsymbol{\theta})$ and $g(x_1, \ldots, x_n)$ is an n-variate function (n元函数). Define random variable

$$T = g(X_1, \ldots, X_n),$$

- then T is called a statistic (统计量) if $g(\,)$ does not involve any unknown parameter.



- Here are some commonly used statistics:
  - Sample mean (样本均值)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

5.1 Population, Sample, Parameters and Statistics

- Commonly used statistics:
  - Sample variance (样本方差) and sample standard deviation (样本标准差)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \qquad S = \sqrt{S^2}.$$

Why not divide by n?

  - Order statistics (顺序统计量)

$$\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right).$$

$X_{(1)} = \min\{X_1, \ldots, X_n\}$,
$X_{(n)} = \max\{X_1, \ldots, X_n\}$

  - Sample p-quantile (样本p分位数), number that exceeds at most $100p\%$ of the sample and is exceeded by at most $100(1-p)\%$ of the sample

$$Q_p = \begin{cases} X_{(\lfloor np+1 \rfloor)}, & \text{if } np \text{ is not an integer} \\ \dfrac{\left(X_{(np)} + X_{(np+1)}\right)}{2}, & \text{if } np \text{ is an integer} \end{cases},$$

  - Specifically, $Q_{0.5}$ is the sample median (样本中位数), $Q_{0.25}$ ($Q_{0.75}$) is the sample lower (upper) quartile (样本下(上)四分位数) or the sample first (third) quartile (样本第一(三)四分位数, also written as $Q_1(Q_3)$). $Q_3 - Q_1$ is the sample interquartile range (sample IQR, 样本四分位距).

5.1 Population, Sample, Parameters and Statistics

- Note that $\sum(X_i - \mu)^2/n$ is not a statistic as it involves unknown parameter μ.
- Typically, statistics measure the location, spread, variability, and other characteristics of the sample, which can be used to estimate the corresponding population parameters.
- Note that, each statistic is a random variable because it is computed from random data. After sampling, the observed value of a statistic can be obtained.
- The distribution of a statistics is called the **sampling distribution (抽样分布)**, which are required to construct confidence interval and perform hypothesis testing latter.
- Despite computing statistics, when it comes to analysis of real data, people often graphically check the distribution of the sample, e.g., through a histogram of the sample. This may help suggest a probability model to be used.

Chapter 5 Basic Concepts in Statistics

- 5.1 Population, Sample, Parameters and Statistics
- 5.2 Parameter Estimation – Point Estimation
  - 5.2.1 Properties of Point Estimation
  - 5.2.2 Method of Moments
  - 5.2.3 Method of Maximum Likelihood
- 5.3 Parameter Estimation – Confidence Interval

5.2.1 Properties of Point Estimation

- Under parametric statistical inference, the form of the population distribution is known but contain unknown parameters, the first task is to estimate the parameters based on a sample.

Example 5.4

- Computer chips produced by a factory have a certain type of rare defects. Suppose that the number of defects on a randomly selected chip X follows a Poisson distribution with unknown parameter λ.
- A simple random sample of computer chips is draw and the number of defects on each chip is recorded, denoted as $X_1, \ldots, X_n$. How to estimate the unknown parameter λ based on the sample?
- Since $\lambda = \mathrm{E}(X)$, should we estimate it with the sample mean $\overline{X}$?
- On the other hand, $\lambda = \mathrm{Var}(X)$, should we use the sample variance S^2 to estimate λ?

Example 5.5

Note that κ is not required to be an integer

- Suppose now that we deal with a $\mathrm{Gamma}(\kappa, \lambda)$ distribution. Its parameters κ and λ does not represent the mean, variance, standard deviation, or other measures discussed before.
- How to estimate these unknown parameters?

5.2.1 Properties of Point Estimation

- Estimating an unknown parameter with a statistic is called **point estimation (点估计)**.

Point Estimation

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$, $\theta_1, \ldots, \theta_k$ are the unknown parameters.
- If $\hat{\theta}_l(X_1, \ldots, X_n)$ is a statistic used to estimate θ_l, then $\hat{\theta}_l$ is called an **estimator (估计量)** of θ_l.
- Plugging int the sample observed values, $\hat{\theta}_l(x_1, \ldots, x_n)$ is called an **estimate** or **estimated value (估计值)** of θ_l.
- Both $\hat{\theta}_l(X_1, \ldots, X_n)$ and $\hat{\theta}_l(x_1, \ldots, x_n)$ can be abbreviated as $\hat{\theta}_l$.

- As stated in Example 5.4, there may be multiple estimators for the same parameter.
- Therefore, before introducing the methods of point estimation, we first introduce some properties used to compare multiple point estimators.

5.2.1 Properties of Point Estimation

- Since an estimator $\hat{\theta}_l$ is a statistic, i.e., a random variable, its value is sometimes greater and sometimes smaller than the true parameter value θ_l.
- So, a natural criterion to evaluate whether $\hat{\theta}_l$ is a good estimator is to see if it underestimate or overestimate θ_l on average.

Unbiasedness

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$ and $\hat{\theta}_l(X_1, \ldots, X_n)$ is an estimator of θ_l. $\boldsymbol{\Theta}$ is the parameter space (参数空间) of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, i.e., the set of values that $\boldsymbol{\theta}$ can take.
- If for $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have

$$\mathrm{E}_{\boldsymbol{\theta}}(\hat{\theta}_l) = \theta_l,$$

The expectation is calculated based on population distribution $f(x; \boldsymbol{\theta})$. It is often simply expressed as $\mathrm{E}(\hat{\theta}_l)$.

- then $\hat{\theta}_l$ is called an unbiased estimator (无偏估计量) of θ_l. Otherwise, it is called a biased estimator (有偏估计量) of θ_l.
- $\mathrm{E}_{\boldsymbol{\theta}}(\hat{\theta}_l) - \theta_l$ is called the bias (偏差) of the estimator. If the bias is not 0, but converge to 0 as $n \to \infty$, then $\hat{\theta}_l$ is called an asymptotic unbiased estimator (渐近无偏估计量) of θ_l.

5.2.1 Properties of Point Estimation

- No matter what is the population distribution, if the population mean $\mu = \mathrm{E}(X)$ and population variance $\sigma^2 = \mathrm{Var}(X)$ exist, then $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = S^2$ are unbiased estimators of μ and σ^2.

Proof:

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \mu.$$

$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator of σ^2, but it is asymptotic unbiased

$$\because (n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}[(X_i - \mu) + (\mu - \bar{X})]^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

$$\therefore (n-1)\mathrm{E}(S^2) = \sum_{i=1}^{n}\mathrm{E}(X_i - \mu)^2 - n\mathrm{E}(\bar{X} - \mu)^2 = \sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2$$

$$\therefore \mathrm{E}(S^2) = \sigma^2.$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}.$$

- This answers the question on Page 12.

5.2.1 Properties of Point Estimation

- When comparing multiple estimators of the same parameter, we prefer those that are unbiased.
- However, unbiased estimator may not be unique.
  - E.g., for the population $X \sim \text{Poisson}(\lambda)$, $\hat{\lambda}_1 = \overline{X}$ and $\hat{\lambda}_2 = S^2$ are both unbiased estimators of λ.
  - Moreover, for any constant c, $c\hat{\lambda}_1 + (1-c)\hat{\lambda}_2$ is also an unbiased estimator of λ.
- Then a natural question is: how to compare multiple unbiased estimators?
- Unbiasedness suggests that an estimator fluctuates around the true parameter, considering the stability of estimation, we would like the magnitude of fluctuation to be as small as possible.

Relative Efficiency

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$, $\hat{\theta}_{l1}$ and $\hat{\theta}_{l2}$ are two unbiased estimators of θ_l. $\boldsymbol{\Theta}$ is the parameter space (参数空间). If for $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have

$$\text{Var}_{\boldsymbol{\theta}}(\hat{\theta}_{l1}) \leq \text{Var}_{\boldsymbol{\theta}}(\hat{\theta}_{l2}),$$

Can be simply expressed as $\text{Var}(\hat{\theta}_{lj})$.

- and $\text{Var}_{\boldsymbol{\theta}}(\hat{\theta}_{l1}) < \text{Var}_{\boldsymbol{\theta}}(\hat{\theta}_{l2})$ for at least one $\boldsymbol{\theta}$, then $\hat{\theta}_{l1}$ is said to be more efficient (更有效) than $\hat{\theta}_{l2}$.

5.2.1 Properties of Point Estimation

Example 5.6

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim \mathrm{U}(0, \theta)$.
- 1. Show that $\hat{\theta}_1 = 2\bar{X}$ and $\hat{\theta}_2 = (n+1)X_{(n)}/n$ are both unbiased estimators of θ.
- 2. Which of these two estimators is more efficient?

Solution

5.2.1 Properties of Point Estimation

Solution

5.2.1 Properties of Point Estimation

- The last property is about the convergence of an estimator as the sample size $n \to \infty$.

Consistency

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$ and $\hat{\theta}_l$ is an estimator of θ_l. $\boldsymbol{\Theta}$ is the parameter space (参数空间). If for $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\forall \varepsilon > 0$, we have

$$\lim_{n \to \infty} \mathrm{P}_{\boldsymbol{\theta}}\left(\left|\hat{\theta}_l - \theta_l\right| > \varepsilon\right) = 0,$$

i.e., $\hat{\theta}_l \xrightarrow{\mathrm{P}} \theta_l$ as $n \to \infty$

- then $\hat{\theta}_l$ is called a **consistent estimator (相合估计量)** of θ_l.

- **Question:** what's the relationship between (asymptotic) unbiasedness and consistency?
- An (asymptotic) unbiased estimator may not be a consistent estimator.
  - E.g., $\hat{\theta} = X_n$ is an unbiased estimator of the population mean $\theta = \mathrm{E}(X)$ since $\mathrm{E}(\hat{\theta}) = \theta$. However, it is not a consistent estimator of θ.
- A consistent estimator may not be an (asymptotic) unbiased estimator.
  - E.g., define $\hat{\theta}$ as $\mathrm{P}(\hat{\theta} = 0) = (n-1)/n$ and $\mathrm{P}(\hat{\theta} = n) = 1/n$, then $\hat{\theta}$ is a consistent estimator of $\theta = 0$. However, it is not an unbiased estimator since $\mathrm{E}(\hat{\theta}) = 1 \neq \theta = 0$.

5.2.1 Properties of Point Estimation

- Actually, if $\hat{\theta}$ is an asymptotic unbiased estimator of θ and $\mathrm{Var}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is a consistent estimator of θ.

Asymptotic unbiasedness + vanishing variance $\Longrightarrow$ consistency

Proof: The proof applies a famous inequality called the **Chebyshev's inequality (切比雪夫不等式)**.

Chebyshev's inequality

- Let X be a random variable with mean $\mu = \mathrm{E}(X)$ and variance $\sigma^2 = \mathrm{Var}(X)$ both exists, then

$$\mathrm{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Therefore, by the Chebyshev's inequality,

$$\mathrm{P}\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(\hat{\theta})}{\varepsilon^2} \to 0 \text{ as } n \to \infty.$$

- The inequality applies to any random variable with mean and variance defined, stating that a minimum of 75% (or 88.89%) of values must lie within 2 (or 3) standard deviation of the mean.

5.2.1 Properties of Point Estimation

- No matter what is the population distribution, if the population mean $\mu = \mathrm{E}(X)$ and population variance $\sigma^2 = \mathrm{Var}(X)$ exist, then $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = S^2$ are consistent estimators of μ and σ^2.

Proof: Based on the weak law of large numbers, we have

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mu \quad \text{as } n \to \infty.$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

$$= \frac{n}{n-1}\cdot\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \frac{n}{n-1}\cdot\bar{X}^2 \xrightarrow{\mathrm{P}} \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \mathrm{Var}(X) = \sigma^2$$

- After introducing the properties of point estimators, we turn to talk about two methods of deriving point estimators: method of moments (矩估计法) and method of maximum likelihood (极大似然估计法)

Chapter 5 Basic Concepts in Statistics

- 5.1 Population, Sample, Parameters and Statistics
- 5.2 Parameter Estimation – Point Estimation
  - 5.2.1 Properties of Point Estimation
  - 5.2.2 Method of Moments
  - 5.2.3 Method of Maximum Likelihood
- 5.3 Parameter Estimation – Confidence Interval

5.2.2 Method of Moments

- **Moments (矩)** are the expectations of powers of a random variable.

Moments

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$.
- The k-th **population moment (总体k阶矩)** and the k-th **sample moment (样本k阶矩)** are defined as

$$\mu_k = \mathrm{E}(X^k), \qquad M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k .$$

- The k-th **population central moment (总体k阶中心矩)** and **sample central moment (样本k阶中心矩)** are defined as ($k \geq 2$)

$$\tilde{\mu}_k = \mathrm{E}[(X - \mu_1)^k], \qquad \tilde{M}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k .$$

Actually, $\tilde{M}_2 = M_2 - M_1^2$

- The method of moments is based on a simple idea: since our sample comes from the distribution $f(x; \theta_1, \ldots, \theta_k)$, we choose the values of $\theta_1, \ldots, \theta_k$ such that the population moments match the sample moments.

5.2.2 Method of Moments

- The population moments can be expressed as functions of the parameters $\theta_1, \ldots, \theta_k$, and we set the population moments to equal to the sample moments:

$$\begin{cases} \mu_1 = h_1(\theta_1, \cdots, \theta_k) = M_1 \\ \cdots\cdots\cdots \\ \mu_k = h_k(\theta_1, \cdots, \theta_k) = M_k \end{cases}.$$

- Then, solving this system of equations, we obtain

$$\begin{cases} \theta_1 = g_1(M_1, \cdots, M_k) \\ \cdots\cdots\cdots \\ \theta_k = g_k(M_1, \cdots, M_k) \end{cases}.$$

Similar procedure applies to the central moments

- For $l = 1, \ldots, k$, $\hat{\theta}_l = g_l(M_1, \cdots, M_k) = \tilde{g}_l(X_1, \ldots, X_n)$ is called the **moment estimator (矩估计)** of θ_l.
- The procedure above of deriving the moment estimators seems to be complicated, but the functions $h_l(\,), g_l(\,)$ are typically simple in practical applications.

5.2.2 Method of Moments

- No matter what is the population distribution, if the population mean μ and population variance σ^2 exist, then the 1st sample moment $\hat{\mu} = M_1 = \bar{X}$ and the 2nd sample central moment $\hat{\sigma}^2 = \tilde{M}_2 = \tilde{S}^2 = (n-1)S^2/n$ are the moment estimators of μ and σ^2, respectively.
- The moment estimator of a parameter may not be unique.

Example 5.4 (Continued)

- The number of defects on a randomly selected chip $X \sim \text{Poisson}(\lambda)$. A simple random sample $X_1, \ldots, X_n$ is draw. Derive the moment estimator of λ.

Solution

5.2.2 Method of Moments

Example 5.7

- Suppose that the population X follows the following discrete distribution

Value	0	1	2	3
Probability	θ^2	$2\theta(1-\theta)$	θ^2	$1-2\theta$

- where $0 < \theta < 0.5$ is an unknown parameter.
- Now we have the observed values of a simple random sample from the population: 3, 1, 3, 0, 3, 1, 2, 3.
- Derive the moment estimator of θ and calculate its estimated value based on the sample.

Solution

Chapter 5 Basic Concepts in Statistics

- 5.1 Population, Sample, Parameters and Statistics
- 5.2 Parameter Estimation – Point Estimation
  - 5.2.1 Properties of Point Estimation
  - 5.2.2 Method of Moments
  - 5.2.3 Method of Maximum Likelihood
- 5.3 Parameter Estimation – Confidence Interval

5.2.3 Method of Maximum Likelihood

- The idea of the method of moments is very straightforward, however, it only uses the moments of a distribution, but not the form of the distribution.
- So, the method of moments may not make full use of all information about the parameter in the population distribution.
- The method of maximum likelihood directly uses the form of the population distribution.
- The central idea of this method is to find the parameter values that maximize the likelihood of observing the given data.
  - For a simple random sample $X_1, \dots, X_n$ from the population $X \sim f(x; \boldsymbol{\theta})$, the sample observed values are $x_1, \dots, x_n$.
  - With different values of $\boldsymbol{\theta}$, the likelihood of observing $X_1 = x_1, \dots, X_n = x_n$ are different.
  - We would estimate $\boldsymbol{\theta}$ with the values that maximize the likelihood of observing $X_1 = x_1, \dots, X_n = x_n$.
- **Question:** why use the word "likelihood", not "probability"?

5.2.3 Method of Maximum Likelihood

Likelihood Function and Maximum Likelihood Estimator (MLE)

- Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a simple random sample from the population $X \sim f(x; \boldsymbol{\theta})$ and the sample observed values are $\boldsymbol{x} = (x_1, \ldots, x_n)$. $\boldsymbol{\Theta}$ is the parameter space.
- The **likelihood function (似然函数)** is a function of $\boldsymbol{\theta}$, measuring the likelihood of observing $\boldsymbol{X} = \boldsymbol{x}$. "Likelihood" means "probability" for the discrete case, and "density" for the continuous case:

$$\mathcal{L}(\boldsymbol{\theta}; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

The joint PMF/PDF of $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$.

- If there exist $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n)$ such that

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}; x_1, \ldots, x_n),$$

Both the estimator and the estimate can be expressed as $\widehat{\boldsymbol{\theta}}$.

- then $\widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n)$ is the **maximum likelihood estimate (极大似然估计值)** of $\boldsymbol{\theta}$, the corresponding estimator $\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$ is the **maximum likelihood estimator (极大似然估计量)** of $\boldsymbol{\theta}$.
- Maximizing $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})$ is equivalent to maximizing the **log-likelihood function (对数似然函数)**:

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}).$$

5.2.3 Method of Maximum Likelihood

Example 5.7 (Continued)

- With the population distribution given below ($0 < \theta < 0.5$) and observed values of a simple random sample: 3, 1, 3, 0, 3, 1, 2, 3,

Value	0	1	2	3
Probability	θ^2	$2\theta(1-\theta)$	θ^2	$1-2\theta$

- Using the method of maximum likelihood, calculate the estimated value of θ based on the sample.

Solution

5.2.3 Method of Maximum Likelihood

Solution

- The general step of obtaining the MLE:

Construct the likelihood function and obtain the log-likelihood function → Take the first derivative of the log-likelihood and set it to zero, solve the equation

This may not work in some cases

5.2.3 Method of Maximum Likelihood

Example 5.8

- $X_1, \ldots, X_n$ is a simple random sample from the population $X \sim \mathrm{N}(\mu, \sigma^2)$. Derive the maximum likelihood estimators of the unknown parameters μ and σ^2.

Solution

5.2.3 Method of Maximum Likelihood

Example 5.9

- $X_1, \ldots, X_n$ is a simple random sample from the population $X \sim \mathrm{U}(0, \theta)$. Derive the maximum likelihood estimator of the unknown parameter θ.

Solution

5.2.3 Method of Maximum Likelihood

- The method of maximum likelihood is a versatile and powerful method in statistics and machine learning for parameter estimation, providing a framework that underpins many of the most common algorithms used in the field.
- Its application spans simple linear models to complex structures like neural networks and hidden Markov models, making it a critical tool in the arsenal of machine learning techniques.
- Under mild regularity conditions, the moment estimators and maximum likelihood estimators are **consistent and asymptotic unbiased estimators**.
- Under mild regularity conditions, for large samples, a maximum likelihood estimator has an approximately normal distribution. This property is called **asymptotic normality (渐近正态性)**.
- The asymptotic normality helps us construct interval estimation for a parameter.

Chapter 5 Basic Concepts in Statistics

- 5.1 Population, Sample, Parameters and Statistics
- 5.2 Parameter Estimation – Point Estimation
  - 5.2.1 Properties of Point Estimation
  - 5.2.2 Method of Moments
  - 5.2.3 Method of Maximum Likelihood
- 5.3 Parameter Estimation – Confidence Interval

5.3 Point Estimation – Confidence Interval

- Given the sample observed values, a point estimate provides a concrete estimated value of the parameter. However, the accuracy of this estimate is not provided by the point estimate itself.
- Interval estimation is proposed to address this limitation.
- Estimating the range of a quantity of interest is very common in daily life, e.g.:
  - During an investigation, the height of a suspect might be estimated to be between 175cm and 180cm.
  - The return rate of a financial product might be estimated to fall between 2.6% and 3.2%.

5.3 Point Estimation – Confidence Interval

- Naturally, we may want to estimate the range of a population parameter based on a sample.
- The general idea of **interval estimation (区间估计)** is
  - Find two statistics $\hat{\theta}_1(X_1, \ldots, X_n) \leq \hat{\theta}_2(X_1, \ldots, X_n)$ and use the random interval $(\hat{\theta}_1, \hat{\theta}_2)$ to estimate the range of θ.
  - Since $(\hat{\theta}_1, \hat{\theta}_2)$ is a random interval while θ is a fixed value, $(\hat{\theta}_1, \hat{\theta}_2)$ may not cover the true value of θ.
  - If the width of the interval is large, then it has higher probability of covering the true value of θ, but its precision is relatively low.
  - Therefore, we need to balance between the coverage probability and precision.
- The **confidence interval (CI, 置信区间)** is a widely used interval estimation proposed by Jerzy Neyman in 1937.

5.3 Point Estimation – Confidence Interval

Confidence Interval

- Let $X_1, \dots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \dots, \theta_k)$.
- For $\forall \alpha \in (0,1)$, if there exists two statistics $\hat{\theta}_{l1}(X_1, \dots, X_n)$ and $\hat{\theta}_{l2}(X_1, \dots, X_n)$ such that

$$P_{\boldsymbol{\theta}}\left(\hat{\theta}_{l1} < \theta_l < \hat{\theta}_{l2}\right) \geq 1 - \alpha, \forall \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

- then $\left(\hat{\theta}_{l1}, \hat{\theta}_{l2}\right)$ is called a confidence interval of θ_l with confidence level $1-\alpha$ (置信水平为$1-\alpha$的置信区间), or simply a $100(1-\alpha)\%$ confidence interval.
- $\hat{\theta}_{l1}$ and $\hat{\theta}_{l2}$ are called the confidence lower limit (置信下限) and confidence upper limit (置信上限), respectively.

- Commonly used values of α are 0.1, 0.05, 0.01, corresponding to 90%, 95%, 99% CIs.
- **Question:** Do you have any idea about how to construct a confidence interval?

5.3 Point Estimation – Confidence Interval

Example 5.10

- $X_1, \ldots, X_n$ is a simple random sample from $X \sim \mathrm{N}(\mu, \sigma^2)$, suppose that the value of σ^2 is known. Derive the $100(1-\alpha)\%$ confidence interval of the unknown parameter μ.

Solution

5.3 Point Estimation – Confidence Interval

Solution

5.3 Point Estimation – Confidence Interval

Example 5.10 (Continued)

- $X_1, \ldots, X_n$ is a simple random sample from $X \sim \mathrm{N}(\mu, \sigma^2)$, given $\sigma^2 = 1$ and sample observed values 4.6, 4.2, 5.0, 3.1, 3.4, 2.4, 4.4, 3.2, 3.9, 4.0, calculate the 95% confidence interval of μ.

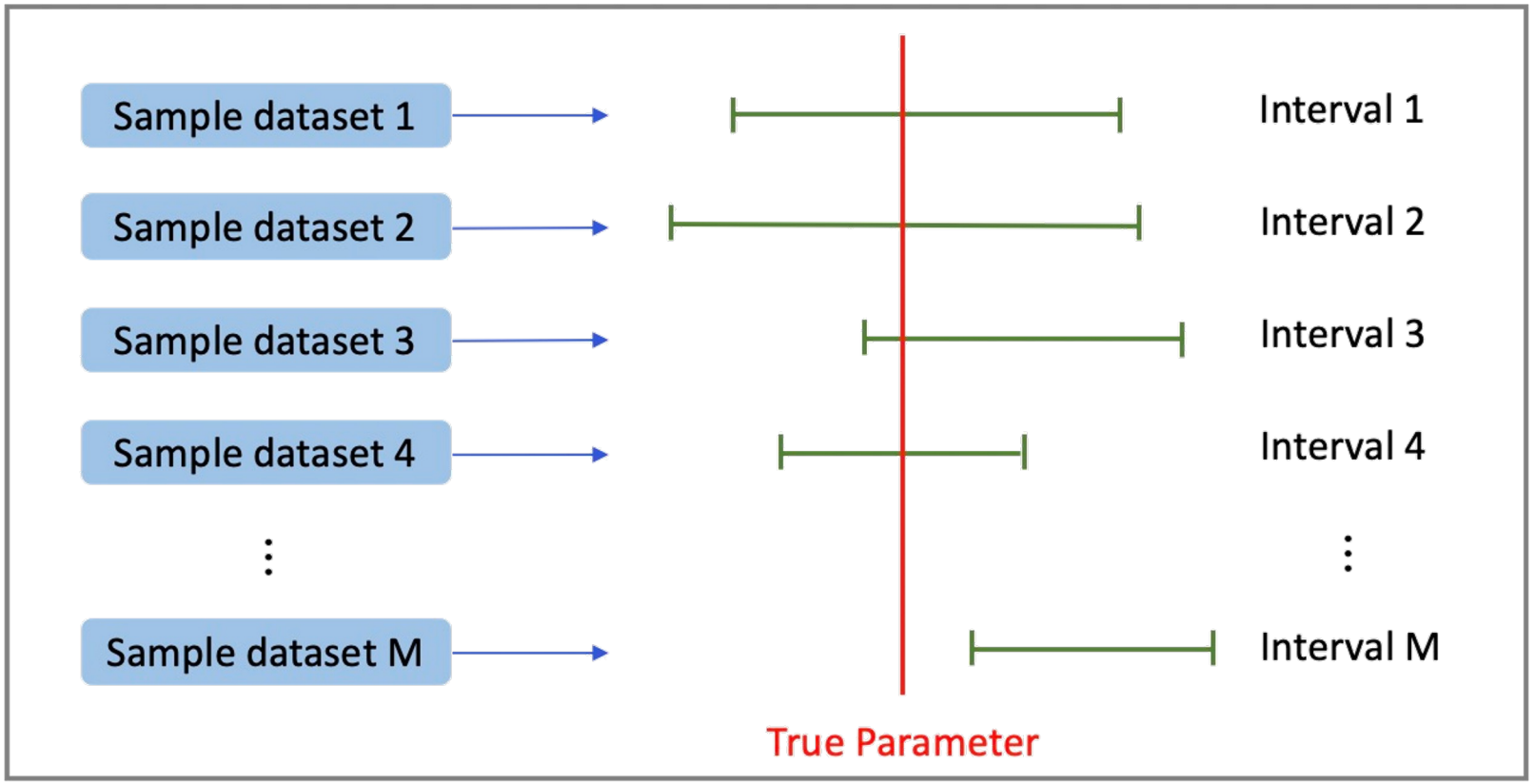
Solution

- **Question:** How to understand the 95% confidence interval (3.20, 4.44)?
  - The probability that (3.20, 4.44) covers the true value of μ is 0.95? ☒
  - 95% of the sample observed value fall into (3.20, 4.44)? ☒

(3.20, 4.44) either covers or fails to cover the true value of μ

5.3 Point Estimation – Confidence Interval

- 95% confidence interval $(\hat{\theta}_1, \hat{\theta}_2)$ is: in repeated sampling, 95% of all the 95% confidence intervals obtained will cover the true value of the parameter. ☑



5.3 Point Estimation – Confidence Interval

- In Example 5.10, we assumed that σ^2 is known. However, σ^2 is typically unknown in reality.
- Under this case, how to obtain the $100(1-\alpha)\%$ confidence interval of μ?

Example 5.10 (Continued)

- Derive the $100(1-\alpha)\%$ confidence interval of the unknown parameter μ if σ^2 is unknown.

Solution

5.3 Point Estimation – Confidence Interval

Solution

5.3 Point Estimation – Confidence Interval

- From Example 5.10, we summarize a general method for the construction of confidence intervals.

Construction of Confidence Intervals: A General Method

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$.
- $\hat{\theta}_l(X_1, \ldots, X_n)$ is an unbiased estimator of θ_l and the standard deviation of $\hat{\theta}_l$ is $\sigma(\hat{\theta}_l) = \mathrm{SD}(\hat{\theta}_l)$ (called the **standard error (标准误差)** of $\hat{\theta}_l$).
- If $\hat{\theta}_l$ exactly follows a normal distribution, i.e., $\hat{\theta}_l \sim \mathrm{N}\left(\theta_l, \sigma^2(\hat{\theta}_l)\right)$, and $\sigma^2(\hat{\theta}_l)$ does not depend on any unknown parameter, then an **exact** $100(1-\alpha)\%$ confidence interval of θ_l is

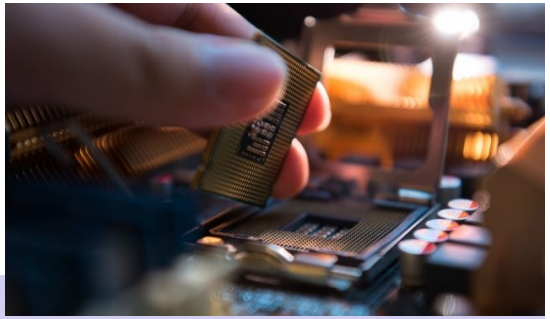
$$\hat{\theta}_l \pm z_{\alpha/2}\sigma(\hat{\theta}_l) = \left(\hat{\theta}_l - z_{\alpha/2}\sigma(\hat{\theta}_l), \hat{\theta}_l + z_{\alpha/2}\sigma(\hat{\theta}_l)\right).$$

- If $\hat{\theta}_l \overset{\text{approx.}}{\sim} \mathrm{N}\left(\theta_l, \sigma^2(\hat{\theta}_l)\right)$ or $\sigma^2(\hat{\theta}_l)$ depends on unknown parameters and $\hat{\sigma}^2(\hat{\theta}_l)$ is a consistent estimator of $\sigma^2(\hat{\theta}_l)$, then a **large sample** $100(1-\alpha)\%$ confidence interval of θ_l is

$$\hat{\theta}_l \pm z_{\alpha/2}\hat{\sigma}(\hat{\theta}_l) = \left(\hat{\theta}_l - z_{\alpha/2}\hat{\sigma}(\hat{\theta}_l), \hat{\theta}_l + z_{\alpha/2}\hat{\sigma}(\hat{\theta}_l)\right).$$

- Commonly used values: $z_{0.10} = 1.282$, $z_{0.05} = 1.645$, $z_{0.025} = 1.960$, $z_{0.01} = 2.326$, $z_{0.005} = 2.576$.

5.3 Point Estimation – Confidence Interval



Example 5.11

- A manager evaluates effectiveness of a major hardware upgrade by running a certain process independently 50 times before the upgrade and 50 times after it.
- Based on the data collected, the average running time is 8.5 minutes (with standard deviation of 1.8 minutes) before the upgrade and 7.2 minutes (with SD of 1.6 minutes) after it.
- 1. Construct a 90% CI of the mean running time of the process before the hardware upgrade.
- 2. Construct a 90% CI of the mean running time of the process after the hardware upgrade.
- 3. Construct a 90% CI showing how much the mean running time of the process changed due to the hardware upgrade.

5.3 Point Estimation – Confidence Interval

Solution

5.3 Point Estimation – Confidence Interval

Solution

5.3 Point Estimation – Confidence Interval

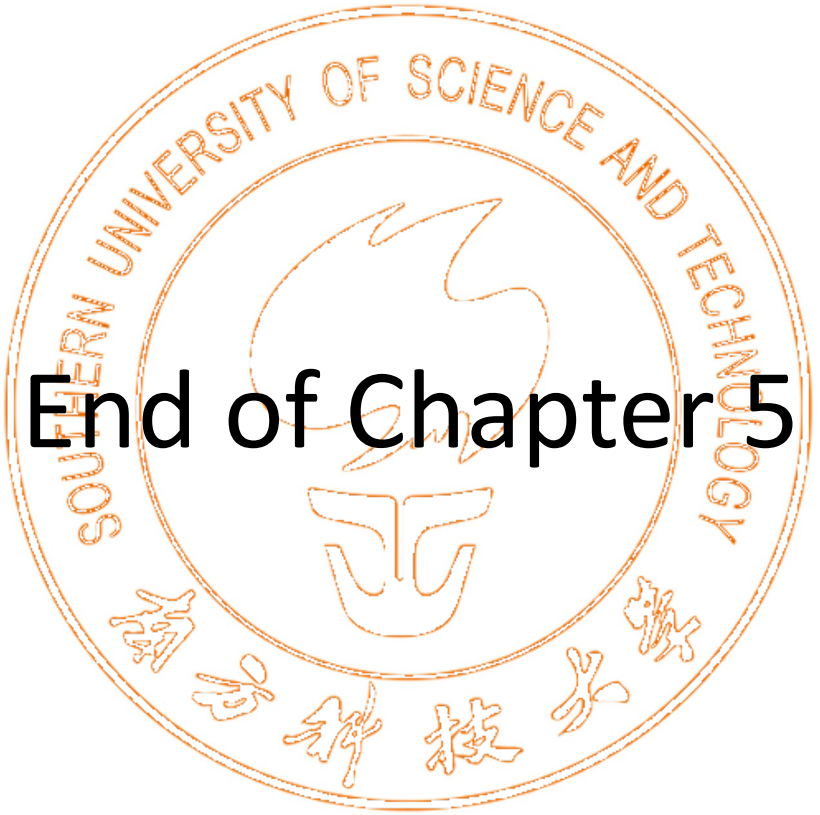
Example 5.12

- A candidate prepares for a local election. During his campaign, 42 out of 70 randomly selected residents in town A and 59 out of 100 randomly selected residents in town B showed they would vote for this candidate.
- 1. Obtain a 95% CI of the proportion of residents in town A who would support this candidate.
- 2. Estimate the difference in support that this candidate gets in towns A and B with 95% confidence.

Solution

5.3 Point Estimation – Confidence Interval

Solution



End of Chapter 5