Chapter 6 Statistics Principles and Sampling Distributions

§ 1 Introduction and Basic principles

(引言和基本概念)

§ 2 Sampling distributions

(抽样分布)

Chapter 6 Statistics Principles and Sampling Distributions

§ 1 Introduction and Basic principles

(引言和基本概念)

§ 2 Sampling distributions

(抽样分布)

What is mathematical statistics（数理统计）?

Mathematical statistics is based on the theory of Probability. It is a mathematical subject related to experimental data **collection, summarizing, analysis and inference**（数据的收集、整理、分析和推断）.

Background

Example: A factory just produced a large batch of products. n products were picked randomly, and m defective products were found. How to evaluate the probability of defective products in those products?

Example: The requirement of some kind of components is that the expected lifespan should be no less than 1000 hours. 25 components were picked randomly, and the average lifespan was found to be 950 hours. Do those components meet the requirement?

Contents of statistical inference

- Point estimation (点估计)
- Interval estimation (区间估计)
- Variance analysis (方差分析)
- Nonparametric hypothesis testing (非参数假设检验)
- Parametric hypothesis testing (参数假设检验)
- Regression analysis (回归分析)

What is experimental data?

The data obtained by carrying out scientific experiments or observing some phenomenon are called **experimental data**.

Feature: Data are affected by random factors.

How to process experimental data?

Collect, summarize, analyze, make inference
（数据的收集、整理、分析和做出推断）

Mathematical statistics is to study these four processes.

How to collect and summarize experimental data?

The mathematical meaning of "collecting and summarizing" data will be discussed in this chapter.

The principles of mathematical statistics

Population (总体): The overall of the research objects is called the Population.

Individual (个体): A specific object from the population is called an individual.

Example: To analyze the English scores of a class. All students in the class is the population and each student in the class is an individual.

Example: To analyze the lifespan of a batch of light bulbs. The batch is the population and each light bulb is an individual.

The above identification of **Population** and **Individual** doesn't fit well into the characteristics of mathematical researches as mathematical researches emphasize more on **abstraction** rather than **specific objects**.

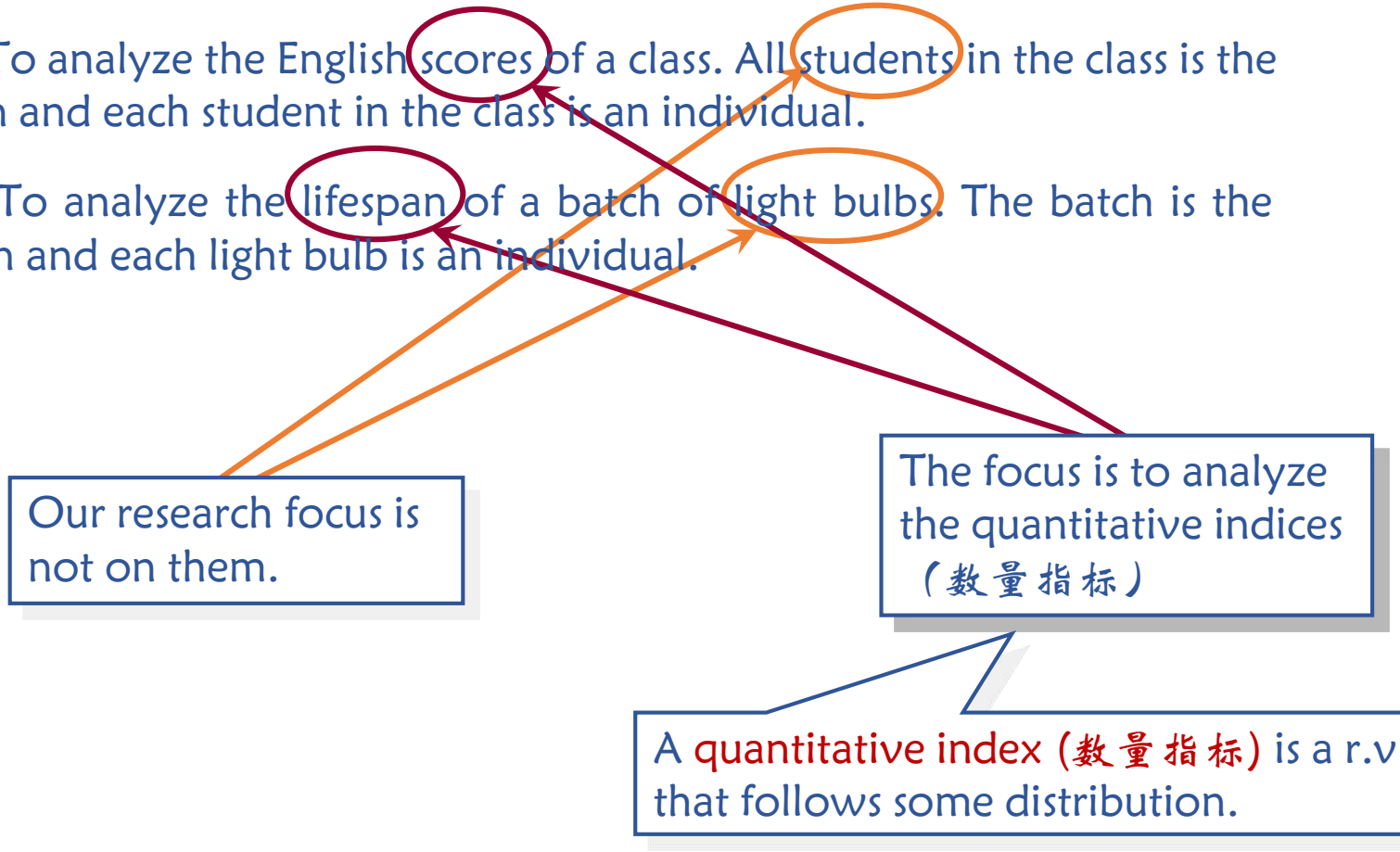
The principles of mathematical statistics

Population (总体): The overall of the research objects is called the population.

Individual (个体): A specific object from the population is called an individual.

Example: To analyze the English scores of a class. All students in the class is the population and each student in the class is an individual.

Example: To analyze the lifespan of a batch of light bulbs. The batch is the population and each light bulb is an individual.



The principles of mathematical statistics

Population (总体): The overall of the research objects is called the population.

Individual (个体): A specific object from the population is called an individual.

Example: To analyze the English scores of a class. All students in the class is the population and each student in the class is an individual.

Example: To analyze the lifespan of a batch of light bulbs. The batch is the population and each light bulb is an individual.

Population: The **quantitative index (数量指标)** $X \sim F(x)$ of research objects.

Individual: A **value (值)** of r.v. X

Example: To analyze the English score X of a class. Since the scores of all students in the class fluctuate around the average score μ. Thus, the population follows

$$X\sim N(\mu,\sigma^2)$$

Example: To analyze the lifespan X of a batch of light bulbs. Since the lifespans of all light bulbs in the batch fluctuate around the average lifespan μ. Thus, the population is

$$X\sim N(\mu,\sigma^2)$$

Exercise: To analyze the probability of defective components from a factory, denote

$$X=\begin{cases}0, & \text{qualified component}\\ 1, & \text{defective component}\end{cases}$$

Then the population is $X\sim b(1,p)$, p is the probability of defective components.

How to collect data ?

Pick n "individuals" from the research objects, to analyse their Quantitative index

$$X_1, X_2, \cdots, X_n$$

The process is called sampling (抽样), $X_1, X_2, \cdots, X_n$ are called a sample (样本) with size (容量) n.

Features of sampling (抽样的特征)

Repeatedly and independently observe X n times under the same condition:

- **Independency (独立性):** Every observation does not affect each other.
- **Representative (代表性):** Every observation has the same distribution as the population.

Features of a sample (样本的特征)

- Before observing: $X_1, X_2, \cdots, X_n$ are i.i.d. r.v.s, and have the same distribution as the population.
- After observing: $x_1, x_2, \cdots, x_n$ are the n specific observed values for $X_1, X_2, \cdots, X_n$.

Sample duality (样本二重性)

Sample observed values (样本观测值)

Example: There is a large batch of light bulbs. Sample 5 of them for testing. The lifespans (days) of them are: 980, 960, 1030, 1300, 850

Analysis: The population is $X \sim N(\mu, \sigma^2)$

Continuous population

The samples are X_1, X_2, X_3, X_4, X_5

The sample observed values are 980, 960, 1030, 1300, 850

} Sample duality (样本二重性)

The sample size (样本容量) is 5.

Exercise: To measure the length of a workpiece 6 times. The measurements (cm) are

29.1, 30.2, 29.3, 29.1, 30.3, 29.5

Analysis: The population is the measurements of the workpiece

$$X \sim N(\mu, \sigma^2)$$

The sample size is 6

The samples are $X_1, X_2, X_3, X_4, X_5, X_6$

The sample observations are 29.1, 30.2, 29.3, 29.1, 30.3, 29.5

} Sample duality

Example: To analyze the quality of products from a factory. Sample 100 products randomly from a batch of products. If a product is qualified, denote the result as 0. If it is defective, denote the result as 1.

Analysis: The population is $X \sim b(1, p)$, representing whether a component is qualified or defective

Discrete population

The population frequency function is

$$P\{X = 1\} = p, \qquad P\{X = 0\} = 1 - p$$

where p is the probability of defective component.

Duality

Sample $X_1, X_2, \cdots, X_{100}$
(they are independent and identically 0-1 distributed random variables)

Sample observed values $x_1, x_2, \cdots, x_{100}$

Understanding of Samples

If $X_1, X_2, \ldots, X_n$ are samples from the population $X \sim F(x)$

1. $X_1, X_2, \cdots, X_n$ are a set of "disordered" data.
2. $X_1, X_2, \cdots, X_n$ contain "information" about the population.
3. $X_1, X_2, \cdots, X_n$ are the basis to infer the population.
4. Before observing, $X_1, X_2, \ldots, X_n$ are i.i.d. random variables.
   After observing, $x_1, x_2, \ldots, x_n$ are the observed values, i.e., specific data.

Distribution of Sample

The samples $X_1, X_2, \ldots, X_n$ can be considered as a ***n*-dimensional random variable**, what distribution does it follow?

1. If the distribution function of the population is $F(x)$, the **joint distribution function** of $(X_1, X_2, \cdots, X_n)$ is

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$$

2. If the **density function** of the population is $f(x)$, the **joint density function** of $(X_1, X_2, \cdots, X_n)$ is

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Exercise: If $X_1, X_2, \ldots, X_n$ are samples from the population $X \sim N(\mu, \sigma^2)$, the joint density function of the sample is

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

A **n-dimensional** normal distribution

Exercise: Assume that $X_1, X_2, \ldots, X_n$ are samples from the population $X \sim b(1, p)(0 < p < 1)$. What is the joint sample distribution?

Answer: The frequency function of the population is

$$P\{X = 1\} = p, \qquad P\{X = 0\} = 1 - p$$

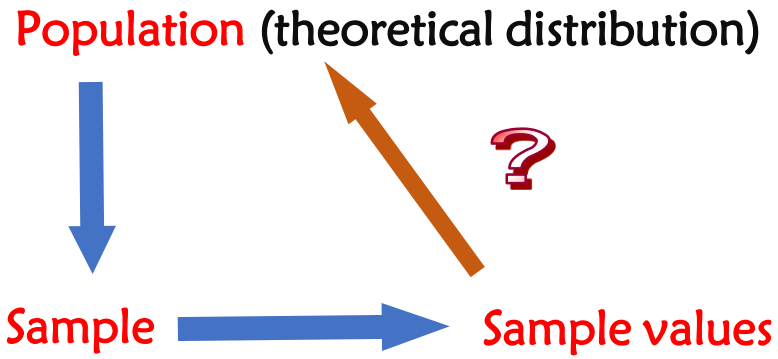
$X_1, X_2, \ldots, X_n$ are the independent and identically distributed r.v.s and each of them follows the (0-1) distribution. Thus, the joint distribution is:

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

Relationship among Population, Sample, Sample value

The information that we get by sampling is the specific observed values. If we pick 10 students to get their height, we get 10 values. The values are sample observed values not samples. We can only observe the values of random variables but not random variables itself.



Statistics is to use the obtained **sample values** to infer the features of the population distribution $F(x)$.

Sample is the bridge that connects sample values and population.

The population distribution decides the probability regularity of samples. It is the regularity of observed values for samples. Therefore, sample values can be used to infer population.

The foundation of statistical inference is data collection

To do sampling from the population $X \sim F(x)$:

$X_1, X_2, \cdots, X_n$

"disordered" data

Contains useful "information"

How to summarize the sample and extract useful information ?

Example: A class had a math exam. The scores of n students are $X_1, X_2, \ldots, X_n$. How to evaluate the study of math for the whole class?

Analysis: The following values could properly represent the math status of the class

$$\frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \max_{1\le i\le n} X_i, \qquad \min_{1\le i\le n} X_i$$

To get useful information by setting up these functions.

The foundation of statistical inference is data collection

To do sampling from the population $X \sim F(x)$:

$$X_1, X_2, \cdots, X_n$$

"Good" statistics can extract useful information from the data.

Summarization of data: Statistics (统计量)

If $X_1, X_2, \ldots, X_n$ are samples from population $X \sim F(x)$ and $g(X_1, X_2, \ldots, X_n)$ is a n-variate function. If r.v. $g(X_1, X_2, \ldots, X_n)$ does not contain any unknown parameter, then $g(X_1, X_2, \ldots, X_n)$ is called a statistic.

Example: if $X_1, X_2, \ldots, X_n$ are samples from population $X \sim N(\mu, \sigma^2)$, and μ, σ^2 are unknown. Which of the following is/are statistic(s)?

$$\frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2, \qquad \frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i}{\sigma}\right)^2, \qquad \max_{1 \le i \le n} X_i$$

Why do we require no unknown parameter in statistics?

The foundation of statistical inference is data collection

To do sampling from the population $X \sim F(x)$:

$$X_1, X_2, \cdots, X_n$$

Summarization of data: Statistics (统计量)

If $X_1, X_2, \ldots, X_n$ are a sample from population $X \sim F(x)$ and $g(x_1, x_2, \ldots, x_n)$ is an n-variate function. If r.v. $g(X_1, X_2, \ldots, X_n)$ does not contain any unknown parameter, then $g(X_1, X_2, \ldots, X_n)$ is called a statistic.

Duality of statistics (统计量)

1. Before trials, $g(X_1, X_2, \ldots, X_n)$ is a random variable.
2. After trials, $g(x_1, x_2, \ldots, x_n)$ is a specific value.

Commonly used statistics

Sample mean
(样本均值)

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

Why not $\frac{1}{n}$?
(talk about this later)

Sample variance
(样本方差)

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

Sample standard deviation
(样本标准差)

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Sample moment of order k
(样本k阶矩)

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad (k = 1, 2, \cdots)$$

Sample central moment of order k
(样本k阶中心矩)

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \quad (k = 1, 2, \cdots)$$

Commonly used statistics

Sort $X_1, X_2, \ldots, X_n$ in ascending order: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$

Order statistic (顺序统计量) $\quad (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$

Minimum value (极小值) $\quad X_{(1)} = \min\{X_1, X_2, \cdots, X_n\}$

Maximum value (极大值) $\quad X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$

Properties of Sample Moment (样本矩)

Assume that $X_1, X_2, \ldots, X_n$ are samples from population $X \sim F(x)$, the population moments of order k (总体k阶矩)

$$\mu_k \triangleq E\left(X^k\right) (k = 1, 2, \cdots).$$

all exist.

① $\because X_1, X_2, \cdots X_n$ are i.i.d. with population X

$\therefore X_1^k, X_2^k, \cdots, X_n^k$ are i.i.d. with population X^k

② $$E(A_k) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \frac{1}{n}\sum_{i=1}^{n} E(X^k) = \mu_k \ \ (k = 1, 2, \cdots)$$

③ By the Khinchine's law of large numbers (辛钦大数定律), for $\forall k = 1, 2, \cdots$

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k\,, \qquad n \to \infty$$

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right\} = 1$$

$$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k), \qquad n \to \infty$$

where g is a continuous function.

Numerical Characteristics of sample mean and sample variance

Assume that the mean and variance of population X as follows do exist:

$$E(X) \triangleq \mu, \qquad D(X) \triangleq \sigma^2.$$

$X_1, X_2, \ldots, X_n$ are samples from population X, then

$$E(\bar{X}) = \mu, \qquad D(\bar{X}) = \frac{\sigma^2}{n}, \qquad E(S^2) = \sigma^2$$

Proof:

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

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

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

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

$2n(\mu - \bar{X})^2$

Numerical Characteristics of sample mean and sample variance

Assume that the mean and variance of population X as follows do exist:

$$E(X) \triangleq \mu, \qquad D(X) \triangleq \sigma^2.$$

$X_1, X_2, \ldots, X_n$ are a sample from population X, then

$$E(\bar{X}) = \mu, \qquad D(\bar{X}) = \frac{\sigma^2}{n}, \qquad E(S^2) = \sigma^2$$

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

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

$$= \sum_{i=1}^{n} \sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2$$

$$\therefore E(S^2) = \sigma^2$$

Summary

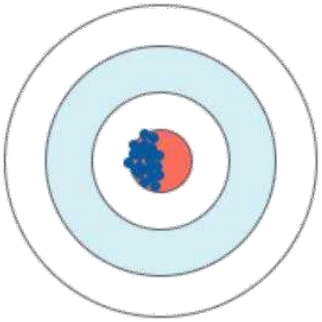
The meaning of Sample Mean and Sample Variance

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ —— is the average of all the data $X_1, X_2, \cdots, X_n$

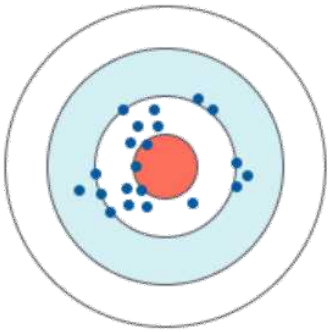
is the center of data $X_1, X_2, \cdots, X_n$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ —— shows the degree of deviation (偏离程度) between the data $X_1, X_2, \ldots, X_n$ and its center. It represents the deviation degree of overall data.

Small variance



Large variance



What do $E(\bar{X}) = \mu$, $D(\bar{X}) = \sigma^2/n$, $E(S^2) = \sigma^2$ suggest?

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

谢谢大家