PROBABILITY AND STATISTICS

CHAPTER 4: CONFIDENCE INTERVALS

Dr. Phan Thi Huong

HoChiMinh City University of Technology
Faculty of Applied Science, Department of Applied Mathematics
Email: huongphan@hcmut.edu.vn

HCM city — 2021.

OUTLINE

1 INTRODUCTION TO STATISTICS

OUTLINE

1. INTRODUCTION TO STATISTICS

2. USING STATISTICS TO SUMMARIZE DATA SETS.

OUTLINE

1. INTRODUCTION TO STATISTICS

2. USING STATISTICS TO SUMMARIZE DATA SETS.

3. CONFIDENCE INTERVALS FOR PARAMETERS OF NORMAL DISTRIBUTION.

OUTLINE

1. INTRODUCTION TO STATISTICS
2. USING STATISTICS TO SUMMARIZE DATA SETS.
3. CONFIDENCE INTERVALS FOR PARAMETERS OF NORMAL DISTRIBUTION.
4. CONFIDENCE INTERVALS FOR OTHER DISTRIBUTIONS

The Science of Data

Definition 1.1

The field of statistics deals with the collection, presentation, analysis, and use of data to make decisions, solve problems, and design products and processes. In simple terms, statistics is the **science of data**. Statistical methods are used to help us describe and understand **variability**.

Population vs. Sample

Definition 1.2

- A population is a collection of objects, items, humans/animals about which information is sought.
- A sample is a part of the population that is observed.
- A parameter is a numerical characteristic of a population, e.g. Vietnamese unemployment rate.
- A statistic is a numerical function of the sampled data, used to estimate an unknown parameter,
  e.g., unemployment rate in a sample.

DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS.

DEFINITION 1.3

- The part of statistics concerned with the description and summarization of data is called descriptive statistics.
- The part of statistics concerned with the drawing of conclusions from data is called inferential statistics.

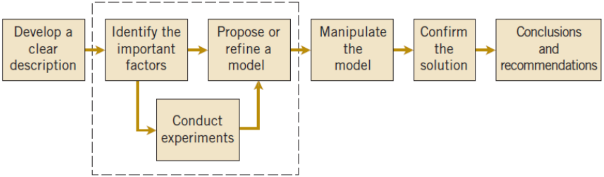

Develop a clear description
Identify the important factors
Propose or refine a model
Conduct experiments
Manipulate the model
Confirm the solution
Conclusions and recommendations

Figure 1-1 The engineering method

SOME USEFUL STATISTICS

- The sample mean is a point estimate of the population mean, which is defined by

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

- The sample variance, denoted by s^2, is used to approximate the population variance σ^2

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$$

s is called the sample standard deviation.

SOME USEFUL STATISTICS

- Once the data values have been listed in order from smallest to largest, the sample median is the middle value in the list, and it divides the list into two equal parts.
- The process of determining the median:
  - When n is odd: the sample median is the single middle value.
  - When n is even: there are two middle values in the ordered list, and we average these two middle values to obtain the sample median.

SOME USEFUL STATISTICS

- sapmle $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these values.
- To find the sample $100p$ percentile of a data set of size n:
    1. Arrange the data in **increasing order**.
    2. If np is **not an integer**, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.
    3. If np is **an integer**, then the average of the values in positions np and $np+1$ is the sample $100p$ percentile.

SOME USEFUL STATISTICS

PROPOSITION 2.1

- The sample 25% is called the first quartile , denoted by $Q1$.
- The sample 50% is called the second quartile (median) , denoted by $Q2$.
- The sample 75% is called the third quartile , denoted by $Q3$.

EXAMPLE 2.1

A sample of scores from a league bowling tournament is given as below.

$126, 150, 177, 133, 149, 157, 162, 188, 166, 175, 177, 122, 183, 199, 212$

(A) Compute the sample mean and sample standard deviation.

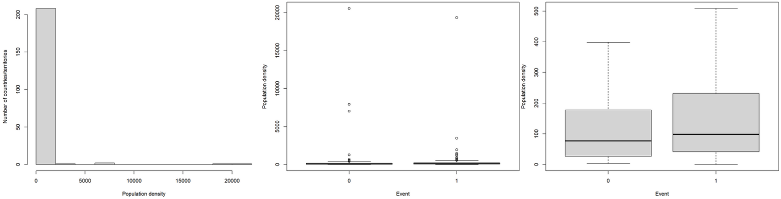
(B) Find the sample quartiles.

(C) Find outliers (if there is any).

OULIERS

DEFINITION 2.1

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.



INDENTIFY THE OUTLIERS BY TUKEY FENCES METHOD

1. Compute Interquartile Range : $IQR = Q3 - Q1$, where $Q1$ is first quartile and $Q3$ third quartile.
2. Outliers are values

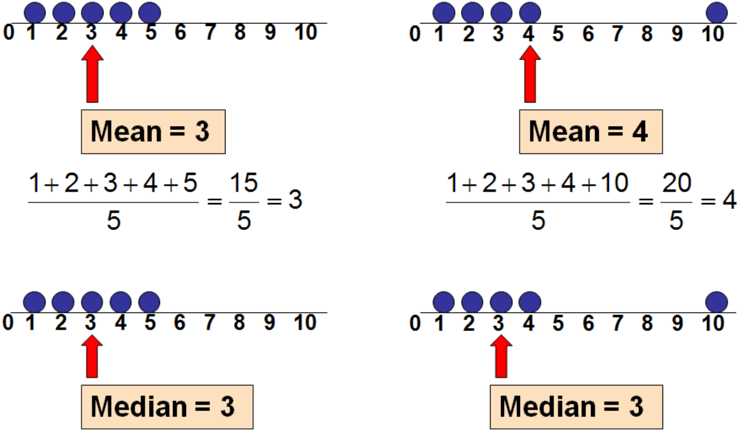
$$x > Q_3 + 1.5IQR \qquad \text{or} \qquad x < Q_1 - 1.5IQR.$$

3. Extreme outlier are values

$$x > Q_3 + 3IQR \qquad \text{or} \qquad x < Q_1 - 3IQR$$

Sample Median vs sample mean

- The median is more robust to outliers.



Mean = 3

$$\frac{1+2+3+4+5}{5}=\frac{15}{5}=3$$

Mean = 4

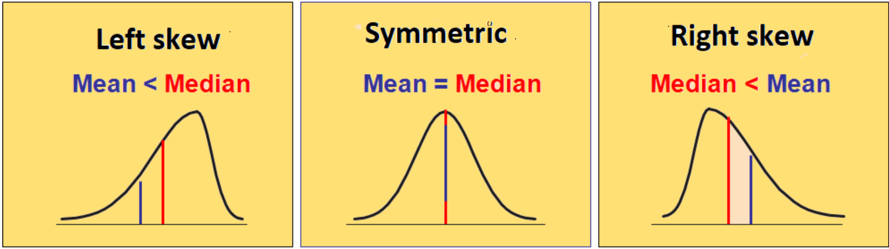
$$\frac{1+2+3+4+10}{5}=\frac{20}{5}=4$$

Median = 3

Median = 3

THE SAMPLE MEDIAN

- Mean and median can tell us the shape of the distribution.



Point estimator vs Point estimate

Definition 2.2

- If X is a random variable with probability distribution $f(x)$, characterized by the unknown parameter θ, and if X_1, X_2, ..., X_n is a random sample of size n from X, the statistic $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$ is called a point estimator of θ.
- Note that $\hat{\Theta}$ is a random variable because it is a function of random variable.
- After the sample has been selected, $\hat{\Theta}$ takes on a particular numerical value $\hat{\theta}$ called the point estimate of θ.

Note: Sample statistics: sample mean, sample variance, sample median, ... are point estimate.

Generally, the point estimate says nothing about how close it is to the true parameter.

Generally, the point estimate says nothing about how close it is to the true parameter.

A way to avoid this is to report the estimate in terms of a range of plausible values called a confidence interval.

Confidence intervals

Definition 2.3

A $100(1-\alpha)\%$ confidence interval for θ is defined by the interval $[l, u]$ such that

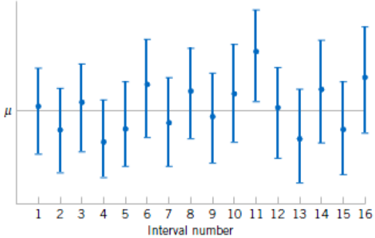
$$\boxed{\mathbb{P}\left(L \le \theta \le U\right) = 1 - \alpha}$$

where l and u are respectively two values of statistics L and U calculated from samples.

- α is called the significance level that we allow ourselves to be wrong when we are estimating a parameter with a confidence interval.
- $\gamma = 1 - \alpha$ is called the confidence level (γ) that is measure of the degree of reliability of the interval.

INTERPRETATION OF CONFIDENCE INTERVALS

If a finite number of random samples are collected and a $100(1-\alpha)\%$ confidence interval for θ is computed from each sample, then $100(1-\alpha)\%$ of these intervals will contain the true value of θ.



ONE-SIDED CONFIDENCE INTERVALS

DEFINITION 2.4 (LEFT-SIDED CONFIDENCE INTERVALS/LIMITS)

- Let S_1 be a statistic: for all values of θ, such that

$$P(S_1 < \theta) = 1 - \alpha$$

(s_1, ∞) is called a left-sided $100(1-\alpha)$ percent CI for θ.

- Let S_2 be a statistic: for all values of θ, such that

$$P(\theta < S_2) = 1 - \alpha$$

$(-\infty, s_2)$ is called a right-sided $100(1-\alpha)$ percent CI for θ.

General formulas of confidence intervals

- General formulas of two-sided confidence intervals:

$$\boxed{\text{point estimate} \pm \text{(reliability factor)(standard divation)}}$$

- General formulas of left-sided confidence intervals:

$$\boxed{\text{point estimate} - \text{(reliability factor)(standard divation)}}$$

- General formulas of right-sided confidence intervals:

$$\boxed{\text{point estimate} + \text{(reliability factor)(standard divation)}}$$

Normal Population + Known σ

CI of population mean

If $X_1, \ldots, X_n$ are iid $\sim N(\mu, \sigma^2)$, $\alpha = 1 - \gamma$ and σ are given, a $100(1-\alpha)\%$ confidence interval of the population mean is

$$\boxed{\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} + \bar{x}},$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution, e.i, $\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$.

Normal Population + Known σ

CI of population mean

If $X_1, \ldots, X_n$ are iid $\sim N(\mu, \sigma^2)$, $\alpha = 1 - \gamma$ and σ are given, a $100(1-\alpha)\%$ confidence interval of the population mean is

$$\boxed{\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} + \bar{x}},$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution, e.i, $\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$.

Sample size

Let $E = \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha/2}$ indicate the error in the confidence interval estimation and ϵ be a specified error. Then $E \leq \epsilon \iff n \geq \left(\frac{\sigma \cdot z_{\alpha/2}}{\epsilon}\right)^2$.

ONE-SIDED CONFIDENCE INTERVAL (NORMAL POPULATION + KNOWN σ)

- A $100(1-\alpha)\%$ upper-confidence bound for μ is

$$\mu \le \bar{x} + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}.$$

- A $100(1-\alpha)\%$ lower-confidence bound for μ is

$$\mu \ge \bar{x} - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}.$$

EXAMPLE 1

In auto racing, a pit stop is where a racing vehicle stops for new tires, fuel, repairs, and other mechanical adjustments. The efficiency of a pit crew that makes these adjustments can affect the outcome of a race. A random sample of 32 pit stop times has a sample mean of 12.9 seconds. Assume that the population distribution is normal and the population standard deviation is 0.19 second.

A. Construct a 99% confidence interval for the mean pit stop time.

B. How many observations must be collected to ensure that the radius of the 99% CI is at most 0.01?

C. Construct a 95% left-sided confidence interval for the mean pit stop time.

STUDENT DISTRIBUTION, $T \sim t_\nu$

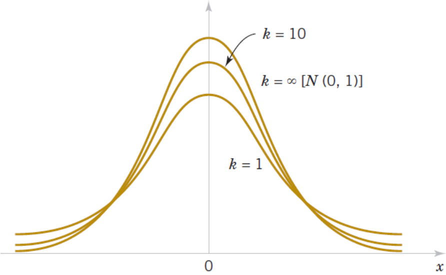
THEOREM 3.1

If $X_1, \dots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$, then the random variable

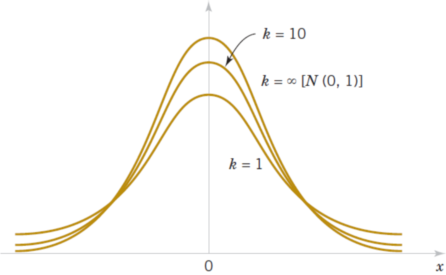
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} is a Student distribution with $n-1$ degrees of freedom.

Probability density functions of several t-distributions.

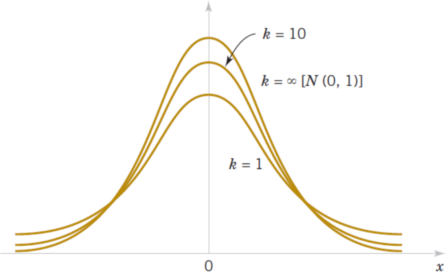


Probability density functions of several t-distributions.



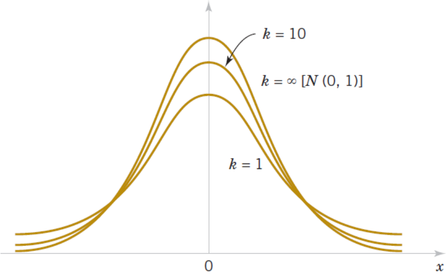
- The general appearance of t-distribution is similar to the standard normal distribution.

Probability density functions of several t-distributions.



- The general appearance of t-distribution is similar to the standard normal distribution.
- t-distribution has heavier tails than the normal.

Probability density functions of several t-distributions.



- The general appearance of t-distribution is similar to the standard normal distribution.
- t-distribution has heavier tails than the normal.
- As the number of degrees of freedom $k \to \infty$, the limiting form of the t-distribution is the standard normal distribution.

Normal Population + Unknown σ

CI of the population mean

If $X_1, \ldots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$ then a $100(1-\alpha)\%$ confidence interval of the population mean is given by

$$\bar{x} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} + \bar{x},$$

where $t_{n-1,\alpha/2}$ is the upper $100\alpha/2$ percentage point of the t-distribution with $n-1$ degrees of freedom, e.i, $\mathbb{P}(T \geq t_{n-1,\alpha/2}) = \alpha/2$.

NORMAL POPULATION + UNKNOWN σ

ONE-SIDED CI OF THE POPULATION MEAN

The $100(1-\alpha)\%$ upper and lower confidence interval of the population mean is respectively given by

$$\mu \leq t_{n-1,\alpha} \cdot \frac{s}{\sqrt{n}} + \bar{x},$$

and

$$\mu \geq \bar{x} - t_{n-1,\alpha} \cdot \frac{s}{\sqrt{n}}$$

where $t_{n-1,\alpha}$ is the upper 100α percentage point of the t-distribution with $n-1$ degrees of freedom.

Normal Population + Unknown σ

Example 2

Eight independent measurements of the point of inflammation of Diesel oil gave the values (in F)

144 147 146 144 142 150 143 141

Assuming normality, determine a 99% confidence interval for the mean.

Large Sample Size

Theorem 4.1

Let $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. If n is large,

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \simeq N(0,1); \quad \text{and,} \quad \frac{\bar{x} - \mu}{s/\sqrt{n}} \simeq N(0,1)$$

CI of population mean - Large sample size

Let $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. If n is large,

Two sided CI: $\boxed{\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \lessapprox \mu \lessapprox \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}}$,

Left-sided CI: $\boxed{\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \lessapprox \mu}$,

Right-sided CI: $\boxed{\mu \lessapprox \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}}$.

Note: If σ is known, we use $sigma$ instead of s.

Large Sample Size

Example 3

A random sample of 110 lighting flashes in a region resulted in a sample average radar echo duration of 0.81s and a sample standard deviation 0.34s. Calculate a 99% (two-sided) CI for the true average echo duration.

POPULATION PROPORTION

CORRALLY 4.1

Suppose that we are interested in the proportion of event $\mathscr{A}$ in a population, denoted by p. Let X be the number of event $\mathscr{A}$ in a sample of size n and assume n is enough large, then

$$\frac{\hat{p}-p}{\sqrt{pq/n}} \simeq N(0,1)$$

where $\hat{p} = \dfrac{X}{n}$ is the sample proportion.

Population Proportion

CIs for population proportion

An approximate $100(1-\alpha)\%$ confidence interval for p is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \lessapprox \mu \lessapprox z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \hat{p}$$

The approximate $100(1-\alpha)\%$ lower and upper confidence bounds are

$$p \gtrsim \hat{p} - z_{\alpha} \cdot \frac{\sqrt{\hat{p}\,(1-\hat{p})}}{\sqrt{n}}$$

and

$$p \lesssim \hat{p} + z_{\alpha} \cdot \frac{\sqrt{\hat{p}\,(1-\hat{p})}}{\sqrt{n}}$$

respectively.

CIS FOR POPULATION PROPORTION

- The error (radius) of $100(1-\alpha)\%$ CIs for population proportion:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

- Since $p(1-p) \le 0.25, \quad 0 < p < 1,$

$$E \le \epsilon \Leftrightarrow n \le \left(\frac{z_{\alpha/2}}{\epsilon}\right)^2 \times 0.25.$$

where ϵ is a specified error.

Population Proportion

Example 4

An article reported that in $n = 45$ trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Let p denote the long-run proportion of all such trials that would result in ignition.

(A) Estimate the 95% confidence interval for p.

(B) How many trials are needed to achieve a maximum error of 0.01 in the calculated CI?