Statistics and Data Analysis

Statistical Estimation

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Road map

- **Statistical estimation**.
- Estimating population mean with known variance.
- Estimating population mean with unknown variance.

Example: average daily consumers

- A retail chain of 3000 stores is going to have a special discount on the next Monday.
  - In the past, the average daily number of consumers on Monday was 700.
  - The marketing manager promises that the average will be above 850 with the discount.
  - The manager wants to know the **average number of daily consumers** entering the stores on that day.
- She decides to do a survey on the next Monday.
  - On that day, there will be some consumers entering each store.
  - For store i, $i = 1, ..., 3000$, let x_i be the number of consumers.
  - It is too costly to collect all x_is and calculate $\mu = \frac{\sum_{i=1}^{3000} x_i}{3000}$.
  - This is a task of **estimating** a **parameter**.
- Her budget is enough for hiring 7 temporary workers to count the number of consumers throughout the day.
  - She decides to randomly draw 7 stores and calculate $\bar{x} = \frac{\sum_{i=1}^{7} x_i}{7}$.
  - We assume that the daily demands of all stores follow the same (population) distribution.

Example: average daily consumers

- On that day, she gets the following sample data:
  - She gets 1026, 932, 852, 1212, 844, 822, and 1032 consumers.
  - The **sample mean** is $\bar{x} = 960$.
- Intuitively, she will think that the population mean μ is "around" 960.
- Suppose she concludes that "μ is within 950 and 970," how much confidence may she have?
- In general, is it okay to conclude that $\mu \in [\bar{x} - 10, \bar{x} + 10]$?

Estimations

- One of the most important statistical tasks is **estimation**.
  - For unknown population **parameters**, we estimate them through **statistics** obtained from samples.
  - For example, when the population mean is unknown, we use sample mean as an estimate.
- We want to go beyond intuitions and conjectures.
  - We need some knowledge about the **sampling distributions**.
  - E.g., we know $\overline{X} \sim \text{ND}(\mu, \frac{\sigma}{\sqrt{n}})$.
- In statistics, we use **confidence intervals** to estimate parameters.
- We will introduce how to estimate the population mean.
  - Estimating other parameters basically follows the same logic.

Notation and terminology

- We have the **population mean** and **sample mean**.
  - The population mean is fixed but unknown.
  - E.g., the average daily demand of the 3000 stores.
  - The sample mean is random.
  - E.g., the average daily demand of the 7 randomly selected stores.
- The population mean is denoted as μ.
- The sample mean is denoted as $\overline{X}$ and $\bar{x}$:
  - Before we observe the outcome, the sample mean is **random** and denoted as $\overline{X}$.
  - After we observe the outcome, the **realized value** of the sample mean is fixed and denoted as $\bar{x}$.
  - $\overline{X}$ is a random variable; $\bar{x}$ is a realized value.

Road map

- Interval estimation.
- **Estimating population mean with known variance**.
- Estimating population mean with unknown variance.

Drawbacks of point estimation

- We may use the sample mean $\bar{x}$ to estimate the population mean μ.
  - "μ should somewhat be close to $\bar{x}$."
  - This is called a **point estimation**.
- However, there are some drawbacks of point estimation:
  - We know that μ is close to $\bar{x}$. But **how close**?
  - More precisely, what is $|\mu - \bar{x}|$?
  - As μ is unknown, we will never know the answer!
- Instead of suggesting a number, we will suggest an **interval**.
  - Then we measure how good the suggested interval is.
  - More precisely, we measure **how likely** the interval contains μ.

Interval estimation: the first illustration

- Consider a population with unknown μ. For simplicity, let's assume:
  - The population variance σ^2 is **known**.
  - The population follows a **normal** distribution.
- Let the sample mean $\overline{X}$ be the **estimator**.
  - $\overline{X}$ as an estimator is random; $\bar{x}$ as a realized value is a constant.
- Suppose that $\sigma^2 = 16$ and the sample size $n = 8$.
- Based on $\overline{X}$, we will choose a **leg length** b and claim that μ lies in the **interval** $[\overline{X} - b, \overline{X} + b]$.
  - We may be either right or wrong.
  - When b increases, we are more confident that we will be right.
  - However, a larger interval means that the estimation is less accurate.
  - What is the **probability** that we are right?

The sampling distribution

- Question: For any given t, find

$$\Pr(\overline{X} - b \leq \mu \leq \overline{X} + b).$$

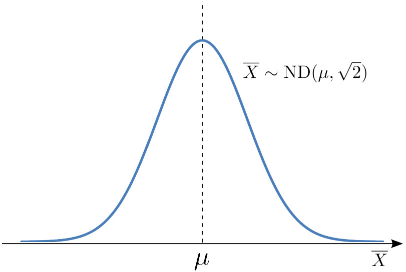
- As the population is normal:

$$\overline{X} \sim \text{ND}\left(\mu, \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{8}} = \sqrt{2}\right).$$

- Suppose someone proposes to set $b = \sqrt{2}$, then the interval will be

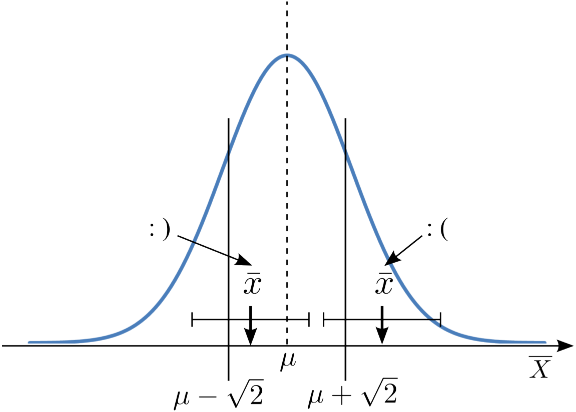
$$\left[\overline{X} - \sqrt{2}, \overline{X} + \sqrt{2}\right].$$

How good the interval is?



How good an interval is?

- If, luckily, $\bar{x}$ is close enough to μ, $[\bar{x} - \sqrt{2}, \bar{x} + \sqrt{2}]$ covers μ.
- If, unluckily, $\bar{x}$ is far from μ, $[\bar{x} - \sqrt{2}, \bar{x} + \sqrt{2}]$ does not cover μ.

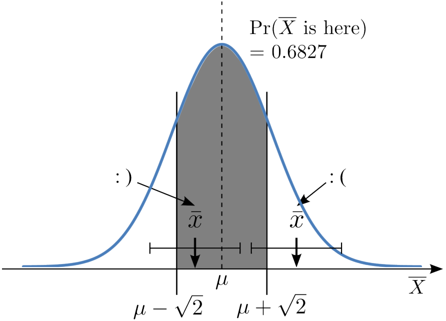


How good an interval is?

- The probability that "we are lucky" can be calculated!
- No matter where μ is, we have

$$\begin{aligned} &\Pr\left(\overline{X} - \sqrt{2} \leq \mu \leq \overline{X} + \sqrt{2}\right) \\ =&\Pr\left(\mu - \sqrt{2} \leq \overline{X} \leq \mu + \sqrt{2}\right) \\ =&0.6827. \end{aligned}$$

- To calculate this, we rely on the fact that $\overline{X} \sim \text{ND}(\mu, \sqrt{2})$.
- This is the probability for a normal random variable to be within **one standard deviation** from its mean.

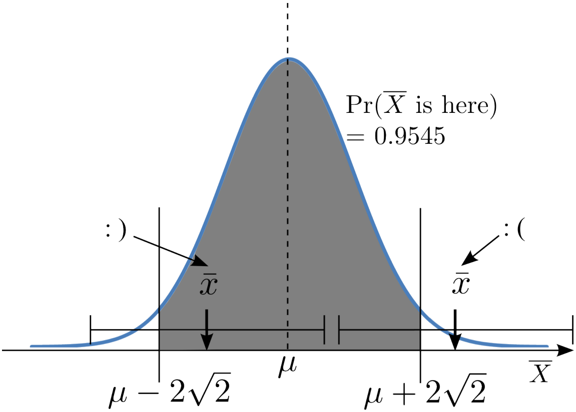


A short summary

- Given **any** realization $\bar{x}$, $[\bar{x} - \sqrt{2}, \bar{x} + \sqrt{2}]$ may or may not covers μ.
- Regarding the random $\overline{X}$, we know $[\overline{X} - \sqrt{2}, \overline{X} + \sqrt{2}]$ covers μ with probability 0.6827.
  - This level of confidence can be calculated as we know $\overline{X} \sim \text{ND}(\mu, \sqrt{2})$.
- The calculation obviously depends on $\frac{\sigma}{\sqrt{n}}$.
  - This quantity $\frac{\sigma}{\sqrt{n}}$ is called the **standard error** of the estimation.
- Instead of having $\sqrt{2}$ as the leg length, let's try $2\sqrt{2}$.

A larger interval

- The probability that "we are lucky" now becomes 0.9545!
  - $\Pr\left(\overline{X} - 2\sqrt{2} \leq \mu \leq \overline{X} + 2\sqrt{2}\right) = \Pr\left(\mu - 2\sqrt{2} \leq \overline{X} \leq \mu + 2\sqrt{2}\right) = 0.9545.$

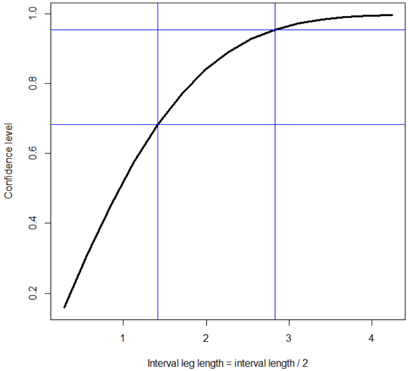


Confidence levels and confidence intervals

- We made two attempts:
  - $\left[\overline{X} - \sqrt{2}, \overline{X} + \sqrt{2}\right]$ results in a covering probability 0.6827.
  - $\left[\overline{X} - 2\sqrt{2}, \overline{X} + 2\sqrt{2}\right]$ results in another covering probability 0.9545.
- In statistics, when we do interval estimation:
  - Such a "covering probability" is called **confidence level**.
  - These intervals are called **confidence intervals** (CI).
- How to choose the interval length?
  - A larger confidence interval results in a higher confidence.
  - There is a **trade-off** between accurate estimation and high confidence.

Confidence levels vs. interval lengths

- To find the relationship:
  - $\Pr(\mu - \sqrt{2} \leq \overline{X} \leq \mu + \sqrt{2}) = 0.68$. $\Pr(\mu - 2\sqrt{2} \leq \overline{X} \leq \mu + 2\sqrt{2}) = 0.95$.
  - Given $b > 0$, we calculate $1 - 2\Pr(\overline{X} \leq \mu - b)$ based on $\overline{X} \sim \text{ND}(\mu, \frac{\sigma}{\sqrt{n}})$.



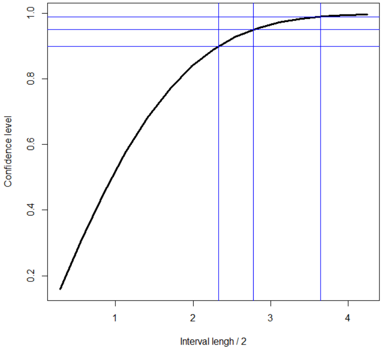
How to choose the interval length?

- In practice, we choose a confidence level first and then the smallest interval that achieves this level.
  - We typically denote the error probability as α.
  - The confidence level is thus $1-\alpha$.
  - Common confidence levels: 90%, 95%, and 99%.
- How to calculate the leg length b?
  - 90%: $1 - 2\Pr(\overline{X} \leq \mu - b) = 0.9$, i.e.,

$$\Pr(\overline{X} \leq \mu - b) = 0.05.$$

  - For a given α, find b such that

$$\Pr(\overline{X} \leq \mu - b) = \frac{\alpha}{2}.$$



Example revisited: average daily consumers

- Recall that we have 3000 stores, each with a number of consumers on a given day.
  - The population consists of 3000 numbers.
  - There is a population mean μ, which is unknown.
- We collected data from 7 stores:
  - The sample data: 1026, 932, 852, 1212, 844, 822, and 1032.
  - The realized sample mean is $\bar{x} = 960$.
- How to do interval estimation with this sample?

Conducting the estimation

- We must know the population variance σ^2.
  - Let's assume that $\sigma = 120$.
- We need either the population is normal or the sample size is large.
  - Let's assume that the population is normal.
- Now we are ready to construct a confidence interval. Let's construct three intervals for $1 - \alpha = 0.9$, 0.95, and 0.99.
  - Step 1: $\bar{x} = 960$.
  - Step 2: The standard deviation of the sample mean is $\frac{\sigma}{\sqrt{n}} = 45.356$.
  - Step 3: The leg lengths are 74.604, 88.896, and 116.829.
  - Step 4: The interval with 90% confidence level is

$$[960 - 74.604, 960 + 74.604] = [885.39, 1034.60].$$

  The other two intervals are [871.10, 1048.90] and [843.17, 1076.82].

Interpreting the estimation

- Consider the interval with 95% confidence level: $[871.10, 1048.90]$.
  - The realized sample mean is $\bar{x} = 960$. The leg length is 88.896.
- What is the business implication?
  - We will claim that the true average daily consumers for all the 3000 stores is within 870 and 1050.
  - We are 95% confident. It is quite unlikely for us to be wrong.
- Recall that the marketing manager has promised that "the average daily consumers will be at least 850."
  - Now we have a strong evidence showing that the target is really achieved.
  - We are 95% confident that this is achieved.
  - Note that the 99% confidence interval is $[843.17, 1076.82]$.
  - We are not 99% confident.
- We will never be 100% confident. However, we now are able to measure how confident we are.

Summary

- Facing an unknown population mean μ (with a known population variance σ^2), we may construct a confidence interval:
  - Centered at the to-be-realized sample mean $\overline{X}$.
  - Will cover μ with a predetermined probability.
- Use the desired confidence level $1 - \alpha$ and the standard error $\frac{\sigma}{\sqrt{n}}$ to calculate the leg length b.
  - Our "plan" is to suggest the interval $[\overline{X} - b, \overline{X} + b]$.
  - Our suggested interval is $[\bar{x} - b, \bar{x} + b]$.
- We need one of the following:
  - The population follows a normal distribution.
  - The sample size $n \geq 30$.

Road map

- Interval estimation.
- Estimating population mean with known variance.
- **Estimating population mean with unknown variance**.

Estimation without the population variance

- Sometimes (actually for most of the time) we **do not** know the population variance σ^2.
- Then we cannot calculate the standard error $\frac{\sigma}{\sqrt{n}}$.
- In this case, intuitively we may try to replace σ by s, the **sample standard deviation**.
  - As an example, for the 7 numbers of consumers 1026, 932, 852, 1212, 844, 822, and 1032, we have

$$s = \sqrt{\frac{(1026-960)^2 + \cdots + (1032-960)^2}{7-1}} = 140.233.$$

  - We then use $\frac{s}{\sqrt{n}}$ to construct an interval.
  - However, $\overline{X} \sim \text{ND}(\mu, \frac{s}{\sqrt{n}})$ is not right!
  - In particular, s can vary from sample to sample.
- We need some adjustments.

The t distribution

- Let S be the sample standard deviation (which is random before sampling) and s be its realization.
- When we replace σ by S, we rely on the following fact:

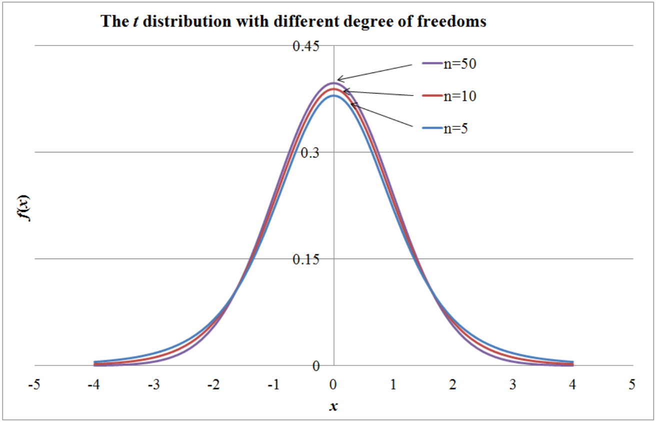
Proposition 1

For a normal population, the quantity $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ follows the t distribution with degree of freedom $n-1$.

  - We know the sampling distribution of T_{n-1} (when the population is normal). We call it **the t distribution**.
  - Its probability density function is known (but we do not care about it). Relevant probabilities may be calculated with software.
  - The only parameter is the **degree of freedom**, which is $n-1$.
  - If X follows a t distribution with degree of freedom $n-1$, we denote this as $X \sim t(n-1)$.

The t distributions

- The t distribution is **symmetric**, **centered at 0**, and **bell-shaped**.
- When n goes up, it approaches the **standard normal distribution**.



Applying the t distribution

- Before sampling, we know we will get the sample mean $\overline{X}$ and sample standard deviation S.
- For any b, we construct an interval $[\overline{X} - b, \overline{X} + b]$. We want to know $\Pr(\overline{X} - b \leq \mu \leq \overline{X} + b)$.
- Now we do not know the distribution of $\overline{X}$; we only know the distribution of $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$. Therefore:

$$\begin{aligned}
&\Pr\left(\overline{X} - b \leq \mu \leq \overline{X} + b\right) = \Pr\left(\mu - b \leq \overline{X} \leq \mu + b\right) \\
=&\Pr\left(\frac{-b}{S/\sqrt{n}} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq \frac{b}{S/\sqrt{n}}\right) = \Pr\left(\frac{-b}{S/\sqrt{n}} \leq T \leq \frac{b}{S/\sqrt{n}}\right).
\end{aligned}$$

- Once we obtain s, we may calculate the probability.

Applying the t distribution

- Consider the example of estimating average daily consumers again.
- Suppose we do not know the population variance σ^2.
  - We know $\bar{x} = 960$ and $s = 140.233$.
- Suppose we propose the interval $[860, 1060]$ with $b = 100$.
  - We calculate

$$\Pr\left(\frac{-b}{S/\sqrt{n}} \leq T_6 \leq \frac{b}{S/\sqrt{n}}\right) = \Pr\left(\frac{-100}{140.233/\sqrt{7}} \leq T_6 \leq \frac{100}{140.233/\sqrt{7}}\right)$$
$$= \Pr(-1.887 \leq T_6 \leq 1.887) = 0.892,$$

    where the last step can be done with any statistical software.
- We are 89.2% confident that the average number of daily consumers lies within 860 and 1060.

From a confidence level to an interval

- How to construct an interval $[\overline{X} - b, \overline{X} + b]$ for us to be 95% confident?
- We have the t distribution; given any value t, we know $\Pr(T_{n-1} \leq t)$.
  - When the degree of freedom is 6, $\Pr(T_{n-1} \leq -2.447) = 0.025$.
  - Statistical software can help us find 2.447.
- Moreover, we have

$$\Pr(T_{n-1} \leq t) = \Pr\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t\right) = \Pr\left(\mu \geq \overline{X} - t\frac{S}{\sqrt{n}}\right).$$

- The leg length is calculated to be $-t\frac{s}{\sqrt{n}} = 2.447 \times \frac{140.233}{\sqrt{7}} = 129.694$.
  - The multiplier $\frac{s}{\sqrt{n}}$ will always be used.
- The desired interval is

$$[960 - 129.694, 960 + 129.694] = [885.40, 1034.60].$$

Finding a confidence interval

- If σ is known, given $\bar{x}$, n, and α, we construct the confidence interval in the following steps:
  - We know $\overline{X} \sim \text{ND}(\mu, \frac{\sigma}{\sqrt{n}})$, i.e., $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \text{ND}(0,1)$.
  - Step 1: Calculate the multiplier $\frac{\sigma}{\sqrt{n}}$.
  - Step 2: Calculate the **critical value** z^* such that $\Pr(Z \leq -z^*) = \frac{\alpha}{2}$.
  - Step 3: The product of the critical z^* and multiplier $\frac{\sigma}{\sqrt{n}}$ is the leg length.
  - Step 4: The interval is $[\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}]$.
- If σ is unknown, given $\bar{x}$, s, n, and α, we construct the confidence interval in the following steps:
  - We know $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.
  - Step 1: Calculate the multiplier $\frac{s}{\sqrt{n}}$.
  - Step 2: Calculate the **critical value** t^* such that $\Pr(T_{n-1} \leq -t^*) = \frac{\alpha}{2}$.
  - Step 3: The product of the critical t^* and multiplier $\frac{s}{\sqrt{n}}$ is the leg length.
  - Step 4: The interval is $[\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}]$.

Remarks

- If the population is normal, the sample size n does not matter.
  - We may use the t distribution anyway.
- If the population is **non-normal** and the sample size is large ($n \geq 30$):
  - The population is non-normal, so we cannot use the t distribution.
  - The sample size is large, so according to the **central limit theorem**, the sample mean is normal.
  - For $n \geq 30$, $t(n-1)$ is very close to $\text{ND}(0,1)$.
  - Using the t distribution as an approximation is acceptable.
- If the population is non-normal and the sample size is small ($n < 30$), using t distribution for estimation is inaccurate.
  - However, the t distribution for estimating the population mean is **robust** to the normal population assumption: Having nonnormal population does not harm a lot.
  - We still suggest one not to use the t distribution in this case.

Summary

- To estimate the population mean μ:

σ^2	Sample size	Population distribution: Normal	Population distribution: Nonnormal
Known	$n \geq 30$	z	z
	$n < 30$	z	Nonparametric
Unknown	$n \geq 30$	t (or z)	t (or z)
	$n < 30$	t	Nonparametric

  - Nonparametric methods are beyond the scope of this course.