Summary---Topic 5: Confidence Interval Estimation

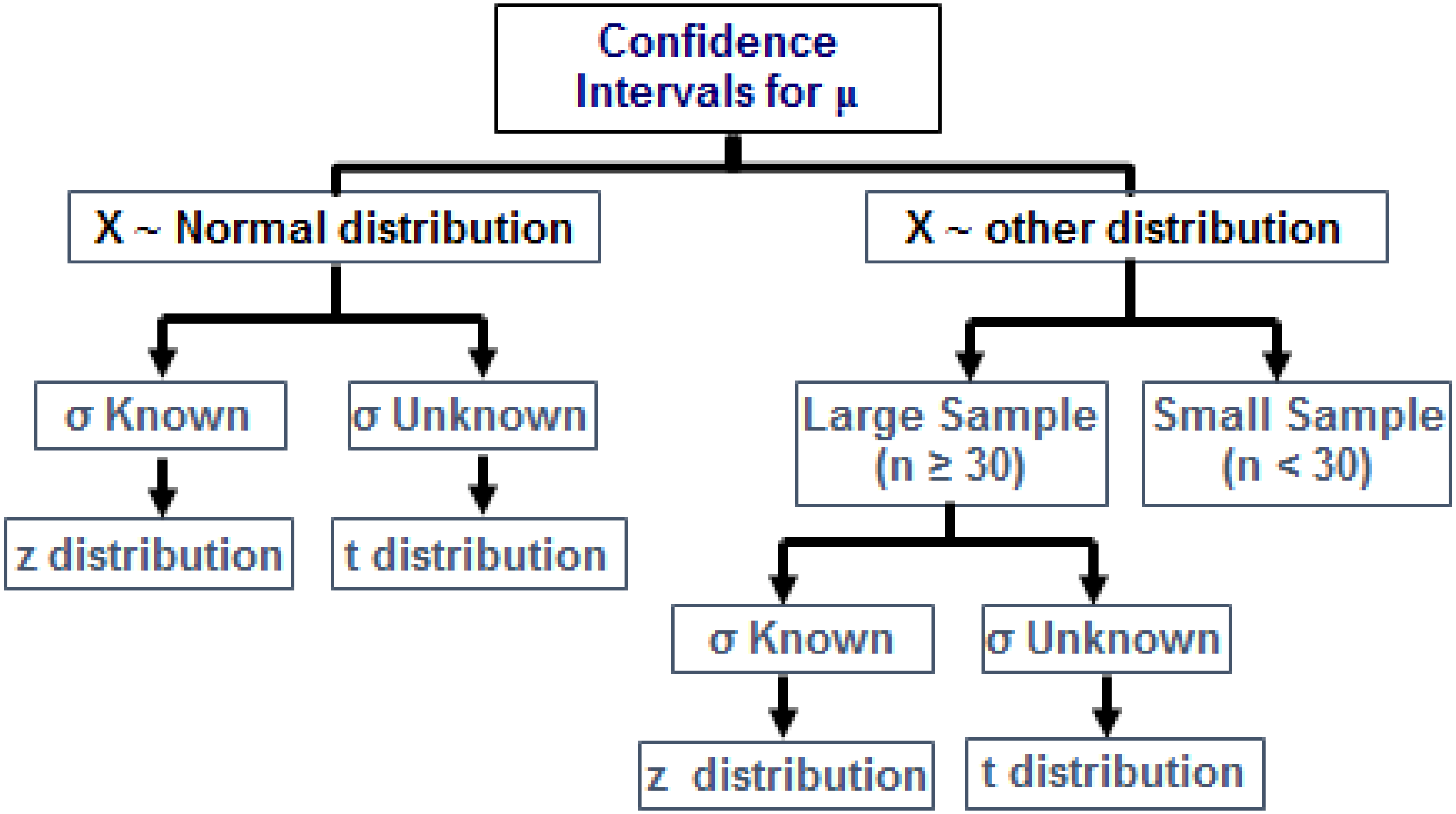
Point Estimation

	Mean	**Variance**	**Proportion**
Population parameter (Population size: N)	μ	σ^2	π
Sample statistics (Sample size: n)	$\bar{X}$	S^2	p

- Without given the information of how likely the sample statistics can truly estimate the population parameter

Interval Estimation for μ

- Confidence interval is an interval estimate of a parameter
- Confidence interval is constructed by data obtained from one sample and a specific confidence level of the estimate



Confidence Interval (CI) Construction

Case	Population Distribution	Sample Size	Sampling Distribution of Mean	Assumption Requirement	Confidence Interval
1	Normal	Whatever	Normal	Nil	- σ known, use Z distribution $$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$ - σ unknown, use t distribution $$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$
2	Unknown / not normal	n ≥ 30	By central limit theorem, the sample mean distribution is approximately normal.	Nil	- σ known, use Z distribution $$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$ - σ unknown, use t distribution $$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$
3	Unknown	n < 30	Not normal	Assume population distribution is normal so that the sample mean also follows normal distribution.	- σ known, use Z distribution $$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$ - σ unknown, use t distribution $$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Degree of Freedom & t Distribution

Degree of freedom

- The number of values that are free to vary in the final calculation of a statistic
- Equal to the **number of observations** used in the analysis **minus** the **number of parameter estimated** during the intermediate steps in calculating the statistic itself
- Before calculating the sample variance, we have to estimate the sample mean, so one degree of freedom is lost (i.e. the degree of freedom of sample variance is n-1)

t distribution: $T \sim t(\nu)$

- Symmetric about mean = 0
- Similar in shape to the standard Normal curve (Bell-shaped)
- Have more area in the tails and less in the center than the Z distribution
- Different degree of freedom will have different t distribution
- Degree of freedom increases, t distribution approaches the Z distribution

Interpretation of Confidence Interval

- For instance → $[\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}]$

1. We are 100 (1- α) % confident that the population mean is between

$$\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \text{ and } \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

2. If all possible samples of size n are taken and the corresponding 100 (1- α) % confidence intervals are constructed, 100 (1- α) % of these intervals include the unknown population mean.

Standard Error and Sampling Error

- Standard error of $\bar{X}$: $\frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$
- Sampling error (E): $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (for Z-distribution, σ is known)

  or $t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$ (for t-distribution, σ is unknown)
- Sampling error increases *(magnitude of sampling error increases)* when

  → 100 (1- α) % ↑ (i.e. α ↓)

  → n ↓

Trade-off between sampling error and confidence level: when you want to obtain a lower sampling error with the same sample size, you have to forgo higher confidence.

- Confidence interval becomes wider when

  → 100 (1- α) % ↑ (i.e. α ↓)

  → n ↓

Determining Sample Size

- n = $(\frac{Z_{\alpha/2}\sigma}{E})^2$, where E = Sampling error

- If only sample standard deviation is given, use

  n = $(\frac{t_{\alpha/2,n-1}S}{E})^2$

- If n is not an integer, remember to round it up to the nearest integer (not round down, as rounding down the sample size will cause the sampling error larger than the desired sampling error).

Exercises and Solutions

Q1. If $\bar{X}$ =120, σ =24 and n=36, construct a 99% confidence interval estimate of the population mean μ.

To construct confidence interval:

1. Check whether the sampling distribution of $\bar{X}$ is normal

- Population distribution is normal→$\bar{X}$ is normally distributed
- Sample size ≥ 30 →$\bar{X}$ is approximately normally distributed(CLT)

If yes, go to step 2;

otherwise, assume the population distribution is normal, and go to step 2.

2. Population standard deviation σ is known or unknown?

- Know → use Z distribution: $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- Unknown → use t distribution: $\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$

Solution:

Since n = 36 > 30 from unknown population distribution, we may use the Central Limit Theorem to conclude that the sampling distribution of $\bar{X}$ is approximately normal. σ is known, so we use the Z-distribution.

Since $\alpha = 1 - 0.99 = 0.01$, $\quad Z_{\alpha/2} = Z_{0.005} = 2.575$,

99% confidence interval for $\mu = \bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 120 \pm 2.575 \times \left(\frac{24}{\sqrt{36}}\right) = [109.7.130.3]$.

We are 99% confident that the population mean is between 109.7 and 130.33.

$Z_{\alpha/2}$ for three most commonly used confidence levels

Confidence Level	90%	95%	99%
α	0.1	0.05	0.01
$Z_{\alpha/2}$	1.644854	1.959964	2.575829

NORM.INV(lower tail probability, mean, s.d.)

$Z_{\alpha/2}$ =NORM.INV(1- α/2, 0,1)
=-NORM.INV(α/2, 0,1)

Q2. A stationery store wants to estimate the mean retail value of greeting cards that it has in its inventory. A random sample of 100 greeting cards indicates a mean value of \$2.65 and a standard deviation of \$0.44.

Assuming a normal distribution, construct a 95% confidence interval estimate of the mean value of all greeting cards in the store's inventory.

Solution:

$$\mathrm{n} = 100, \qquad \bar{X} = 2.65, \qquad \mathrm{s} = 0.44, \qquad \alpha = 1 - 0.95 = 0.05$$

Since population of the retail value of greeting cards is normally distributed, we can conclude that the sampling distribution of $\bar{X}$ is also normally distributed.

σ is unknown, so we use the t-distribution.

95% confidence interval for $\mu = \bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} = 2.65 \pm t_{0.025,99} \frac{0.44}{\sqrt{100}}$

$= 2.65 \pm 1.9842 \frac{0.44}{\sqrt{100}}$

$= [2.5627, 2.7373]$ in dollar

=T.INV(lower tail probability, df)
=T.INV(1-α/2, n-1)
=-T.INV(α/2, n-1)

We are 95% confident that the true population mean value of all greeting cards in the store's inventory is between \$2.5627 and \$2.7373.

Q3*. The U.S. Department of Transportation requires tire manufacturers to provide tire performance information on the sidewall of the tire to better inform prospective customers when making purchasing decisions. One very important measure of tire performance is the tread wear index, which indicates the tire's resistance to tread wear compared with a tire graded with a base of 100. This means that a tire with a grade of 200 should last twice as long, on average, as a tire graded with a base of 100. A consumer organization wants to estimate the actual tread wear index of a brand name of tires that claims "graded 200" on the sidewall of the tire. A random sample of $n=18$ indicates a sample mean tread wear index of 195.3 and a sample standard deviation of 21.4.

a) Assuming that the population of tread wear indexes is **normally distributed**, construct a 95% confidence interval estimate of the population mean tread wear index for tires produced by this manufacturer under this brand name.

b) Do you think that the consumer organization should accuse the manufacturer of producing tires that do not meet the performance information provided on the sidewall of the tire? Explain.

c) Explain why an observed tread wear index of 210 for a particular tire is not unusual, even though it is outside the confidence interval developed in (a).

Q3*. Solution:

a) Since population of tread wear indexes is normally distributed, we can conclude that the sampling distribution of of $\bar{X}$ is also normal. σ is unknown, so we use the t-distribution.

95% confidence interval for $\mu = \bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 195.3 \pm t_{0.025,17}\frac{21.4}{\sqrt{18}}$

$$= 195.3 \pm 2.1098 \times \frac{21.4}{\sqrt{18}}$$

$$= [184.6581{,}205.9419]$$

We are 95% confident that the population mean tread wear index for tires produced by this manufacturer under this brand name is between 184.6581 and 205.9419.

b) No, a grade of 200 is in the interval

c) By comparing this value to the sample mean and sample standard deviation: It is not unusual. A tread-wear index of 210 for a particular tire is only (210-195.3)/21.4 = 0.69 standard deviation above the sample mean of 195.3

Q4. An advertising agency that serves a major radio station wants to estimate the mean amount of time that the station's audience spends listening to the radio on a daily basis. From past studies, the standard deviation is estimated as 45 minutes. Assume the population has the normal distribution.

a) What sample size is needed if the agency wants to be 90% confident of being correct to within $\pm$ 10 minutes?

b) If 99% confidence is desired, how many listeners need to be selected?

Solution:

a) Since population distribution is normal, the sampling distribution of of $\bar{X}$ is also normal. As σ is known, we use the Z-distribution.

"Correct to within $\pm$ 10 minutes" means sampling error E=10.

$$\mathrm{E} = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{Z_{(1-0.9)/2}45}{10}\right)^2 = \left(\frac{1.645\times45}{10}\right)^2 = 54.797 \approx 55\,(\mathit{round\ up})$$

Hence, the agency needs at least 55 observations in order to be 90% confident that the error of estimation is within $\pm$ 10 minutes.

b) $n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{Z_{(1-0.99)/2}45}{10}\right)^2 = \left(\frac{2.575\times45}{10}\right)^2 = 134.2702 \approx 135\,(\mathit{round\ up})$

Hence, the agency needs at least 135 observations in order to be 95% confident that the error of estimation is within $\pm$ 10 minutes.

Q5. The personal director of a large corporation wants to study absenteeism among clerical workers at the corporation's central office during the year. A random sample of 25 clerical workers reveals the following:

Absenteeism: $\bar{X}$=9.7 days, s=4.0 days

a) Set up a 95% confidence interval estimate of the average number of absences for clerical workers. Give a practical interpretation of the interval obtained.

b) What assumption must hold in order to perform the estimation in (a)?

Solution:

a) Assume that population is normal, the sampling distribution of of $\bar{X}$ is also normal. As σ is unknown, we use the t-distribution.

95% confidence interval for $\mu = \bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 9.7 \pm t_{0.025,24}\frac{4}{\sqrt{25}}$

$$= 9.7 \pm 2.0639 \times \frac{4}{\sqrt{25}}$$

$$= [8.0489, 11.3511]$$

We are 95% confident that the average number of absences is between 8.049 and 11.351 days.

Q5. The personal director of a large corporation wants to study absenteeism among clerical workers at the corporation's central office during the year. A random sample of 25 clerical workers reveals the following:

Absenteeism: $\bar{X}$=9.7 days, s=4.0 days

b) What assumption must hold in order to perform the estimation in (a)?

c)* If the personnel director also wants to take a survey in a branch office, what sample size is needed if the director wants to be 95% confident of being correct to within ±1.5 days and the population standard deviation is assumed to be 4.5 days?

Solution:

b) Assume that the number of absences is normally distributed, so that the sampling distribution of $\bar{X}$ is normal.

c) $\alpha = 1 - 0.95 = 0.05$;

"Correct to within $\pm$ 1.5 minutes" means sampling error E=1.5;

$\sigma = 4.5.$

$$\text{n} = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{Z_{0.05/2}4.5}{1.5}\right)^2 = \left(\frac{1.96\times4.5}{1.5}\right)^2 = 34.57 \approx 35\mathit{(round\ up)}$$

Hence, the personal director needs at least 35 observations in order to be 95% confident that the error of estimation is within $\pm$ 1.5 days.

Q6*. A food inspector, examining 10 bottles of a certain brand of honey, obtained the following percentages of impurities:

23.5 19.8 21.3 22.6 19.4 18.2 24.7 21.9 20.0 21.1

a) What are the mean and standard deviation of this sample?

b) With 95% confidence, what is the sampling error if the inspector used the sample mean to estimate the mean percentage of impurities in this brand of honey?

Solution:

a) The sample mean: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{23.5+19.8...+20+21.1}{10} = 21.25$

The sample std: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} = \sqrt{\frac{(23.5-21.25)^2 + \cdots + (21.1-21.25)^2}{9}} = 1.9896$

b) Assume the population distribution is normal.

So the sampling distribution of $\bar{X}$ is normal. As σ is unknown, we use the t-distribution.

Sampling error, $E = t_{\alpha/2,n\text{-}1}\frac{s}{\sqrt{n}} = t_{0.05/2,9}\frac{1.9896}{\sqrt{10}} = 2.2622 \times \frac{1.9896}{\sqrt{10}} = 1.4233$

Q7. A sample of 12 observations is obtained from an infinite population with normal distribution. Based on the sample data, the 95% confidence interval for the population mean is calculated to be [20, 30]. Form this 95% confidence interval, determine the mean and standard deviation of the sample.

Solution:

The population distribution is normal, so the sampling distribution of $\bar{X}$ is normal.
As σ is unknown, we use the t-distribution.

95% confidence interval for $\mu = \bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = \bar{X} \pm t_{0.05/2,11}\frac{s}{\sqrt{12}}$

$$= \bar{X} \pm 2.2010\frac{s}{\sqrt{12}} = [20,30]$$

$$\rightarrow \begin{cases} \bar{X} - 2.2010\frac{s}{\sqrt{12}} = 20 \\ \bar{X} + 2.2010\frac{s}{\sqrt{12}} = 30 \end{cases} \qquad \rightarrow \begin{cases} \bar{X} = 25 \\ s = 7.8694 \end{cases}$$

Q8. From past experience, the numbers of vitamin supplements sold per day in a health food store well fit a normal distribution with variance 9. The number of vitamin supplements sold per day in a sample of 11 days is obtained:

19	19	20	20	20	22	23	25	26	27	30

a) Construct a 95% confidence interval for population average number of vitamin supplements sold per day in the store.

b) Someone suggests constructing the confidence interval by t-distribution. Do you agree? Explain your opinion.

Solution:

a) Since the population is a normal distribution with known standard deviation (σ =3), we use Z-distribution.

95% confidence interval for $\mu = \bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$$= \frac{19+\cdots+30}{11} \pm Z_{0.05/2} \times \left(\frac{3}{\sqrt{11}}\right) = 22.8182 \pm 1.96 \times \left(\frac{3}{\sqrt{11}}\right)$$

$$= [21.0453, 24.5911].$$

We are 95% confident that the population average number of vitamin supplements sold per day in the store is between 21.0453 and 24.5911.

b) No, although n is small, the sampling distribution follows normal distribution as the population distribution is normal. In addition, the population variance (s.d.) is given, normal distribution, instead of t-distribution, should be used to construct the confidence interval.

Q8. From past experience, the numbers of vitamin supplements sold per day in a health food store well fit a normal distribution with variance 9. The number of vitamin supplements sold per day in a sample of 11 days is obtained:

19	19	20	20	20	22	23	25	26	27	30

c) The data analyst found that there is a data-entry mistake. The last observation should be smaller than 30. How will this affect the confidence interval?

d)* If the company would like to limit the sampling error within ±0.5, what is the least sample size needed for a 90% confidence interval?

Solution:

c) 95% confidence interval for $\mu = [\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$

- The new observation is smaller, the new sample mean will be smaller as well. $\bar{X} \downarrow$
- The confidence interval will shift to left with center at the new sample mean.
- However, the width of the interval will not be changed as the width of the confidence interval is related to level of confidence (α), population standard deviation (σ) and sample size (n). While these three values keep unchanged, the width of interval will not change.

d) sampling error E=0.5;

$$\text{n} = (\frac{Z_{\alpha/2}\sigma}{E})^2 = (\frac{Z_{(1-0.9)/2}3}{0.5})^2 = (\frac{1.645\times 3}{0.5})^2 = 97.4169 \approx 98 \ (round\ up)$$

Hence, at least 98 observations is required in order to be 90% confident that the sampling error is within $\pm$ 0.5.