GE2262 Business Statistics

Topic 5
Confidence Interval Estimation

Reference

Levine, D.M., Krehbiel, T.C. and Berenson, M.L., *Business Statistics: A First Course*, Pearson Education Ltd, Chapter 8

Outline

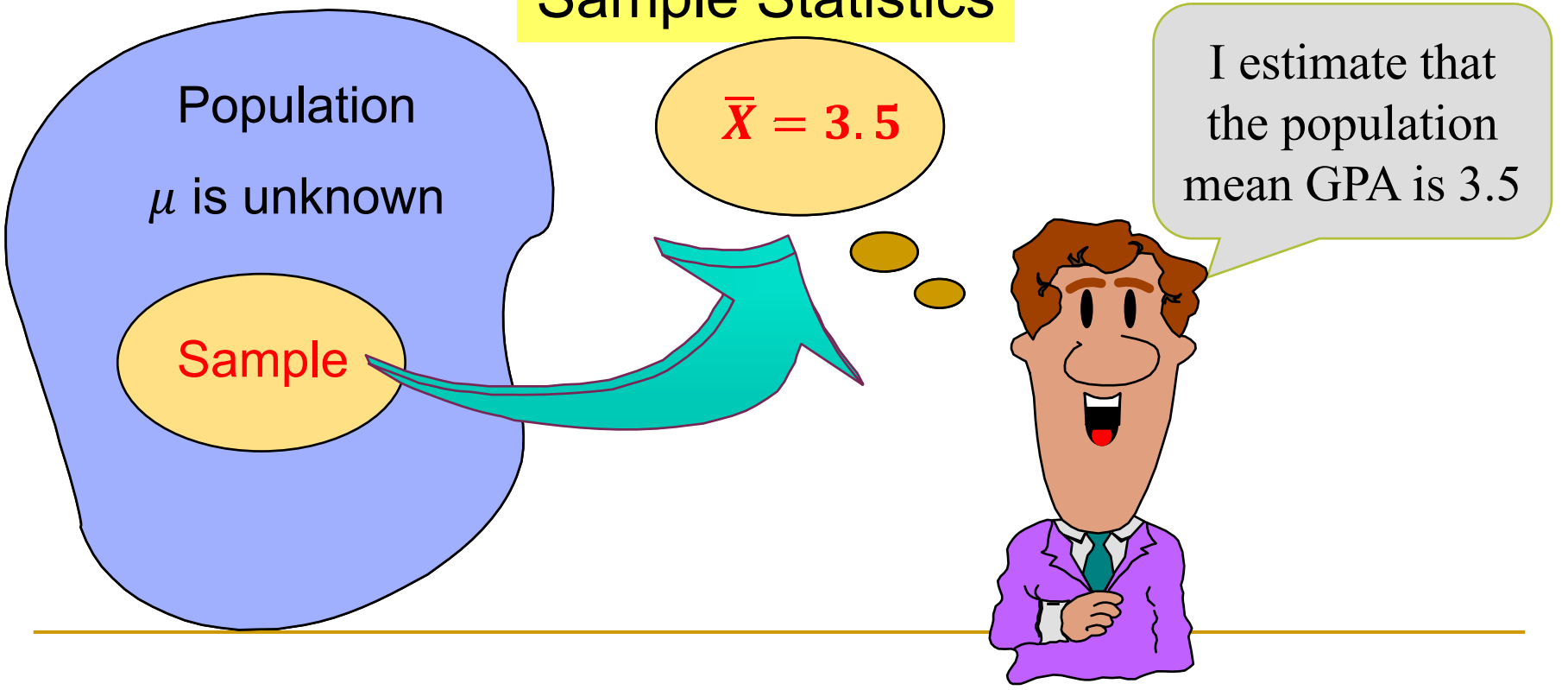
- Point Estimates
- Confidence Interval Estimation for the Population Mean
  - Using Z Distribution
  - Using t Distribution
- Determining Required Sample Size for Estimating Mean

Point Estimates

1. Define the Population

2. Select a Random Sample and Obtain the Sample Statistics

3. Make an Estimation



Point Estimates

Cont'd

	Population Parameters	Sample Statistics
Mean	μ	$\bar{X}$
Variance	σ^2	S^2
Proportion	π	p

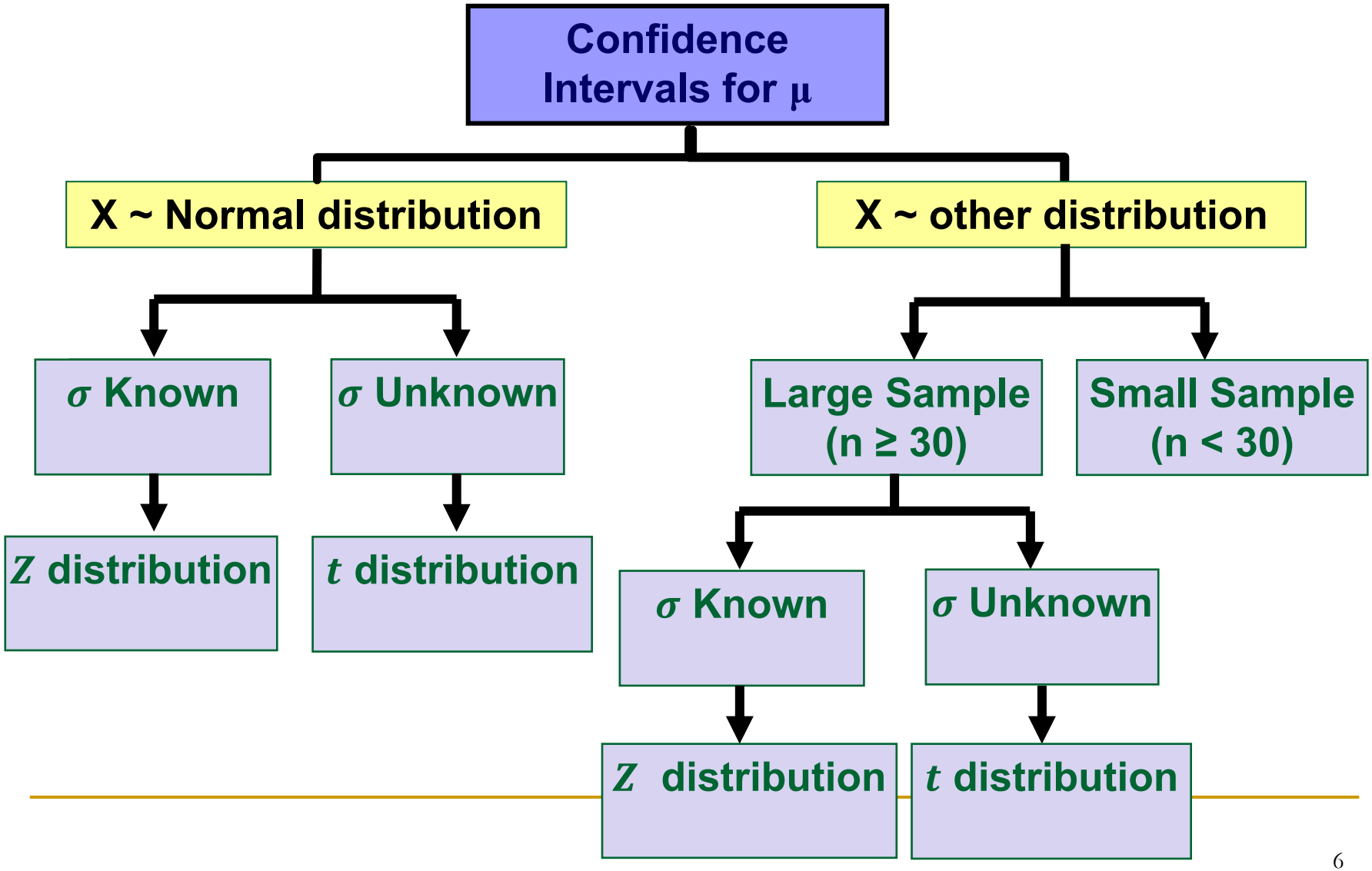
← To be discussed in Topic 7

- How likely the sample statistics can truly estimate the population parameter?
  - i.e. What is $P(\mu = \bar{X})$?
- We will see how to assess the accuracy and reliability of this point estimation

What is a Confidence Interval?

- Provides a range of values
  - Based on one sample from the population
  - Taken into consideration the variation in sample statistics from sample to sample, i.e. standard error
  - Stated in terms of level of confidence, $100(1-\alpha)\%$
    - Gives the assurance that, if the statistical model is correct, then taken over all data that might have been obtained, the range estimates that obtained using the sample statistic would then included the true population parameter the proportion of time set by the confidence level $(1-\alpha)$
    - Can never be 100% sure!

Confidence Interval Estimation



Confidence Interval for μ (σ Known)

- Conditions
  - Population standard deviation (σ) is known
  - Population is normally distributed → $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$
  - If population is not normal, but with a large sample ($n \geq 30$), by Central Limit Theorem → $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$

- $100(1-\alpha)\%$ Confidence interval estimate

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Standard Error, $\sigma_{\bar{X}}$

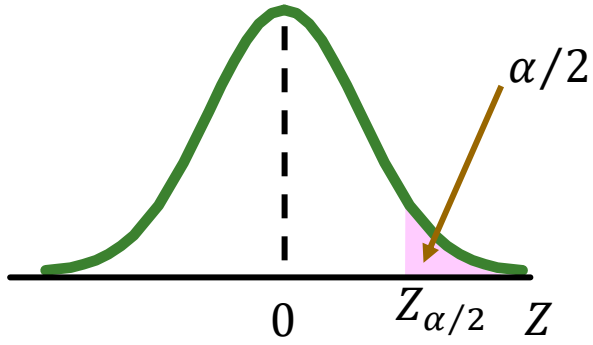
Sampling Error, E
(Also called margin of error)

Critical Value

Confidence Interval for μ (σ Known)

Cont'd

- Level of Confidence $100(1-\alpha)\%$
  - Confidence that the interval will cover the unknown population mean
- Z-value (Critical Value)
  - $Z_{\alpha/2}$ is the value corresponding to an upper-tail probability of $\alpha/2$ from the standardized normal distribution
- Sampling Error (Margin of Error)
  - $E = Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

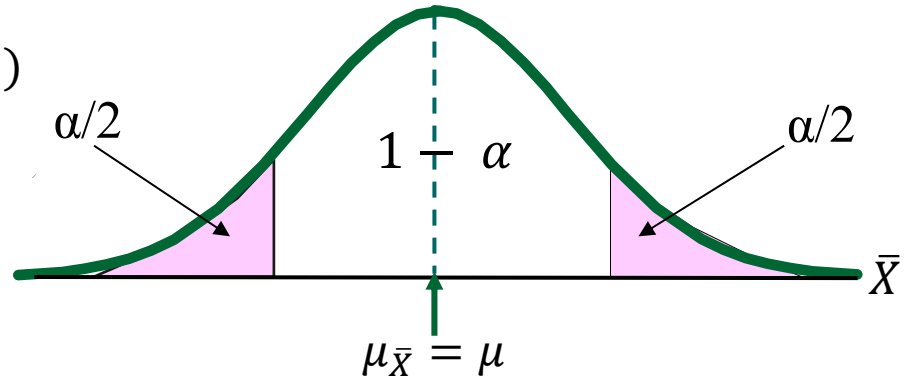


Confidence Interval for μ (σ Known)

Cont'd

- Derivation

$\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$



$$P\left(-Z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right) = 1-\alpha$$

$$\rightarrow P\left(-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\rightarrow P\left(Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \geq \mu-\bar{X} \geq -Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

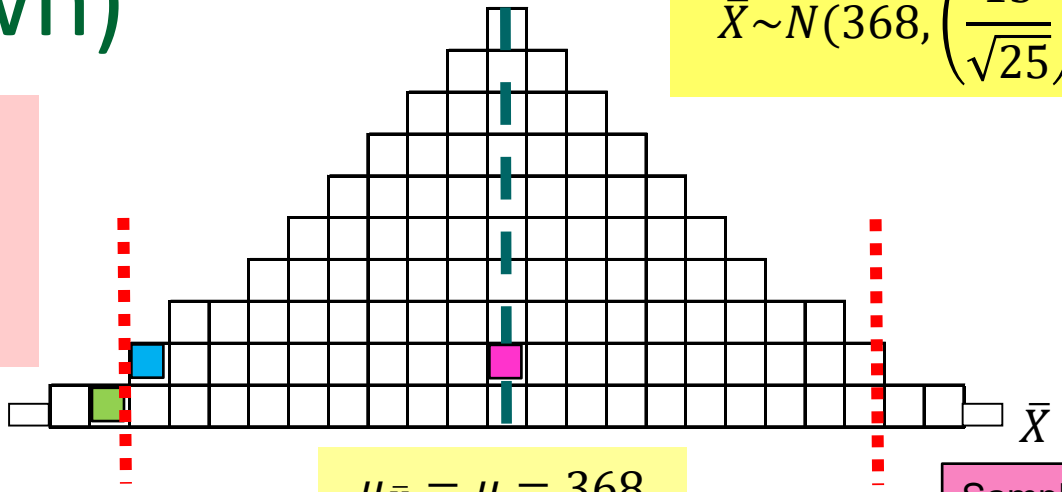
$$\rightarrow P\left(\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Confidence Interval for μ (σ Known)

Cont'd

$$\bar{X} \sim N\left(368, \left(\frac{15}{\sqrt{25}}\right)^2\right)$$

Construct 95% Confidence Intervals from $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ to $\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$



$\mu_{\bar{X}} = \mu = 368$

362.12 — 378.88

356.42 — 368.18

355.12 — 366.88

Sample 1: $\bar{X} = 368$
95% CI: $368 \pm 1.96 \times \frac{15}{\sqrt{25}} = [362.12, 378.88]$

Sample 2: $\bar{X} = 362.3$
95% CI: $362.3 \pm 1.96 \times \frac{15}{\sqrt{25}} = [356.42, 368.18]$

Sample 3: $\bar{X} = 361$
95% CI: $361 \pm 1.96 \times \frac{15}{\sqrt{25}} = [355.12, 366.88]$

If you repeat the sampling by 100 times, you will find that 95% of intervals constructed cover μ; 5% do not.

Confidence Interval for μ (σ Known)

Cont'd

- A relative frequency interpretation
  - In the long run, $100(1-\alpha)\%$ of all the confidence intervals that can be constructed will cover the unknown population parameter

- A conventional interpretation
  - We are $100(1-\alpha)\%$ confident that the unknown population parameter lies between $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
    - In other words, you got this interval by a method that gives correct results $100(1-\alpha)\%$ of the time

Confidence Interval for μ (σ Known)

Cont'd

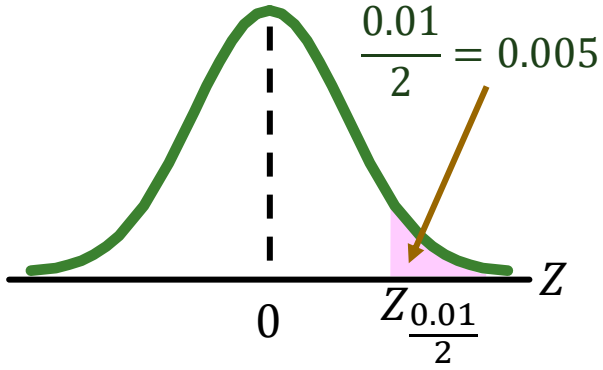
- A specific interval will either cover or not cover the population parameter
  - No probability involved in a specific interval
  - As you cannot know whether your sample is one of the $100(1-\alpha)\%$ for which the interval catches μ, or one of the unlucky 5%

Confidence Interval for μ (σ Known) – Exercise

- A random sample of 15 stocks traded on the Hang Seng Index showed an average shares traded to be 215,000
- From the past experience, it is believed that the population standard deviation of shares traded is 195,000 and the shares traded are very close to a Normal distribution
- Construct a 99% confidence interval for the average shares traded on the Hang Seng Index. Interpret the result.

Confidence Interval for μ (σ Known) – Exercise

Cont'd



Confidence Interval for μ (σ Known) – Exercise

Cont'd

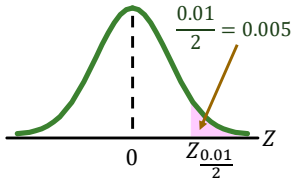
Since the population number of shares traded (X) follows Normal distribution, the distribution of sample means also follows Normal distribution, i.e. $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$

With known population standard deviation (σ), Z distribution is used

99% confidence interval (C.I.) for μ

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 215000 \pm Z_{0.01/2}\frac{195000}{\sqrt{15}}$$

$$= 215000 \pm 2.575\frac{195000}{\sqrt{15}} = [85351.88, 344648.12]$$



Confidence Interval for μ (σ Unknown) – Exercise

Cont'd

Let X be the amount spent in the store and $\bar{X}$ be the sample mean amount spent

The samples are drawn from an unknown distribution, but at a large size n = 200, by CLT $\bar{X}$ follows normal distribution approximately, but σ is unknown, and t distribution is used

95% confidence interval (C.I.) for μ

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} = 21.34 \pm 1.96\frac{9.22}{\sqrt{200}}$$

$$= [20.0622, 22.6179]$$

We are 95% confident that the population mean amount spent is between \$20.0622 and \$22.6179

Determining Sample Size – Exercise

Cont'd

2. If we want to increase our confidence to 95%, how many individuals should we interview? Keep all other factors remain unchanged.

$$n = (\frac{Z_{\alpha/2}\sigma}{E})^2 = (\frac{1.96 \times 28.84}{2.5})^2 = 511.24 \cong 512$$

Confidence Interval for μ (σ Known) – Exercise

Cont'd

- Interpretation

- If all possible samples of size 15 are taken and the corresponding 99% confidence intervals are constructed, 99% of these intervals will cover the unknown population mean

- We are 99% confident that the population average number of shares traded on the Hang Seng Index is between 85351.88 and 344648.12

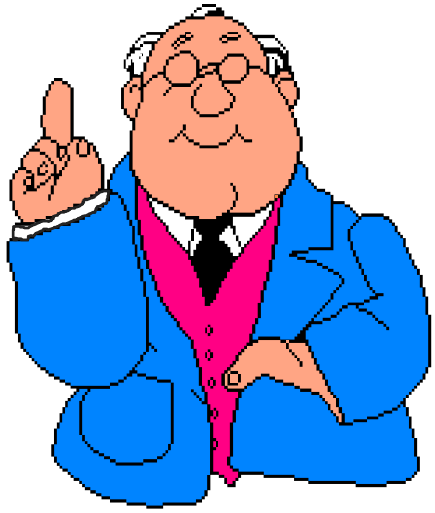
- There is 99% chance that the unknown population mean will be in between 85351.88 and 344648.12

Factors Affecting Interval Width (Precision)

- Data variation
  - Measured by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
  - $n \uparrow \rightarrow \sigma_{\bar{X}} \downarrow \rightarrow$ width of interval $\downarrow$

Intervals extend from $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ to $\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

- Level of confidence
  - Measured by $100(1-\alpha)\%$
  - $(1-\alpha) \uparrow \rightarrow |Z\text{-value}| \uparrow \rightarrow$ width of interval $\uparrow$



- σ can never be changed
- $\bar{X}$ affects the location of the interval, but not the width

http://www.rossmanchance.com/applets/ConfSim.html

Confidence Interval for μ (σ Unknown)

- Recall that since $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, then $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

  This transformation holds as long as the population standard deviation σ is known

- If σ is unknown, it needs to be estimated by the sample standard deviation S from a sample containing n values, $X_1, X_2, \ldots, X_n$ such that

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

Student's t-distribution

Cont'd

- The variable $\mathrm{T} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ no longer follows the standardized normal distribution
- Mathematically, T follows a distribution called Student's t-distribution
  - Also simply called t-distribution
  - Often denoted as $T \sim t(\nu)$
- The probability density function of t-distribution is

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad \nu > 0$$

where Γ is the gamma function and ν is the parameter of the function which often called the degrees of freedom

Student's t-distribution

Cont'd

- Degrees of freedom in t-distribution
  - The number of values in the final calculation of a statistic that are free to vary
  - Equals to the total number of observations used in the analysis minus the number of parameters estimated as intermediate steps in the estimation of the parameter itself
- When σ is an unknown, S needs to be estimated in order to construct confidence interval for μ
  - From the sample with size n, 1 degree of freedom is loss as to estimate $\bar{X}$ embedding in S
  - Therefore, we have $(n-1)$ degrees of freedom when constructing the necessary confidence interval

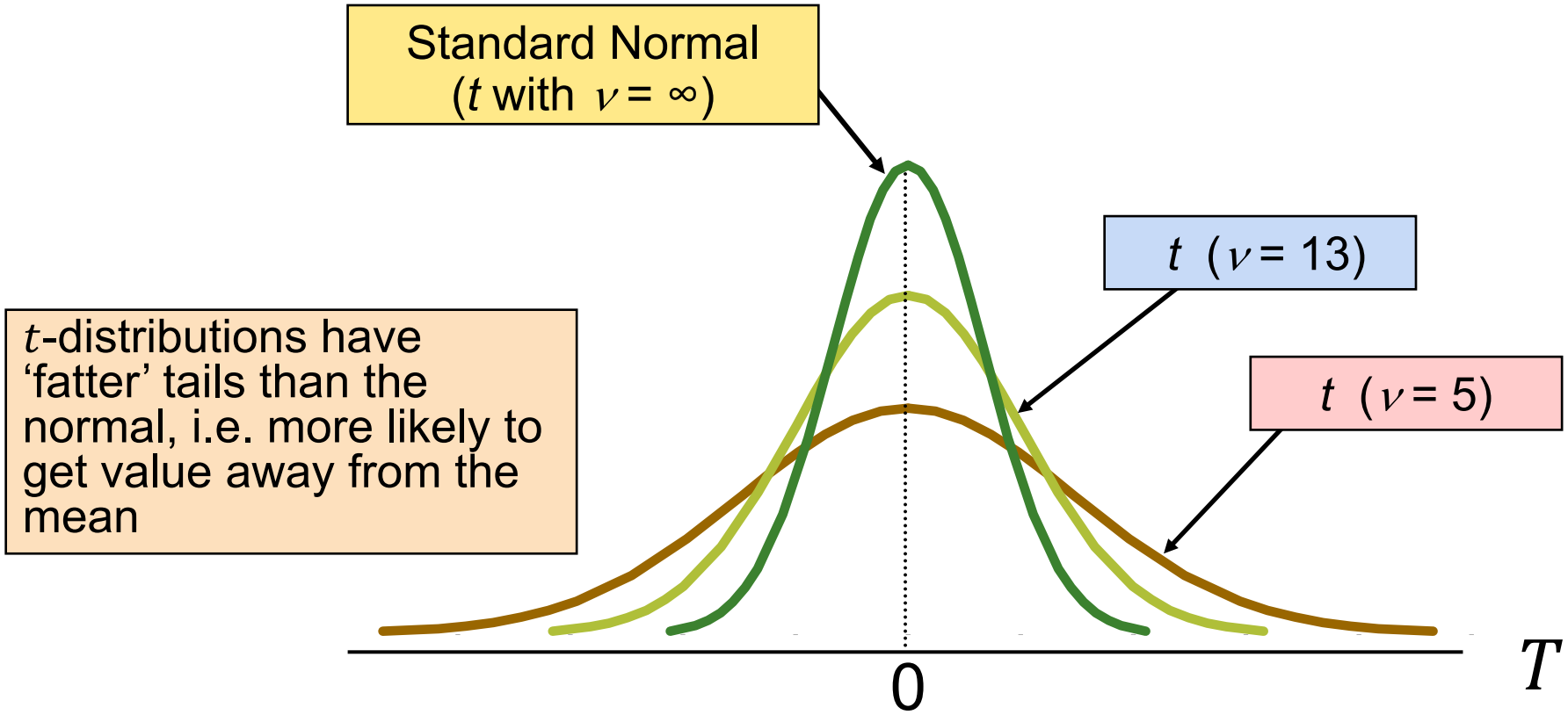
Student's t-distribution

Cont'd

- Properties of t distribution
  - For $T \sim t(\nu)$
    - Mean & Standard Deviation
      - Mean = 0 for ν>1, otherwise it is undefined
      - Standard deviation = $\nu/(\nu-2)$ for ν> 2, = ∞ for $1 < \nu \leq 2$, otherwise undefined
    - The shape of the density function
      - The theoretical range of T is infinite, i.e. $-\infty$ to $+\infty$
      - Bell shaped
      - Symmetric about T = 0
      - Median = mode = 0
      - As ν increases, the density curve approaches the $N(0, 1)$ curve

Student's *t*-distribution

Cont'd



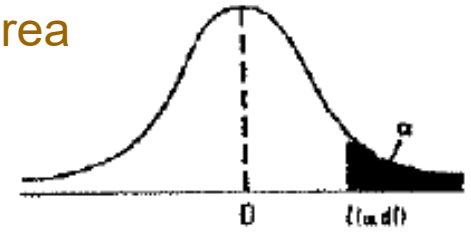
Student's t-distribution

Cont'd

The column gives the upper tail area

Critical Values of t

For a particular number of degrees of freedom, entry represents the critical value of t corresponding to a specified upper-tail area (α)



	Upper-Tail Areas					
Degrees of Freedom	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995

The row shows the degrees of freedom

The value within the table gives the t-value corresponding to a particular degrees of freedom and upper-tail area

At 7 degrees of freedom, $P(t > 2.3646) = 0.025$

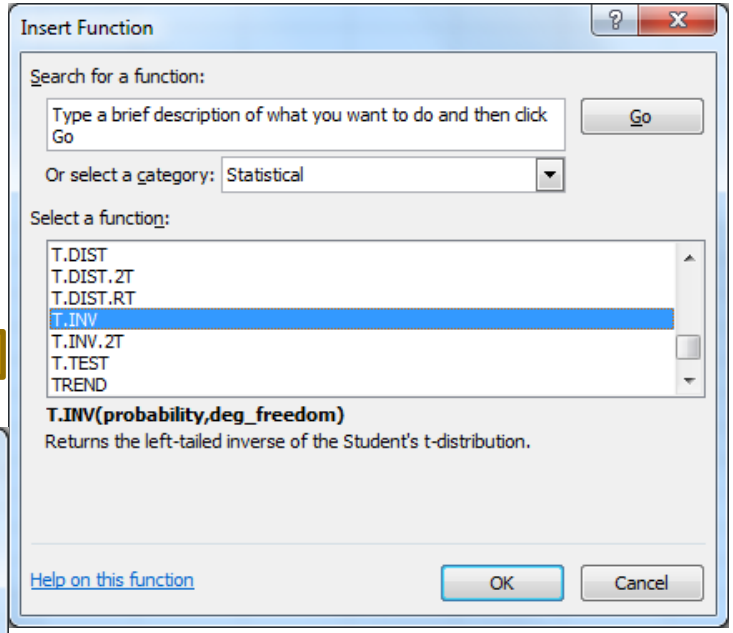
Finding t-Value in Excel

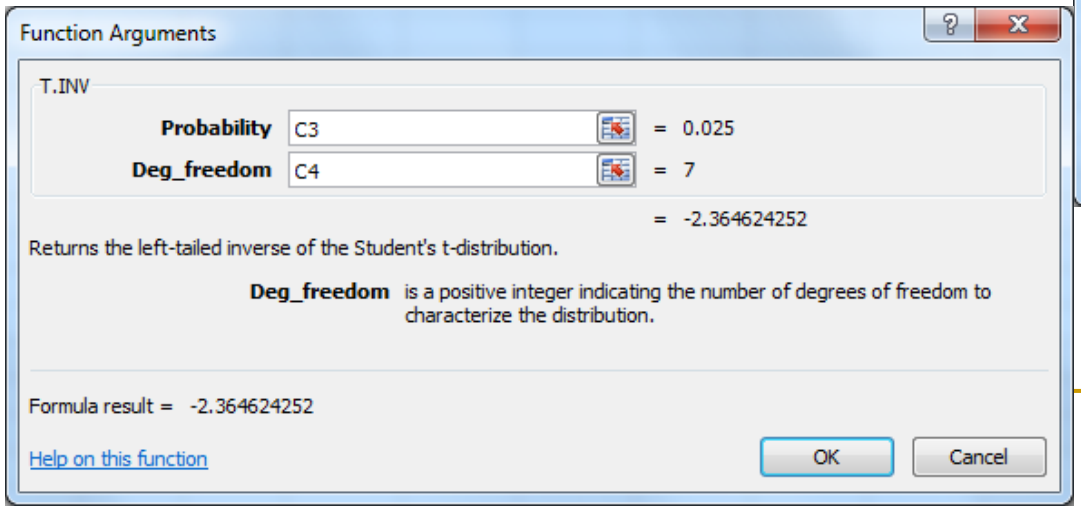
Step 1: Type the given information $(\frac{\alpha}{2}, df)$

Step 2: Insert the "T.INV" function

	A	B	C
1	t Distribution		
2			
3	Lower Tail Probability	α/2 =	0.025
4	Degrees of Freedom	df =	7
5			
6	t-Value =	-2.3646	

=T.INV(lower tail probability, df)





Confidence Interval for μ (σ Unknown)

Cont'd

- Conditions
  - Population standard deviation (σ) is unknown
  - Population is normally distributed → $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$
  - If population is not normal, but with a large sample ($n \geq 30$), by Central Limit Theorem → $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$
- $100(1-\alpha)\%$ Confidence interval estimate

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

  - With the use of t-distribution with $(n-1)$ degrees of freedom in this context

Confidence Interval for μ (σ Unknown) – Example

- The monthly salary of brokers is found to be Normally distributed
- A random sample of 25 brokers has a mean monthly salary HK$80K and a standard deviation of HK$16K
- Set up a 95% confidence interval estimation for the population mean

Confidence Interval for μ (σ Unknown) – Example

Cont'd

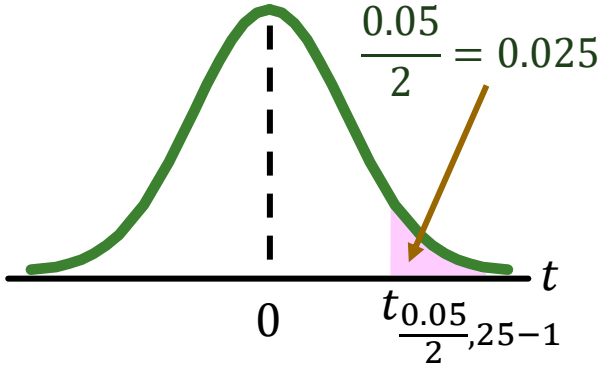
Since the monthly salary of brokers (X) follows Normal distribution, the distribution of sample means also follows Normal distribution, i.e. $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$

However, σ is unknown, t distribution is used

95% confidence interval (C.I.) for μ

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} = 80 \pm t_{0.05/2,25-1}\frac{16}{\sqrt{25}}$$

$$= 80 \pm 2.0639\frac{16}{\sqrt{25}} = [73.396, 86.604]$$



We are 95% confident that the population mean monthly salary of brokers is between HK$73.396K and HK$86.604K

Confidence Interval for μ (σ Unknown) – Exercise

Cont'd

- The branch manager of an outlet of a nationwide chain of pet supply stores want to study characteristics of her customers. In particular, she would like to estimate the population mean amount spent in the pet supply store. A random sample of 200 customers is selected. The sample mean of amount of money spent is $21.34, and the sample standard deviation is $9.22. Construct a 95% confidence interval estimate for the population mean amount spent in the pet supply store.

Confidence Interval for μ (σ Unknown) – Exercise

Cont'd

Confidence Interval for μ (σ Known) – Exercise

Cont'd

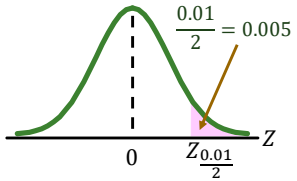
Since the population number of shares traded (X) follows Normal distribution, the distribution of sample means also follows Normal distribution, i.e. $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$

With known population standard deviation (σ), Z distribution is used

99% confidence interval (C.I.) for μ

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 215000 \pm Z_{0.01/2}\frac{195000}{\sqrt{15}}$$

$$= 215000 \pm 2.575\frac{195000}{\sqrt{15}} = [85351.88, 344648.12]$$



Confidence Interval for μ (σ Unknown) – Exercise

Cont'd

Let X be the amount spent in the store and $\bar{X}$ be the sample mean amount spent

The samples are drawn from an unknown distribution, but at a large size n = 200, by CLT $\bar{X}$ follows normal distribution approximately, but σ is unknown, and t distribution is used

95% confidence interval (C.I.) for μ

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} = 21.34 \pm 1.96\frac{9.22}{\sqrt{200}}$$

$$= [20.0622, 22.6179]$$

We are 95% confident that the population mean amount spent is between \$20.0622 and \$22.6179

Determining Sample Size – Exercise

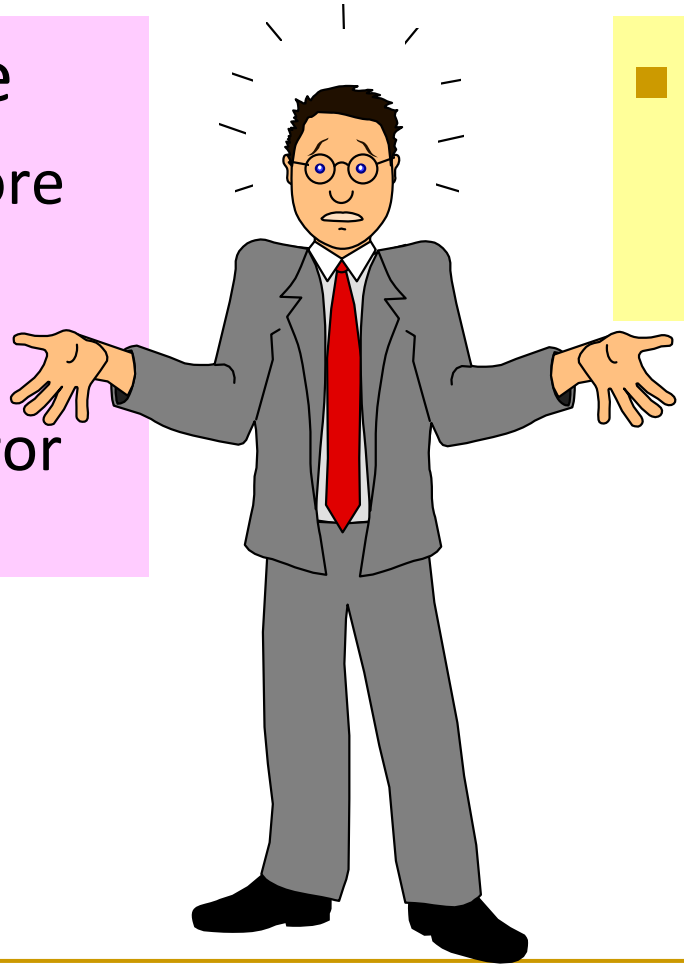
Cont'd

2. If we want to increase our confidence to 95%, how many individuals should we interview? Keep all other factors remain unchanged.

$$n = (\frac{Z_{\alpha/2}\sigma}{E})^2 = (\frac{1.96 \times 28.84}{2.5})^2 = 511.24 \cong 512$$

Determining Sample Size

- Large sample
  - Requires more resources
  - Smaller standard error

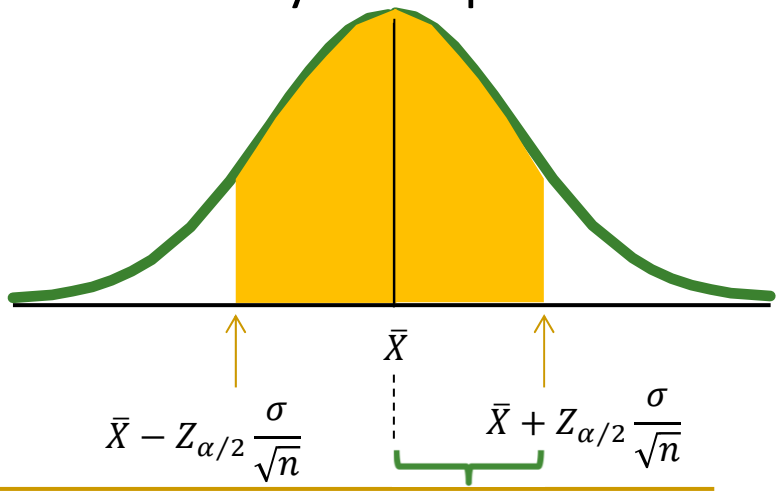


- Small sample
  - Less precise estimation

Determining Sample Size

Cont'd

- Statisticians have control over the sampling error (E) by choosing appropriate sample sizes that are large enough to make the results appear credible
  - Sampling error measures how far off the estimation results are likely to be from the result that they would have gotten if the entire population are surveyed instead of merely a sample
  - $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- The required sample size can be found to reach desired sampling error with a specified level of confidence $(1-\alpha)$



Determining Sample Size

Cont'd

- What sample size is needed to be $100(1-\alpha)\%$ confidence of being correct to within $\pm E$?
- Assume σ is known

$$P(\mu - E \leq \bar{X} \leq \mu + E) = 1 - \alpha$$

$$\rightarrow P(-E \leq \bar{X} - \mu \leq E) = 1 - \alpha$$

$$\rightarrow E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\rightarrow n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2$$

Determining Sample Size – Example

Source: Hong Kong Free Press, 30 October 2019

HONG KONG **POLITICS & PROTEST**

Hong Kong leader Carrie Lam's popularity rating drops by over 2% to another historic low – poll

30 October 2019 16:20 · Kris Cheng · 3 min read

Donate

Chief Executive Carrie Lam's popularity rating has dropped to a record low of 20.2 points out of 100, a recent public opinion survey has found.

The rating is the lowest among any post-colonial chief executive, according to the Hong Kong Public Opinion Research Institute (HKPORI), the crowdfunded successor to the University of Hong Kong's Public Opinion Programme. The institute interviewed 1,038 people between October 17 and 23 over the phone.

Only 11 per cent of respondents said they would support Lam as the chief executive if they had a right to vote, representing a drop of three per cent from the last survey conducted in early October, whilst 82 per cent said they would not support her, representing an increase of three per cent – a record low.

Since June, Hong Kong has seen large-scale protests against the now-withdrawn extradition bill which would have enabled extraditions to China. After 21 weeks, they have evolved into a wider movement calling for democracy, with sometimes violent displays of dissent over alleged Beijing's encroachment and police brutality.

Lam's popularity has declined rapidly since the start of the protests, according to the HKPORI.

Chief Executive Carrie Lam enters the Chamber of the Legislative Council Complex in order to deliver her 2019 Policy Address. Photo: inmediahk.net.

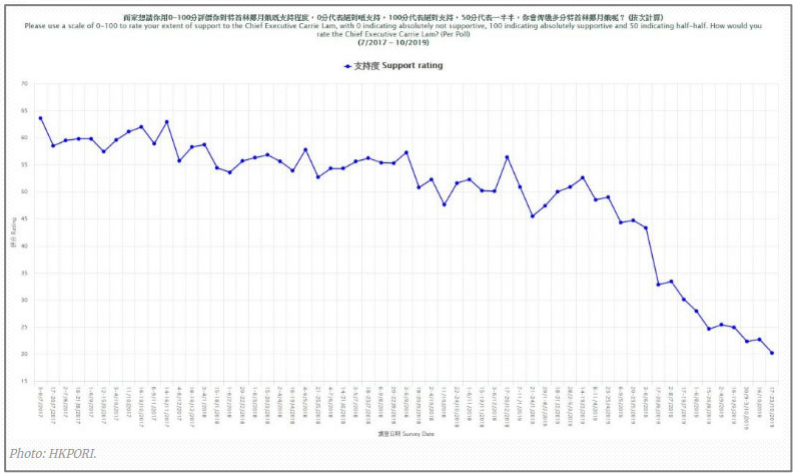


Photo: HKPORI.

Determining Sample Size – Example

Cont'd

http://hkupop.hku.hk/english/index.html https://www.pori.hk/

- POP Poll is a regularly survey aims to monitor to what extend does the general public support to the Chief Executive Mrs. Carrie Lam
- The target population is Cantonese speakers in Hong Kong of age 18 or above
- Randomly selected respondents are asked to rate their extend of support by using a scale of 0 – 100
  - 0 indicating absolutely not supportive
  - 100 indicating absolutely supportive

Determining Sample Size – Example

Cont'd

- According to the POP Poll conducted between 17-23 October 2019, the 1038 random sampled respondents give a mean score is 20.23 with standard deviation 28.84

1. In order to be 90% confident of being correctly reflecting the population opinion to within ± 2.5 points, what sample size is needed?

Use S to replace σ when σ is unknown

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.645 \times 28.84}{2.5}\right)^2 = 360.12 \cong 361$$

Round Up

Determining Sample Size – Exercise

Cont'd

2. If we want to increase our confidence to 95%, how many individuals should we interview? Keep all other factors remain unchanged.

Confidence Interval for μ (σ Known) – Exercise

Cont'd

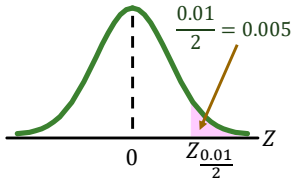
Since the population number of shares traded (X) follows Normal distribution, the distribution of sample means also follows Normal distribution, i.e. $\bar{X} \sim N(\mu_{\bar{X}}, (\frac{\sigma}{\sqrt{n}})^2)$

With known population standard deviation (σ), Z distribution is used

99% confidence interval (C.I.) for μ

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 215000 \pm Z_{0.01/2}\frac{195000}{\sqrt{15}}$$

$$= 215000 \pm 2.575\frac{195000}{\sqrt{15}} = [85351.88, 344648.12]$$



Confidence Interval for μ (σ Unknown) – Exercise

Cont'd

Let X be the amount spent in the store and $\bar{X}$ be the sample mean amount spent

The samples are drawn from an unknown distribution, but at a large size n = 200, by CLT $\bar{X}$ follows normal distribution approximately, but σ is unknown, and t distribution is used

95% confidence interval (C.I.) for μ

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} = 21.34 \pm 1.96\frac{9.22}{\sqrt{200}}$$

$$= [20.0622, 22.6179]$$

We are 95% confident that the population mean amount spent is between \$20.0622 and \$22.6179

Determining Sample Size – Exercise

Cont'd

2. If we want to increase our confidence to 95%, how many individuals should we interview? Keep all other factors remain unchanged.

$$n = (\frac{Z_{\alpha/2}\sigma}{E})^2 = (\frac{1.96 \times 28.84}{2.5})^2 = 511.24 \cong 512$$

Determining Sample Size

Cont'd

- If the population standard deviation is unknown, we need to guess the value of standard deviation based on some prior information such as the sample standard deviation in earlier similar studies
- If no previous sample data are available, one practical approach is to develop an estimate of the range of the data and then estimate the standard deviation as range/4

Determining Sample Size – Example

Cont'd

- Suppose you want to estimate the mean GPA (μ) of all the students at your university at a margin of error of 0.3 and 95% confidence. How many students should be sampled?
  - For 95% confidence level, α = 0.05, then $Z_{\alpha/2}$ = 1.96
  - Since the range of GPA is 4.3, we estimate the standard deviation as 4.3/4
  - The required sample size is

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.96 \times \left(\frac{4.3}{4}\right)}{0.3}\right)^2 = 49.33 \cong 50$$

Determining Sample Size

Cont'd

- Warning:
  - The value of n does not depend on the size of the population. This is true as long as the population is much larger than the sample
  - The derived sample size should only be taken as a rough indicator for the desired margin of error
  - The true required sample size, which is unknown to us in practice, might be larger or smaller than the computed value
  - Many statistical studies report the margin of error of involved estimates. This should not be taken as the error of a study. The margin of error only reports the possible error size of an estimate due to sampling, at a specified confidence level and under a "perfect environment"